# Nonlinear Pricing and Misallocation *

Gideon Bornstein†
Alessandra Peter ‡

February 18, 2022.

**Abstract**

This paper studies misallocation across heterogeneous firms and consumers. Contrary to the typical model, we do not restrict firms to charging linear prices. We show that when firms are allowed to set a pricing schedule that depends on quantity, markup heterogeneity is no longer a sign of misallocation. Although larger firms charge higher markups, the allocation of resources across firms is efficient. Further, we point to a new source of misallocation. In general equilibrium, high-taste consumers are allocated too much of each good, low-taste consumers too little. When labor supply is elastic, firms' market power depresses aggregate labor, but this effect is independent of the level of the aggregate markup in the economy. Using micro data from the retail sector, we show that nonlinear pricing is prevalent and quantify the model. We find that the welfare losses from misallocation across consumers under nonlinear pricing are twice as large as those from misallocation across firms under linear pricing.

*We would like to thank Hugo Hopenhayn, Pete Klenow, Virgiliu Midrigan, Ivan Werning, and EAGLS, as well as participants at VMACS, IIES Macro Lunch, Stony Brook, Wharton Macro Lunch, World Bank Research Seminar, Insper, USC Marshall Macro day, the NYU Macro Lunch, and MIT for insightful discussions and comments. We also thank Tanvi Jindal, who provided excellent research assistance. Researchers' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researchers and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

†The Wharton School, University of Pennsylvania; gideonbo@wharton.upenn.edu

‡Department of Economics, New York University; alessandra.peter@nyu.edu

# 1 Introduction

Should the dominance of the largest firms in the economy be restricted? Whether or not the presence of giant firms and their market power calls for government intervention is an ongoing policy debate. The study of misallocation of resources across firms robustly concludes that firms with high market power should be subsidized. That is, the economy's largest firms should be even larger. We show that this seemingly counterintuitive result hinges on a single, commonly made assumption: firms are restricted to offering linear pricing schedules—that is, firms must choose a single per-unit price and sell any quantity at that price.

In this paper, we allow firms to set nonlinear pricing schedules, and we study the implications for misallocation across heterogeneous firms and across consumers with idiosyncratic tastes. In our first key finding, we show that when firms are not restricted to charging linear prices, there is no misallocation of production across firms, even though larger firms charge higher markups. Our second key finding is that under nonlinear pricing, a new type of misallocation arises. For each good, consumers with a high taste consume too much of it, consumers with a low taste too little. When we extend the environment to allow for endogenous labor supply, we show that nonlinear pricing breaks the link between the aggregate markup and the undersupply of labor.

In the final part of the paper, we estimate the model using data from the retail sector. We find that the welfare losses from misallocation across consumers are twice as large as the standard welfare losses from misallocation across firms that arise with linear pricing. If a social planner were to implement the taxes and subsidies that would restore efficiency in an economy in which firms are restricted to charging linear prices, they would induce larger welfare losses than the ones the planner set out to correct.

The model we develop features firms that produce differentiated goods and are heterogeneous in their marginal cost of production. Consumers differ in their idiosyncratic taste for each good. We allow for variable elasticities of substitution in preferences, which, together with cost heterogeneity, gives rise to variable markups. Firms can offer a pricing schedule to consumers—that is, a set of prices that is potentially nonlinear in the quantity purchased. The only restriction we place on firms' pricing behavior is that they must offer the same schedule to all consumers. This assumption reflects legal or practical constraints as well as the possibility that individual consumer preferences are not fully observable to the firm.

Conditional on the aggregate price index, the optimal allocation features the familiar result from the micro theory literature: *no distortion at the top* and *quantity rationing at the bottom*. That is, the high-taste consumer's allocation equates marginal utility with marginal cost, and the low-taste consumer is sold too little of the good. We extend this result by studying a general equilibrium with a continuum of firms that all engage in second-degree price discrimination. Instead of assuming a quasi-linear or outside good, equilibrium is sustained by an aggregate price index that adjusts to clear the labor market. As a result, the allocation of high-taste consumers is also distorted, and there is misallocation across consumers of the same firm: high-taste consumers are allocated too much of the good, whereas those with low taste consume too little.

We next analyze the allocation of production across firms. To do so, we define a condition on preferences: *constant elasticity of taste differential* (CETD). Under this condition, the difference in the efficient allocation between consumer types is proportional to firm productivity. Most commonly used utility functions fall into this class, including CES, CARA, HARA, and quadratic preferences à la Melitz and Ottaviano (2008). We show that under CETD, there is no misallocation of production across firms. In general equilibrium, the oversupply to high-taste consumers exactly offsets the undersupply to low-taste ones. While all firms distort allocations across their consumers, the total production of each firm is identical to the first best.

With nonlinear pricing, there is perfect allocative efficiency across firms, even though larger firms charge higher markups. This result highlights that without the restrictive assumption of linear pricing, the tight link between markups and misallocation breaks. Relatedly, there is no rationale for a social planner to subsidize large, high-markup firms—contrary to the robust conclusion from models that assume linear pricing. In fact, a social planner who has access to a set of fully flexible firm-level subsidies and taxes would choose not to use these. Intuitively, since the amount of over- and undersupply to the different types of consumers is constant across firms, there is no benefit of reallocating labor from one firm to another.

When labor supply is elastic, firms' pricing behavior leads to an inefficiently low level of aggregate labor in equilibrium. Unlike in the standard linear pricing environment, however, the distortion in labor supply does not depend on the aggregate markup. Rather, it is a result of the downward distortion in consumption by low-taste consumers. When choosing the optimal subsidies, a social planner trades off higher sales to low-taste consumers against inefficiently higher sales to high-taste consumers. The resulting optimal subsidy is uniform across firms yet leads to a disproportionately higher increase in employment for smaller firms that charge low markups. Conversely, in the market equilibrium, the employment share of large, high-markup firms is too large.

To explore the quantitative importance of misallocation across consumers, we calibrate the baseline model to micro data on consumer packaged goods from the Nielsen Retail Scanner dataset. Using the same moments, we also calibrate a model that is identical except for the restriction that firms must charge linear prices. The Nielsen dataset has the key advantage that we can see prices paid for different quantities (i.e., package sizes). In this data, we document that nonlinear pricing is prevalent and quantitatively important. Around 90% of sales are accounted for by multi-size products. On average, a 10% increase in package size is associated with a 6% decline in per-unit prices.

Misallocation across consumers of the same firm amounts to welfare losses of 0.8% of permanent consumption. This loss is about twice as large as the welfare losses one would infer from the same data through a standard model with linear pricing. If one were to implement the optimal taxes and subsidies implied by the linear pricing model, this policy would lead to additional welfare losses of 0.4%. Firm-level subsidies do not correct misallocation across consumers of the same firm. Moreover, this policy introduces misallocation across firms by subsidizing large, high-markup firms at the expense of smaller ones.

Finally, we quantify the effect of firms' pricing on aggregate labor supply. In the baseline model, labor is undersupplied by 7%, which leads to additional welfare losses of 0.25%. Strikingly, a standard

linear pricing model would conclude that labor is 47% lower than in the first-best allocation. This is true even though the aggregate markup is actually higher in the baseline model with nonlinear pricing. When implementing optimal linear pricing subsidies, welfare losses would now be nearly 20% because of the massive oversupply of labor induced by the policy.

