Discussion Paper Series – CRC TR 224

Discussion Paper No. 510
Project B 07

Optimal Internality Taxation of Product Attributes

Andreas Gerster[1]
Michael Kramm[2]

February 2024

[1] University of Mannheim & RWI – Leibniz Institute for Economic Research, Email: gerster@uni-mannheim.de
[2]University of Cologne & Technical University of Dortmund, Email: kramm@wiso.uni-koeln.de

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Optimal Internality Taxation of Product Attributes

Andreas Gerster* and Michael Kramm†

October 11, 2023

Abstract

This paper explores how a benevolent policy maker should optimally tax (or subsidize) product attributes when consumers are behaviorally biased. We demonstrate that market choices are informative about biases, which can be exploited for targeting biased consumers via a non-linear tax schedule. We show that its properties depend on few parameters of the joint distribution of consumer valuations and biases. Furthermore, we provide a novel justification for behaviorally motivated product standards and derive when a combination of taxes and standards is optimal. We illustrate our findings based on a numerical example from the lightbulb market.

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

JEL codes: H21, D82, D04, Q58.

Acknowledgements: We are grateful for valuable comments and suggestions by Jason Abaluck, Felix Bierbrauer, Sebastian Fuchs, Lorenz Götte, Andreas Haufler, Eckhard Janeba, Duk Gyoo Kim, Wolfgang Leininger, Jörg Peter Lingens, Lars Metzger, Alex Rees-Jones, Adam Sanjurjo, Rene Saran, Arthur Seibold, Thomas Tröger, and three anonymous referees. This paper also benefited from discussions with participants of the Early-Career Behavioral Economics Conference (Princeton, 2021), the ZEW Public Finance Conference (Mannheim, 2021), the IIPF Annual Congress (Reykjavik, 2020), the Conference on Mechanism and Institution Design (Klagenfurt, 2020), the Public Economic Theory Conference (Strasbourg, 2019), the European Winter Meeting of the Econometric Society (Naples, 2018), the CESifo Area Conference on Public Sector Economics (Munich, 2019), and the Behavioral Economic Policy Symposium (Maastricht, 2019), as well as a seminar at the University of Mannheim. A previous version of this paper was titled "Optimal Non-Linear Taxation of Internalities". Support by the German Research Foundation (DFG) through CRC TR 224 (Project B07) is gratefully acknowledged. Copyright American Economic Association; reproduced with permission.

*University of Mannheim, L7 3-5, Mannheim, Germany, phone: +49 621 181-1791, and RWI - Leibniz Institute for Economic Research, Essen, Germany; mail: gerster@uni-mannheim.de

†University of Cologne, Sibille-Hartmann-Straße 2-8, Cologne, Germany, and Technical University of Dortmund, Vogelpothsweg 87, Dortmund, Germany, phone: +49 221 470 1741, mail: kramm@wiso.uni-koeln.de

1 Introduction

A growing body of literature has demonstrated that consumers may make decisions that are not in their best interest. For example, present focused consumers undervalue the future cost from consuming "sin goods", such as cigarettes or sugar (Allcott et al., 2019; O'Donoghue and Rabin, 1999, 2006). Consumers are often inattentive to opaque product attributes, such as the energy efficiency of an appliance (Allcott and Taubinsky, 2015a) or the expected out-of-pocket costs of a health insurance plan (Abaluck and Gruber, 2011). Furthermore, consumers hold biased beliefs concerning 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.

In practice, the notion that consumers make mistakes has led policy makers to impose taxes, product bans or standards, and combinations thereof. For example, policy makers have taxed cigarettes and sugar, and banned unhealthy foods, drugs, and risky financial contracts (see, e.g., Chaloupka et al., 2019). Another important application is the regulation of energy efficiency. Policy makers in the US and the EU have imposed minimum standards for appliances and buildings (DOE, 2017; European Commission, 2015) and initiated large-scale subsidy programs (BMF, 2019; EIA, 2018). To influence car ownership decisions, governments have levied car registration taxes that are non-linear in fuel efficiency, imposed fleet-wide emission standards, and subsidized the scrappage of inefficient vehicles. While taxes (or subsidies), standards, and combinations thereof are widely used in practice, there is little theoretical guidance on the optimal regulation of internalities to increase consumer welfare.

In this paper, we derive optimal internality taxes in a non-linear commodity taxation framework. The main application of our approach is the taxation (or subsidization) of product attributes, such as the energy efficiency of an electrical appliance, the fuel efficiency of a car, and the sugar content of a beverage. In these cases, non-linear taxation is feasible as all consumers have to pay the same price for a product variety, which rules out problems of resale. We employ a general model of biases that encompasses a broad class of behavioral failures driving a wedge between "normative" and "decision utility" (Bernheim and Taubinsky, 2018), such as present focus, limited attention, and biased beliefs. We also allow for arbitrary consumer heterogeneity in preferences and biases. Based on this generic specification of a behavioral bias, we derive the optimal non-linear commodity tax. Using data by Allcott and Taubinsky (2015a), we apply our results to the light bulb market and determine the optimal non-linear subsidy for energy efficiency.

Our paper puts forward a simple, but so far little appreciated, economic logic whereby behavioral consumers partly reveal their bias through their position on the demand curve. The policy maker uses this information to target internality taxes towards biased consumers via a possibly non-linear tax schedule, taking consumers' self-selection into account. Our framework also allows us to explore the optimality of uniform product standards that mandate a particular attribute level, as well as minimum and maximum standards that ban products with low or high attribute levels.

Our results show that the optimal tax rate is determined by two components. The first term captures the corrective motives and corresponds to the average bias of consumers who are marginal to a local change in the tax rate, while the second term captures redistributive motives. We explore the properties of the optimal tax schedule and demonstrate that it hinges crucially on what we denote as the "local bias heterogeneity". It measures the degree to which consumers with marginally different perceived valuations vary in their average bias. Based on a first-order approximation, we demonstrate that local bias heterogeneity depends on the correlation between biases and valuations, as well as the ratio of their standard deviations. We find that knowledge of these two parameters is sufficient to derive whether the optimal tax schedule is convex, concave, or linear.

Furthermore, we provide a novel rationale for product standards. We show that the optimal internality tax scheme involves standards if there is a fundamental misalignment between consumers' perceived preferences and the normative stance of the policy maker. More precisely, standards become necessary as soon as higher valuations, as perceived by consumers, coincide with lower normative valuations, as inferred by the policy maker. When preferences are fundamentally misaligned, the use of price instruments for correcting behaviorally biased consumers leads to a failure of incentive compatibility. The policy maker must then optimally resort to standards, i.e., she must restrict the set of product varieties on the market. Furthermore, we find that the degree of local bias heterogeneity determines the relative merits of linear taxes ("price regulation") relative to standards ("quantity regulation"). As we show, standards yield higher expected welfare when local bias heterogeneity is high, whereas linear taxes are better suited for situations where the local bias heterogeneity is low.

In addition, we investigate mixed policies that combine taxes or subsidies with minimum or maximum standards. We demonstrate that their optimality hinges crucially on the concept of coskewness, which has previously been applied in finance (see, e.g., Kraus and Litzenberger, 1976). The coskewness of perceived valuations $\hat{v}$ and biases b is defined as their standardized third cross central moment: $cosk(\hat{v}, b) = E\left((\hat{v} - \mu_{\hat{v}})^2(b - \mu_b)\right) / (\sigma_{\hat{v}}^2 \sigma_b)$. It measures how outliers in $\hat{v}$ coincide with values of b. We show that imposing a maximum standard, i.e., banning products with attribute

levels above the standard, is optimal if two conditions are met. First, perceived valuations must be positively skewed, so that positive outliers in $\hat{v}$ exist. Hence, some individuals would demand product varieties with extreme attribute levels in the absence of regulation. Second, the skew in $\hat{v}$ must originate predominantly from biases, as measured by the coskewness of $\hat{v}$ and b. Conversely, a minimum standard is optimal when the presence of a negative skew in the distribution of perceived valuations can be attributed to the bias b. These insights demonstrate when mixed policies are optimal and provide a novel rationale for their widespread use in practice.

To illustrate our findings, we determine the optimal non-linear tax for energy efficiency in the light bulb market. We show that the optimal non-linear tax increases welfare significantly beyond the optimal linear tax. In this empirical setting, we find that the optimal policy involves minimum standards that ban very inefficient light bulbs. Yet, we also find that implementing the optimal internality tax based on our first-order approximation realizes similar welfare gains without restricting consumers' choice sets.

Our paper contributes to several strands of the literature. First and foremost, it is closely related to a literature on the optimal taxation of behaviorally biased consumers. Allcott and Taubinsky (2015a) derive the optimal corrective tax for a binary choice and show that it corresponds to the average bias of consumers who are indifferent between both options at market prices. O'Donoghue and Rabin (2006) show that linear corrective taxes on sin goods are welfare-improving when some consumers are present-biased, while others are unbiased. More recently, Farhi and Gabaix (2020) provide a general framework for assessing optimal non-linear income taxation and linear commodity taxation with behavioral agents. Furthermore, Allcott et al. (2019) investigate the interplay between redistributive and corrective motives by allowing for non-linear income taxes and linear commodity taxes. Ferey et al. (2021) explore optimal non-linear income and commodity taxes under two-dimensional heterogeneity in abilities and preferences for the commodity. They also investigate a corrective taxation motive by allowing for biased consumers and one-dimensional consumer heterogeneity.[1]

We extend this body of literature by deriving the optimal internality tax schedule for behaviorally biased consumers who choose from a spectrum of quality-differentiated products. By allowing for non-linear tax schedules, our approach generalizes previous research on optimal binary subsidies for energy efficiency (Allcott and Taubinsky, 2015a) and optimal linear commodity taxation (e.g., O'Donoghue and Rabin 2006, Allcott et al. 2019). In comparison to Ferey et al. (2021), we explore corrective taxation

[1] Furthermore, Allcott et al. (2014) and Carlsson and Johansson-Stenman (2019) analyze the use of multiple policy instruments in the presence of internalities and externalities.

in a setting where biases may be heterogeneous conditional on an individual's type. Such heterogeneity has important implications for the shape of the optimal non-linear tax schedule and the question whether price instruments or standards are welfare-optimal.

Our paper also relates to studies on the relative merits of price versus quantity regulation and extends them to behavioral consumers. In a seminal contribution, Weitzman (1974) showed that quantity regulation may yield higher welfare than price regulation when uncertainty about the cost and benefits of a good exists. Beyond that, economic justifications for bans are limited. Their widespread use in practice is commonly explained by normative views that certain behaviors are undesirable, as well as practical considerations that the enforcement of bans may be less costly (e.g., Glaeser and Shleifer, 2001). More recently, Farhi and Gabaix (2020) provide a behavioral rationale for quantity regulation. Assuming that valuations and biases are uncorrelated, they show that a large variance of the bias favors quantity instruments relative to linear tax instruments. Similarly, Houde and Myers (2019) find that minimum energy efficiency standards yield higher welfare than a Pigouvian tax on energy when consumers differ strongly in their misperception of operating cost.[2]

Our results add to this literature by formalizing when bans and standards are optimal among all feasible (possibly non-linear) price instruments. They provide a novel rationale for product standards by showing that standards are optimal when there is a fundamental misalignment of preferences between the consumers and the regulator. We also provide conditions under which it is optimal to regulate behaviorally biased consumers via a mix of taxes and standards. Furthermore, we compare linear tax instrument with standards allowing for any joint distribution of perceived preferences and biases.

More broadly, our paper is related to a literature that has explored how firms can use price discrimination to maximize profits (see, e.g., Mussa and Rosen, 1978; Bergemann et al., 2015; and 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. As part of the optimal pricing problem, firms may also decide to engage in product line pruning, i.e., to offer only a subset of all possible product varieties (Johnson and Myatt, 2003). We build on the same basic intuition, but explore the potential of a differentiated internality tax schedule to correct choices of behaviorally biased consumers.

[2]More generally, our results relate to Moser and Olea de Souza e Silva (2019), who allow for two-dimensional heterogeneity in present focus and abilities to show that optimal savings policies for present-biased workers involve bunching for low-ability types.

From a methodological angle, we build on Mirrlees's (1971) seminal 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 the optimal income taxation of workers who misjudge the benefit of working because of present focus (Lockwood, 2020) or because of biases more broadly (Gerritsen, 2016). These papers have shown that workers' behavioral biases provide a corrective motive for subsidizing work through lower tax rates if consumers undervalue the benefit of working, and for increasing tax rates in case of an overvaluation. We add to this literature strand by exploring the rationale of a policy maker to impose bans or product standards and by showing that the shape of the optimal non-linear internality tax schedule can be derived from two parameters of the joint distribution of valuations and biases.

While the main application of our paper is the taxation of product attributes, its results also carry over to the case when a policy maker taxes quantities and issues consumption bans. Non-linear taxes on quantities are feasible if purchases can be tracked and resale is forbidden or difficult. This holds true for cannabinoids that must be purchased in officially recognized stores, for instance. Other examples include the purchase of cars, guns, and houses, where policy makers can observe all purchases based on official registers.

The remainder of the paper is structured as follows. In Section 2, we introduce our model. Section 3 derives the analytical characterization of the optimal tax schedule and the conditions that ensure its implementability. In Section 4, we investigate the curvature of the optimal tax schedule and the optimality of standards. In Section 5, we provide a numerical example, which we use to illustrate our findings and to quantify the optimal non-linear subsidy scheme for energy efficiency in the light bulb market. Section 6 discusses our findings and concludes.

2 Model

In this section, we present a model that allows us to analyze how the policy maker (*she*) implements a welfare maximizing tax schedule in an economy with behaviorally biased consumers (*he*).[3] The policy maker commits to a (possibly non-linear) tax regime $t : Q \to \mathbb{R}$, where $Q \subseteq [0, \infty)$ is the consumer's choice set. Subsequently, a consumer chooses $q \in Q$, his consumption of the product attribute level (or "quality"). We begin by presenting the decision problem of the consumer and then describe the problem of the policy maker.

[3] To simplify notation and without loss of generality, we assume that the mass of consumers is one.

2.1 Consumer

The choice variable of a consumer is the attribute level q of a quality-differentiated product, where $q \in Q \subseteq [0, \infty)$. The consumer has unit demand for the good. His normative per-unit valuation of the benefits of consuming a good with attribute level q is captured by the random variable v, which is distributed according to the cumulative distribution function F_v with support $supp(F_v) := [\underline{v}, \overline{v}] \subseteq (-\infty, \infty)$, density function f_v, a finite expected value μ_v, and variance σ_v. The bias b, reflecting a misperception of the valuation of q, is distributed according to the cumulative distribution function F_b with support $supp(F_b) := \left[\underline{b}, \overline{b}\right] \subseteq (-\infty, \infty)$, density function f_b, a finite expected value μ_b, 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 attribute level q 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 normative valuation v and the bias b. The perceived valuation $\hat{v}$ is induced by the distributions of v and b. It is distributed according to the cumulative distribution function F and the density f and its mean is denoted by $\mu_{\hat{v}}$.[4]

Let $z \in \mathbb{R}$ denote the money (numeraire good) a consumer spends for the consumption of other goods. For simplicity, we assume a quasilinear utility function that abstracts from income effects.[5] The increasing and weakly convex cost function of producing q is given by $c : Q \to \mathbb{R}$. We assume that the good with product attribute level q is produced on competitive markets so that the cost function corresponds to the price of consuming q net of taxes. The exogenous, real-valued scalar $m > 0$ denotes the initial endowment with the numeraire. Therefore, the budget constraint is given by $z \leq m - c(q) - t(q)$.

Under the standard assumption that the consumer fully utilizes his budget for consumption, decision utility can be written as:

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

and normative utility as:

$$u^n(q, t, v) = z + vq = m + vx - t(q) - c(q).$$

[4] More formally, the density f is determined by the joint distribution $\mathfrak{f}(v, b)$ of v and b according to $f(\hat{v}) = \int_v \mathfrak{f}(v, \hat{v} - v)\mathrm{d}v$ and its support $supp(F) := \left[\underline{\hat{v}}, \overline{\hat{v}}\right]$ is determined by the support of the distributions F_v and F_b.

[5] This assumption is particularly plausible for applications such as the subsidization of appliances with high energy efficiency or the taxation of beverages with high sugary content.

As common in the behavioral public finance literature, we model the bias as the wedge between marginal decision and normative utility (see, e.g., Allcott et al. 2019 and Farhi and Gabaix 2020). The bias is defined as $b = u_q^d(q, t, \hat{v}) - u_q^n(q, t, v) = \hat{v} - v$, where u_q^d and u_q^n denote the first derivatives of decision and normative utility with respect to q. Hence, a consumer with bias $b = 0$ is unbiased, while $b < 0$ ($b > 0$) imply underestimation (overestimation) of the marginal utility of consumption. Intuitively, the bias is a money-metric measure of the difference between a consumer's perceived and normative valuation. We write q^d as shorthand notation for the choice of a biased consumer given by:

$$q^d(\hat{v}, t) := \arg\max_q u^d(q, t, \hat{v}).$$

2.2 Policy Maker

The policy maker's objective is to find the welfare-maximizing tax schedule $t : Q \to \mathbb{R}$, based on information about the distributions F, F_v, and F_b. She can condition taxes on the consumer's choice of q, but not on a realization of the random variables v, b, and $\hat{v}$. This assumption reflects that obtaining information about individual-level valuations and biases prior to a purchase may be infeasible or associated with prohibitive cost. We model the policy maker's social welfare function $W : \mathbb{R} \to \mathbb{R}$ using non-negative Pareto weights $\alpha(\hat{v})$, so that $u^n \mapsto \alpha(\hat{v})u^n$ with $\alpha : \left[\underline{\hat{v}}, \overline{\hat{v}}\right] \to \mathbb{R}_+$.[6] Furthermore, we assume that the policy maker considers the implications of tax revenues for financing a government budget of B: $\int_{\hat{v}} t(q^d) f(\hat{v}) d\hat{v} \geq B$.[7]

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

$$\max_{t \in \mathbb{T}} \int_{\hat{v}} \alpha(\hat{v}) \cdot E\left[u^n\left(q^d, t, v\right) | \hat{v}\right] f(\hat{v}) d\hat{v} + \lambda \left(\int_{\hat{v}} t(q^d) f(\hat{v}) d\hat{v} - B\right), \tag{1}$$

where $\lambda \in \mathbb{R}$ denotes the multiplier of the budget constraint.

2.3 Internality Revelation and Local Bias Heterogeneity

In the policy maker's optimization problem, we can rewrite $E\left[u^n\left(q^d, t, v\right) | \hat{v}\right] = u^d\left(q^d, t, \hat{v}\right) - E\left[b | \hat{v}\right] q^d$. A crucial term in this expression is the conditional expectation $E\left[b | \hat{v}\right]$, which captures the information a policy maker can learn about individual

[6] Modeling Pareto weights as a function of v would introduce an additional updating problem about $E(\alpha(v)|\hat{v})$, without adding substantive insights. In particular, the optimal tax formula from Equation (2) would then contain the covariance between biases and welfare weights, $cov\left(b(\hat{v}_q), \hat{g}(\hat{v}_q)\right)$, as an additional summand.

[7] We make this assumption for consistency with the optimal taxation literature. Abstracting from a government budget constraint leaves our results unaffected.

biases from observing his perceived valuation $\hat{v}$. We start by defining "internality revelation" and "local bias heterogeneity" as follows.

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

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

For a differentiable conditional expectation function $E\left[b|\hat{v}\right]$, internality revelation exists if and only if local bias heterogeneity $A(\hat{v}) = \partial E[b|\hat{v}]/\partial\hat{v}$ is non-zero at some point, i.e., $\exists \hat{v} \in \left[\underline{\hat{v}}, \overline{\hat{v}}\right] : A(\hat{v}) \neq 0$.

Under internality revelation, a policy maker can learn about the local heterogeneity in consumers' biases by observing their perceived valuations. Local bias heterogeneity, which we define as $A(\hat{v}) = \partial E[b|\hat{v}]/\partial\hat{v}$, is important from two perspectives. First, it has an information value because it captures the extend to which the policy maker can infer a higher or lower bias from observing a higher perceived valuation. Second, it describes the degree of the local (mis)alignment of preferences between the consumer and the policy maker. Differentiating the definition $E[\hat{v}|\hat{v}] = E[b|\hat{v}] + E[v|\hat{v}]$ with respect to $\hat{v}$ shows that if local bias heterogeneity is too large, i.e., $\partial E[b|\hat{v}]/\partial\hat{v} > 1$, there is a fundamental misalignment of preferences among consumer and policy maker. In that case, a consumer with a higher $\hat{v}$ would like to consume a product of higher attribute level, but the normative stance of the policy maker implies that he should consume a product of lower attribute level ($\partial E[v|\hat{v}]/\partial\hat{v} < 0$).