**Related literature** Our paper is most closely related to the macro literature on markups and misallocation. Recent evidence on the size of markups and their dispersion across firms (De Loecker, Eeckhout, and Unger (2020)) has renewed attention to this topic. On the theoretical side, a robust conclusion emerges: firms that charge high markups are inefficiently small. This is true irrespective of whether markups are modeled as reduced-form distortions (Restuccia and Rogerson (2008); Hsieh and Klenow (2009); Baqaee and Farhi (2020)), arising from oligopolistic competition (Atkeson and Burstein (2008)) or limit pricing (Peters (2020)), or as a result of preferences featuring variable elasticity of substitution (Edmond, Midrigan, and Xu (2021); Boar and Midrigan (2019)). While most of the models do not exactly fit into the Dhingra and Morrow (2019) framework, the conclusion that well-behaved preferences lead to an inverse relationship between markups and size distortions carries through. In this paper, we show that one crucial assumption—linear pricing—is driving all of these results and that relaxing it entirely flips the welfare implications of markup heterogeneity.

The starting point of our analysis is a classic model of second-degree price discrimination that is commonly used in theoretical IO (see Spence (1977); Mussa and Rosen (1978); Maskin and Riley (1984); Tirole (1988); Wilson (1993) and references therein). Relative to this literature, our main contribution is to take the analysis to general equilibrium without relying on a quasi-linear good to close the model. We show that as long as the elasticity of labor supply is finite, the classic result of "no distortion at the top" (as initially discovered by Mirrlees (1971)) no longer holds: consumers with the highest taste for each good are allocated too much in equilibrium.

We explore the quantitative importance of misallocation across consumers using detailed micro data on prices and quantities. Other papers that use micro data to study the behavior of firm-level prices and derive macroeconomic implications include Argente, Lee, and Moreira (2019), Burstein, Carvalho, and Grassi (2020), Bornstein (2021), Afrouzi, Drenik, and Kim (2021), and Einav, Klenow, Levin, and Murciano-Goroff (2021). Contrary to this set of papers, we focus on price heterogeneity within a firm and location—a feature unique to nonlinear pricing.

**Organization** The remainder of the paper is structured as follows. Section 2 lays out the baseline model and defines the market equilibrium and the planner's allocation. Section 3 discusses the main misallocation results of the paper and compares them to a setup with linear pricing. Section 4 extends the model to endogenous labor supply. In Section 5, we introduce the data and quantify the model. Finally, Section 6 concludes.

# 2 Model

## 2.1 Environment

**Households.** The economy is populated by a measure 1 of households $i \in [0,1]$ who supply one unit of labor inelastically. Labor is chosen as the numéraire. Households have idiosyncratic tastes over a measure 1 of varieties of consumption goods $j \in [0,1]$. The level of taste consumer $i$ has for variety $j$ is denoted by $\tau_{ij}$, where a higher $\tau_{ij}$ indicates that household $i$ derives higher utility from good $j$:

$$U_i = \int_0^1 \tau_{ij} u(q_{ij}) dj, \tag{2.1}$$

where $q_{ij}$ denotes the quantity consumed of variety $j$ by household $i$. The utility function $u(\cdot)$ is continuously differentiable, strictly increasing and concave for all $q_{ij} \geq 0$, and satisfies $u(0) = 0$.

For simplicity, the taste shifters $\tau_{ij}$ can take one of two values: 1 or $\tau > 1$.[1] Each consumer has a high preference $\tau$ for a random subset of goods of measure $\pi$. Taste shifters are iid across households and varieties, and therefore all households are identical in their aggregate consumption and utility. All firms in the economy are jointly owned by households. Household income consists of labor earnings as well as any profits rebated by firms.

**Firms.** There is a measure 1 of firms who each produce one of the differentiated varieties $j \in [0,1]$. Firms produce with a linear technology using labor as the only input. They are heterogeneous in their labor cost per unit produced, denoted by $c_j$.

Contrary to the typical assumption in the literature, firms are not restricted to offering a linear pricing schedule (i.e., a commitment to sell any quantity for a constant per-unit price). Firms maximize profits by offering a single menu of prices $p(q)$ to all households. That is, firms engage in second-degree price discrimination. Firms might offer a single menu to all households, rather than tailoring the price schedule $p(q)$ to each individual consumer, for many reasons. For example, household tastes may be unobservable to firms. Alternatively, legal or practical requirements could make it impossible to charge different consumers different prices for the same quantity purchased.[2]

Given the price schedule, each consumer chooses the quantity that maximizes her utility. The firm's problem is given by

$$\begin{aligned} \max_{\{p_j(\cdot), q_{1j}, q_{\tau j}\}} \quad & \pi \left(p_j(q_{\tau j}) - c_j\right) q_{\tau j} + (1-\pi) \left(p_j(q_{1j}) - c_j\right) q_{1j} \\ \text{s.t.} \quad & q_{\tau j} \in \operatorname*{argmax}_{q \geq 0} \ \tau u(q) - \frac{p_j(q) q}{P} \\ & q_{1j} \in \operatorname*{argmax}_{q \geq 0} \ u(q) - \frac{p_j(q) q}{P} \end{aligned} \tag{2.2}$$

[1] Neither the theoretical nor the quantitative results of the paper are sensitive to this assumption. In Online Appendix A, we repeat the analysis for an environment with a continuum of tastes.

[2] In the absence of taste heterogeneity, or if we were to assume that firms can tailor prices to each individual consumer, the model boils down to a model of perfect price discrimination. In this environment, the link between markups and misallocation also breaks. In fact, all allocations—including labor supply—are efficient, irrespective of the level and dispersion of markups.

where $q_{\tau j}$ denotes the quantity purchased by households with a high taste for the good and $q_{1j}$ that of households with low taste. Households evaluate the cost of each quantity, $p_j(q)q$, using the price index $P$. The price index $P$ is an equilibrium outcome that measures the cost of purchasing an additional unit of utility. In Appendix A.3, we show how the aggregate price index $P$ is supported despite preferences not featuring the standard quasi-linearity.

Each of the two constraints in (2.2) is an infinite set of inequalities: the surplus from the quantity the household chooses must be larger than that from any other quantity. To solve (2.2), we use standard tools from mechanism design (see, e.g., Mussa and Rosen (1978)). It is straightforward to show that, in the optimal solution, only two constraints bind: (1) the consumer with a low taste must have non-negative surplus from her bundle—the individual rationality constraint of the low type ($IR_1$); and (2) the high-taste consumer must weakly prefer her bundle to the one tailored toward the low-taste consumer—the incentive compatibility constraint of the high type ($IC_\tau$). The firm's problem can therefore be written as

$$\begin{aligned} \max_{\{q_{1j},q_{\tau j},p_{1j},p_{\tau j}\}} \quad & \pi q_{\tau j}\left(p_{\tau j}-c_j\right)+(1-\pi)q_{1j}\left(p_{1j}-c_j\right) && (2.3)\\ \text{s.t} \quad & u(q_{1j})-\frac{p_{1j}q_{1j}}{P}=0 && [IR_1]\\ & \tau u(q_{\tau j})-\frac{p_{\tau j}q_{\tau j}}{P}=(\tau-1)u(q_{j1}) && [IC_\tau] \end{aligned}$$

where $p_{1j}\equiv p(q_{1j})$ and $p_{\tau j}\equiv p(q_{\tau j})$. The second constraint uses the fact that the individual rationality constraint of the low-taste consumer holds with equality.[3]

**Firm-level optimal prices and quantities.** Conditional on the aggregate price index $P$, which firms take as given, the quantities offered to high- and low-taste consumers respectively solve

$$\tau\, u'(q_{\tau j})=\frac{c_j}{P}, \tag{2.4}$$

$$u'(q_{1j})=\frac{c_j}{P}+\frac{\pi}{1-\pi}(\tau-1)u'(q_{1j})=\frac{1-\pi}{1-\tau\pi}\,\frac{c_j}{P}. \tag{2.5}$$

Equation (2.4) pins down the optimal quantity sold to high-taste consumers. Firms always choose a bundle that equates the marginal revenue to the marginal cost of production. Marginal revenue is equal to the marginal utility, as that is the consumer's willingness to pay for an additional unit. The fact that marginal utility equals marginal cost for the high-taste consumer is reminiscent of the standard result of "no distortion at the top" in models of second-degree price discrimination. Since low-taste consumers have no incentive to choose the larger quantity designated for the high-taste consumer, there is no need to distort the allocation at the top. In our setup, however, the "no distortion at the top" result only holds *conditional* on the aggregate price index, which, as we show below, is not equal to the price index prevailing in the efficient allocation.