3 Optimal Internality Taxation

In this section, we derive the optimal tax schedule. We proceed in two steps. First, we derive the optimal non-linear tax schedule and discuss its properties. Second, we determine the conditions under which the optimal non-linear tax schedule can be implemented. We also explore the consequences of a failure of implementability, which arises if the smooth tax schedule is not incentive compatible, i.e., if consumers do not reveal their perceived valuations via their choices. We solve for the optimal tax schedule using a mechanism design approach (Mirrlees, 1971) as it provides a unified framework for analyzing both the optimal tax schedule and the conditions for its implementability. A derivation of the optimal non-linear tax schedule using the so-called perturbation approach (Saez, 2001) is provided in Appendix A.A.2.

3.1 The Optimal Smooth Non-Linear Tax Schedule

We first assume that the optimal tax schedule, i.e., the solution to the policy maker's problem in Equation (1), satisfies incentive compatibility and characterize it in Proposition 1.[8]

Proposition 1 (Optimal Non-linear Commodity Tax)**.** *If the solution to the policy maker's problem in Equation* (1) *is characterized by first-order conditions, the optimal non-linear commodity tax rate is given by:*

$$t'(q) = \hat{g}(\hat{v}_q)\ E[b|\hat{v}_q] + (1 - \hat{G}(\hat{v}_q))\frac{1 - F(\hat{v}_q)}{f(\hat{v}_q)} \quad \forall q \in Q, \tag{2}$$

where $\hat{v}_q$ is the type of a consumer that yields the allocation q under the optimal tax schedule, $\hat{g}(\hat{v}) := \frac{\alpha(\hat{v})}{\lambda}$ is the social marginal welfare weight, and $\hat{G}(\hat{v}) := \frac{\int_{\hat{v}}^{\infty} \hat{g}(m) f(m) dm}{1 - F(\hat{v})}$.

Equivalently, the optimal tax schedule can be written as:

$$t'(q) = g(q)E[b|q] + \frac{1 - G(q)}{e(q)a(q)}\left(t'(q) + c'(q)\right) \quad \forall q \in Q, \tag{3}$$

where $g(q) := \hat{g}(\hat{v}_q)$, $G(q) := \hat{G}(\hat{v}_q)$, $a(q) := \frac{qh(q)}{1 - H(q)}$, and h and H are the density and distribution of attribute levels q under the optimal tax schedule. The term e denotes the price elasticity of demand for q evaluated at the net-price $t' + c'$.

Proof. See Appendix A.A □

To convey the intuition of the optimal tax rate, we analyze the two summands separately. We first examine the first summand of Equation (2), which embodies a corrective taxation motive. It is proportional to the expected bias of a consumer conditional on his perceived valuation, which represents the information a policy maker can infer about internalities from observing consumer choices. The policy maker uses her potential to correct the choice of a consumer to the extent that she can infer his internality. The expected bias of a type is weighted by his social marginal welfare weight g, which captures the policy maker's valuation of increasing the utility of the respective consumer, measured in terms of public funds λ. This weight reflects that the rationale of corrective taxation is to make an individual consumer better off. It differs from the rationale of externality taxation, where the degree of correction depends on the damage suffered by other consumers.

[8] Participation constraints do not matter as a consequence of the assumption of unit demand. This assumption is common in the literature (e.g., Mussa and Rosen 1978). In our case, it can also be motivated by products from which a policy maker would not want to exclude participants by taxation, such as electric appliances and real-estate.

If the policy maker does not have redistributive preferences ($g = G = 1$), the marginal tax rate from Equation (2) simplifies to the expected bias conditional on the type, $t'(q) = E[b|\hat{v}_q]$. This result bears close similarity with the findings of Allcott and Taubinsky (2015a) 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. Yet, the optimal non-linear tax differs from that result by its dependence on $\hat{v}$ rather than a fixed market price. This dependence has important implications for the optimal tax schedule, which we analyze in Section 4. Furthermore, the optimal non-linear tax improves upon a constant per-unit tax by exploiting additional information on consumers' perceived valuations, as revealed under the non-linear tax schedule. By contrast, the optimal linear tax corresponds to an average bias, weighted by the consumer's demand responsiveness. Because it uses only an aggregate measure of bias, a linear tax has a lower potential to target behaviorally biased consumers (for derivations, see Appendix A.B).

The second summand of Equation (2) captures a redistributive motive of taxation. It only matters if the government has preferences for redistribution, as reflected by social margial welfare weights that are unequal to one ($g \neq 1$). Without corrective motives ($E[b|\hat{v}_q] = 0$), the policy maker encounters a classical equity-efficiency trade-off (for analogous results in the income taxation literature, see Mirrlees 1971; Diamond 1998). The optimal tax rates increase in the product attribute q (are "progressive") if the welfare weight of a consumer choosing at least attribute level q is lower than the average welfare weight ($\hat{G} < 1$). If the reverse holds true, the tax rates decrease in q (are "regressive"). The term $(1 - \hat{G})$ is weighted more strongly in the optimal tax formula if the probability mass above $\hat{v}_q$ as measured by $1 - F(\hat{v}_q)$ is high and if the probability density $f(\hat{v}_q)$ is low. In that case, an increase of the tax rate at q is effective in raising additional revenue while causing only modest distortions for marginal consumers.

In Equation (3), we express the optimal taxation formula in terms of consumers' product choices q. Its first term captures the corrective rationale of the tax, which depends on the average bias of consumers who consume an attribute level of q under the optimal tax schedule. It bears similarity to findings by Gerritsen (2016) and Lockwood (2020), who show that the optimal income tax rate includes an additive corrective term, which captures the average degree of misoptimization of consumers at a given income level, weighted by their (average) social marginal welfare weight.

The second term captures redistributive motives. As in the optimal income taxation literature (e.g., Saez 2001), the curvature of the tax depends on three factors. First, it increases in the strength of preferences for redistribution ($\partial|t'|/\partial|1 - G| > 0$). Second, it decreases in the elasticity e of the demand for q ($|\partial t'|/\partial e < 0$). Third, if

the distribution of consumption is thin at the top, as measured by a low value of the Pareto parameter a, the distortive effects of a tax are small relative to the effects of raising revenue from all consumers choosing higher product attribute levels. Hence, a low value of a increases the curvature of the tax ($\partial|t'|/\partial(-a) > 0$). Our findings differ from optimal income taxation literature in that the optimal tax rate is proportional to the marginal net-price of the attribute level q, $t' + c'$, rather than the net-of-tax rate $1 - t'$ (see e.g. Saez 2001). This difference reflects that consumers' marginal rate of substitution between numeraire consumption and q in our model is not identical to the respective rate between numeraire consumption and gross income in income taxation models.

3.2 Failure of Incentive Compatibility and Implementation via Bunching

We now allow for a situation where the smooth optimal tax schedule from Proposition 1 does not satisfy incentive compatibility and hence cannot be implemented. We first investigate the conditions that guarantee the implementability of the smooth optimal internality tax schedule and then discuss the implications for policy when these conditions are not met.

Proposition 2 (Internality Tax Implementability)**.** *A necessary and sufficient condition for the implementability of the optimal tax schedule from Proposition 1 is satisfied at $\hat{v}$ if:*

$$\underbrace{1 - \hat{g}(\hat{v})\frac{\partial E[b|\hat{v}]}{\partial \hat{v}}}_{\substack{\textcircled{1} \\ \geq 0 \text{ if "no fundamental misalignment"}}} \underbrace{- \frac{\partial \hat{g}(\hat{v})}{\partial \hat{v}} E[b|\hat{v}]}_{\substack{\textcircled{2} \\ = 0 \text{ if "welfare weights uniform"}}} \underbrace{- \frac{\partial}{\partial \hat{v}} \left[(1 - \hat{G}(\hat{v})) \cdot \frac{1 - F(\hat{v})}{f(\hat{v})} \right]}_{\substack{\textcircled{3} \\ \geq 0 \text{ if "no excess redistribution"}}} \geq 0. \tag{4}$$

In the absence of redistributional preferences ($\hat{g} = 1$), the expression simplifies to a "no fundamental misalignment" condition between the normative stance of the policy maker and the consumer.

Proof. See Appendix A.C □

To give an intuition for Proposition 2, we start by considering its first term ("no fundamental misalignment"). In the absence of redistributive motives ($\hat{g} = 1$), it can be rewritten as $\partial E[v|\hat{v}]/\partial \hat{v} \geq 0$, using that $\hat{v} = v + b$. This term requires that there exists a minimum alignment between the normative stance of a policy maker and the perceived valuations of the consumer. The condition is violated if higher valuations as perceived by the consumer are valued less by the policy maker ($\partial E[v|\hat{v}]/\partial \hat{v} < 0$). This condition follows from the distinction between perceived and normative valuations and, to the best of our knowledge, is novel to the literature on optimal taxation and

mechanism design. It is violated when when equilibrium normative utility should optimally be decreasing in $\hat{v}$, while incentive compatibility requires that decision utility is non-decreasing in $\hat{v}$.

Since the policy maker must ensure a minimum alignment of preferences, there is a maximum degree of bias heterogeneity that she can exploit through corrective taxation. Higher values of local bias heterogeneity $\partial E[b|\hat{v}]/\partial\hat{v}$ imply that a policy maker can extract much information about consumer biases from observing perceived valuations. In that case, she would like to strongly differentiate tax rates (Proposition 1). Yet, tax differentiation is only feasible to the extent that a fundamental misaligment between the normative stance of the policy maker and the consumer does not materialize.

When we allow for redistributive tastes of the government, i.e., $g(\hat{v}) \neq 1$, the implementability condition from Proposition 2 involves a second and a third term. The second term captures that a policy maker would like to differentiate corrective taxes if the social marginal welfare weights differ across $\hat{v}$. The third term reflects that the policy maker may want to differentiate tax rates for redistributional motives alone. It is equivalent to the implementability condition in the optimal income taxation literature (e.g. Mirrlees 1971; Diamond 1998) and depends on the hazard rate and on $1 - \hat{G}(\hat{v})$, i.e., the deviation of the average marginal welfare weight of all consumers above $\hat{v}$, $\hat{G}(\hat{v})$, from the mean of all weights (which equals one).

The rationale of the implementability condition from Proposition 2 is to set an upper limit for the differentiation of tax rates. To see that link, it is useful to consider the optimal tax formula from Equation (1). The implementability condition requires that $\partial t'/\partial\hat{v} \leq 1$, i.e., it limits the increase of the tax rate in the attribute level. Intuitively, if a tax schedule increases too steeply in the attribute level, it disincentivizes the consumption of higher attribute levels to such an extent that a violation of incentive compatibility occurs.

Proposition 3 (Standards as Part of the Optimal Policy)**.** *If the solution to the policy maker's problem in Equation* (1) *does not satisfy Internality Tax Implementability, the optimal policy includes bunching. In particular, if Internality Tax Implementability...*

1. *... is violated globally, i.e., for every $\hat{v}$, the optimal policy consists of setting a uniform standard;*

2. *... is violated at the lower bound of the perceived valuation distribution, i.e., for every $\hat{v} \in [\underline{\hat{v}}, \hat{v}^l]$, the optimal policy consists of setting a minimum standard;*

3. *... is violated at the upper bound of the perceived valuation distribution, i.e., for every $\hat{v} \in [\hat{v}^h, \overline{\hat{v}}]$, the optimal policy consists of setting a maximum standard.*

Proof. See Appendix A.D. □

If Internality Tax Implementability from Proposition 2 is not met, incentive compatibility can be ensured by applying the so-called ironing approach (Myerson, 1981; Guesnerie and Laffont, 1984). The approach achieves implementability by bunching, i.e., by assigning the same commodity bundle to individuals with different (perceived) valuations. Proposition 3 demonstrates that if implementability fails globally, the only way to ensure incentive compatibility is to impose a uniform standard. Minimum or maximum standards, i.e., bans of high or low attribute levels, are part of the optimal non-linear tax schedule if failure of implementability occurs at the bounds of the distribution of perceived valuations. In all three cases, the policy maker restricts the set of product varieties available to consumers in order to approximate the optimal non-linear tax schedule, while taking implementability constraints into account.

4 Properties of the Optimal Non-Linear Tax Schedule

In this section, we investigate the properties of the optimal internality tax schedule. For the ease of exposition, we assume that the government has no preference for redistribution, which implies that marginal social welfare weights equal one for every consumer.[9] In that case, according to Proposition 1, the optimal tax rate equals the expected bias conditional on the perceived valuation $\hat{v}$ of consumers consuming q; that is, $t'(q) = E[b|\hat{v}_q]$.

The optimal tax rate can be decomposed into two components. The first component is $\hat{v}_q$, i.e., the relation between $\hat{v}$ and q, which is endogenous to the tax schedule. It mirrors that product choices of a consumer can be understood as a signal of her perceived valuation $\hat{v}$. The second component is the expected bias conditional on a type $\hat{v}$, $E[b|\hat{v}]$. It reflects that the policy maker faces a signal extraction problem. She receives a signal $\hat{v}$ from a consumer's position on the demand curve. Based on this signal, she makes inferences about $E[b|\hat{v}]$ in order to determine the optimal tax schedule. Such inferences are independent of the tax schedule and depend only on the joint distribution of valuations and biases in the population of consumers. In the following, we first formalize the signal extraction problem and explore how it relates to parameters of the joint distribution of $\hat{v}$ and b. We then derive how the shape of the optimal tax schedule can be inferred from these parameters.

[9] In the absence of income effects, constant Pareto weights $\alpha(\hat{v}) = \alpha$ imply that marginal social welfare weights g are equal to one (see, e.g., Piketty and Saez 2013).

We start by assuming that biases and valuations follow a joint normal distribution. This assumption allows us to describe the global relationship between b and $\hat{v}$ in terms of its first two moments. The conditional expectation $E[b|\hat{v}]$ is then given by:

$$E\left[b|\hat{v}\right] = E[b|\mu_{\hat{v}}] + \underbrace{\frac{(\sigma_b/\sigma_v)+\rho}{(\sigma_b/\sigma_v)+(\sigma_v/\sigma_b)+2\rho}}_{=:A} \cdot (\hat{v}-\mu_{\hat{v}})\,, \tag{5}$$

where the measure of bias heterogeneity, $A = \partial E[b|\hat{v}]/\partial\hat{v}$, is constant across $\hat{v}$. If valuations and biases are not jointly normal distributed, Equation (5) corresponds to a first-order approximation to the conditional expectation $E[b|\hat{v}]$. As shown in Appendix A.E, it minimizes the Mean Squared Error between b and its linear prediction $\tilde{b}$, $E(b-\tilde{b})^2$.

As described in Equation (5), the magnitude of updating depends crucially on two factors. First, it increases in the difference between a perceived valuation $\hat{v}$ and its expected value $\mu_{\hat{v}}$. Hence, the expected bias deviates most strongly from the mean bias for consumers with extreme perceived valuations.[10] Second, it increases in the absolute value of the local bias heterogeneity A, which reflects an information value of $\hat{v}$ about consumer biases. The information problem of the policy maker consists of decomposing the information contained in the "truth-plus-noise" signal $\hat{v}$ into its components $E\left[v|\hat{v}\right]$ and $E\left[b|\hat{v}\right]$. In this decomposition problem, the term A has the role of a decomposition weight. It captures how much of a one-unit increase in $\hat{v}$ can be attributed to a change in $E(b|\hat{v})$. The remaining part, $(1-A)$, is attributed to a change in $E(v|\hat{v})$.[11]

The local bias heterogeneity A measures the extent to which the average bias varies across consumers with different perceived valuations. In addition, it represents the degree to which a policy maker would optimally like to differentiate the tax schedule. Local bias heterogeneity depends only on two statistics of the joint distribution of v and b: their correlation coefficient ρ and the ratio σ_b/σ_v. When σ_b/σ_v is equal to one, variation in the perceived valuations is equally likely to stem from variation in the bias and in the valuation, which implies $A = 1-A = \frac{1}{2}$. When it is equal to zero, observing $\hat{v}$ is uninformative about the bias, but very informative about the valuation. The reverse holds true when σ_b/σ_v becomes infinitely large.

[10] As a consequence, the optimal non-linear tax implies "no distortion at the top and at the bottom" when the distribution functions F_v and F_b both have a bounded support and the condition for internality tax implementability (Proposition 2) is satisfied. In this case, an extreme signal resolves all uncertainty about the bias of the respective consumer, which allows the policy maker to fully correct for it.

[11] This can be seen from the first-order approximation for $E[v|\hat{v}]$, which is given by $\hat{E}\left[v|\hat{v}\right] = E[v|\mu_{\hat{v}}] + (1-A)\cdot(\hat{v}-\mu_{\hat{v}})$.

4.1 Curvature of the Optimal Non-Linear Tax Schedule

We continue by exploring the curvature of the optimal tax schedule, which is given by $t''(q) = \partial E[b|\hat{v}]/\partial\hat{v} \cdot \partial\hat{v}/\partial q$. As a consequence of incentive compatibility, an allocation must be non-decreasing in perceived valuations ($\partial\hat{v}/\partial q \geq 0$). Hence, the sign of $t''(q)$ is determined by the local bias heterogeneity $A = \partial E[b|\hat{v}]/\partial\hat{v}$. Following that insight, Proposition 4 characterizes the curvature of the optimal tax schedule in terms of the two determinants of A: the correlation between normative valuations and biases, ρ, and the ratio of their standard deviations σ_b/σ_v. We provide a graphical overview in Figure 1.

Proposition 4. *The curvature of the optimal tax schedule is determined by the correlation between normative valuations and biases, ρ, and the ratio of their standard deviations, σ_b/σ_v. The optimal tax is ...*

1. *... concave in q if and only if $A < 0 \Leftrightarrow \rho < -(\sigma_b/\sigma_v)$;*
2. *... linear in q if and only if $A = 0$, i.e., if either $(\sigma_b/\sigma_v) = 0$ or $\rho = -(\sigma_b/\sigma_v)$;*
3. *... convex in q if and only if $A \in (0,1) \Leftrightarrow \rho > -(\sigma_b/\sigma_v)$ and $\rho > -(\sigma_v/\sigma_b)$.*

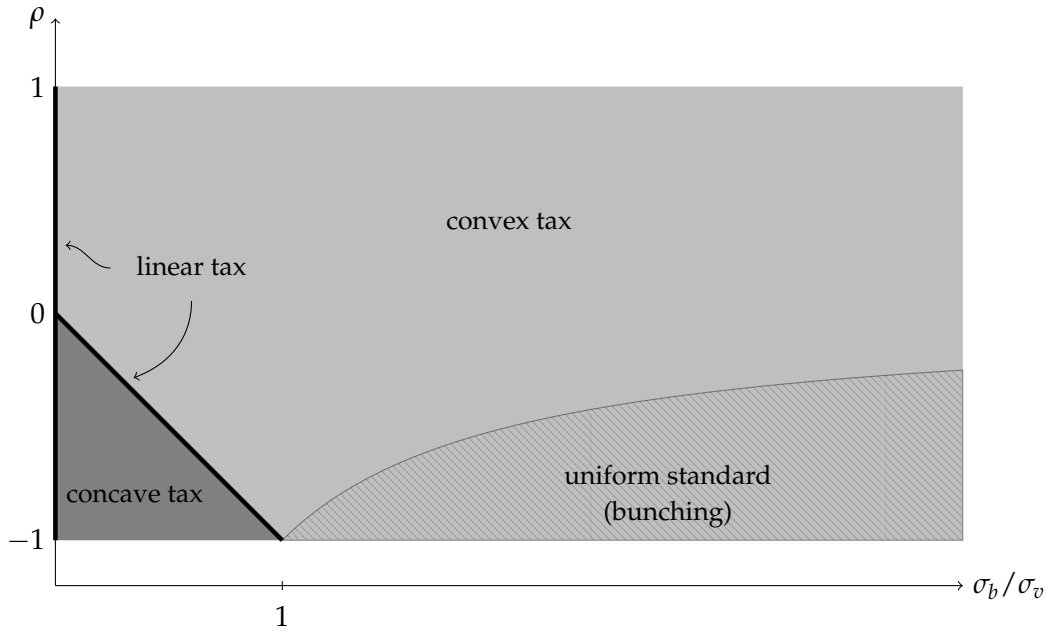
Proof. Follows from Proposition 1 (for $g = 1$) and Equation 5. □

Proposition 4 clarifies that a linear tax is optimal when local bias heterogeneity is zero ($A = 0$). In that case, nothing can be learned about the bias from obtaining information about consumers' perceived valuations $\hat{v}$, which eliminates the benefit from targeting consumers via non-linear taxes. If A is between zero and one, the expected bias increases in $\hat{v}$, which implies a convex optimal tax schedule. With a negative A, larger perceived valuations are associated with a lower expected bias and the optimal tax schedule is concave.