[3] While the firm's problem pins down $p_{1j}$ and $p_{\tau j}$, the prices the firm charges for quantities that are not purchased in equilibrium are indeterminate. Firms can charge arbitrary prices for $q_j \notin \{q_{1j}, q_{\tau j}\}$ as long as neither of the two consumer types wants to deviate and purchase that quantity.

Equation (2.5) pins down the optimal quantity sold to low-taste consumers.[4] The optimal quantity again equates marginal revenue with marginal cost. However, the marginal cost now includes not only the real production cost $c_j/P$ but also the *shadow cost* of ensuring separation between the two types. For each additional unit sold to the low-taste consumer, firms need to lower the charges to high-taste ones by $(\tau - 1)u'(q_{1j})$ in order for them to remain indifferent between the two bundles. This corresponds to the increase in consumer surplus for the high type: each additional unit of $q_{1j}$ increases the high-taste consumer's utility by $\tau u'(q_{1j})$, whereas the cost of this bundle can go up only by $u'(q_{1j})$—the low type's additional utility. This shadow cost is weighted by the relative share of high-taste consumers, $\pi/(1-\pi)$. Overall, the distortion creates a wedge between the marginal utility of the low-taste consumer and the marginal production cost equal to $(1-\pi)/(1-\tau\pi) > 1$.

Firms are able to extract the full consumer surplus of low-taste customers. Customers with a high taste, on the other hand, have a positive consumer surplus. Self-selection of each type into their respective bundles is achieved by distorting the allocation of the low-taste consumer and reducing the price charged to the high-taste consumer. The quantity distortion—the wedge in the marginal utility of the low-taste consumer—is increasing in both the share of high-taste consumers $\pi$ and the taste difference $\tau$. The more high-taste consumers the firm faces, and the higher the relative taste for the good, the more costly it is for firms to achieve separation by lowering the price charged to high-taste consumers, and the more they choose to distort the low-taste consumer's allocation.

Define the markups charged to low-taste consumers as $\mu_{1j} \equiv \frac{p_{1j}}{c_j}$ and that charged to high-taste consumers as $\mu_{\tau j} \equiv \frac{p_{\tau j}}{c_j}$. The equilibrium markups charged by firms can be written as

$$\mu_{1j} = \frac{1-\pi}{1-\tau\pi}\,\psi\left(q_{1j}\right), \tag{2.6}$$

$$\mu_{\tau j} = \left(1 - (\tau - 1)\frac{u(q_{1j})}{\tau u(q_{\tau j})}\right)\,\psi(q_\tau j), \tag{2.7}$$

where $\psi(q)$ is defined by

$$\psi(q) \equiv \frac{u(q)}{qu'(q)}. \tag{2.8}$$

The term $\psi(q)$ is the *social markup*, a term coined by Dhingra and Morrow (2019). It is equal to the utility per unit produced, $u(q)/q$, relative to the resource cost of producing a unit in the efficient allocation. In the efficient allocation, the planner equates marginal utility with marginal cost, so $u'(q)$ is equal to the resource cost of producing one unit.

If firms could perfectly price discriminate, they would extract the full consumer surplus from each of their consumers. The markup charged from each consumer would be equal to the social markup $\psi(q_{ij})$. This is not the case with nonlinear pricing.

The markup the firm charges to low-taste consumers, (2.6), is higher than the social markup. Low-taste consumers are willing to pay this higher markup because the quantity offered to them is

[4] Note we have assumed that preferences as well as the distribution of $c_j$ are such that all firms are active and optimally choose to serve both types.

distorted downward. Since utility is concave, the average utility of a unit consumed is higher.

For high-taste consumers, the chosen markup of the firm, (2.7), is lower than the social markup. The markup has to be lower than the social markup; otherwise, the high-taste consumer would choose the low-taste bundle instead. From Equation (2.4), we know that the quantity sold to the high-taste consumer is identical to the case of a perfectly price discriminating monopolist. Therefore, if the firm were to charge the social markup, it would extract the entire consumer surplus, violating the incentive compatibility constraint. The chosen markup makes high-taste consumers exactly indifferent between their own bundles and those of low-taste consumers.

**Equilibrium** Given a distribution of production costs across firms, $F(c_j)$, an equilibrium is a set of firm-level prices $\{p_{1j}, p_{\tau j}\}_{j=0}^{1}$ and quantities $\{q_{1j}, q_{\tau j}\}_{j=0}^{1}$ as well as an aggregate price index $P$, such that prices and quantities solve the firm's problem and the labor market clears:

$$\int_0^\infty \left[\pi q_{\tau j} + (1-\pi) q_{1j}\right] \; c_j \, dF(c_j) = 1. \tag{2.9}$$

## 2.2 Efficient allocation

In this section, we derive the efficient allocation by solving the problem of a utilitarian social planner who chooses allocations subject to the same production technology. The planner solves

$$\begin{aligned} \max_{q_{ij}} \quad & \int_i \int_j \tau_{ij} u(q_{ij}) dj di \\ \text{s.t.} \quad & \int_i \int_j q_{ij} c_j dj di = 1 \end{aligned} \tag{2.10}$$

The optimal allocations are given by

$$u'(q_{ij}^{\text{FB}}) = \frac{c_j}{\tau_{ij}} \frac{1}{P^{\text{FB}}}, \tag{2.11}$$

where $P^{\text{FB}}$ is the inverse Lagrange multiplier on the aggregate resource constraint.

The equation above implies that in the optimal allocation, the marginal utilities of all consumers of a given variety are equalized:

$$\frac{\tau u'(q_{\tau j}^{\text{FB}})}{u'(q_{1j}^{\text{FB}})} = 1. \tag{2.12}$$

Further, the relative marginal utility of two different varieties is equal to the relative marginal costs of production:

$$\frac{u'(q_{\tau j}^{\text{FB}})}{u'(q_{\tau k}^{\text{FB}})} = \frac{u'(q_{1j}^{\text{FB}})}{u'(q_{1k}^{\text{FB}})} = \frac{c_k}{c_j}. \tag{2.13}$$

# 3 Misallocation within and across firms

In this section, we analyze allocative efficiency along two dimensions: within firms across consumers and across firms. We then solve for the optimal firm-level taxes and subsidies a social planner would set and compare the misallocation results to a version of the model in which firms are restricted to set linear prices.

## 3.1 Misallocation within firms

We start by analyzing the allocation of consumption across different types of consumers within a firm. Comparing the efficient allocation, (2.12) and (2.13), to the market allocation, (2.5) and (2.4), we obtain the following relationship:

$$\frac{1-\tau\pi}{1-\pi} = \frac{\tau u'(q_{\tau j})}{u'(q_{1j})} < \frac{\tau u'(q_{\tau j}^{\mathrm{FB}})}{u'(q_{1j}^{\mathrm{FB}})} = 1. \tag{3.1}$$

Compared to the efficient benchmark, the relative marginal utilities are distorted in the market allocation. The distortion comes from the wedge in marginal utilities discussed in the previous section: firms distort low-taste quantities downward in order to extract more from high-taste consumers. As a result, the marginal utility of low-taste consumers is higher than that of high-taste consumers.

This result is familiar from the micro theory literature.[5] In partial equilibrium, i.e., conditional on the aggregate price index $P$, we know that the distortion in relative marginal utilities comes from a combination of no distortion at the top and quantity rationing at the bottom. However, under this aggregate price index, all firms employ less labor than in the efficient allocation, and the labor market cannot clear.

In general equilibrium, the aggregate price index $P$ must therefore be higher than in the efficient allocation in order to induce all firms to produce more and hire more workers. The resulting allocation features not only too little consumption by low-taste consumers, whose quantity is directly distorted downward, but also too much consumption by high-taste consumers. The standard result of "no distortion at the top" no longer holds in general equilibrium, as formalized in the following proposition.

Proposition **1.** *In equilibrium, households consume too much of the goods for which they have a high taste and too little of the goods for which they have a low taste.*

All proofs are relegated to Appendix A.

## 3.2 Misallocation across firms

Next we turn to the question of misallocation across firms. Here, we proceed in two steps. The first is positive: we compare *overall* production, and hence employment, of firm $j$ to the efficient allocation. The second is normative: we consider the problem of a social planner who has access to firm-specific production taxes and subsidies.

[5] See Mirrlees (1971) or Tirole (1988) and references therein.

Firm-level production is equal to a weighted average of quantities sold to each type of consumer, $q_j = \pi q_{1j} + (1-\pi)q_{\tau j}$. From Proposition 1, we know that, relative to the efficient allocation, $q_{1j}$ is too small and $q_{\tau j}$ too large. The *aggregate* labor employed by all firms together is identical to the first best, as guaranteed by the general equilibrium price index $P$. Here, we are interested in each *individual* firm's production relative to the first best—that is, for which firms the undersupply to the low type outweighs the oversupply to the high type and vice versa. To this end, it is helpful to define the following property of preferences.

Definition **1** (Elasticity of taste differential)**.** Let $q_{ij}^{FB}$ be consumer $i$'s allocation of good $j$ in the first best. Define the taste differential of good $j$ as $q_{\tau j}^{FB} - q_{1j}^{FB}$. We then define the elasticity of taste differential as

$$\eta(\mathrm{mc}_j, \tau) := \frac{\partial \log(q_{\tau j}^{FB} - q_{1j}^{FB})}{\partial \log(\mathrm{mc}_j)},$$

where $\mathrm{mc}_j$ is the real marginal cost of firm $j$, $c_j/P^{FB}$.

The elasticity of taste differential measures how the optimal consumption difference between high- and low-taste consumers of a particular good varies with the cost of producing that good.[6] If the elasticity is equal to zero, then the optimal consumption difference between the two types of consumers is equal across all goods. High-taste consumers are always allocated a constant extra quantity. When the elasticity is negative, the optimal consumption difference is lower for high-cost goods.

Since $q_{ij}^{FB} = (u')^{-1}\left(\frac{1}{\tau_{ij}}\frac{c_j}{P^{FB}}\right)$, the elasticity of taste differential is ultimately a function of the inverse marginal utility:

$$\eta(x, \tau) = \frac{\partial \log(u'^{-1}(x/\tau) - u'^{-1}(x))}{\partial \log x}. \tag{3.2}$$

**Firm-level production relative to the efficient allocation.** In general, the oversupply to high types and undersupply to low types could lead to arbitrary patterns of firm-level output relative to the first best as a function of productivity. In Proposition 2, we show one of the key results of the paper: for a large class of preferences, defined formally in Assumption 1, the two effects exactly offset one another. That is, all firms produce precisely the same total quantity using the same amount of labor as in the efficient allocation.

Assumption **1.** Preferences $u(\cdot)$ exhibit constant elasticity of taste differential. That is:

$$\eta(\mathrm{mc}_j, \tau) = \eta, \quad \forall\{\mathrm{mc}_j, \tau\}.$$

We further assume that $\eta > 1$ so that optimal markups are finite.

Proposition **2.** *Suppose preferences satisfy Assumption 1. Then, the equilibrium levels of firm-level production and employment are identical to the efficient allocation.*

[6] While we defined the elasticity in terms of *taste* differential, one could equally interpret it as an elasticity of *price* differential. Under the latter interpretation, $\eta(x, \delta)$ measures the elasticity of the difference in consumption between two varieties whose marginal costs differ by a factor of $\delta$.

Note that Assumption 1 nests a large class of utility functions: CES, quadratic preferences à la Melitz and Ottaviano (2008), constant absolute risk aversion (CARA), as well as preferences in the hazard analysis and risk assessment class (HARA). When preferences feature constant elasticity of taste differential (henceforth, CETD), the difference in consumption between high- and low-taste consumers is proportional to firm productivity.

Since this result is one of the main results of the paper, we sketch its proof as well as the intuition behind it in the main text. Consider first an economy with only one firm that has a given marginal cost $c_j$. In general equilibrium, it must be that the total labor employed by this firm equals the total labor employed for production in the planner's allocation. The aggregate price index, $\tilde{P}_j$, that guarantees labor market clearing in equilibrium is implicitly defined as

$$\pi \left[ q_{\tau j}(\tilde{P}_j) - q_{\tau j}^{\mathrm{FB}} \right] - (1-\pi) \left[ q_{1j}^{\mathrm{FB}} - q_{1j}(\tilde{P}_j) \right] = 0. \tag{3.3}$$

The core of the argument lies in showing that this price index $\tilde{P}_j$ does not depend on firm productivity. That is, whichever aggregate price index guarantees that the oversupply to high types exactly offsets the undersupply to low types for a firm with a given $c_j$ will equate the two for all firms.

Note that Assumption 1 implies that $\partial \log(q_{\tau j} - q_{\tau j}^{FB}) / \partial \log(\mathrm{c}_j) = \eta$. The reasoning is that the market allocation $q_{ij}$ depends on the inverse marginal utility in the same way as the first-best allocation. Relabeling the arguments in Equation (3.2) as $x = c_j/(\tau \tilde{P}_j)$ and $\tau = \tilde{P}_j / P^{FB}$, the result follows. Relative to the planner allocation, the market behaves *as if* preferences of the high type were shifted by $\tilde{P}_j / P^{FB}$. Since there is constant elasticity of taste differential, the elasticity of the difference between planner and market allocation is also constant. Similarly, $\partial \log(q_{1j}^{FB} - q_{1j}) / \partial \log(\mathrm{c}_j) = \eta$.

Now consider a firm with $c_k = (1+\Delta) c_j$. Using Assumption 1,

$$\begin{aligned} &\pi \left( q_{\tau,k}(\tilde{P}_j) - q_{\tau,k}^{FB} \right) \qquad\qquad - (1-\pi) \left( q_{1,k}^{FB} - q_{1,k}(\tilde{P}_j) \right) = \\ &\pi (1+\Delta)^{\eta} \left( q_{\tau,j}(\tilde{P}_j) - q_{\tau,j}^{FB} \right) - (1-\pi)(1+\Delta)^{\eta} \left( q_{1,j}^{FB} - q_{1,j}(\tilde{P}_j) \right) = 0. \end{aligned}$$

When the difference in quantities sold to the two types of consumers scales proportionately with costs, under- and oversupply relative to the first best are also proportional to cost. Therefore, in order for overall labor to be neither too high nor too low, each firm's overall labor demand must be identical to the first best.