Furthermore, Proposition 4 has direct relevance for policy makers with qualitative information about the source of a bias. The curvature of the optimal tax schedule crucially depends on the sign of the correlation between valuations and biases, ρ. That sign can be derived from knowing the specific behavioral bias at work, even in the absence of data. For example, exogenous inattention implies that valuations are uncorrelated with biases ($\rho = 0$) and rational addiction models imply a positive correlation ($\rho > 0$).[12] As shown by Proposition 4 and Figure 1, the optimal tax schedule is convex in both cases (except for $\sigma_b = 0$, in which case it would be linear). The rationale for a convex schedule is that higher biases b translate into higher perceived valuations

[12] This follows from the fact that consumers with large valuations for an addictive good consume it more than consumers with low valuations.

Figure 1: Optimal Internality Taxation (see Proposition 4 and 5)



$\hat{v} = v + b$. Hence, the conditional expectation of the bias $E(b|\hat{v})$ and the optimal tax rate increases in $\hat{v}$ and q, respectively.

4.2 Taxation vs. Standards

We now turn to the question when behaviorally motivated bans of certain product attribute levels, i.e., product standards, are optimal. In our model, standards imply that consumers with different perceived preferences choose the same attribute level under the optimal tax schedule. Proposition 5 and Figure 1 show that a uniform standard is optimal when ρ is negative and sufficiently small.

Proposition 5 (Uniform Standard)**.** *The optimal tax corresponds to a uniform standard if and only if* $A > 1 \Leftrightarrow \rho < -(\sigma_v/\sigma_b)$.

Proof. Follows from Proposition 3 (Case 1) and Equation 5. □

In this case, there exists a large degree of local bias heterogeneity A, which the policy maker would like to correct. Yet, she cannot differentiate corrective taxes to the extent required because incentive compatibility would fail if she did. It fails because $A > 1$ implies a fundamental disagreement of preferences between the policy maker and the consumer, which cannot be corrected via a price instrument (Proposition 2). When A is constant, incentive compatibility fails globally for every $\hat{v}$. "Ironing" the optimal smooth tax schedule then results in a uniform standard (Proposition 3).

Next, we compare the welfare implications of linear taxes with those of a uniform standard. This analysis extends Weitzman's (1974) investigation of price and quantity regulation to a setting with behavioral consumers who choose among a variety of quality-differentiated products.

Proposition 6 (Linear Taxation vs. Uniform Standard)**.** *There exists an* $\hat{A} \in [0, 1]$ *such that for all* $A < \hat{A}$ $(A > \hat{A})$ *as defined in Equation* (5) *price regulation via linear taxes is welfare-superior (welfare-inferior) compared to regulation that sets a uniform standard.*

Proof. See Appendix A.F. □

Proposition 6 clarifies the rationale of a benevolent policy maker to employ either linear taxes or standards. Ultimately, the policy maker is interested in using the instrument that better approximates the slope of the expected bias function, which is given by the local bias heterogeneity A. When A is sufficiently small, a linear tax approximates the expected bias function better than a standard, while the reverse holds true for large values of A. This logic is similar to the rationale by Weitzman (1974), where a policy maker chooses the instrument that better approximates the marginal damage function from an externality.

Our analysis extends the behavioral analysis of linear taxes and quantity regulations by Farhi and Gabaix (2020). In a setting with uncorrelated normative valuations and biases, they show that quantity regulation yields a higher expected welfare than a linear tax if the variation in the bias is sufficiently large relative to the variation in normative valuations. In our model, this finding corresponds with the result that A increases in σ_b/σ_v, when ρ equals zero. Our approach demonstrates that this finding does not necessarily hold true when normative valuations are correlated with biases. When $\rho < 0$, an increase in σ_b/σ_v may reduce the local bias heterogeneity A and hence the expected welfare of a standard relative to a linear tax.[13] Furthermore, our results show when a uniform standard is welfare-superior to all linear and non-linear taxes rather than to linear taxes alone. As shown by Figure 1, a uniform standard is never welfare-superior to all price instruments when $\rho = 0$, which corresponds to the case investigated by Farhi and Gabaix (2020).

In the following, we analyze when a policy maker should optimally set minimum or maximum standards. So far, our analysis of optimal policies has relied on the conditional expected bias function from Equation 5, which corresponds to a first-order approximation of $E(b|\hat{v})$ with a constant local bias heterogeneity A. To assess minimum and maximum standards, we turn to a second-order approximation. This ap-

[13] For $\rho \geq 0$, we have that $\partial A/\partial\left(\sigma_b/\sigma_v\right) > 0$. For $\rho < 0$, we have that $\partial A/\partial\left(\sigma_b/\sigma_v\right) < 0$ if either $\sigma_b/\sigma_v < -\frac{1}{\rho} - \sqrt{\frac{1}{\rho^2} - 1} \leq 1$ or $\sigma_b/\sigma_v > -\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} - 1} \geq 1$. Furthermore, $\partial A/\partial\rho < 0$ if and only if $\sigma_b/\sigma_v > 1$.

proximation allows A to vary across perceived valuations $\hat{v}$. We use it to investigate how higher-order moments of the joint distribution of perceived valuations and biases affect the shape of the optimal non-linear tax schedule.

Proposition 7 (Minimum and Maximum Standards)**.** *The optimal non-linear tax schedule involves a maximum standard, i.e., a ban of high product varieties with $\hat{v} \in (\hat{v}^h, \bar{\hat{v}}]$ from the market, where $\hat{v}^h > \underline{\hat{v}}$, if:*

$$\frac{cosk\left(\hat{v}, b\right)}{sk\left(\hat{v}\right)} > \rho\left(\hat{v}, b\right) \qquad \text{and} \qquad sk(\hat{v}) > 0, \tag{6}$$

and it involves a minimum standard, i.e., a ban of low product varieties with $\hat{v} \in [\underline{\hat{v}}, \hat{v}^l]$ from the market, where $\hat{v}^l < \bar{\hat{v}}$, if:

$$\frac{cosk\left(\hat{v}, b\right)}{sk\left(\hat{v}\right)} > \rho\left(\hat{v}, b\right) \qquad \text{and} \qquad sk(\hat{v}) < \frac{cosk\left(\hat{v}, v\right)}{\rho\left(\hat{v}, v\right)} < k < 0, \tag{7}$$

where $sk(\cdot)$ denotes the skewness, $cosk(\hat{v}, y) = E\left((\hat{v} - \mu_{\hat{v}})^2 (y - \mu_y)\right) / (\sigma_{\hat{v}}^2 \sigma_y)$ is the coskewness of $\hat{v}$ and a random variable y, $\rho\left(\cdot\right)$ denotes the correlation, and $k := -2(\mu_{\hat{v}}/\sigma_{\hat{v}}) - \sigma_{\hat{v}^2}/\sigma_{\hat{v}}^2$ is a negative constant. Both conditions require that $[\underline{\hat{v}}, \bar{\hat{v}}) := [0, \infty)$ and that $A < 1$, so that bunching all individuals is not optimal (as stated in Proposition 5).

Proof. See Appendix A.G □

Proposition 7 states when a mix of taxation with bans for either high or low attribute levels is optimal. To provide intuition, let us first assume that perceived valuations $\hat{v}$ are positively skewed as required by the second of the two inequalities in (6). In this case, the policy maker observes some consumers with very high perceived valuations who would demand product varieties with a very high attribute level q in the absence of regulation. If such extreme preferences are mostly due to a large bias, a regulator should intervene to prevent welfare losses for these consumers. How extreme perceived valuations relate to the bias is captured by the term $cosk\left(\hat{v}, b\right) / sk\left(\hat{v}\right)$. A large coskewness implies that extreme values of $\hat{v}$ are associated with a large bias. Normalizing the coskewness by the skewness yields a measure of the extent to which extreme observations of $\hat{v}$ are explained by the bias.[14] If this term is sufficiently large, the policy maker should regulate choices of consumers with very high perceived valuations because they originate from high biases, as reflected by the first of the two inequalities in (6).

[14]The skewness of $\hat{v}$ can be expressed as a weighted mean of the coskewnesses: $sk(\hat{v}) = (\sigma_b/\sigma_{\hat{v}})cosk(\hat{v}, b) + (\sigma_v/\sigma_{\hat{v}})cosk(\hat{v}, v)$. As the weights are non-negative, $cosk(\hat{v}, b)/sk(\hat{v})$ measures the extent to which the coskewness between $\hat{v}$ and b explains the overall skewness of $\hat{v}$.

To regulate the choices of these consumers, the policy maker must rely on bans if tax instruments cannot be implemented due a violation of incentive compatibility. As shown in Proposition 2, implementability fails if higher perceived valuations are associated with lower normative valuations as assigned by the policy maker. This can occur locally for consumers with very high perceived valuations if their choices are much more informative about biases than the choices of the average consumer. The average informativeness of perceived valuations about biases is measured by $\rho(\hat{v}, b)$. If it is large, choices are very informative on average and local bunching becomes more unlikely, as shown by the right hand side of the first of the two inequalities in (6).[15]

The same rationale applies for a ban of low attribute levels. As described by Proposition 7, such a ban is part of the optimal non-linear tax if perceived valuations are negatively skewed and if the bias explains a sufficiently large share of the negative skew, as measured by the term $cosk(\hat{v}, b) / sk(\hat{v})$. As shown by the second of the inqualities in (7), the skew must be larger (in absolute terms) than a constant k and the (variance-weighted) coskewness of $\hat{v}$ and b.[16]

5 Optimal Internality Taxation in the Light Bulb Market

In this section, we illustrate our findings based on a numerical example from the light bulb market. First, we discuss consumer choices in the light bulb market and calibrate the joint distribution of perceived valuations and biases based on data elicited by Allcott and Taubinsky (2015a). Second, we derive the local bias heterogeneity A in our setting and discuss how it determines the shape of the optimal tax schedule. Third, we quantify the optimal tax schedule.

In the light bulb market, consumers face a trade-off between paying a higher purchase price for a more energy efficient bulb or bearing higher operating and replacement cost (ORC) for a less efficient one. The regulator is concerned that consumers are only imperfectly informed about and inattentive towards ORC savings at the time of the purchase, thereby underinvesting into energy efficiency. We measure the attribute level of a light bulb in terms of its energy efficiency. In particular, we define product attribute level q as the ORC savings relative to the most energy inefficient light bulb on the market over a time period of eight years.[17]

[15] Note that bunching over the entire support of $\hat{v}$ becomes more likely as $\rho(\hat{v}, b)$ increases.

[16] The latter follows from: $cosk(\hat{v}, v)/\rho(\hat{v}, v) = (\sigma_{\hat{v}} sk(\hat{v}) - \sigma_b cosk(\hat{v}, b)) / (\rho(\hat{v}, v)\sigma_v) < k < 0$.

[17] For simplicity, we assume that bulbs are identical in all other attribute level dimensions (e.g., luminosity and light color). In practice, consumers choose among four main light bulb technologies that differ in their energy efficiency: incandescent light bulbs, halogen bulbs, compact fluorescent bulbs (CFL), and LED bulbs. Within each category, small differences in energy efficiency exist (see Appendix Figure 2).

We start by discussing how information about the joint distribution of biases and valuations can be elicited in practice. For the purpose of quantifying biases, the literature has suggested the use of surveys to elicit experts' decision-making as a rational consumer benchmark (e.g., Handel and Kolstad 2015), to measure beliefs directly (e.g., Rees-Jones and Taubinsky 2019), and to debias consumers in survey-based experiments (e.g., Allcott and Taubinsky 2015a). Perceived preferences can also be collected in surveys or obtained from observational data through calibration so that they reflect the distribution of actual behaviour. As an alternative to survey-based approaches, the joint distribution of biases and preferences can be obtained from observational studies that use semi-parametric estimators for discrete choice models (see, e.g., Houde and Myers 2019).

In our example, we draw upon data by Allcott and Taubinsky (2015b) to quantify biases and valuations. After a baseline elicitation of the relative willingness-to-pay (WTP) for the more energy efficient CFL bulb, consumers in a treatment group obtain information on the lifetime costs of both bulbs. We use the difference in consumers' relative WTP before and after information provision as a measure of their bias. We divide this measure by the ORC difference between both bulbs, which yields a per unit bias b (in terms of q), whose distribution is presented in Appendix Figure 3a. We then construct a measure for the normative valuation v of one unit of ORC savings from individual-specific discount rates (see Appendix Section B for details and Appendix Figure 3b for the distribution). Knowledge of v and b allows us to determine the joint distribution of biases and perceived valuations by calculating $\hat{v}$ as the sum of v and b.

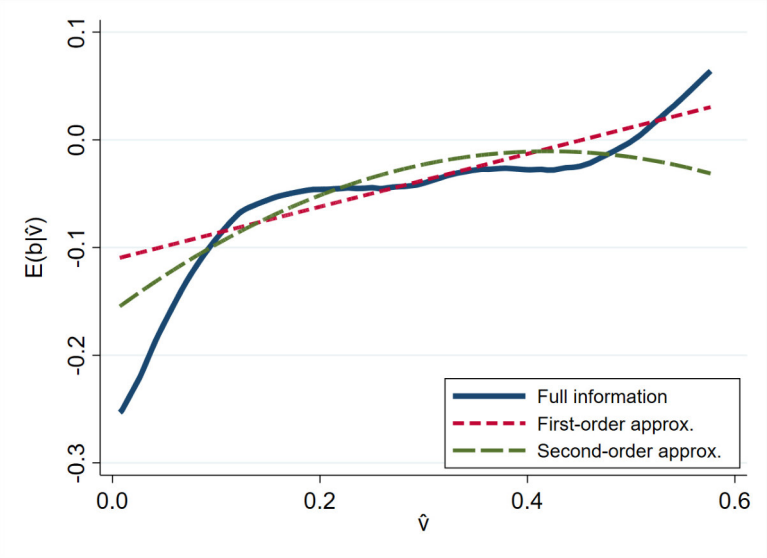
In a second step, we determine the conditional expectation of the bias, $E[b|\hat{v}]$, and the local bias heterogeneity $A(\hat{v}) = \partial E[b|\hat{v}]/\partial \hat{v}$ using three approximations of the conditional expectation of the bias.[18] First, a non-parametric approximation using local linear regressions that requires full knowledge of the joint distribution of $\hat{v}$ and b. Second, the first-order approximation from Equation (5), which requires only knowledge of the first two moments of the joint distribution and yields a constant A. Third, the second-order approximation from Proposition 7, which allows us to capture how A varies with $\hat{v}$ by additionally taking information about third moments into account.

We start by discussing the qualitative insights on the optimal tax schedule under full information. As shown by Figure 2a, the conditional expectation $E[b|\hat{v}]$ increases in perceived valuations $\hat{v}$. Hence, the optimal tax rate under a smooth tax schedule must be increasing in q, which implies a convex tax schedule. Yet, for low perceived valuations $\hat{v}$, the local bias heterogeneity A exceeds one. Hence, there is a fundamen-

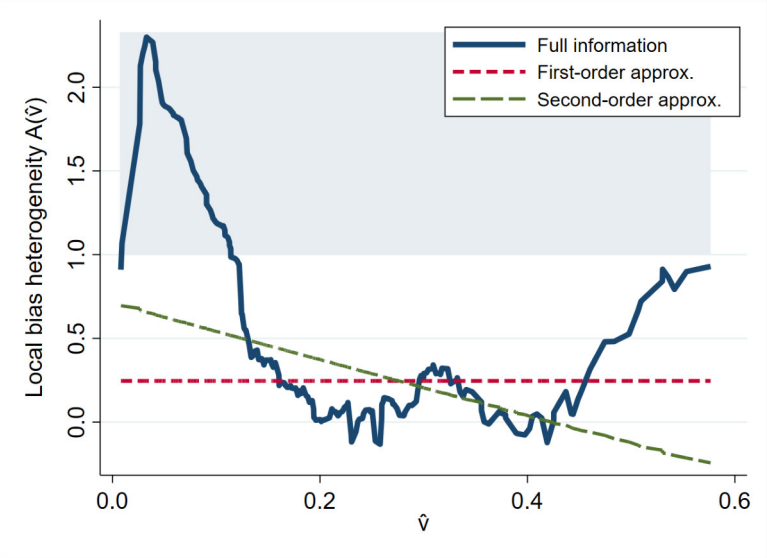
[18] We explore $E(b|\hat{v})$ rather than $E(b|q)$ for two reasons: First, incentive compatibility requires that q increases in $\hat{v}$. Hence, both functions carry the same qualitative information. Second, $E(b|q)$ is not directly related to our implementability condition from Proposition 2 and endogenous to the tax schedule.

Figure 2: Optimal Non-Linear Taxation in the Light Bulb Market

(a) Conditional expectation of the bias



(b) Local bias heterogeneity



(c) Summary Statistics

	Min	Max	μ	p50	σ	$\rho(.,b)$	Skew	Coskew$(\cdot,b)$
Biases b	-0.41	0.12	-0.05	-0.02	0.09	–	-1.53	–
Norm. valuations v	0.10	0.59	0.29	0.27	0.13	-0.38	0.35	-0.09
Perc. valuations $\hat{v}$	0.01	0.58	0.24	0.21	0.13	0.34	0.58	0.01

Notes for Figure a): The conditional expectation of the bias under full information is estimated via local linear regression (Epanechikov kernel, bandwidth: 0.056, respectively). The first- and second-order approximations minimize the Minimum Mean Squared Error (MMSE) between the bias and its linear and quadratic prediction, respectively.
Notes for Figure b): Under full information, the local bias heterogeneity $A(\hat{v})$ is calculated as the numerical gradient of the conditional expectation function from Figure 2a. The blue shaded area highlights values of $A(\hat{v})$ that exceed one and thus indicate a violation of Internality Tax Implementability (Proposition 2).

tal misalignment of preferences between the consumer and the policy maker and a smooth tax schedule does not satisfy incentive compatibility in this range. Consumers with low perceived valuations are thus bunched to ensure incentive compatibility of the tax schedule, which implies a ban of low attribute levels (Proposition 3). Under full information, the optimal policy will thus consist of a minimum standard, coupled with a convex tax schedule above the bunching range.

If only information about the first two moments of the joint distribution of b and $\hat{v}$ are used to determine the optimal non-linear tax, the properties of the optimal non-linear tax schedule follow from Proposition 4. In our data, the correlation between v and b is $\rho = -0.38$, and the standard deviation of b and v are $\sigma_b = 0.09$ and $\sigma_v = 0.13$, respectively (Figure 2c). These values confirm that the conditional expectation $E[b|\hat{v}]$ increases in $\hat{v}$ and the optimal corrective tax schedule is convex, as $\rho > -\sigma_b/\sigma_v$ (Proposition 4). As shown in Figure 2b, A estimated under the first-order approximation is constant and smaller than one. Hence, implementing the optimal non-linear tax based on that approximation would not lead to a violation of incentive compatibility and thus not require setting a standard.

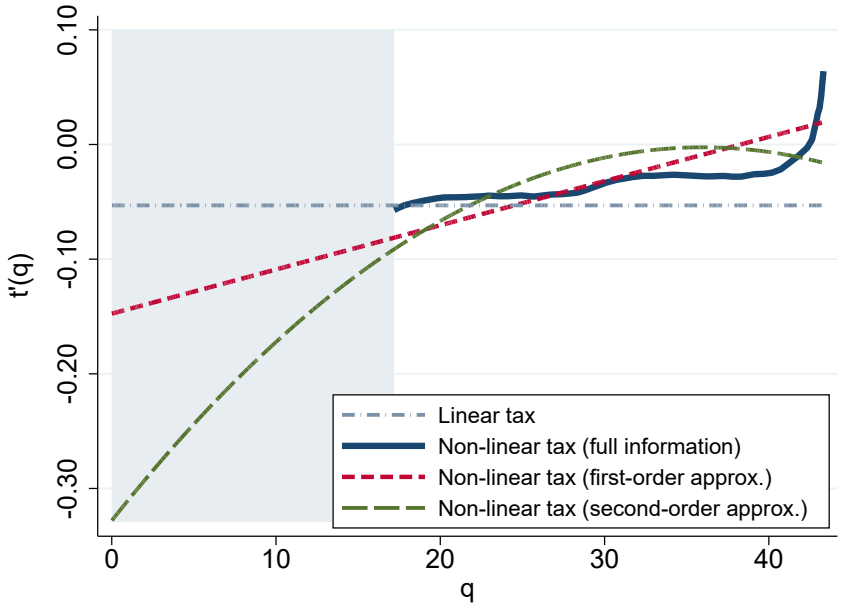
Next, we approximate the conditional expectation of the bias using a second-order approximation. Again, $A(\hat{v})$ is smaller than one for the entire support of perceived valuations. Hence, the implementation of the optimal non-linear tax under this approximation is feasible and a minimum standard is not necessary. This is also reflected by the fact that the sufficient conditions for a minimum standard from Proposition 7 are not fulfilled in our setting ($sk(\hat{v}) = 0.58$, $cosk(\hat{v}, b) = 0.01$, $\rho(\hat{v}, b) = 0.34$). Furthermore, Figure 2b demonstrates that A decreases in $\hat{v}$, which implies that the policy maker would like to differentiate taxes more strongly compared to the first-order approximation.