**Taxes and Subsidies.** Consider the problem of a social planner who maximizes utilitarian social welfare.[7] Instead of directly choosing allocations, she has access to *firm-specific* taxes and subsidies $t_j$. She cannot, however, levy consumer-specific taxes. We consider this restricted problem for two reasons. The first is realism, as the implementation of a firm- and consumer-specific subsidy might not be feasible. The second is that, with access to fully flexible tax instruments at the consumer level, the planner could simply implement the efficient allocation.

We model the taxes (or subsidies) set by the social planner as production taxes. When the planner

[7] Given that taste shifters are iid across firms and consumers, this is akin to maximizing the utility of a representative consumer.

levies a tax $t_j$ on firm $j$, its marginal cost becomes $c_j(1+t_j)$. We allow the planner to impose lump-sum taxes on households, so that they can uniformly subsidize or tax all firms while maintaining a balanced budget.

The planner chooses firm-level taxes $t_j$ as well as both consumers' allocation from each firm $q_{ij}$ to maximize welfare, anticipating the resulting bundles firms will offer to consumers:

$$\begin{aligned} \max_{\{t_j, q_{1j}, q_{\tau j}, P\}} \quad & \int_0^1 \pi\tau u(q_{\tau j}) + (1-\pi)u(q_{1j})dj \\ \text{s.t.} \quad & q_{\tau j} = (u')^{-1}\left(\frac{c_j(1+t_j)}{\tau P}\right), \quad \forall j \\ & q_{1j} = (u')^{-1}\left(\frac{1-\pi}{1-\tau\pi}\frac{c_j(1+t_j)}{P}\right), \quad \forall j \\ & 1 = \int_0^1 c_j\left(\pi q_{\tau j} + (1-\pi)q_{1j}\right) \end{aligned} \tag{3.4}$$

From Proposition 2, we know that the total production of the firm is equal to its level in the efficient allocation. However, there is misallocation across consumers of each firm, as Proposition 1 shows. The question facing the social planner is therefore whether firm-level taxes or subsidies can be used to indirectly affect misallocation across consumers.

Whenever the planner subsidizes a firm, its production increases for both the low type, who was consuming too little, and the high type, who was already consuming too much. At the same time, in order for the planner's budget to balance, the planner must levy a tax on another firm, which will further decrease sales to the low-taste consumer. As long as the benefits from such a subsidy are heterogeneous across firms, the social planner will be able to use firm-level taxes to alleviate some of the misallocation across consumers.

We show in Proposition 3 below, however, that under the maintained assumption on preferences, within-firm misallocation cannot be mitigated using firm-level taxes. In fact, the optimal taxes set by the social planner are identically zero.[8]

PROPOSITION **3.** *Suppose preferences satisfy Assumption 1. Then, imposing no subsidies and taxes at the firm level is optimal.*

To understand why the planner does not want to impose any taxes, consider the welfare gains of adding a single unit of labor to a firm with cost $c_j$, starting from the equilibrium allocation—that is, from an environment without taxes. We will show that the welfare benefits of adding one more worker to a firm are equal across all firms in general equilibrium. Since there is no heterogeneity in benefits across firms, the planner has no incentive to impose taxes and reallocate labor across firms.

Adding an additional unit of labor to firm $j$ results in an additional $1/c_j$ units of production for firm $j$. While the planner can choose firm-level production, she has no control over how the firm allocates consumption across the two types of consumers. Denote by $\gamma_j$ the share of additional production

[8] Since labor supply is inelastic, scaling up the lump-sum transfer and all firm-level taxes in a budget-neutral way does not affect allocations. We therefore assume that the planner chooses the implementation with zero lump-sum transfers.

allocated to high-taste consumers. The welfare gains can then be expressed as

$$\frac{\partial \left[\int U_i di\right]}{\partial l_j} = \gamma_j \frac{1}{c_j} \tau u'(q_{\tau j}) + (1-\gamma_j) \frac{1}{c_j} u'(q_{1j}). \tag{3.5}$$

The term $\gamma_j \frac{1}{c_j}$ denotes the additional consumption units allocated to high-taste consumers of firm $j$, and $\tau u'(q_{\tau j})$ is their marginal utility of an additional consumption unit. In general equilibrium we know that

$$\tau u'(q_{\tau j}) = \frac{c_j}{P}, \tag{3.6}$$

$$u'(q_{1j}) = \frac{1-\pi}{1-\tau\pi} \frac{c_j}{P}, \tag{3.7}$$

so that the welfare benefits of adding an additional unit of labor to firm $j$ are equal to

$$\left.\frac{\partial \left[\int U_i di\right]}{\partial l_j}\right|_{GE} = \left(\gamma_j + (1-\gamma_j)\frac{1-\pi}{1-\tau\pi}\right) \frac{1}{P}. \tag{3.8}$$

If $\gamma_j$ varies across firms, the planner has an incentive to shift production toward firms with a low $\gamma_j$. These firms would allocate more of their additional production to low-taste consumers, whose allocation is distorted downward. However, in the proof to Proposition 3 we show that when preferences feature constant elasticity of taste differential, $\gamma_j$ is constant across firms. Firms therefore have no incentive to reallocate production, and their taxes and subsidies are zero.

### 3.3 Comparison to linear pricing

In this section, we compare our main results to a model in which firms offer linear prices, as is standard in the literature. All other elements of the model remain as laid out previously. Firms are now restricted to offering a single per-unit price $p_j$, and they commit to selling any quantity $q_{ij}$ to consumers at that price. Firms therefore solve the following problem:

$$\begin{aligned} \max_{\{p_j, q_{1j}, q_{\tau j}\}} \quad & \left(\pi q_{\tau j} + (1-\pi) q_{1j}\right)(p_j - c_j) \\ \text{s.t.} \quad & \tau u'(q_{\tau j}) = \frac{p_j}{P}, \qquad (3.9) \\ & u'(q_{1j}) = \frac{p_j}{P}. \qquad (3.10) \end{aligned}$$

where $P$ is the aggregate price index. Equilibrium in this economy is defined analogously to the nonlinear pricing case and relegated to Appendix B.

**No misallocation within firms.** From the two demand curves (3.9)–(3.10), it follows directly that marginal utilities are equal across the two types of consumers. That is, there is *no misallocation within firms*, and a social planner cannot improve welfare by reallocating production of a firm across its consumers. This result is the first main difference relative to the nonlinear pricing economy. Recall

that Proposition 1 states that under nonlinear pricing, reallocating a firm's production from high-taste to low-taste consumers raises welfare.

The intuition for the difference is straightforward. With linear pricing, both types of consumers equate their marginal utility with the real price of the good. Since both types of consumers face the same price, their marginal utilities are equal. With nonlinear pricing, firms ensure separation between the two types of consumers by restricting the quantity sold to the low type, increasing its marginal utility relative to the high type.[9]

**Misallocation across firms.** Under the linear pricing assumption, allocative efficiency is closely tied to markup heterogeneity. In the efficient allocation, the ratio of marginal utility to production costs, $\tau_{ij}u'\left(q_{ij}^{\text{FB}}\right)/c_j$, is equated across all goods and consumers. In the linear pricing equilibrium, we have that

$$\left(\frac{\tau_{ij}u'\left(q_{ij}\right)}{c_j}\right) \Big/ \left(\frac{\tau_{lk}u'\left(q_{lk}\right)}{c_k}\right) = \frac{\mu_j}{\mu_k}, \qquad \forall\left\{(i,l) \in \{1,\tau\}, (j,k) \in [0,1]\right\} \tag{3.11}$$

where $\mu_j \equiv \frac{p_j}{c_j}$ is the markup of firm $j$. We obtain equation (3.11) by using the demand function of each consumer together with the definition of markups.