In a third step, we simulate the optimal non-linear tax in the light bulb market under either information requirement and compare it with the outcomes under the optimal linear tax. Following the optimal income taxation literature, we base our simulations on the distribution of types rather than consumption choices, which are endogenous to the tax system (see, e.g., Saez 2001 for a discussion). We estimate a quadratic cost function based on market data (for details, see Appendix Section B).

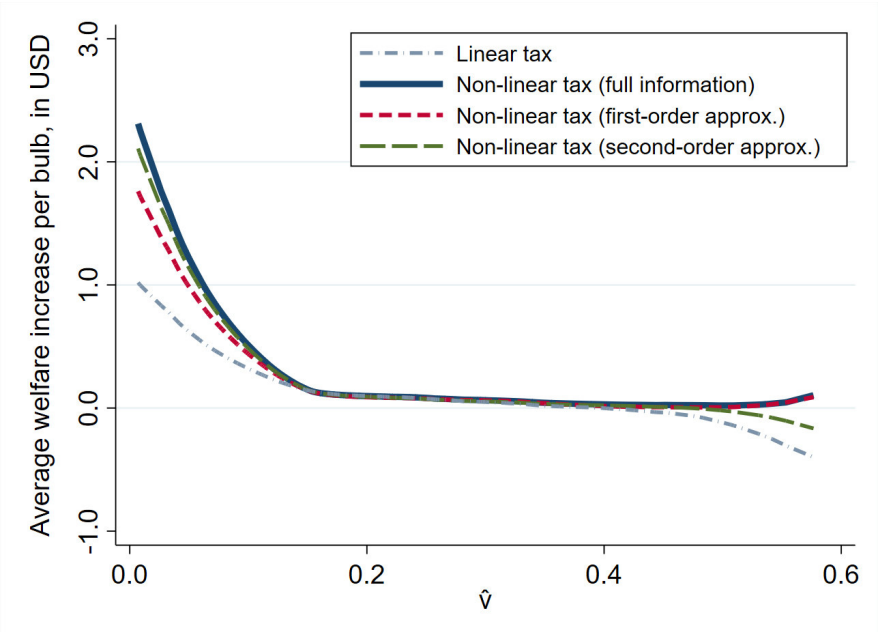
Based on the conditional expectation function of the bias and the cost function c, we then quantify the optimal tax schedule. To do so, we exploit that $t'(q) = E[b|\hat{v}_q]$ (Equation 3). For every consumer $\hat{v}$, we determine the marginal tax rate as $t'(q(\hat{v})) = E[b|\hat{v}]$, as well as his consumption choice when facing the tax rate. To obtain consumption choices, we exploit the first-order condition of consumers' utility maximization problem, which requires that $\hat{v} = c'(q(\hat{v})) + t'(q(\hat{v}))$. Because $t'(q(\hat{v})) = E[b|\hat{v}]$, consumers' consumption choices under the tax schedule $t(q)$ are equal to $q(\hat{v}) =$

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

(a) Optimal tax rates as a function of the attribute level q (ORC savings in USD)



(b) Average welfare increase per bulb as a function of perceived valuations $\hat{v}$.



Notes for Figure a): Optimal non-linear and linear tax rates are determined by using the formulas from Proposition 1 and Appendix Section A.B, respectively. The blue shaded area denotes the range of restricted product qualities that fall below the minimum standard for "Non-linear tax (full information)".
Notes for Figure b): The average welfare weight is smoothed using local linear regressions (Epanechikov kernel, bandwidth = 0.05)

$c'^{-1}(\hat{v} - E[b|\hat{v}])$. Knowledge of tax rates and consumption choices q for every consumer type $\hat{v}$ then allows us to express the tax rates as a function of q. When implementability of a smooth tax schedule fails, we additionally "iron" the tax schedule in order to determine the minimum standard, as explained in Appendix A.D.

The optimal tax rates are depicted in Figure 3a. As a reference, we first discuss the optimal linear tax. In this case, the optimal tax rate is constant and corresponds to a subsidy of 0.05 U.S. Dollar (USD) per USD of ORC savings over the lifetime of a light bulb.[19] The optimal non-linear tax under full information implies an increasing tax rate and hence a convex tax schedule. The tax rates are negative for bulbs with ORC savings smaller than about 42 USD and thus correspond to marginal subsidies for energy efficiency in that range. The marginal subsidies decrease in q, thereby reflecting that consumers who purchase products with a higher attribute level tend to have lower negative biases (Figure 2a). For individuals with perceived valuations of less than 18 USD of ORC savings, the regulator would like to differentiate the tax rates in a manner that is not consistent with incentive compatibility. Hence, the optimal non-linear tax results in bunching at the bottom and a minimum standard that forbids product qualities with ORC savings of less than 18 USD, as depicted by blue shaded area. For high values of ORC savings above about 42 USD, an increase in energy efficiency is taxed at the margin. This is optimal as the individuals who are indifferent between very energy efficient product varieties overvalue energy efficiency on average (Figure 2a).

The optimal tax rates based on a first-order approximation increase in q, which again reflects a convex tax schedule. The subsidization of energy efficiency at the margin amounts to 0.15 USD per unit of ORC savings for product with the lowest attribute level and thus exceeds the respective rate under the linear tax. Furthermore, marginal subsidies decrease in q and eventually become a marginal tax for product varieties with the highest attribute level, thereby reflecting that consumers with demand for a higher product attribute level have a lower bias in absolute terms. As the optimal tax schedule does not violate incentive compatibility, it does not require setting a standard and the marginal tax rates span the entire support of q. The optimal tax rates based on the second-order approximation are qualitatively very similar, yet they are differentiated more strongly. In this case, marginal subsidies amount up to 0.30 USD per unit of ORC savings for low energy efficiency levels.

Figure 3b displays the welfare increase of the optimal linear tax, the optimal non-linear tax under full information, as well as the first- and second-order approximation,

[19] Assuming for illustration that the subsidy level for the product with the lowest energy efficiency is zero, the total subsidy payment for a bulb with the maximum energy efficiency of about 50 USD in ORC savings would equal about $0.05 \times 50 = 2.5$ USD.

each compared to no taxation. It shows that linear taxes increase welfare for individuals with low and medium perceived valuations, but decrease it for individuals with high valuations. Non-linear taxes are better suited to target individuals. They more than double the welfare gain for individuals with low perceived valuations and reduce welfare losses for individuals with high perceived valuations. On aggregate, the welfare gain over the status quo of no subsidy amounts to 0.12 USD per bulb for the optimal linear tax, compared to 0.16 and 0.19 USD per bulb for the optimal non-linear tax under its first-order approximation and full information, respectively (Appendix Table 1). In relative terms, implementing the non-linear schedule rather than the linear tax increases welfare by 36% under a first-order approxmation and by 57% under full information.

Hence, a first-order approximation with modest information requirements is able to reap a substantial part of the welfare gains of non-linear taxation. Furthermore, it does not involve restrictions on consumer choice, which might increase its political acceptability. Our empirical example also uncovers an apparent paradox: when information about the joint distribution of biases and valuations is perfectly known, a policy maker may find it optimal to set standards, but not when information availability is restricted. This is because more information may imply that the policy maker would like to differentiate tax rates more strongly than she can without violating incentive compatibility. In such a case, price instruments are infeasible and the policy maker must resort to standards.

6 Discussion and Conclusion

In this paper, we derive the optimal non-linear tax to correct behaviorally biased consumers. We show that consumers' consumption choices contain information about their biases, which can be exploited for targeting via a non-linear internality tax. We also demonstrate how a policy maker can make use of qualitative insights about the underlying behavioral bias to determine whether the optimal non-linear tax is convex, linear, or concave.

Our results also provide a novel rationale for behaviorally motivated standards. When choosing between levying a tax and setting a standard, a policy maker should use the instrument that better approximates the gradient of the expected bias function, which we denote as the local bias heterogeneity. It captures the informativeness of differences in perceived valuations about differences in the average bias and thus measures the degree to which a benevolent policy maker should differentiate tax rates. Yet, tax rates that increase too steeply in the attribute levels reflect a fundamental misalignment of preferences between the policy maker and the consumer. In this case,

the tax schedule cannot be implemented because it violates incentive compatibility and the policy maker must optimally resort to standards. This rationale results in a uniform standard when the fundamental misalignment is global, and to minimum and maximum standards, when it arises at the bounds of the distribution of perceived valuations.

In a numerical example from the light bulb market, we demonstrate that optimal non-linear taxation increases welfare beyond the optimal linear tax. Furthermore, we illustrate that the informational requirements for implementing a first-order approximation of the optimal non-linear tax are modest and require knowledge of only a few statistics of the joint distribution of valuations and biases.

In practice, non-linear tax schedules can be implemented based on existing attribute level measures such as the energy efficiency ratings used for energy performance certificates and energy labels, for example. In the context of energy efficiency investments and hybrid car purchases, previous studies have shown that perceived valuations are positively correlated with the bias (Allcott et al., 2015). In such cases, the optimal non-linear tax schedule is convex, giving the largest marginal subsidies to participants with low perceived valuations. This optimal tax schedule contrasts with many subsidy schedules employed in practice. For example, the German government grants subsidies for energy efficiency in housing only if a newly built (or retrofitted) house meets predefined minimum efficiency levels. In other words, marginal subsidies are essentially zero for consumers with low perceived valuations. Hence, the most heavily biased consumers are not subsidized at the margin. Optimal non-linear tax and subsidy schedules avoid such shortcomings.

For the application of our results to different contexts, we see at least three valuable model extensions. First, the assumption of unit demand could be relaxed to explore how the availability of extensive margin effects changes the rationale of a policy maker to impose minimum and maximum standards. Second, in some settings, consumers may choose both the attribute level and the quantity of a product. Hence, extending our model to include both choice margins would be useful. Third, when analyzing large-scale internality tax policies, allowing for income effects and the interplay with other redistributive policies such as income taxes is important. These extensions may prove useful to establish solid theoretical (and empirical) foundations for the use of behaviorally motivated taxation and product regulation in practice.

References

Abaluck, Jason, and Jonathan Gruber. 2011. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program." *American Economic Review* 101 (4): 1180–1210.

Allcott, Hunt, Christopher Knittel, and Dmitry Taubinsky. 2015. "Tagging and Targeting of Energy Efficiency Subsidies." *American Economic Review* 105 (5): 187–191.

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

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

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

Allcott, Hunt, and Dmitry Taubinsky. 2015b. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market: Dataset. TESS3 125 Allcott_Fielding2_Weighted Data_111813." 10.1257/aer.20131564.

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

Attari, Shahzeen Z., Michael L. DeKay, Cliff I. Davidson, and Wändi Bruine de Bruin. 2010. "Public Perceptions of Energy Consumption and Savings." *Proceedings of the National Academy of Sciences of the United States of America* 107 (37): 16054–16059.

Bergemann, Dirk, Benjamin Brooks, and Stephen Morris. 2015. "The Limits of Price Discrimination." *American Economic Review* 105 (3): 921–957.

Bernheim, B Douglas, and Dmitry Taubinsky. 2018. "Behavioral Public Economics." *Handbook of Behavioral Economics: Applications and Foundations 1* 1 381–516.

BMF. 2019. "8. EKF-Bericht: Bericht des Bundesministeriums der Finanzen über die Tätigkeit des Energie- und Klimafonds (EKF; Kap. 6092) im Jahr 2018 und über die im Jahr 2019 zu erwartende Einnahmen- und Ausgabenentwicklung." `https://www.bundesfinanzministerium.de/Content/DE/Standardartikel/Themen/Oeffentliche_Finanzen/Bundeshaushalt/Energie-und-Klimafond/2019-05-27-EKF-Bericht-2019-download.pdf`, Bundesministerium der Finanzen.

Bohrnstedt, George W., and Arthur S. Goldberger. 1969. "On the Exact Covariance of Products of Random Variables." *Journal of the American Statistical Association* 64 (328): 1439–1442.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen. 2011. "Calorie Posting in Chain Restaurants." *American Economic Journal: Economic Policy* 3 (1): 91–128.

Carlsson, Fredrik, and Olof Johansson-Stenman. 2019. "Optimal Prosocial Nudging." *Working Papers in Economics No. 757, University of Gothenburg, Department of Economics*.

Chaloupka, Frank J., Lisa M. Powell, and Kenneth E. Warner. 2019. "The Use of Excise Taxes to Reduce Tobacco, Alcohol, and Sugary Beverage Consumption." *Annual Review of Public Health* 40 (1): 187–201.

Diamond, Peter A. 1998. "Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates." *American Economic Review* 88 (1): 83–95.

DOE. 2017. "Saving Energy and Money with Appliance and Equipment Standards in the United States." `https://www.energy.gov/sites/prod/files/2017/01/f34/Appliance%20and%20Equipment%20Standards%20Fact%20Sheet-011917_0.pdf`, U.S. Department of Energy (DOE).

EIA. 2018. "Annual Electric Power Industry Report, Form EIA-861 detailed data files." `https://www.eia.gov/electricity/data/eia861/zip/f8612018.zip`, U.S. Energy Information Administration (EIA).

European Commission. 2015. "Trends in Energy Consumption and Energy Efficiency 2000 - 2012." JRC Science and Policy Report Nr. JRC95238.

Farhi, Emmanuel, and Xavier Gabaix. 2020. "Optimal Taxation with Behavioral Agents." *American Economic Review* 110 (1): 298–336.

Ferey, Antoine, Benjamin Lockwood, and Dmitry Taubinsky. 2021. ""Sufficient Statistics for Nonlinear Tax Systems with General Across-Income Heterogeneity"." Working Paper 29582, National Bureau of Economic Research. 10.3386/w29582.

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

Gibbard, Allan. 1973. "Manipulation of Voting Schemes: A General Result." *Econometrica* 41 (4): 587–601.

Glaeser, Edward L., and Andrei Shleifer. 2001. "A Reason for Quantity Regulation." *American Economic Review* 91 (2): 431–435.

Guesnerie, Roger, and Jean-Jacques Laffont. 1984. "A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm." *Journal of Public Economics* 25 (3): 329–369.

Handel, Benjamin R, and Jonathan T Kolstad. 2015. "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare." *American Economic Review* 105 (8): 2449–2500.

Houde, Sébastien, and Erica Myers. 2019. "Heterogeneous (Mis-) Perceptions of Energy Costs: Implications for Measurement and Policy Design." Working Paper 25722, National Bureau of Economic Research.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *Quarterly Journal of Economics* 125 (2): 515–548.

Johnson, Justin P., and David P. Myatt. 2003. "Multiproduct Quality Competition: Fighting Brands and Product Line Pruning ." *American Economic Review* 93 (3): 748–774. 10.1257/000282803322157070.

Kraus, Alan, and Robert H Litzenberger. 1976. "Skewness Preference and the Valuation of Risk Assets." *The Journal of Finance* 31 (4): 1085–1100.

Lockwood, Benjamin B. 2020. "Optimal Income Taxation with Present Bias." *American Economic Journal: Economic Policy* 12 (4): 298–327.

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

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

Mussa, Michael, and Sherwin Rosen. 1978. "Monopoly and Product Quality." *Journal of Economic Theory* 18 (2): 301–317.

Myerson, Roger B. 1981. "Optimal Auction Design." *Mathematics of Operation Research* 6 (1): 58–73.

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review* 89 (1): 103–124.

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

Piketty, T, and E Saez. 2013. "Optimal Income Labor Taxation." *Handbook of Public Economics* 5 391–474.

Rees-Jones, Alex, and Dmitry Taubinsky. 2019. "Measuring "Schmeduling"." *The Review of Economic Studies* 87 (5): 2399–2438.

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

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

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

Online Appendix

Optimal Internality Taxation of Product Attributes

Andreas Gerster and Michael Kramm

A Proofs

A.A Proof of Proposition 1: Optimal Non-Linear Tax

We first use the mechanism design approach to characterize the optimal tax in terms of types and then employ the perturbation approach to characterize the optimal tax as a function of the attribute level. Finally, we show that the results we obtain under either approach are equivalent.

A.A.1 Equation (2) in Types (Mechanism Design Approach)

For our mechanism design approach, we employ the Revelation Principle for dominant-strategy implementation (Gibbard, 1973) to solve for a direct mechanism, where the consumer truthfully reveals information about his perceived valuation.

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. Even though a consumer may or may not know his bias, i.e., be sophisticated or naive, we can, without loss of generality, neglect that distinction and restrict our analysis to one-dimensional mechanisms in the perceived valuation $\hat{v}$. Naive consumers are unaware of their bias and thus cannot report it, so that a social planner can only employ a one-dimensional mechanism in $\hat{v}$ to correct them. Sophisticated consumers know their bias and can, in principle, report it. Yet, as biases do not influence decision utility, truth-telling in a two-dimensional mechanism is not incentive-compatible.

The formal argument of the above is as follows. We want to show that two-dimensional mechanisms, where sophisticated consumers report both their perceived valuation and their bias, lead to a violation of truth-telling. 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 policy maker 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 $\hat{v}_1$, truth-telling is violated.

Hence, we can apply 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 policy maker confines herself to designing a direct mechanism $(\xi,\tau) : \left[\underline{\hat{v}},\overline{\hat{v}}\right] \to Q \times \mathbb{R}$ under truth-telling in order to implement the welfare maximizing outcome. Based on the consumer's strategical report $\tilde{v}$, the allocation rule of the direct mechanism assigns the consumed attribute level, $\xi(\tilde{v}) \in Q$, and transfer rule the amount of taxes to be paid, $\tau(\tilde{v}) \in \mathbb{R}$. Because the consumer has unit demand for the good, participation constraints are not relevant in our setting.

Under the direct mechanism, the decision utility for report $\tilde{v}$ for a given 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, truth-telling must be ensured by implementing an incentive compatible mechanism. This implies that the tax schedule must satisfy:

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

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:

$$\hat{v}\xi'(v^*) - \tau'(v^*) - \xi'(v^*)c'(\xi(v^*)) \stackrel{!}{=} 0. \tag{8}$$

As incentive compatibility requires 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 normative utility is given by $\hat{u}^n(\hat{v},b) := u^n(\xi(\hat{v}),\tau(\hat{v})|v) = \hat{u}^d(\hat{v}) - b\xi(\hat{v})$.

Put differently, incentive compatibility implies that, for all $\hat{v} \in \left[\underline{\hat{v}}, \overline{\hat{v}}\right]$, the following equation has to hold:

$$\frac{\partial \hat{u}^d(\hat{v})}{\partial \hat{v}} = \xi(\hat{v}) + \hat{v}\xi'(\hat{v}) - \tau'(\hat{v}) - \xi'(\hat{v})c'(\xi(\hat{v})) \overset{(8)}{=} \xi(\hat{v}). \tag{9}$$

To determine the optimal tax schedule, the policy maker solves an 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 policy maker is given by:

$$\max_{\xi \in \mathbb{Q}} \int_{\hat{v}} \alpha(\hat{v}) \cdot E\left[\hat{u}^n\left(\hat{v}, b\right) | \hat{v}\right] dF(\hat{v}) + \lambda \left(\int_{\hat{v}} \tau\left(\hat{v}\right) dF(\hat{v}) - B \right), \tag{10}$$

subject to the condition from Equation (9), where $\mathbb{Q} := \{f | f : \left[\underline{\hat{v}}, \overline{\hat{v}}\right] \to Q\}$ is the function space containing all functions with domain $\left[\underline{\hat{v}}, \overline{\hat{v}}\right]$ and codomain Q. 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 given by Equation (9).

Using the definition of decision utility to replace the tax and rewriting equilibrium normative utility in terms of equilibrium decision utility, the Hamiltonian for the problem stated in Equation (10) for all $\hat{v} \in \left[\underline{\hat{v}}, \overline{\hat{v}}\right]$ is given by:

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

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.[20]

$$\text{FOC on control: } \frac{\partial \mathfrak{H}}{\partial \xi} = \left[-E[b|\hat{v}] \cdot \alpha(\hat{v}) + \lambda\left(\hat{v} - c'(\cdot)\right)\right] f(\hat{v}) + \mu\left(\hat{v}\right) \overset{!}{=} 0, \quad (\text{FOC}_\text{q})$$

$$\text{FOC on state: } \frac{\partial \mathfrak{H}}{\partial \hat{u}^d} = \left[\alpha(\hat{v}) - \lambda\right] f(\hat{v}) \overset{!}{=} -\mu'\left(\hat{v}\right), \quad (\text{FOC}_\text{u})$$

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

[20] 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.

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

$$\left.\frac{\partial u^d(q,t,\hat{v})}{\partial q}\right|_{q=q^d} = \hat{v} - c'(q^d) - t'(q^d) \stackrel{!}{=} 0 \Leftrightarrow c'(q^d) = \hat{v} - t'(q^d). \tag{11}$$

The second order condition is satisfied if $c''(q) + t''(q) \geq 0$ for all $q \in Q$. Since the costs are convex in q by assumption, this condition is satisfied if the optimal tax schedule is convex in q as well. We find that convex optimal tax schedules are optimal for many behavioral biases (see discussion after Proposition 4). More generally, this condition also holds if the optimal tax schedule is concave. In that case, the cost function may not be "too concave" in the sense that $c''(q) + t''(q) \geq 0$ holds.

In our setting, we abstract from participation constraints. Hence, we can w.l.o.g. assume 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 $\mu\left(\overline{\hat{v}}\right) = 0$. Integrating Equation ($\text{FOC}_\text{u}$) and using $\mu\left(\overline{\hat{v}}\right) = 0$, we obtain

$$\int_{\hat{v}}^{\overline{\hat{v}}} -\mu'(n)dn = -\mu\left(\overline{\hat{v}}\right) - \left[-\mu\left(\hat{v}\right)\right] = \mu\left(\hat{v}\right) \stackrel{!}{=} \int_{\hat{v}}^{\overline{\hat{v}}} \left[\alpha(m) - \lambda\right] f(m)dm. \tag{12}$$

Using the Equations (11) and (12), we rearrange Equation (FOC_q), to obtain the result:

$$\begin{aligned}
\lambda\left(\hat{v} - c'(\cdot)\right) &\stackrel{!}{=} -\frac{\mu\left(\hat{v}\right)}{f(\hat{v})} + E[b|\hat{v}] \cdot \alpha(\hat{v}) \\
\stackrel{(11)}{\Leftrightarrow} \quad t'(q) &= -\frac{\mu\left(\hat{v}_q\right)}{\lambda f(\hat{v}_q)} + E[b|\hat{v}_q] \cdot \frac{\alpha(\hat{v}_q)}{\lambda} \\
\stackrel{(12)}{\Leftrightarrow} \quad t'(q) &= \frac{\int_{\hat{v}_q}^{\overline{\hat{v}}} \left[1 - \hat{g}(m)\right] f(m)dm}{f(\hat{v}_q)} + E[b|\hat{v}_q] \cdot \hat{g}(\hat{v}_q).
\end{aligned}$$

The average marginal social welfare weight for consumers in the upper part of the distribution (in comparison to $\hat{v}$) is defined as:

$$\hat{G}(\hat{v}) := \frac{\int_{\hat{v}}^{\overline{\hat{v}}} \hat{g}(m)dF(m)}{1 - F(\hat{v})} = E[g(\theta)|\theta \geq \hat{v}]. \tag{13}$$

Using this we get

$$\begin{aligned}
t'(q_{\hat{v}}) &= \hat{g}(\hat{v})E\left[b|\hat{v}\right] + \frac{\int_{\hat{v}}^{\overline{\hat{v}}} \left[1 - \hat{g}(n)\right] f(n)dn}{f(\hat{v})} \\
&= \hat{g}(\hat{v})E\left[b|\hat{v}\right] + \frac{\int_{\hat{v}}^{\overline{\hat{v}}} f(n)dn - \int_{\hat{v}}^{\overline{\hat{v}}} \hat{g}(n)f(n)dn}{f(\hat{v})}
\end{aligned}$$

$$\overset{(13)}{=} \hat{g}(\hat{v})E\left[b|\hat{v}\right] + \frac{\left[1-F(\hat{v})\right] - \hat{G}(\hat{v})\left[1-F(\hat{v})\right]}{f(\hat{v})}$$
$$= \hat{g}(\hat{v})E\left[b|\hat{v}\right] + \frac{\left[1-F(\hat{v})\right]\left[1-\hat{G}(\hat{v})\right]}{f(\hat{v})}$$

A.A.2 Equation (3) in Attribute Level q (Perturbation Approach)

The perturbation approach uses insights from a local perturbation of the tax schedule to derive the optimal tax. In particular, it examines a change in the slope of the consumer's budget set in a small band $(q, q+dq)$ from $1-c'(q)-t'(q)$ to $1-c'(q)-t'(q)-d\tau$. Such a change has four effects on welfare. The mechanical effect represents an increase in collected tax money, since all consumers with consumption levels higher than $q+dq$ are affected by the increased tax rate. This effect is captured by

$$d\tau dq\left[1-H(q)\right].$$

The consumer welfare effect captures the social value of the decrease in consumption of the numeraire due to the local tax perturbation:

$$-d\tau dq\left[1-H(q)\right]G(q).$$

Since we abstract from income effects, there is no change of the consumption of q for these consumers. This is different for the consumers inside the small band $(q, q+dq)$. Here, relative prices are changed and thus consumers change their consumption of the attribute level by

$$\delta q = \frac{dq}{d\left(c'+t'\right)}d\tau = -e\frac{q}{c'+t'}d\tau. \tag{14}$$

where the elasticity is defined as

$$e := -\frac{dq}{d\left(t'+c'\right)}\frac{t'+c'}{q}. \tag{15}$$

We define the elasticity at the net price $t'+c'$, since the slope of the consumer's budget constraint is determined by $t'+c'$.

The fiscal externality captures the decrease in collected taxes

$$\delta q t'(q)h(q)dq,$$

while the bias correction effect

$$-g(q)\delta q E\left[b|q\right]h(q)dq$$

captures the change in welfare due to the induced change in consumer utility, which is caused by the potential correction of an internality.

In optimum, a local perturbation of the tax may not change social welfare. Hence, the four effects have to sum to zero, so that

$$\begin{aligned}
0 &= d\tau dq\left[1-H(q)\right]\left[1-G(q)\right]+\delta q\left[t'(q)-E\left[b|q\right]g(q)\right]h(q)dq \\
\Leftrightarrow t'(q) &= g(q)E\left[b|q\right]-\frac{d\tau\left[1-H(q)\right]\left[1-G(q)\right]}{\delta q h(q)} \\
&\overset{(14)}{=} g(q)E\left[b|q\right]-\frac{d\tau\left[1-H(q)\right]\left[1-G(q)\right]}{-e(q)\frac{q}{t'(q)+c'(q)}d\tau h(q)} \\
&= g(q)E\left[b|q\right]+\frac{1-G(q)}{e(q)a(q)}\left(t'(q)+c'(q)\right).
\end{aligned}$$

A.A.3 Equivalence of Equations (2) and (3)

We first linearize the budget constraint of the consumer in the tax so that, using the virtual income R, it can be rewritten as $z = R - c(q) - \tau q$. Then, the decision utility is given by $u = \hat{v}q + R - c(q) - \tau q$.

The goal is now to write the change of consumption q in type $\hat{v}$ using elasticity e, which is the elasticity of q with respect to net-price $\tau + c'$. The first-order condition of the consumer with respect to consumption gives

$$0 \overset{!}{=} \hat{v} - \tau - c'(q). \tag{16}$$

Applying the implicit function theorem twice to (16) yields

$$\begin{aligned}
\frac{dq}{d\hat{v}} &= -\frac{1}{-c''(q)} = \frac{1}{c''(q)} \\
\frac{dq}{d\left(\tau+c'\right)} &= -\frac{-1}{-c''(q)} = -\frac{1}{c''(q)} \\
\Rightarrow \frac{dq}{d\hat{v}} &= -\frac{dq}{d\left(\tau+c'\right)}.
\end{aligned} \tag{17}$$

We use the above results and the definition of the elasticity from (15):

$$e = -\frac{dq}{d\left(\tau+c'\right)}\frac{\tau+c'}{q}$$

$$\overset{(17)}{=} \frac{dq}{d\hat{v}} \frac{\tau + c'}{q}$$
$$\Leftrightarrow \frac{dq}{d\hat{v}} = \frac{qe}{\tau + c'}. \tag{18}$$

For density $h(q)$ the following must hold

$$h(q_{\hat{v}})dq_{\hat{v}} = f(\hat{v})d\hat{v}$$
$$\Leftrightarrow h(q_{\hat{v}})\frac{dq(\hat{v})}{d\hat{v}} = f(\hat{v})$$
$$\overset{(18)}{\Leftrightarrow} h(q_{\hat{v}})\frac{qe}{t' + c'} = f(\hat{v}). \tag{19}$$

The average marginal social welfare weight for consumers in the upper part of the distribution (in comparison to q) can be rewritten using a change of variables (Ch.o.V.):

$$\hat{G}(\hat{v}) := \frac{\int_{\hat{v}}^{\overline{\hat{v}}} \hat{g}(m)dF(m)}{1 - F(\hat{v})} \overset{Ch.o.V}{=} \frac{\int_{q_{\hat{v}}}^{\overline{q}} g(m)dH(q_m)}{1 - H(q_{\hat{v}})} =: G(q_{\hat{v}}). \tag{20}$$

The thinness of the top tail (individuals with consumption above q in relation to those at q) is given by

$$a(q) := \frac{h(q)q}{1 - H(q)}. \tag{21}$$

Using the above results we get

$$t'(q_{\hat{v}}) = \hat{g}(\hat{v})E\left[b|\hat{v}\right] + \frac{\left[1 - F(\hat{v})\right]\left[1 - \hat{G}(\hat{v})\right]}{f(\hat{v})}$$
$$\overset{(19),C.o.V.}{=} g(q)E\left[b|q\right] + \frac{\left[1 - H(q)\right]\left[1 - G(q)\right]}{h(q)\frac{qe(q)}{t'(q) + c'(q)}}$$
$$= g(q)E\left[b|q\right] + \frac{\left(t'(q) + c'(q)\right)\left[1 - H(q)\right]\left[1 - G(q)\right]}{h(q)qe(q)}$$
$$\Leftrightarrow t'(q) \overset{(21)}{=} g(q)E\left[b|q\right] + \frac{1 - G(q)}{a(q)e(q)}\left(t'(q) + c'(q)\right).$$

A.B Derivation of the Optimal Linear Tax

The policy maker's problem of setting the optimal linear tax can be written as

$$\max_{t \in \mathbb{R}} \int_v \int_b \alpha(v,b)u^n(q^d, t, v)dF_b(b|v)dF_v(v) + \lambda\left(\int_v \int_b t \cdot q^d dF_b(b|v)dF_v(v) - B\right) =: V(t).$$

We evaluate the derivative with respect to the linear tax t:

$$\frac{\partial V(t)}{\partial t} = \int_v \int_b \alpha(v,b) \left[-q^d + \big(v - t - c'(\cdot)\big) \frac{\partial q^d}{\partial t} \right] dF_b(b|v) dF_v(v) + \lambda \int_v \int_b \left[q^d + t \cdot \frac{\partial q^d}{\partial t} \right] dF_b(b|v) dF_v(v)$$
$$= \int_v \int_b \left[\alpha(v,b) \big(v - c'(\cdot)\big) \frac{\partial q^d}{\partial t} - (\alpha(v,b) - \lambda)(q^d + t \frac{\partial q^d}{\partial t}) \right] dF_b(b|v) dF_v(v).$$

The individually optimal consumption is again characterized by Equation (11), i.e., $c'(\cdot) = \hat{v} - t'(q) = (v + b) - t$, where the last equality holds since t is linear. Thus,

$$\frac{\partial V}{\partial t} = \int_v \int_b \left[\alpha(v,b)(t - b) \frac{\partial q^d}{\partial t} - (\alpha(v,b) - \lambda)(q^d + t \frac{\partial q^d}{\partial t}) \right] dF_b(b|v) dF_v(v)$$
$$= \int_v \int_b \left[(\lambda t - \alpha(v,b) b) \frac{\partial q^d}{\partial t} - (\alpha(v,b) - \lambda) q^d \right] dF_b(b|v) dF_v(v).$$

Using that t is constant, we can rewrite the equation as follows:

$$\frac{\partial V}{\partial t} = \lambda t \frac{\partial \bar{q}^d}{\partial t} - \int_v \int_b \left[\alpha(v,b) b \frac{\partial q^d}{\partial t} + (\alpha(v,b) - \lambda) q^d \right] dF_b(b|v) dF_v(v),$$

where the change in average demand $\bar{q}^d$ in response to a tax increase is given by $\frac{\partial \bar{q}^d}{\partial t} = \int_v \int_b \left[\frac{\partial q^d}{\partial t} \right] dF_b(b|v) dF_v(v)$. The optimal tax t^* is given by $\frac{\partial V}{\partial t}|_{t=t^*} \stackrel{!}{=} 0$, which gives:

$$\frac{\partial V}{\partial t} = t \frac{\partial \bar{q}^d}{\partial t} - \int_v \int_b \left[\hat{g}(v,b) b \frac{\partial q^d}{\partial t} - (1 - \hat{g}(v,b)) q^d \right] dF_b(b|v) dF_v(v) = 0$$

Abstracting from redistributive motives ($\hat{g}(v,b) = 1$ for all (v,b)) implies

$$t^* = \int_v \int_b b \left[\frac{\partial q^d}{\partial t} / \frac{\partial \bar{q}^d}{\partial t} \right] dF_b(b|v) dF_v(v), \tag{22}$$

where $\frac{\partial q^d}{\partial t} / \frac{\partial \bar{q}^d}{\partial t}$ denotes the relative responsiveness of a consumer type (v,b), i.e., the change in demand for that consumer type in response to a tax increase, relative to change in total demand.

If $c'''(\cdot) = 0$, the optimal linear tax simplifies to $t^* = E[b]$. To see this, differentiate Equation (11) with respect to t, which yields $\frac{\partial q^d}{\partial t} = -\frac{1}{c''(q^d)}$. This term is constant if $c'''(\cdot) = 0$, so that $\partial q^d / \partial t = \partial \bar{q}^d / \partial t$. Hence, the optimal linear tax is $t^* = E[b]$.

A.C Proof of Proposition 2: Condition for Implementability

Definition 2 (Internality Tax Implementability). *The allocation function* $\xi : \left[\underline{\hat{v}}, \overline{\hat{v}}\right] \to Q$ *is internality-tax implementable if there exists a tax function* $\tau : \left[\underline{\hat{v}}, \overline{\hat{v}}\right] \to \mathbb{R}$ *such that* $\left\{(\xi(\hat{v}), \tau(\hat{v}))|\hat{v} \in \left[\underline{\hat{v}}, \overline{\hat{v}}\right]\right\}$ *satisfy incentive compatibility according to*

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

Implementability hinges on two necessary conditions, which concern the *consumer preferences* or directly stem from them. First, the consumer's utility function must satisfy a single-crossing condition. Second, ξ must be monotonic in $\hat{v}$. Since, in our setting, the single-crossing condition is satisfied via $\partial\left(u_q^d/u_t^d\right)/\partial\hat{v} < 0$, a necessary condition for implementability is that the allocation is non-decreasing in perceived types, i.e., $\partial\xi/\partial\hat{v} \geq 0$ (Proof: See Appendix A.C.1).

Implementability of an incentive compatible mechanism additionally hinges on conditions, which stem from the *policy maker's preferences*. In our setting, these involve (paternalistic) corrective and redistributive motives. In a setting without corrective motives, a sufficient condition for this requirement involves the hazard rate of F, that is, $f(\hat{v})/(1-F(\hat{v}))$, which is a measure of the thinness of the tail of the distribution. For our setting, Proposition 2 shows that the condition is more restrictive involving terms that stem from the corrective motive (Proof: See Appendix A.C.2).

A.C.1 Proof: Single-Crossing and Monotonicity

Incentive compatibility requires

$$\hat{v} = \arg\max_{\tilde{v}} u^d\left(\xi(\tilde{v}), \tau(\tilde{v}), \hat{v}\right) \quad \forall \tilde{v} \in \hat{V}$$

The first-order condition implies

$$\begin{aligned} &\left.\frac{\partial u^d\left(\xi(\tilde{v}), \tau(\tilde{v}), \hat{v}\right)}{\partial \tilde{v}}\right|_{\tilde{v}=\hat{v}} =: \left.\frac{\partial u^d\left(\tilde{v}, \hat{v}\right)}{\partial \tilde{v}}\right|_{\tilde{v}=\hat{v}} = 0 \\ \Leftrightarrow\; &\frac{\partial u^d\left(\tilde{v}, \hat{v}\right)}{\partial \xi} \cdot \frac{\partial \xi(\tilde{v})}{\partial \tilde{v}} + \frac{\partial u^d\left(\tilde{v}, \hat{v}\right)}{\partial \tau} \cdot \frac{\partial \tau(\tilde{v})}{\partial \tilde{v}} = 0 \\ &\Leftrightarrow \frac{\partial \tau(\tilde{v})}{\partial \tilde{v}} = -\frac{\partial u^d\left(\tilde{v}, \hat{v}\right)/\partial \xi}{\partial u^d\left(\tilde{v}, \hat{v}\right)/\partial \tau} \cdot \frac{\partial \xi(\tilde{v})}{\partial \tilde{v}}. \end{aligned} \tag{23}$$

Differentiating the first-order condition with respect to $\hat{v}$ yields

$$\left.\frac{\partial^2 u^d\left(\tilde{v}, \hat{v}\right)}{\partial \tilde{v}^2}\right|_{\tilde{v}=\hat{v}} + \left.\frac{\partial^2 u^d\left(\tilde{v}, \hat{v}\right)}{\partial \tilde{v} \partial \hat{v}}\right|_{\tilde{v}=\hat{v}} = 0. \tag{24}$$

The second-order condition implies that

$$\left.\frac{\partial^2 u^d\left(\tilde{v},\hat{v}\right)}{\partial \tilde{v}^2}\right|_{\tilde{v}=\hat{v}} \leq 0. \tag{25}$$

Equations (25) and (24) imply

$$\begin{aligned}
&\left.\frac{\partial u^d\left(\tilde{v},\hat{v}\right)}{\partial \tilde{v}\partial \hat{v}}\right|_{\tilde{v}=\hat{v}} \geq 0 \\
\Leftrightarrow\ & \frac{\partial\left(\frac{\partial u^d(\tilde{v},\hat{v})}{\partial \xi}\right)}{\partial \hat{v}} \cdot \frac{\partial \xi(\tilde{v})}{\partial \tilde{v}} + \frac{\partial\left(\frac{\partial u^d(\tilde{v},\hat{v})}{\partial \tau}\right)}{\partial \hat{v}} \cdot \frac{\partial \tau(\tilde{v})}{\partial \tilde{v}} \geq 0 \\
\overset{(23)}{\Leftrightarrow}\ & \frac{\partial\left(\frac{\partial u^d}{\partial \xi}\right)}{\partial \hat{v}} \cdot \frac{\partial \xi(\tilde{v})}{\partial \tilde{v}} - \frac{\partial\left(\frac{\partial u^d}{\partial \tau}\right)}{\partial \hat{v}} \cdot \frac{\partial u^d/\partial \xi}{\partial u^d/\partial \tau} \cdot \frac{\partial \xi(\tilde{v})}{\partial \tilde{v}} \geq 0 \\
\Leftrightarrow\ & \frac{\frac{\partial\left(\frac{\partial u^d}{\partial \xi}\right)}{\partial \hat{v}} \partial u^d/\partial \tau - \frac{\partial\left(\frac{\partial u^d}{\partial \tau}\right)}{\partial \hat{v}} \cdot \partial u^d/\partial \xi}{\partial u^d/\partial \tau} \cdot \frac{\partial \xi(\tilde{v})}{\partial \tilde{v}} \geq 0 \\
\Leftrightarrow\ & \frac{\frac{\partial\left(\frac{\partial u^d}{\partial \xi}\right)}{\partial \hat{v}} \partial u^d/\partial \tau - \frac{\partial\left(\frac{\partial u^d}{\partial \tau}\right)}{\partial \hat{v}} \cdot \partial u^d/\partial \xi}{\left(\partial u^d/\partial \tau\right)^2} \cdot \frac{\partial u^d}{\partial \tau} \cdot \frac{\partial \xi(\tilde{v})}{\partial \tilde{v}} \geq 0 \\
\Leftrightarrow\ & \frac{\partial\left(\frac{\partial u^d/\partial \xi}{\partial u^d/\partial \tau}\right)}{\partial \hat{v}} \cdot \frac{\partial u^d}{\partial \tau} \cdot \frac{\partial \xi(\tilde{v})}{\partial \tilde{v}} \geq 0,
\end{aligned}$$

where the second-but-last step follows from the quotient rule. Since we know that $\frac{\partial u^d}{\partial \tau} < 0$ and via single crossing $\frac{\partial\left(\frac{\partial u^d/\partial \xi}{\partial u^d/\partial \tau}\right)}{\partial \hat{v}} = \frac{\partial\left(\frac{\hat{v}-c'-t'}{-1}\right)}{\partial \hat{v}} = -1 < 0$, it follows that we need monotonicity of the form $\frac{\partial \xi(\tilde{v})}{\partial \tilde{v}} \geq 0$.