When all firms charge the same markup, the ratios of marginal utility to cost are equal across consumers and goods, and the equilibrium allocation coincides with the efficient allocation.[10] When markups are heterogeneous, there is misallocation across firms. Firms that charge higher markups are underproducing, whereas firms with relatively lower markups are overproducing. As a result, a planner needs to subsidize high-markup firms and tax low-markup firms in order to implement the efficient allocation.

The equilibrium markups across firms depend on consumer preferences. The optimal markup charged by firm $j$ is a function of the effective demand elasticity faced by the firm, which we denote by $\epsilon_j$:

$$\mu_j = \frac{\epsilon_j}{\epsilon_j - 1}. \tag{3.12}$$

The effective demand elasticity, $\epsilon_j$, is a weighted average of the demand elasticities of the two consumers:

$$\epsilon_j = \alpha_j\, \epsilon(q_{\tau j}) + (1-\alpha_j)\, \epsilon(q_{1j}), \tag{3.13}$$

where $\epsilon(q) \equiv -\frac{u'(q)}{qu''(q)}$ is the inverse elasticity of marginal utility and $\alpha_j \equiv \frac{\pi q_{\tau j}}{\pi q_{\tau j} + (1-\pi) q_{1j}}$ is the high-taste consumers' share of sales.

As in Dhingra and Morrow (2019), there is misallocation across firms if and only if the elasticity of demand varies with the quantity sold (i.e., $\epsilon(q)$ is not constant). This result is summarized in the

[9] Note that this result does not rely on the two-types setup, in which consumers make a discrete choice instead of a marginal one. With a continuum of types, consumers would equate the marginal utility with the marginal price of an additional unit, similarly to the linear pricing case. However, firms would set non-constant marginal prices, leading to misallocation across types.

[10] All formal derivations are relegated to Appendix A.

following proposition.

PROPOSITION **4.** *If preferences exhibit variable elasticity of substitution, there is misallocation across firms in the linear pricing equilibrium. In particular, if the elasticity is decreasing in the quantity consumed:*

1. *Firms with higher productivity ($\frac{1}{c_j}$) charge higher markups.*
2. *Firms that charge high markups sell too little relative to the efficient allocation.*
3. *The optimal firm-level subsidies are increasing in productivity.*

Proposition 12 confirms that the classic result of the macro literature on markups and misallocation also holds in our setup with consumer heterogeneity. The relationship between demand elasticity and quantity consumed is common to macro models with variable markups (e.g., Baqaee and Farhi, 2020; Edmond et al., 2021). In our setup, Assumption 1 implies that the demand elasticity is weakly decreasing in quantity. As a result, there is a positive relationship between firm size and markups.

High-productivity firms are larger, face a lower demand elasticity, and charge higher markups. Because they charge higher markups, these large, highly productive firms are *too small* relative to the efficient allocation. Another way to view this result is that they restrict supply in order to keep prices high. A welfare-maximizing social planner would tax small and medium-sized firms, which charge relatively low markups, and use the revenues to subsidize the largest firms in the economy.

This classic result is in stark contrast to the economy we study in this paper, in which firms are not restricted to linear pricing. Note that in the baseline economy with nonlinear pricing, there is markup heterogeneity as well. The higher a firm's productivity, the more it sells and the higher the markup it charges to consumers (Proposition 9 in Appendix A formalizes this). There is, however, no misallocation across firms. As a result, observing large firms that charge high markups does not imply that these firms should be subsidized. In fact, as long as pricing is not artificially restricted to be linear, subsidizing large, high-markup firms increases misallocation and leads to welfare losses.

# 4 Elastic labor supply

In the previous sections, we assumed that aggregate labor supply is inelastic (i.e., the Frisch elasticity is zero). This implied that the aggregate price index had to adjust to clear the labor market. In this section, we extend our analysis to the case of elastic labor supply, in which the deviation from perfect competition also has the potential to distort the overall production level in the economy. We analyze the polar opposite case of perfectly elastic labor supply, in which the real wage and hence the aggregate price index are constant.

**Setup.** The disutility of labor is linear in the amount of labor supplied, so that household utility is given by

$$U_i = \int_0^1 \tau_{ij} u(q_{ij}) dj - \nu l_i, \tag{4.1}$$

where $l_i$ is the amount of labor supplied by household $i$ and $\nu$ governs the degree of disutility of labor. The linear disutility of labor implies that the aggregate price index ($P$) must be equal to $1/\nu$ in equilibrium. Any other aggregate price index would not be consistent with an interior solution. We denote the aggregate level of labor by $L \equiv \int_0^1 l_i di$.

Given the aggregate price index, $P$, the firm's problem and equilibrium allocations are identical to the ones presented in Section 2.

**Equilibrium** Given a distribution of production costs across firms, $F(c_j)$, an equilibrium is a set of firm-level prices $\{p_{1j}, p_{\tau j}\}_{j=0}^1$ and quantities $\{q_{1j}, q_{\tau j}\}_{j=0}^1$ as well as an aggregate price index $P$ and labor $L$ such that $P = \frac{1}{\nu}$, prices and quantities solve the firm's problem, and the labor market clears:

$$\int_0^\infty (q_{\tau j} + q_{1j})\ c_j\ dF(c_j) = L.$$

## 4.1 Efficient allocation

The social planner solves the following problem:

$$\begin{aligned} \max_{l_i, q_{ij}} \quad & \int_i \int_j \tau_{ij} u(q_{ij}) dj di - \nu \int_i l_i di \\ \text{s.t.} \quad & \int_i \int_j q_{ij} c_j dj di = \int_i l_i di. \end{aligned}$$

The optimal allocations are given by

$$u'(q_{ij}^{\text{FB}}) = \frac{c_j}{\tau_{ij}}\ \nu. \tag{4.2}$$

As in the case of inelastic labor supply, the equation above implies that in the optimal allocation, the marginal utilities of all consumers of a given variety are equalized and that the relative marginal utility of two different varieties is equal to the relative marginal costs of production.

## 4.2 Undersupply of aggregate labor

We now revisit our analysis with elastic labor supply. Note that the relative misallocation within firms remains unchanged, since Equation (3.1) is identical under elastic labor supply. We start by comparing the aggregate level of labor in equilibrium and in the efficient allocation.

PROPOSITION **5.** *The aggregate level of labor in equilibrium, $L$, is lower than the aggregate level of labor in the efficient allocation.*

With an infinite Frisch elasticity, aggregate labor is determined purely by labor demand. The aggregate price index $P$ must equal the disutility of labor $1/\nu$. This implies that we can directly compare the efficient allocation, (4.2), with the firm's optimality conditions from Section 3, (2.4)–(2.5). Firms sell the efficient quantity to high-taste consumers and distort downward the quantity

sold to low-taste consumers. Overall firm production, and hence labor demand, are therefore lower in equilibrium than in the efficient allocation.[11]

The decline in aggregate labor leads to a change in the distribution of market shares across firms. When labor supply is inelastic, Proposition 2 implies that the equilibrium market share of all firms is identical to their market share in the efficient allocation. This is no longer the case with elastic labor supply.

PROPOSITION **6.** *Let a firm's* excess employment share *be the ratio between its equilibrium employment share and the employment share in the efficient allocation. Suppose preferences satisfy Assumption 1. Then, excess employment shares are increasing in firm productivity.*

As discussed in the previous paragraph, firms sell the efficient quantity to high-taste consumers but distort their sales to low-taste consumers downward in order to achieve separation between types. The intuition behind Proposition 6 is closely related to the magnitude of this distortion across different firms.