A.C.2 Proof: Internality Tax Implementability

Inserting the optimal smooth tax from Proposition 1, the consumer's first-order condition can be rewritten as

$$c'[\xi(\hat{v})] \overset{!}{=} \underbrace{\hat{v} - \hat{g}(\hat{v})E[b|\hat{v}] - (1-\hat{G}(\hat{v}))\frac{1-F(\hat{v})}{f(\hat{v})}}_{=:\phi(\hat{v})}. \tag{26}$$

An incentive compatible mechanism must guarantee that ξ is non-decreasing in $\hat{v}$ (see Appendix A.C.1). Since c is convex, ξ is non-decreasing in $\hat{v}$ if and only if the left-hand side of Equation (26) is non-decreasing in $\hat{v}$. Accordingly, the right-hand side of Equation (26) must be non-decreasing in $\hat{v}$ as well, which implies

$$1 - \frac{\partial \hat{g}(\hat{v})}{\partial \hat{v}} E[b|\hat{v}] - \frac{\partial E[b|\hat{v}]}{\partial \hat{v}} \hat{g}(\hat{v}) - \frac{\partial[(1-\hat{G}(\hat{v}))\cdot(\{1-F(\hat{v})\}/f(\hat{v}))]}{\partial \hat{v}} \geq 0. \tag{27}$$

A.D Proof of Proposition 3: Bunching

In this section, we show that a failure of Internality Tax implementability leads to bunching and discuss how bunching at the top or at the bottom is equivalent to a ban for high or low attribute levels. To determine the optimal policy when bunching occurs, we apply the approach by Guesnerie and Laffont (1984) to conduct "ironing" using optimal control theory.

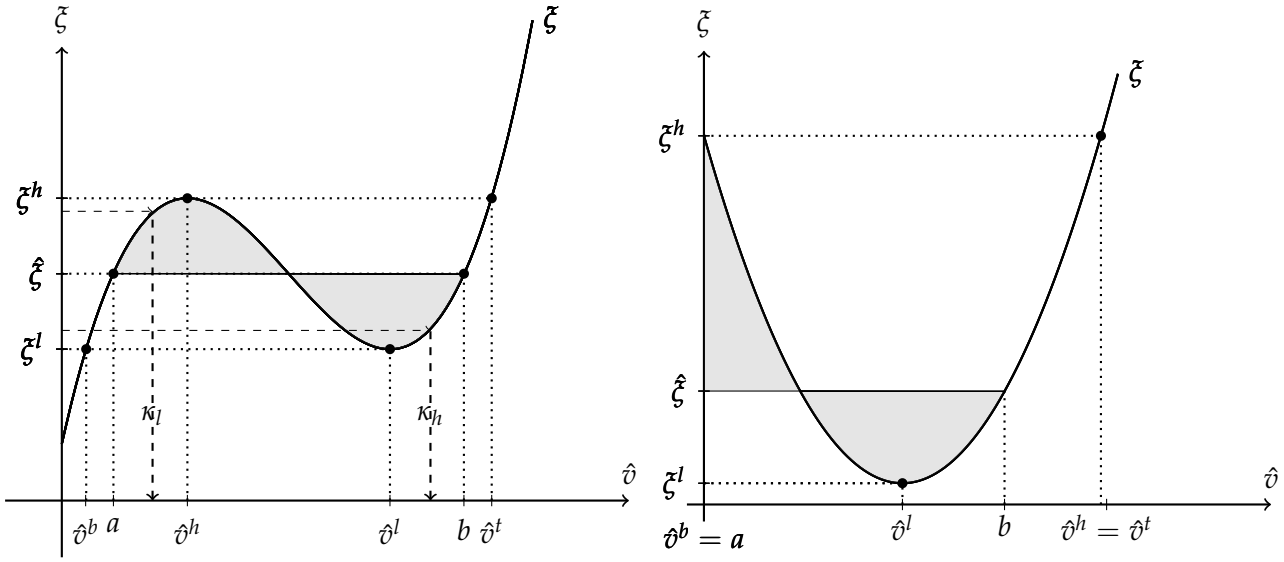
Before we present the formal method of ironing, we illustrate the intuition of this procedure and explain why it implies different forms of standards (uniform, maximum, and minimum). For the ease of exposition we abstract from redistributive motives ($g = 1 = G$). Suppose the smooth tax derived in Proposition 1 implies that the allocation rule ξ is decreasing in the interval $[\hat{v}^h, \hat{v}^l]$, where $\hat{v}^h$ denotes the level of perceived valuation that leads to a higher ξ, while $\hat{v}^l$ leads to a lower ξ ($\hat{v}^h < \hat{v}^l$, but $\xi(\hat{v}^h) > \xi(\hat{v}^l)$). The goal of ironing is to determine over which interval $[a, b]$ of the type space bunching will take place.

Recall that an allocation rule that decreases in $\hat{v}$ implies a fundamental misalignment of preferences violating the condition in Proposition 2: the policy maker would like to allocate a lower allocation level to a consumer with a higher perceived valuation. However, incentive compatibility requires a non-decreasing allocation rule (see Appendix A.C.1): a consumer will not reveal a higher perceived valuation, if he obtains a lower attribute level than a consumer with a lower perceived valuation.

To ensure incentive compatibility, the policy maker resorts to bunching and assigns the same product attribute $\xi^* = \hat{\xi}$ to a subset of consumers. Suppose to the contrary that the implemented allocation rule ξ were partly increasing in the interval $[\hat{v}^h, \hat{v}^l]$. Then, the welfare of types that underconsume compared to ξ can be increased by increasing the allocation, while the welfare of types that overconsume can be increased by decreasing the allocation. This is true up to the point where the consumption of all types $\hat{v} \in [a, b]$ is equal to $\hat{\xi}$.

The standard $\hat{\xi}$ is set such that it minimizes the loss in welfare compared to the solution candidate ξ. Intuitively, the standard $\hat{\xi}$ lies in between the extremes of ξ^l and ξ^h so that some consumers overconsume compared to candidate ξ and some underconsume. If all consumers underconsume, i.e., $\hat{\xi} = \xi^l$, then increasing the standard marginally would raise the welfare of all consumers that still underconsume after the increase, while it would only decrease the welfare of the consumers who overconsume after the increase, that is, for types with an ideal consumption $\xi = \xi^l$. An analogous statement holds for the case when all consumers overconsume, i.e., $\hat{\xi} = \xi^h$.

The exact procedure of determining the types that over- or underconsume compared to ξ is described below for the general case of interior bunching, which sub-



(a) Interior bunching

(b) Bunching at the bottom: Banning low attribute levels

Figure 1: Ironing yields the optimal bunching regions

sumes all cases relevant to our scenario (uniform, minimum and maximum standards). Figure 1 b) visualizes the specific case of minimum standards, i.e., bunching at the bottom.

We now present the formal method of ironing. Bunching occurs if ξ (or equivalently, the term $\phi(\hat{v})$ as defined in Equation (26)) is decreasing over an interval $[\hat{v}^h, \hat{v}^l] \subseteq [\underline{\hat{v}}, \overline{\hat{v}}]$. Let ξ denote the non-ironed "solution candidate" and let ξ^* denote the correct solution involving ironing. Since we cannot guarantee that the allocation is strictly increasing, write

$$\nu(\hat{v}) := \tfrac{\partial \xi}{\partial \hat{v}} \geq 0. \tag{28}$$

We now have a control problem with an inequality constraint on the control ν. The Hamiltonian $\mathfrak{H}(\hat{v}, \xi, \nu, \delta)$ of the general optimization problem without implicitly assuming that the incentive constraints hold, can be written as

$$\mathfrak{H}(\cdot) = \left[\alpha(\hat{v}) \cdot \underbrace{\left(\hat{u}^d(\hat{v}) - E[b|\hat{v}]\xi(\hat{v}) \right)}_{=E[\hat{u}^n(\hat{v},b)|\hat{v}]} + \lambda \underbrace{\left(m + \hat{v}\xi(\hat{v}) - \hat{u}^d(\hat{v}) - c(\xi(\hat{v})) \right)}_{=\tau(\hat{v})} \right] f(\hat{v}) + \delta(\hat{v})\nu(\hat{v}),$$

where δ denotes the multiplier for the constraint given by Inequality (28). The first-order condition on the state ξ yields

$$\frac{-\delta'(\hat{v})}{\lambda} \stackrel{!}{=} \left[\hat{v} - \hat{g}(\hat{v}) E[b|\hat{v}] - c'(\xi^*(\hat{v})) \right] f(\hat{v}). \tag{29}$$

Without redistributive motives ($\hat{G} \equiv 1$), the term $\phi(\hat{v})$ as defined in Equation (26) is given by $\phi(\hat{v}) = \hat{v} - \hat{g}(\hat{v})E[b|\hat{v}]$. For the ease of exposition, we discuss the case without redistributive motives. Thus, Equation (29) can be rewritten as

$$\frac{-\delta'(\hat{v})}{\lambda} \stackrel{!}{=} \left[\phi(\hat{v}) - c'(\xi^*(\hat{v}))\right] f(\hat{v}). \tag{30}$$

For simplicity, let us assume there is one unique interval $[a, b]$ of bunching.[21] Then, integrating (30) between a and b, and using the transversality conditions $\delta(a) = \delta(b) = 0$, which hold since the monotonicity constraint is non-binding at the boundaries, yields

$$\int_a^b \left[\phi(\theta) - c'(\xi^*(\theta))\right] f(\theta)d\theta \stackrel{!}{=} \frac{1}{\lambda}\int_a^b -\delta'(\theta)f(\theta)d\theta = \frac{-\delta(b) + \delta(a)}{\lambda} = 0, \tag{31}$$

which implies that the average difference between the ϕ and the marginal costs (i.e., the average distortion of the "virtual surplus") is zero over the bunching interval.[22] Equation (31) together with the fact that $\xi^*(a) = \xi^*(b) = \xi(a) = \xi(b)$ characterizes the allocation $\xi^*(\hat{v}) = \hat{\xi}$ given to types $\hat{v} \in [a, b]$ in the optimal mechanism.

We now use the above characterization of the bunching region to determine its boundaries a and b and the allocation $\hat{\xi}$. The procedure is illustrated in Figure 1. We first consider a case where the bunching region is in the middle of the range of $\hat{v}$. If $a > \underline{v}$ and $b < \overline{v}$, then there exists $\hat{v}^l := \arg\min_{\hat{v}} \xi(\hat{v})$ s.t. $\hat{v} \in [a, b]$ and $\hat{v}^h := \arg\max_{\hat{v}} \xi(\hat{v})$ s.t. $\hat{v} \in [a, b]$. Denote $\xi^h := \xi(\hat{v}^h)$ and $\xi^l := \xi(\hat{v}^l)$. Let κ_l denote the inverse function of the increasing part of ξ defined on $[\hat{v}^b, \hat{v}^h]$ with $\hat{v}^b := \min_{\hat{v}} \hat{v}$ s.t. $\xi(\hat{v}) = \xi^l$. Analogously, κ_h denotes the inverse function of the increasing part of ξ defined on $[\hat{v}^l, \hat{v}^t]$ with $\hat{v}^t := \max_{\hat{v}} \hat{v}$ s.t. $\xi(\hat{v}) = \xi^h$. For $\tilde{\xi} \in [\xi^h, \xi^l]$, define

$$\Delta(\tilde{\xi}) := \int_{\kappa_l(\tilde{\xi})}^{\kappa_h(\tilde{\xi})} \left[\phi(\theta) - c'(\tilde{\xi})\right] f(\theta)d\theta. \tag{32}$$

Since $\xi^h > \xi(\hat{v})$, $\forall \hat{v} \in (\hat{v}^h, \hat{v}^t)$, it holds that $\Delta(\xi^h) < 0$. Analogously, since $\xi^l < \xi(\hat{v})$, $\forall \hat{v} \in (\hat{v}^b, \hat{v}^l)$, it holds that $\Delta(\xi^l) > 0$. Therefore, by the intermediate value theorem, there must exists some $\hat{\xi}$, such that $\Delta(\hat{\xi}) = 0$ as required by Equation (31). Thus, $a = \kappa_l(\hat{\xi})$ and $b = \kappa_h(\hat{\xi})$. As a consequence, all individuals with $\hat{v} \in [a, b]$ will be assigned the same level of the product attribute $\hat{\xi}$.

[21] The extension to a setting with several bunching regions is only subject to minor caveats. See Guesnerie and Laffont (1984) for details.

[22] Note that in a mechanism which optimally does not involve bunching, the difference between ϕ and the marginal costs is zero at each point as can be seen in Equation (26).

If $a = \underline{\hat{v}}$ (analogous reasoning applies for $b = \overline{\hat{v}}$), bunching occurs at the bottom of the type distribution (see Panel b of Figure 1). In that case, we need to define and evaluate

$$\Delta(\tilde{\xi}) := \int_{\underline{\hat{v}}}^{\kappa_h(\tilde{\xi})} \left[\phi(\theta) - c'(\tilde{\xi})\right] f(\theta) d\theta. \tag{33}$$

As a consequence, all individuals with $\hat{v} \in [\underline{\hat{v}}, b]$ will be assigned the same level of the product attribute $\hat{\xi}$. As bunching occurs at the bottom, this outcome corresponds to a minimum standard, i.e., a ban of all realizations of the product attribute below $\hat{\xi}$. The rationale is equivalent for bunching at the top.

A.E Proof of Equation (5): First-Order Approximation

As a consequence of the Regression Conditional Expectation Function Theorem (e.g., 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:

$$\hat{E}\left[b|\hat{v}\right] = E(b|\hat{v} = \mu_{\hat{v}}) + \frac{cov(b,\hat{v})}{\sigma^2_{\hat{v}}} \cdot \left[\hat{v} - \mu_{\hat{v}}\right]. \tag{34}$$

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:

$$\hat{E}\left[b|\hat{v}\right] = E\left[b|\mu_{\hat{v}}\right] + \frac{\rho + (\sigma_b/\sigma_v)}{(\sigma_b/\sigma_v) + (\sigma_v/\sigma_b) + 2\rho} \cdot \left[\hat{v} - \mu_{\hat{v}}\right]. \tag{35}$$

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

In the bivariate normal case, (v, b) is jointly normal distributed with:

$$(v,b) \sim N\left(\left[\begin{array}{c} \mu_v \\ \mu_b \end{array}\right], \left[\begin{array}{cc} \sigma^2_v & \rho\sigma_v\sigma_b \\ \rho\sigma_v\sigma_b & \sigma^2_b \end{array}\right]\right). \tag{36}$$

Hence, the perceived valuation $\hat{v} = v + b$ has the following normal distribution

$$\hat{v} \sim N\left(\mu_v + \mu_b, \sigma^2_v + \sigma^2_b + 2\rho\sigma_v\sigma_b\right) =: \left(\mu_{\hat{v}}, \sigma^2_{\hat{v}}\right). \tag{37}$$

The conditional expectation $E[b|\hat{v}]$ of the bias can be calculated as

$$E\left[b|\hat{v}\right] = \frac{cov(b,\hat{v})}{\sigma^2_{\hat{v}}} \cdot \hat{v} + \left[\mu_b - \frac{cov(b,\hat{v})}{\sigma^2_{\hat{v}}} \cdot \mu_{\hat{v}}\right] = \mu_b + \frac{cov(b,\hat{v})}{\sigma^2_{\hat{v}}} \cdot \left[\hat{v} - \mu_{\hat{v}}\right], \tag{38}$$

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

$$E\left[b|\hat{v}\right] = \mu_b + \frac{\rho + (\sigma_b/\sigma_v)}{(\sigma_b/\sigma_v) + (\sigma_v/\sigma_b) + 2\rho} \cdot \left[\hat{v} - \mu_{\hat{v}}\right]. \tag{39}$$

Hence, the approximation error $R(\hat{v})$ equals zero for all $\hat{v}$ if v and b follow a bivariate normal distribution.

A.F Proof of Proposition 6: Prices Vs. Quantities

The proof proceeds in two steps. In a first step, we explore the welfare gains from non-linear taxation relative to linear taxes and to standards. The corresponding results are summarized in Proposition 8. In a second step, we then use these results to compare linear taxes with standards. We analyze a setting in which the local bias heterogeneity A is constant.

Proposition 8 (Welfare Gain from Non-Linear Taxation)**.** *The expected welfare gain of the optimal non-linear tax relative to the optimal linear tax and the optimal standard is weakly positive. It depends on the local bias heterogeneity A as follows:*

a) The expected welfare gain over the optimal linear tax is zero when $A = 0$ and increases in the absolute value of A;

b) The expected welfare gain over the optimal standard is zero when $A \geq 1$ and decreases in A.

We now provide the proof of Proposition 8. We analyze the impact of a change in the local bias heterogeneity A on expected welfare under the optimal non-linear tax, the optimal linear tax, and the optimal standard. To separate the impact of changes in A from changes in the population of consumers, we hold the distribution of $\hat{v}$ as well as the first moments of v and b constant. In addition, we consider a scenario in which redistribution does not matter, i.e., $\hat{g}(\hat{v}) = 1$.

To evaluate welfare implications of a change in A, we need to evaluate the derivative with respect to A of the expected equilibrium normative utility (net of taxes) for consumers with $\hat{v}$ (see also the Hamiltonian of the policy maker's problem):

$$\int_{\hat{v}} \hat{u}^n(A) dF(\hat{v}) = \int_{\hat{v}} E[v|\hat{v}](A) q^M(A) - c\left(q^M(A)\right) dF(\hat{v}), \tag{40}$$

where q^M denotes the allocation under the non-linear, **p**rice (linear tax) or **s**tandard (consumption ban) mechanism, $q^M(A) \in \{\xi(A), q^P(A), q^S\}$, and $\hat{u}^n(A, q^M)$ denotes the equilibrium normative utility in the respective mechanism.

We first consider the welfare implications of a change in A under the optimal non-linear tax. We have that:

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

By an envelope theorem argument, the second summand is zero since $E[v|\hat{v}] = \hat{v} - E[b|\hat{v}] = c'(\xi)$ due to optimal consumer behavior (see Equation (11)).

The optimal standard q^S can be calculated by solving the following utility maximization problem: $\max_{q\in\mathbb{R}} \int_v u^n(q,v)dF_v(v)$. Inserting $u^n(q,v) = m + vx - c(q)$ and solving for the maximum yields the following implicit solution: $c'(q^S) = E[v]$. Hence, q^S does not depend on A and we obtain:

$$\frac{\partial \hat{u}^n(A,q^S)}{\partial A} = \frac{\partial E[v|\hat{v}]}{\partial A}q^S.$$

As shown in Section A.B, the optimal linear tax is given by the following expression: $t^*(A) = \int_{\hat{v}} E(b|\hat{v})(A)\left[\frac{\partial q^d}{\partial t} / \frac{\partial \bar{q}^d}{\partial t}\right] dF(\hat{v})$. From the first order condition of consumer maximization, we know that $c'(q^d) = \hat{v} - t^*(A) = E[v|\hat{v}] + E[b|\hat{v}] - t^*(A)$. Using this equation, we obtain:

$$\frac{\partial \hat{u}^n(A,q^P)}{\partial A} = \frac{\partial E[v|\hat{v}]}{\partial A}q^P + \frac{\partial q^P}{\partial A}\{E[v|\hat{v}] - c'(q^P)\} = \frac{\partial E[v|\hat{v}]}{\partial A}q^P + \frac{\partial q^P}{\partial A}\{E[b|\hat{v}] - t^*(A)\}.$$

The second summand vanishes when taking the integral over all types, which can be seen as follows:

$$\begin{aligned}&\int_{\hat{v}} \frac{\partial q^P}{\partial A}\{E[b|\hat{v}] - t^*(A)\}dF(\hat{v}) = \frac{\partial t^*(A)}{\partial A}\int_{\hat{v}} \frac{\partial q^P}{\partial t^*}\{E[b|\hat{v}] - t^*(A)\}dF(\hat{v})\\ =&\frac{\partial t^*(A)}{\partial A}\left(\int_{\hat{v}} \frac{\partial q^P}{\partial t^*}E[b|\hat{v}]dF(\hat{v}) - t^*(A)\int_{\hat{v}} \frac{\partial q^P}{\partial t^*}dF(\hat{v})\right) = 0,\end{aligned}$$

where the last equality follows from the definition of $t^*(A)$ and the fact that $\frac{\partial \bar{q}^d}{\partial t^*} = \int_{\hat{v}} \frac{\partial q^P}{\partial t^*}dF(\hat{v})$. Hence, we obtain:

$$\frac{\partial \hat{u}^n(A,q^P)}{\partial A} = \frac{\partial E[v|\hat{v}]}{\partial A}q^P.$$

We now evaluate how a change in A changes the relative advantage of non-linear taxation relative to a standard in terms of expected welfare:

$$\frac{\partial \Delta u^n_{NL,S}}{\partial A} = \int \frac{\partial E[v|\hat{v}]}{\partial A}\left(\xi - q^S\right) dF(\hat{v}),$$

where $\Delta u^n_{NL,S} := \int \hat{u}^n(A,\xi) - \hat{u}^n(A,q^S)dF(\hat{v})$. Using $\frac{\partial E[v|\hat{v}]}{\partial A} = (\mu_{\hat{v}} - \hat{v})$, we obtain:

$$\begin{aligned}
\frac{\partial \Delta u^n_{NL,S}}{\partial A} &= \int (\mu_{\hat{v}} - \hat{v}) \left[\xi(\hat{v}) - q^S\right] dF(\hat{v}) \\
&= \mu_{\hat{v}} \int \xi(\hat{v}) dF(\hat{v}) - \int \hat{v}\xi(\hat{v}) dF(\hat{v}) - q^S \left[\mu_{\hat{v}} - \int \hat{v} dF(\hat{v})\right] \\
&= E[\hat{v}] \cdot E\left[\xi(\hat{v})\right] - E\left[\hat{v} \cdot \xi(\hat{v})\right] - 0 \\
&= E[\hat{v}] \cdot E\left[\xi(\hat{v})\right] - E[\hat{v}] \cdot E\left[\xi(\hat{v})\right] - cov\left(\hat{v}, \xi(\hat{v})\right) \leq 0,
\end{aligned}$$

where the last inequality holds since ξ is increasing in $\hat{v}$ for an incentive compatible mechanism, so that $cov\left(\hat{v}, \xi(\hat{v})\right)$ is positive.

The relative advantage in terms of expected welfare of non-linear taxation relative to the optimal linear tax is given by:

$$\frac{\partial \Delta u^n_{NL,P}}{\partial A} = \int \frac{\partial E[v|\hat{v}]}{\partial A} \left(\xi(\hat{v}) - q^P(\hat{v})\right) dF\hat{v}.$$

We approximate $\xi - q^P$ by a first-order Taylor approximation in the marginal tax rate t' around $t' = t^*$, which yields:

$$\xi - q^P \approx \left.\frac{dq^P}{dt}\right|_{t'=t^*} \left[t'(\hat{v}) - t^*\right],$$

where $t'(\hat{v}) = E[b|\hat{v}]$ is the optimal non-linear tax rate from Proposition 1 and t^* is the optimal linear tax rate from Equation (22).

Using that $E[b|\hat{v}] = E[b|\mu_{\hat{v}}] + A(\hat{v} - \mu_{\hat{v}})$, rearranging, and using $W(\hat{v}) := \frac{dq(\hat{v})/dt}{d\bar{q}/dt}$ with $\frac{d\bar{q}}{dt} := \int_{\hat{v}} \frac{dq(\hat{v})}{dt} dF(\hat{v})$ to denote the relative responsiveness of a consumer type $\hat{v}$ we obtain:

$$\xi - q^P \approx \left.\frac{dq^P}{dt}\right|_{t'=t^*} A\left(\hat{v} - \int \hat{v} W(\hat{v}) dF(\hat{v})\right).$$

Denoting $v^P = \int \hat{v} W(\hat{v}) dF(\hat{v})$, we have:

$$\begin{aligned}
\frac{\partial \Delta u^e_{NL,P}}{\partial A} &= A \int (\mu_{\hat{v}} - \hat{v})(\hat{v} - v^P) \left.\frac{dq^P}{dt}\right|_{t'=t^*} dF(\hat{v}) \\
&= A \int (\mu_{\hat{v}}\hat{v} - \mu_{\hat{v}} v^P - \hat{v}^2 + \hat{v} v^P) \left.\frac{dq^P}{dt}\right|_{t'=t^*} dF(\hat{v}) \\
&= A \left((\mu_{\hat{v}} + v^P) \int \hat{v} \left.\frac{dq^P}{dt}\right|_{t'=t^*} dF(\hat{v}) - (\mu_{\hat{v}} v^P) \int \left.\frac{dq^P}{dt}\right|_{t'=t^*} dF(\hat{v}) - \int \hat{v}^2 \left.\frac{dq^P}{dt}\right|_{t'=t^*} dF(\hat{v}) \right) \\
&= A \frac{d\bar{q}^P}{dt} \left((\mu_{\hat{v}} + v^P) \int \hat{v} W(\hat{v}) dF(\hat{v}) - \mu_{\hat{v}} v^P - \int \hat{v}^2 W(\hat{v}) dF(\hat{v}) \right)
\end{aligned}$$

$$= A\frac{d\bar{q}^P}{dt}\left((v^P)^2 - \int \hat{v}^2 W(\hat{v}) dF(\hat{v})\right)$$
$$= \underbrace{A\frac{d\bar{q}^P}{dt}}_{\leq 0 \text{ if } A \geq 0} \underbrace{\left(\left(\int \hat{v} W(\hat{v}) dF(\hat{v})\right)^2 - \int \hat{v}^2 W(\hat{v}) dF(\hat{v})\right)}_{\leq 0} \geq 0,$$

where $\frac{d\bar{q}^P}{dt} = \int \frac{dq^P}{dt}\big|_{t'=t^*} dF(\hat{v})$. The fact that the third factor is non-positive follows from Jensen's inequality because $d(\hat{v}) = W(\hat{v}) f(\hat{v})$ satisfies the properties of a density function, with $d(\hat{v}) \geq 0 \; \forall \, \hat{v}$ and $\int d(\hat{v}) d\hat{v} = \int W(\hat{v}) dF(\hat{v}) = 1$.

Comparison of Linear Price Instruments with Standards

In Proposition 8, we have shown that a linear price instrument is welfare-optimal for $A = 0$, a standard is optimal for $A = 1$, and the advantage of a standard compared to a linear price instrument increases in A. Hence, the existence of $\hat{A}$ from Proposition 6 is guaranteed by applying the intermediate value theorem.

To complete the proof of Proposition 6, we need to show that the welfare advantage of a price instrument over a quantity instrument becomes more pronounced as A decreases below zero. For that purpose, it is sufficient to show that the difference in the expected equilibrium normative utility between the linear tax and the quantity regulation, $\partial \Delta u^n_{P,S} / \partial A = \int \hat{u}^n(A, q^P) - \hat{u}^n(A, q^S) dF(\hat{v})$, decreases in A. Let $\bar{v}$ be the $\hat{v}$ such that $q^P(\hat{v}) = q^S$. We approximate the demand function by a first-order Taylor approximation at $\bar{v}$, which yields $q^P(\hat{v}) \approx q^P(\bar{v}) + \frac{\partial q^P}{\partial \hat{v}}\big|_{\bar{v}} (\hat{v} - \bar{v})$. As the first-order conditions of utility maximization imply $\hat{v} - t^* = c'\left(q^P(\hat{v})\right)$ for linear taxation and $c'\left(q^S\right) = E[v]$ for quantity regulation, we have that $\bar{v} = E[v] + t^*$, where $t^* = \int_{\hat{v}} E[b|\hat{v}] \, W(\hat{v}) \, dF(\hat{v})$. We can now show that the welfare difference between prices and quantities decreases in A, since

$$\begin{aligned}
\frac{\partial \Delta u^n_{P,S}}{\partial A} &= \int_{\hat{v}} \frac{\partial E[v|\hat{v}]}{\partial A} \left[q^P(\hat{v}) - q^S\right] dF(\hat{v}) \\
&= \int_{\hat{v}} (\mu_{\hat{v}} - \hat{v}) \left[q^P(\hat{v}) - q^S\right] dF(\hat{v}) \\
&= -\frac{\partial q^P}{\partial \hat{v}}\big|_{\hat{v}=\bar{v}} \int_{\hat{v}} (\mu_{\hat{v}} - \hat{v})(\hat{v} - \bar{v}) \, dF(\hat{v}) \\
&= -\frac{\partial q^P}{\partial \hat{v}}\big|_{\hat{v}=\bar{v}} \int_{\hat{v}} (\mu_{\hat{v}} - \hat{v}) \left[(v - E[v])(b - t^*)\right] dF(\hat{v}) \\
&= \underbrace{-\frac{\partial q^P}{\partial \hat{v}}\big|_{\hat{v}=\bar{v}}}_{<0} \underbrace{\left[E\left[\hat{v}^2\right] - \mu^2_{\hat{v}}\right]}_{>0} < 0,
\end{aligned}$$

where the evaluation of the second factor in the last line follows from Jensen's inequality.

A.G Proof of Proposition 7: Minimum and Maximum Standards

The proof proceeds as follows. From Proposition 3, we know that $\partial E[v|\hat{v}]/\partial\hat{v} < 0$ for some $\hat{v}$ implies that a policy maker will use bunching as part of the optimal policy. Furthermore, we know from Appendix A.D that bans are part of the optimal policy mix if bunching occurs at the top or the bottom of the perceived valuation distribution. We use a second-order Taylor approximation to approximate $E[v|\hat{v}]$ as a function of $\hat{v}$ and $\hat{v}^2$. This allows to derive the conditions under which mixed policies optimally involve bans for high or low levels of the product attribute in terms of higher-order moments of the (joint) distributions of the random variables v, b, and $\hat{v}$.

The second-order Taylor approximation of $E\left[v|\hat{v}\right]$ is:

$$\hat{E}\left[v|\hat{v}\right] = c_0 + c_1\hat{v} + c_2\hat{v}^2. \tag{41}$$

Bunching occurs whenever $\partial E\left[v|\hat{v}\right]/\partial\hat{v} = c_1 + 2c_2\hat{v} \leq 0$ for some $\hat{v}$. For the purpose of this proof, we are interested in bunching at the top and at the bottom. With $\hat{v} \in [0, \infty)$, we have:

1. Bunching at the Top if $c_1 > 0$ and $c_2 < 0$.
2. Bunching at the Bottom if $c_1 < 0$ and $c_2 > 0$.

Otherwise, bunching takes place over the entire support of $\hat{v}$ ($c_1 < 0, c_2 \leq 0$) or not at all ($c_1 > 0, c_2 \geq 0$). We proceed by first calculating c_1 and c_2. In a subsequent step, we derive sufficient conditions for bunching at the top and at the bottom.

The Frisch-Waugh-Lovell Theorem demonstrates that c_r for $r \in \{1, 2\}$, as defined by Equation (41), can be calculated as

$$c_r = \frac{cov\left(\epsilon_r, \hat{\epsilon}_r\right)}{var\left(\hat{\epsilon}_r\right)}, \tag{42}$$

where $\hat{\epsilon}_r$ is the "residual" of $\hat{v}^r$ after regressing it on all covariates except $\hat{v}^r$, and ϵ_r is the "residual" of v after regressing it on all covariates except $\hat{v}^r$.

We first calculate c_1. Regressing $\hat{v}$ on $\hat{v}^2$ and a constant yields:

$$\begin{aligned}\hat{v} &= d_1 + d_2\hat{v}^2 + \hat{\epsilon}_1 \\ &= E\left[\hat{v}|\hat{v}^2 = 0\right] + \frac{cov\left(\hat{v}, \hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2 + \hat{\epsilon}_1.\end{aligned}$$

The residual $\hat{\epsilon}_1$ is then given by:

$$\hat{\epsilon}_1 = \hat{v} - \left\{ E\left[\hat{v}|\hat{v}^2 = 0\right] + \frac{cov\left(\hat{v}, \hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2 \right\}.$$

Regressing v on $\hat{v}^2$ and a constant yields:

$$\begin{aligned} v &= e_0 + e_1\hat{v}^2 + \epsilon_1 \\ &= E\left[v|\hat{v}^2 = 0\right] + \frac{cov\left(v, \hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2 + \epsilon_1 \end{aligned}$$

The residual ϵ_1 is then given by:

$$\epsilon_1 = v - \left\{ E\left[v|\hat{v}^2 = 0\right] + \frac{cov\left(v, \hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2 \right\}.$$

Therefore, by Equation (42), the coefficient c_1 is given by:

$$\begin{aligned} c_1 &= \frac{cov\left(\epsilon_1, \hat{\epsilon}_1\right)}{var\left(\hat{\epsilon}_1\right)} \\ &= \frac{cov\left(v - \left\{ E\left[v|\hat{v}^2 = 0\right] + \frac{cov\left(v,\hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2 \right\}, \hat{v} - \left\{ E\left[\hat{v}|\hat{v}^2 = 0\right] + \frac{cov\left(\hat{v},\hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2 \right\}\right)}{var\left(\hat{\epsilon}_1\right)} \\ &= \frac{cov\left(v, \hat{v} - \frac{cov\left(\hat{v},\hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2\right)}{var\left(\hat{\epsilon}_1\right)} + \frac{cov\left(-\frac{cov\left(v,\hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2, \hat{v} - \frac{cov\left(\hat{v},\hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2\right)}{var\left(\hat{\epsilon}_1\right)} \\ &= \frac{cov\left(v, \hat{v} - \frac{cov\left(\hat{v},\hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2\right)}{var\left(\hat{\epsilon}_1\right)} - \frac{\frac{cov\left(\hat{v},\hat{v}^2\right)}{var\left(\hat{v}^2\right)}cov\left(v, \hat{v}^2\right)}{var\left(\hat{\epsilon}_1\right)} + \frac{\frac{cov\left(\hat{v},\hat{v}^2\right)cov\left(v,\hat{v}^2\right)}{\left[var\left(\hat{v}^2\right)\right]^2}cov\left(\hat{v}^2, \hat{v}^2\right)}{var\left(\hat{\epsilon}_1\right)} \\ &= \frac{cov\left(v, \hat{v} - \frac{cov\left(\hat{v},\hat{v}^2\right)}{var\left(\hat{v}^2\right)}\hat{v}^2\right)}{var\left(\hat{\epsilon}_1\right)}. \end{aligned} \tag{43}$$

Next, we calculate c_2. Regressing $\hat{v}^2$ on $\hat{v}$ and a constant yields:

$$\begin{aligned} \hat{v}^2 &= d_0 + d_1\hat{v} + \hat{\epsilon}_2 \\ &= E\left[\hat{v}^2|\hat{v} = 0\right] + \frac{cov\left(\hat{v}^2, \hat{v}\right)}{var\left(\hat{v}\right)}\hat{v} + \hat{\epsilon}_2. \end{aligned}$$

The residual $\hat{\epsilon}_2$ is then given by:

$$\hat{\epsilon}_2 = \hat{v}^2 - \left\{ E\left[\hat{v}^2 | \hat{v} = 0\right] + \frac{cov\left(\hat{v}^2, \hat{v}\right)}{var\left(\hat{v}\right)} \hat{v} \right\}.$$

Regressing v on $\hat{v}$ and a constant yields:

$$\begin{aligned} v &= d_0 + d_1 \hat{v} + \epsilon_2 \\ &= E\left[v | \hat{v} = 0\right] + \frac{cov\left(v, \hat{v}\right)}{var\left(\hat{v}\right)} \hat{v} + \epsilon_2. \end{aligned}$$

The residual ϵ_2 is then given by:

$$\epsilon_2 = v - \left\{ E\left[v | \hat{v} = 0\right] + \frac{cov\left(v, \hat{v}\right)}{var\left(\hat{v}\right)} \hat{v} \right\}.$$

Therefore, by Equation (42), the coefficient c_2 is given by:

$$\begin{aligned} c_2 &= \frac{cov\left(\epsilon_2, \hat{\epsilon}_2\right)}{var\left(\hat{\epsilon}_2\right)} \\ &= \frac{cov\left(v - \left\{ E\left[v | \hat{v} = 0\right] + \frac{cov(v,\hat{v})}{var(\hat{v})} \hat{v} \right\}, \hat{v}^2 - \left\{ E\left[\hat{v}^2 | \hat{v} = 0\right] + \frac{cov\left(\hat{v}^2,\hat{v}\right)}{var(\hat{v})} \hat{v} \right\} \right)}{var\left(\hat{\epsilon}_2\right)} \\ &= \frac{cov\left(v, \hat{v}^2\right) - \frac{cov(v,\hat{v})}{var(\hat{v})} cov\left(\hat{v}^2, \hat{v}\right)}{var\left(\hat{\epsilon}_2\right)} + \frac{cov\left(-\frac{cov(v,\hat{v})}{var(\hat{v})} \hat{v}, \hat{v}^2 - \frac{cov\left(\hat{v}^2,\hat{v}\right)}{var(\hat{v})} \hat{v} \right)}{var\left(\hat{\epsilon}_2\right)} \\ &= \frac{cov\left(v, \hat{v}^2\right) - \frac{cov(v,\hat{v})}{var(\hat{v})} \left[cov\left(\hat{v}^2, v\right) + cov\left(\hat{v}^2, b\right)\right]}{var\left(\hat{\epsilon}_2\right)} + \frac{-\frac{cov(v,\hat{v})}{var(\hat{v})} cov\left(\hat{v}, \hat{v}^2\right) + cov\left(v, \hat{v}\right) \frac{cov\left(\hat{v}^2,\hat{v}\right)}{var(\hat{v})}}{var\left(\hat{\epsilon}_2\right)} \\ &= \frac{cov\left(v, \hat{v}^2\right) - \frac{cov(v,\hat{v})}{var(\hat{v})} \left[cov\left(\hat{v}^2, v\right) + cov\left(\hat{v}^2, b\right)\right]}{var\left(\hat{\epsilon}_2\right)} \\ &= \frac{cov\left(v, \hat{v}^2\right) cov\left(b, \hat{v}\right) - cov\left(v, \hat{v}\right) cov\left(\hat{v}^2, b\right)}{var\left(\hat{v}\right) var\left(\hat{\epsilon}_2\right)} \\ &= \frac{cov\left(\hat{v}, \hat{v}^2\right) cov\left(b, \hat{v}\right) - cov\left(\hat{v}, \hat{v}\right) cov\left(\hat{v}^2, b\right)}{var\left(\hat{v}\right) var\left(\hat{\epsilon}_2\right)} \\ &\overset{(*)}{=} \frac{\left[2\mu_{\hat{v}}\sigma_{\hat{v}}^2 + sk(\hat{v})\sigma_{\hat{v}}^3\right] cov\left(b, \hat{v}\right) - \sigma_{\hat{v}}^2 cov\left(\hat{v}^2, b\right)}{var\left(\hat{v}\right) var\left(\hat{\epsilon}_2\right)} \\ &= \frac{\sigma_{\hat{v}}^2 \overbrace{\left\{\left[2\mu_{\hat{v}} + sk(\hat{v})\sigma_{\hat{v}}\right] cov\left(b, \hat{v}\right) - cov\left(\hat{v}^2, b\right)\right\}}^{=:K}}{var\left(\hat{v}\right) var\left(\hat{\epsilon}_2\right)}, \end{aligned} \tag{44}$$

where the denominator is always positive and $(*)$ follows from the fact that, for every random variable Y, we have:[23]

$$cov\left(Y, Y^2\right) = \left[sk(Y)\sigma_Y + 2\mu_Y\right]\sigma_Y^2. \tag{45}$$

We now analyze the term K:

$$\begin{aligned}
K &\overset{\text{BG1969}}{=} \left[2\mu_{\hat{v}} + sk(\hat{v})\sigma_{\hat{v}}\right] cov\left(b, \hat{v}\right) - \left[2\mu_{\hat{v}} cov\left(\hat{v}, b\right) + E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^2 \left(b - \mu_b\right)\right]\right] \\
&= sk(\hat{v})\sigma_{\hat{v}} cov\left(b, \hat{v}\right) - E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^2 \left(b - \mu_b\right)\right] \\
&= E\left[\left(\frac{\hat{v} - \mu_{\hat{v}}}{\sigma_{\hat{v}}}\right)^3\right] \sigma_{\hat{v}} E\left[\left(\hat{v} - \mu_{\hat{v}}\right)\left(b - \mu_b\right)\right] - E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^2 \left(b - \mu_b\right)\right] \\
&= \frac{E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^3\right] E\left[\left(\hat{v} - \mu_{\hat{v}}\right)\left(b - \mu_b\right)\right]}{\sigma_{\hat{v}}^2} - E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^2 \left(b - \mu_b\right)\right] \\
&= \frac{E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^3\right] E\left[\left(\hat{v} - \mu_{\hat{v}}\right)\left(b - \mu_b\right)\right]}{\sigma_{\hat{v}}^2} - E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^2 \left(b - \mu_b\right)\right] \\
&= \sigma_{\hat{v}}^2 \sigma_b \left\{ \frac{E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^3\right] E\left[\left(\hat{v} - \mu_{\hat{v}}\right)\left(b - \mu_b\right)\right]}{\sigma_{\hat{v}}^4 \sigma_b} - \frac{E\left[\left(\hat{v} - \mu_{\hat{v}}\right)^2 \left(b - \mu_b\right)\right]}{\sigma_{\hat{v}}^2 \sigma_b} \right\} \\
&= \sigma_{\hat{v}}^2 \sigma_b \left\{ sk\left(\hat{v}\right) \rho\left(\hat{v}, b\right) - cosk\left(\hat{v}, b\right) \right\},
\end{aligned}$$

where BG1969 refers to Bohrnstedt and Goldberger (1969). Hence, a sufficient (and necessary) condition for $c_2 < 0$ is:

$$c_2 < 0 \iff \rho\left(\hat{v}, b\right) sk\left(\hat{v}\right) < cosk\left(\hat{v}, b\right) \tag{46}$$

and a sufficient (and necessary) condition for $c_2 > 0$ is:

$$c_2 > 0 \iff \rho\left(\hat{v}, b\right) sk\left(\hat{v}\right) > cosk\left(\hat{v}, b\right). \tag{47}$$