Consider the firm's optimality condition for the quantity sold to low-taste consumers:

$$u'\left(q_{1j}\right) = \frac{1-\pi}{1-\pi\tau}\nu c_j, \tag{4.3}$$

which we can compare to production in the first-best allocation:

$$u'\left(q_{1j}^{FB}\right) = \nu c_j. \tag{4.4}$$

Equations (4.3)–(4.4) imply that with endogenous labor supply, allocations offered by the firm feature a constant distortion in marginal utilities relative to the efficient allocation. The distortion in quantities that corresponds to a specific distortion in marginal utilities is directly related to the demand elasticity:

$$\frac{\partial \log(u'(q))}{\partial \log(q)} = -\frac{1}{\epsilon(q)} \tag{4.5}$$

When demand is highly elastic, as is the case when consumers buy small quantities, then the marginal utility changes slowly with the quantity consumed. Therefore, firms distort sales to the low type by a greater amount. Productive firms, on the other hand, sell large quantities and face relatively inelastic demand. Marginal utilities decline faster, and hence only a small distortion in quantities is necessary to achieve the same ratio of marginal utilities.

In the previous section, with inelastic labor supply, we showed that under linear pricing, the employment shares of high-productivity firms are too small relative to the efficient allocation. We compared it to the nonlinear pricing equilibrium, in which the employment shares of high-productivity firms are identical to the efficient allocation. With elastic labor supply, Proposition 6 takes a step further. The employment share of high-productivity firms is *higher* relative to the efficient allocation.

[11] With a finite Frisch elasticity, the aggregate price index would adjust upward in general equilibrium. As long as the Frisch elasticity is strictly positive, the adjustment would be lower than in the inelastic labor supply case studied in the previous section, and the equilibrium level of aggregate labor would be lower than in the efficient allocation.

**Optimal taxes and sfubsidies.** Consider the problem of a planner who can impose firm-level taxes and subsidies and use lump-sum transfers. Proposition 6 implies that the planner would like to set taxes and subsidies to not only increase aggregate labor supply but also allocate relatively more new workers to the smaller firms. We show that, in order to achieve this increase in the employment share of small firms, the planner optimally imposes a *uniform* subsidy.

PROPOSITION **7.** *Suppose preferences satisfy Assumption 1. Then, the optimal firm-level subsidies are positive and constant across firms.*

We start by discussing the planner's incentive to impose a subsidy and then explain why she does so in a homogeneous way. In the market equilibrium, firms sell the optimal quantity to high-taste consumers but sell too little to low-taste ones. As the planner starts subsidizing production, firms increase their sales to both types of consumers. Initially, this change increases welfare, since the benefits of increasing the consumption of distorted low-taste consumers outweigh the costs of increasing the quantity sold to high-taste consumers, whose allocation was efficient. The optimal level of subsidy trades off these benefits and costs. The resulting aggregate level of labor is therefore higher than in the market equilibrium but lower than in the efficient allocation.

The uniform subsidy the planner imposes increases the employment share of small firms, whereas that of large firms goes down. The uniform subsidy partially offsets the uniform wedge between the marginal utility and the marginal production cost that the nonlinear pricing equilibrium introduces. As aggregate employment in the constrained efficient allocation is smaller than in the efficient allocation, employment shares across firms are also not identical to the efficient allocation. The employment share of small firms remains lower than in the efficient allocation.

Similar to the case of inelastic labor supply (Proposition 3), the planner has no incentive to impose heterogeneous subsidies. The share of additional production induced by a subsidy that goes to low-taste consumers is identical across firms. Therefore, reallocating a worker from one firm to another does not raise welfare.Although employment shares in the market equilibrium are different from those in the efficient allocation (Proposition 6), the planner does not have an incentive to impose heterogeneous subsidies.

## 5 Quantitative analysis

In this section, we explore the magnitude of welfare losses from misallocation under nonlinear pricing. We focus on retail sector goods since these data allow us to observe how prices of the same product vary by quantity sold: the package size. We start by showing that nonlinear pricing is abundant and quantitatively significant. Goods that are offered for more than one size account for more than 90% of sales, and the price per unit declines on average by 6% when product size is 10% larger.

We then use product-level data on sales and purchases to calibrate the structural parameters of our model. We compare the size of misallocation to a counterfactual environment in which firms are restricted to linear pricing schedules. We find that the welfare costs of misallocation under nonlinear

Table 1: Summary Statistics

| | |
|---|---|
| Number of products | 165,053 |
| Number of product modules | 552 |
| Number of product lines | 41,950 |
| Share of sales in multi-size product lines | 90.5% |
| Share of UPCs in multi-size product lines | 71.3% |

**Notes:** This table reports summary statistics of the dataset. Products are at the UPC level. Product line is the collection of products of the same brand sold in the same product module.

pricing are about twice as large as those under linear pricing. Moreover, implementing a tax system that would eliminate misallocation under linear pricing significantly worsens misallocation under nonlinear pricing.

Finally, we study the inefficiency from the distortion in aggregate labor supply in the nonlinear and linear pricing environments. While both environments feature a large average markup, the distortion in aggregate labor is an order of magnitude larger under linear pricing. Nonlinear pricing breaks the link between aggregate markups and the aggregate labor supply.

### 5.1 Data and descriptive statistics

We use Nielsen Retail Scanner Data provided by Kilts Center at the University of Chicago to conduct our analysis. The dataset contains information on average weekly product-level pricing and sales in over 35,000 stores.[12] We focus on core grocery goods, which include the departments of dry groceries, frozen food, and dairy.[13] We use data from a single week in 2017.[14]

In addition to data on pricing and sales, the dataset includes information on product characteristics. In particular, for each product, we observe its product module (e.g., "popcorn - popped"), the brand (e.g., "Skinny Pop"), and its size (e.g., 4.4 oz). We define a product line to be a set of products that share the same brand and product module. For example, products of different sizes under the brand "Skinny Pop" in the "popcorn - popped" product module are all of the same product line.

Before turning to the calibration of our model, we show that nonlinear pricing is abundant in the data. We document two features of the data. First, the vast majority of product lines contain more than one size: 90.5% of sales and 71.3% of products are in product lines that offer at least two size options. Table 1 presents these statistics along with other summary statistics.

Second, within product lines, the price per unit declines significantly with product size. We run the following regression:

$$\ln p_{ujs} = \beta \ln q_{ujs} + \Gamma \mathbf{X}_{ujs} + \epsilon_{ujs}, \tag{5.1}$$

where $p_{ujs}$ is the price per unit of product $u$ in product line $j$ sold at store $s$, $q_{ujs}$ is the package size of that product, and $\mathbf{X}_{ujs}$ is a set of additional controls. For additional controls, we include both product

[12] The weekly price of a product in a store is defined as the weekly revenues from selling that specific product in the store over the quantity sold. A product is at the barcode (UPC) level.

[13] We exclude products in other departments, such as lightbulbs, as the Nielsen dataset may not be representative of their respective markets.

[14] We chose the week of October 16 for our analysis.