[23] This can be seen as follows:

$$\begin{aligned}
cov\left(Y, Y^2\right) &= E\left(Y^3\right) - \mu_{Y^2}\mu_Y = E\left(Y^3\right) - \mu_Y\left[E\left(Y^2\right) - \mu_Y^2 + \mu_Y^2\right] = E\left(Y^3\right) - \mu_Y\sigma_Y^2 - \mu_Y^3 \\
&= \frac{\left[E\left(Y^3\right) - \mu_Y^3 - 3\mu_Y\sigma_Y^2 + 3\mu_Y\sigma_Y^2\right]\sigma_Y^3}{\sigma_Y^3} - \mu_Y\sigma_Y^2 = sk(Y)\sigma_Y^3 + 3\mu_Y\sigma_Y^2 - \mu_Y\sigma_Y^2 = sk(Y)\sigma_Y^3 + 2\mu_Y\sigma_Y^2.
\end{aligned}$$

Sufficient Conditions for Bunching at the Top

To obtain sufficient conditions for Bunching at the Top, we need to find sufficient conditions for $c_1 > 0$, given that $c_2 < 0$. Using the results for c_1 from Equation (43), we obtain:

$$\begin{aligned} & c_1 > 0 \\ \Leftrightarrow & \frac{cov\left(v, \hat{v} - \frac{cov\left(\hat{v}, \hat{v}^2\right)}{var(\hat{v}^2)}\hat{v}^2\right)}{var\left(\hat{\epsilon}_1\right)} > 0 \\ \Leftrightarrow & \underbrace{cov\left(v, \hat{v}\right)}_{(1)} - \underbrace{cov\left(v, \hat{v}^2\right)}_{(2)} \underbrace{\frac{cov\left(\hat{v}, \hat{v}^2\right)}{var\left(\hat{v}^2\right)}}_{(3)} > 0. \end{aligned} \tag{48}$$

Term (1) of Inequality (48) is positive by assumption because we restrict ourselves to settings where partial bans become relevant (note that $A < 1$ from Proposition 5 is equivalent to $cov(v, \hat{v}) > 0$). A sufficient condition for Term (3) to be positive is that the skewness is positive, which follows from:

$$\begin{aligned} cov\left(\hat{v}, \hat{v}^2\right) > 0 & \Leftrightarrow 2\mu_{\hat{v}}\sigma_{\hat{v}}^2 + sk(\hat{v})\sigma_{\hat{v}}^3 > 0 \\ & \Leftrightarrow sk(\hat{v}) > -\frac{2\mu_{\hat{v}}}{\sigma_{\hat{v}}} \\ & \Leftarrow \boxed{sk(\hat{v}) > 0}. \end{aligned}$$

Next, we show that Inequality (48) holds under the assumptions made so far. When $cov\left(v, \hat{v}^2\right) \leq 0$, the inequality holds because Term (1) and Term (3) are positive. When $cov\left(v, \hat{v}^2\right) > 0$, we can derive an upper bound for Term (2) by rearranging Equation (44):

$$c_2 < 0 \iff cov\left(v, \hat{v}^2\right) < cov\left(v, \hat{v}\right) \frac{cov\left(\hat{v}^2, \hat{v}\right)}{var\left(\hat{v}\right)}. \tag{49}$$

Then, Inequality (48) holds if it holds after substituting the upper bound from Inequality (49) for Term (2):

$$\begin{aligned} cov\left(v, \hat{v}\right) - cov\left(v, \hat{v}\right) \frac{cov\left(\hat{v}^2, \hat{v}\right)}{var\left(\hat{v}\right)} \frac{cov\left(\hat{v}, \hat{v}^2\right)}{var\left(\hat{v}^2\right)} > 0 \\ \Leftrightarrow cov\left(v, \hat{v}\right)\left(1 - \rho^2_{\hat{v}^2, \hat{v}}\right) > 0, \end{aligned}$$

where $\rho^2_{\hat{v}^2, \hat{v}} \in (-1, 1)$ is the square of the correlation between $\hat{v}^2$ and $\hat{v}$. As we consider settings with $cov(v, \hat{v}) > 0$, this inequality holds.

To close the proof, note that $sk(\hat{v}) > 0$ in combination with Equation (46) implies that:

$$\boxed{\frac{cosk\left(\hat{v}, b\right)}{sk\left(\hat{v}\right)} > \rho\left(\hat{v}, b\right)}.$$

Hence, this condition and the condition that $\boxed{sk(\hat{v}) > 0}$ are sufficient for Bunching at the Top if $\hat{v} \in [0, \infty)$ and $cov(v, \hat{v}) > 0$.

Sufficient Conditions for Bunching at the Bottom

To obtain sufficient conditions for Bunching at the Bottom, we need to find sufficient conditions for $c_1 < 0$, given that $c_2 > 0$. Using the results for c_1 from Equation (48), we obtain:

$$c_1 < 0 \iff cov\left(v, \hat{v}\right) - cov\left(v, \hat{v}^2\right) \frac{cov\left(\hat{v}, \hat{v}^2\right)}{var\left(\hat{v}^2\right)} < 0. \tag{50}$$

Using Equation (45), we obtain that

$$cov\left(\hat{v}, \hat{v}^2\right) = \left[sk(\hat{v})\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right]\sigma_{\hat{v}}^2.$$

Furthermore,

$$\begin{aligned}
cov\left(v, \hat{v}^2\right) &= E\left[\left(v - \mu_v\right)\left(\hat{v}^2 - \mu_{\hat{v}^2}\right)\right] \\
&= E\left(v\hat{v}^2\right) - \mu_{\hat{v}^2}\mu_v - \mu_v E\left(\hat{v}^2\right) + \mu_v\mu_{\hat{v}^2} \\
&= \frac{\left[E\left(v\hat{v}^2\right) - \mu_v E\left(\hat{v}^2\right) - 2E\left(v\hat{v}\right)\mu_{\hat{v}} + 2\mu_{\hat{v}}^2\mu_v + 2E\left(v\hat{v}\right)\mu_{\hat{v}} - 2\mu_{\hat{v}}^2\mu_v\right]\sigma_{\hat{v}}^2\sigma_v}{\sigma_{\hat{v}}^2\sigma_v} \\
&\overset{(*)}{=} cosk(\hat{v}, v)\sigma_{\hat{v}}^2\sigma_v + 2E\left(v\hat{v}\right)\mu_{\hat{v}} - 2\mu_{\hat{v}}^2\mu_v \\
&= cosk(\hat{v}, v)\sigma_{\hat{v}}^2\sigma_v + 2\mu_{\hat{v}}\left[E\left(v\hat{v}\right) - \mu_{\hat{v}}\mu_v + \mu_{\hat{v}}\mu_v - \mu_{\hat{v}}\mu_v\right] \\
&= cosk(\hat{v}, v)\sigma_{\hat{v}}^2\sigma_v + 2\mu_{\hat{v}}cov\left(v, \hat{v}\right) \\
&= \left[cosk(\hat{v}, v)\sigma_{\hat{v}} + 2\mu_{\hat{v}}\rho\left(v, \hat{v}\right)\right]\sigma_{\hat{v}}\sigma_v,
\end{aligned}$$

where in $(*)$ we use that $cosk(\hat{v}, v) = E\left(v\hat{v}^2\right) - 2E\left(v\hat{v}\right)\mu_{\hat{v}} - 2E\left(\hat{v}^2\right)\mu_v + 2\mu_{\hat{v}}^2\mu_v$. Hence, we can rewrite Inequality (50) as follows

$$\begin{aligned}
cov\left(v, \hat{v}\right) - cov\left(v, \hat{v}^2\right)\frac{cov\left(\hat{v}, \hat{v}^2\right)}{var\left(\hat{v}^2\right)} &< 0 \\
\Leftrightarrow cov\left(v, \hat{v}\right) - \left[cosk(\hat{v}, v)\sigma_{\hat{v}} + 2\mu_{\hat{v}}\rho\left(v, \hat{v}\right)\right]\sigma_{\hat{v}}\sigma_v\frac{\left[sk(\hat{v})\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right]\sigma_{\hat{v}}^2}{var\left(\hat{v}^2\right)} &< 0
\end{aligned}$$

$$\Leftrightarrow \left[cosk(\hat{v},v)\sigma_{\hat{v}} + 2\mu_{\hat{v}}\rho\left(v,\hat{v}\right)\right]\sigma_{\hat{v}}\sigma_{v}\frac{\left[sk(\hat{v})\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right]\sigma_{\hat{v}}^{2}}{var\left(\hat{v}^{2}\right)} > cov\left(v,\hat{v}\right)$$

$$\Leftrightarrow \left[cosk(\hat{v},v)\sigma_{\hat{v}} + 2\mu_{\hat{v}}\rho\left(v,\hat{v}\right)\right]\frac{\left[sk(\hat{v})\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right]\sigma_{\hat{v}}^{2}}{var\left(\hat{v}^{2}\right)} \overset{(*)}{>} \rho\left(v,\hat{v}\right)$$

$$\Leftrightarrow \left[\frac{cosk(\hat{v},v)}{\rho\left(v,\hat{v}\right)}\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right]\frac{\left[sk(\hat{v})\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right]\sigma_{\hat{v}}^{2}}{var\left(\hat{v}^{2}\right)} > 1$$

$$\Leftrightarrow \left[\frac{cosk(\hat{v},v)}{\rho\left(v,\hat{v}\right)}\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right]\left[sk(\hat{v})\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right] > \frac{var\left(\hat{v}^{2}\right)}{\sigma_{\hat{v}}^{2}}$$

$$\Leftrightarrow \left[\frac{cosk(\hat{v},v)}{\rho\left(v,\hat{v}\right)}\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right]\left[sk(\hat{v})\sigma_{\hat{v}} + 2\mu_{\hat{v}}\right] - \frac{\sigma_{\hat{v}^2}^{2}}{\sigma_{\hat{v}}^{2}} > 0$$

$$\Leftrightarrow \underbrace{\left[\frac{cosk(\hat{v},v)}{\rho\left(v,\hat{v}\right)}\sigma_{\hat{v}} + 2\mu_{\hat{v}} + \frac{\sigma_{\hat{v}^2}}{\sigma_{\hat{v}}}\right]}_{(1)}\underbrace{\left[sk(\hat{v})\sigma_{\hat{v}} + 2\mu_{\hat{v}} - \frac{\sigma_{\hat{v}^2}}{\sigma_{\hat{v}}}\right]}_{(2)} + \underbrace{\sigma_{\hat{v}^2}\left[\frac{cosk(\hat{v},v)}{\rho\left(v,\hat{v}\right)} - sk(\hat{v})\right]}_{(3)} > 0,$$

where in $(*)$ we use that $cov\left(v,\hat{v}\right) > 0$. The last inequality holds if $(1) < 0$, $(2) < 0$, and $(3) > 0$. Note that $(3) > 0$ follows from $c_2 > 0$. Furthermore, $(1) < 0$ if:

$$\frac{cosk(\hat{v},v)}{\rho\left(v,\hat{v}\right)} < -2\frac{\mu_{\hat{v}}}{\sigma_{\hat{v}}} - \frac{\sigma_{\hat{v}^2}}{\sigma_{\hat{v}}^{2}} =: k, \tag{51}$$

where $k < 0$. From $(3) > 0$ and $(1) < 0$ it also follows that $(2) < 0$, which can be seen as follows:

$$sk(\hat{v}) \overset{(3)>0}{<} \frac{cosk(\hat{v},v)}{\rho\left(v,\hat{v}\right)} \overset{(1)<0}{<} -2\frac{\mu_{\hat{v}}}{\sigma_{\hat{v}}} - \frac{\sigma_{\hat{v}^2}}{\sigma_{\hat{v}}^{2}} < -2\frac{\mu_{\hat{v}}}{\sigma_{\hat{v}}} + \frac{\sigma_{\hat{v}^2}}{\sigma_{\hat{v}}^{2}}. \tag{52}$$

To close the proof, note that $sk(\hat{v}) < 0$ in combination with Equation (47) implies that:

$$\boxed{\frac{cosk\left(\hat{v},b\right)}{sk\left(\hat{v}\right)} > \rho\left(\hat{v},b\right)}.$$

Hence, this condition and the condition that $\boxed{sk(\hat{v}) < \frac{cosk(\hat{v},v)}{\rho\left(v,\hat{v}\right)} < k < 0}$ are sufficient for Bunching at the Bottom if $\hat{v} \in [0,\infty)$ and $cov(v,\hat{v}) > 0$.

B Optimal Non-Linear Taxation in the Light Bulb Market

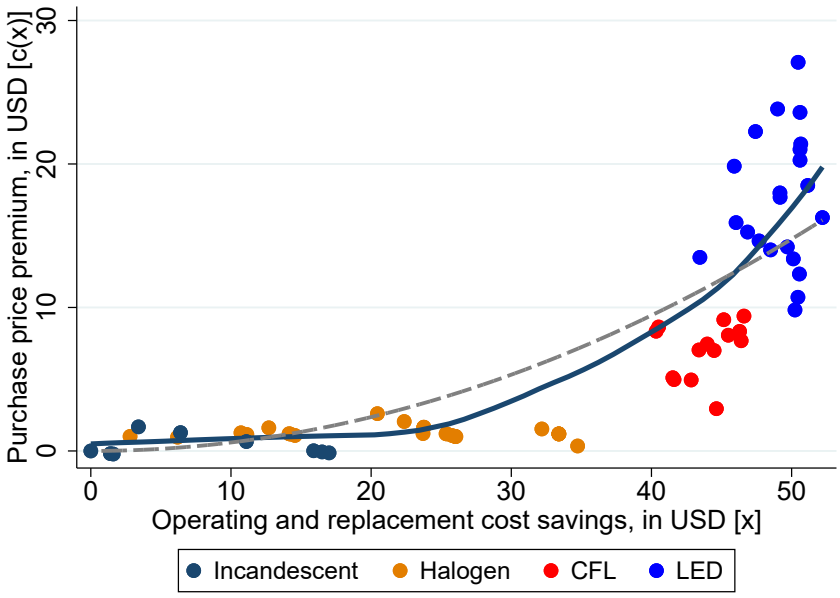
We first explain how we approximate the cost function for energy efficiency and consumers' normative valuations v. In a subsequent step, we present the aggregate welfare effects of the optimal linear and non-linear subsidy on energy efficiency, relative to no taxation.

Our supply data stems from a price comparison service, *geizhals.de*, which reports the cheapest price of a product offered on an online website. 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 two large manufacturers, *Osram* and *Philips*, that offer bulbs both in the European Union and the United States. As in Allcott and Taubinsky (2015a), we express all prices in 2012 US dollars (USD) and collect 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 their aggregate annual price trends, which 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 (ORC) 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, 2015a).

Based on this data, we determine the purchase price premiums and ORC savings relative to the most electricity-intensive bulb. In the following, we use ORC savings as the measure of attribute level q, i.e., of energy efficiency. Figure 2 plots the price premiums against ORC savings, which corresponds to the cost function $c(q)$ in our model. The least energy inefficient, yet cheapest, bulbs are incandescent bulbs, followed by halogen, CFL and LED bulbs. The cost curve is convex, which reflects that a one unit increase 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 about 50 USD over the course of eight years, compared to the most energy inefficient incandescent bulbs.

We use the elicitation of time preferences by Allcott and Taubinsky (2015a) to determine individual-specific discount factors. We assume that all other factors that influence normative valuations do not vary by participant and thus merely constitute a scaling factor. This assumption allows us to calibrate valuations to match the supply function from Figure 2. In particular, we set valuations to $v = s \cdot D(\delta)$, where s is a scaling factor that ensures that consumers demand every product variety offered on the market. To illustrate, consider a consumer with a discount rate of $\delta = 20\%$ and annual operating cost of $1/8$ USD for eight years, which results in ORC savings of 1

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 (2015a). The solid line plots predictions from a local linear regression (bandwidth: 9), while the dashed line plots predictions from a regression on quadratic terms.

USD. For that consumer, the normative valuation of 1 USD in ORC savings is then $D(\delta) = (1 + 1/(1+\delta) + \cdots + 1/(1+\delta)^7) \cdot (1/8) = 0.58$ USD. We consider this approach as a useful approximation of individuals' normative preferences v that isolates one source of heterogeneity in v and is consistent with observable market behavior.[24]

We assume a quadratic cost function and estimate it based on the data from Figure 2. We derive consumers' choices in five scenarios: the absence of a corrective tax, the presence of the optimal linear tax, the optimal non-linear tax a) under full information, b) under a first-order approximation of the conditional bias, and c) under a second-order approximation of the conditional bias. The aggregate welfare effects are presented in Table 1.

[24]We set the scaling factor to 0.6, which ensures that the largest perceived valuations are equal to the highest gradient of the cost function from Figure 2. In addition, we impose some consistency restrictions on the data. Starting with a sample of 633 individuals with non-censored valuations in the treatment groups, we drop all observations with missing values on biases and discount rates (23 observations). We also drop observations where the elicitation of time preferences does not yield a discount factor between 0 and 1 (50 observations) and where biases or perceived valuations are above below the 1 or above the 99 percentile (14 observations). In addition, we drop all observations where perceived valuations would be negative (57 observations), which leaves us with 489 observations for our numerical example.

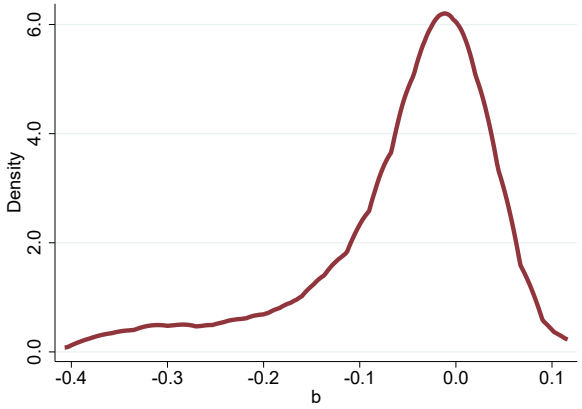
Table 1: Welfare Implications of Taxation

	Mean welfare, in USD/bulb	Mean welfare gain over status quo, in EUR/bulb	Mean welfare gain relative to linear tax, in %
Status quo (no tax)	3.79	0.00	–
Linear tax	3.91	0.12	0
Non-linear tax (first-order approx.)	3.96	0.16	36
Non-linear tax (second-order approx.)	3.97	0.17	44
Non-linear tax (full information)	3.98	0.19	57
Welfare effects under a first-order approximation, (wrongly) setting $\rho = 0$			
Non-linear tax (first-order approx, $\rho = 0$)	3.95	0.16	31

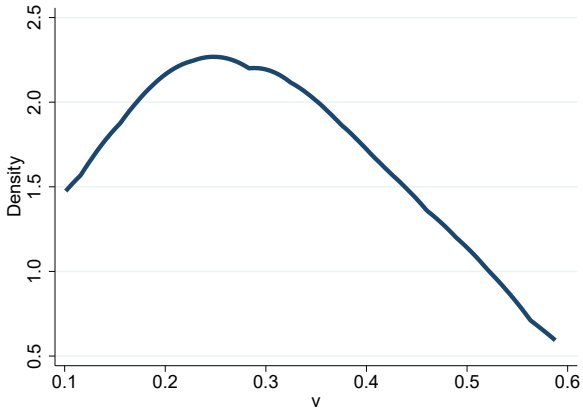
Note: Mean welfare is calculated under the optimal linear and non-linear tax schedules using the joint distribution of perceived valuations and biases, as well as the cost function estimated in Section 5 and Appendix Section B. "Non-linear tax (first-order approx, $\rho = 0$)" implements the optimal non-linear tax based on the first-order approximation to the expected bias, (wrongly) setting $\rho = 0$.

Figure 3: Distribution of bias and normative valuations

(a) Distribution of bias b



(b) Distribution of normative valuation v



Notes for Figure a) and b): Densities estimated via kernel density estimation (Epanechikov kernel, bandwidth: 0.03 and 0.1, respectively).