Table 2: Nonlinear Pricing in the Retail Sector

| *Dependent variable:* | price per unit | |
|---|---|---|
| | (1) | (2) |
| Size (ln) | -0.61 | -0.64 |
| (s.e.) | (0.0001) | (0.0001) |
| Product line & store f.e. | ✔ | |
| Product line × store f.e. | | ✔ |
| Observations | 88.3M | 69.7M |

**Notes:** This table reports the results of regression (5.1). The first column contains both product line and store-level fixed effects and includes about 88 million observations. The second column includes product line by store-level fixed effects and includes about 70 million observations.

line and store fixed effects, or product line by store fixed effects. Table 2 presents the results. The first specification includes product line and store-level fixed effects. The second specification includes product line by store fixed effects. The estimates suggest that a 10% larger package size is sold at a 6% lower per-unit price .

Recall that our model consists of two types of households and, therefore, two different sizes offered to consumers. To map the data into our model, we split products in each product line into two categories: small and large. All products smaller than the median product are assigned to the small size category, and the ones larger than the median are assigned to the large size category. For product lines with odd numbers of products, the sales of the median product are split equally between the two size categories. We then define the price and size of each category in each product line to be the sales-weighted average of prices and sizes, respectively, within each category in that product type. The purchases of each category are defined as the sum of purchases of all products within that category.[15]

All the statistics we report are averages across product types, where weights are equal to total sales in the corresponding product type. On average, 51% of purchases are of the large package. The large package is, on average, about 90% larger but only 25% more expensive. We also use the data to compute the market share distribution across firms. The market is highly concentrated, as the top 5% of firms control, on average, a market share of 73%.

## 5.2 Calibration

We assume that preferences satisfy Assumption 1. In Lemma 1 in Appendix A, we show that this assumption implies the following form for the inverse marginal utility:

$$u'(q) = -\beta_0 + \beta_1 q^{-\eta}, \tag{5.2}$$

where $\beta_0$, $\beta_1$, and $\eta$ are structural parameters. In the theoretical part of the paper, we analyze the two extreme cases of a zero or infinite Frisch elasticity. For the quantification, we allow the Frisch

[15] Using this method, multiplying the purchases by the price yields the sum of sales within each size category in every product line.

Table 3: Calibrated Moments and Parameters

A. Moments

| Moment | Data | Model | |
|---|---|---|---|
| | | Benchm. | Linear pricing |
| Fraction buying large $q$ | 51% | 51% | 51% |
| $\mathbb{E}[\ln q_{j\tau} - \ln q_{j1}]$ | 0.65 | 0.65 | 0.65 |
| Sales share top 5% | 73% | 77% | 79% |
| Sales share top 10% | 86% | 84% | 84% |
| Sales share top 25% | 97% | 92% | 90% |
| Sales share top 50% | 99.6% | 96.9% | 95.4% |

B. Parameters

| Parameter | | Model | |
|---|---|---|---|
| | | Benchm. | Linear pricing |
| $\pi$ | Share of high-taste consumers | 0.51 | 0.51 |
| $\tau$ | High-taste demand shifter | 1.17 | 1.17 |
| $\eta$ | Elasticity of taste differential | 1.84 | 2.16 |
| $\theta$ | Pareto shape | 0.84 | 1.17 |
| Externally Set | | | |
| $\varphi$ | Inverse Frisch elasticity | 0.5 | 0.5 |

**Notes:** Panel A presents the model fit. The data column presents the moments we target in our estimation procedure. The second column presents the model moments of our benchmark specification in which firms can offer nonlinear pricing schedules. The third column presents the model moments for a specification in which firms are restricted to linear pricing schedules. Panel B presents the set of calibrated parameters for the two model specifications.

elasticity to take on an intermediate value. Household preferences are given by

$$U_i = \int_0^1 \tau_{ij} u(q_{ij}) dj - \nu \frac{l_i^{1+\varphi}}{1+\varphi}. \tag{5.3}$$

Firm productivity is assumed to follow a Pareto distribution with shape parameter $\theta$.[16]

Following the macro literature, we set the Frisch elasticity to 2 ($\varphi = 0.5$). We normalize $\beta_0 = \beta_1 = 1$. This normalization is without loss of generality.[17] The disutility of labor, $\nu$, is calibrated such that in the market equilibrium, aggregate labor supply is equal to 1.[18]

We calibrate the four structural parameters—the elasticity of taste differential $\eta$, the taste shifter $\tau$, the share of high-taste consumers $\pi$, and the Pareto tail $\theta$—to match six key moments in the data. We set the fraction of high-taste consumers, $\pi$, to match the share of purchases of the large size in the data. The mapping between this data moment and $\pi$ is independent of the rest of the model. The remaining three parameters are then calibrated to match the average difference in package size as well as four quantiles of the distribution of sales across firms. Parameters are chosen to minimize the sum of squared deviations of model to data moments. Table 3 presents the results of the calibration.

In addition to calibrating our benchmark model, we quantify a version in which firms are restricted to offering linear pricing schedules, as discussed in Section 3.3. The last column of Table 3 shows the calibration of the linear pricing model. We calibrate the same set of parameters to match the same set of moments, the purpose of which is twofold. First, this approach allows us to compare the magnitude of misallocation to what researchers would conclude if they used a standard linear pricing model calibrated to the same data. Second, we analyze the welfare effects of implementing the subsidy schedule that would be optimal if the data were generated by firms posting linear prices.

Both models match the data well. However, only our baseline model, in which firms are allowed to price nonlinearly, is able to generate significant dispersion in unit prices within the same product line. In the data, the price per unit charged for large packages is, on average, 43 log points lower than the price per unit charged for small packages. The model accounts for a substantial portion of this

[16] We choose the scale parameter to ensure that all firms produce.

[17] See Appendix A.4 for a formal argument.

[18] Note that the estimation of the structural parameters $\{\pi, \tau, \theta, \eta\}$ does not depend on the calibrated value of the Frisch elasticity. This is because $\nu$ adjusts so that aggregate labor is equal to 1 regardless of the level of $\varphi$.

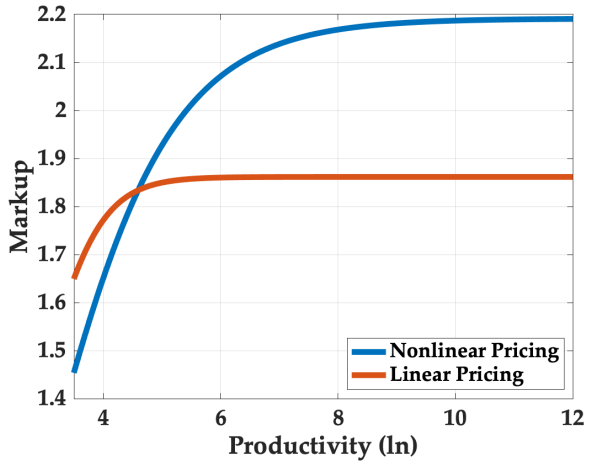
key non-targeted moment: the markup charged on the large bundle is, on average, 29 log points lower than the markup on the small bundle.[19]

### 5.3 Misallocation

We first study the welfare costs of misallocation. We consider both the misallocation of production across firms and of consumption across households. That is, in this section, we hold the aggregate labor supply constant.

In both models, firm-level markups are increasing in firm size, as illustrated in Figure 1, which plots markups against firm productivity. More productive, and hence larger, firms charge higher markups in both environments. Yet, only with the assumption that firms are restricted to linear pricing is this a sign of misallocation across firms.

Figure 1: Firm-level Markups




**Notes**: This figure plots firm-level markups as a function of log productivity ($1/c_j$). In the linear pricing model, $\mu_j = p_j/c_j$, whereas in the nonlinear pricing model, we define firm-level markups as sales-weighted average markups, which is identical to the ratio of total sales to total costs of the firm.

Table 4 reports the welfare implications of misallocation for both models. When firms are not restricted to linear pricing and optimally charge different markups to different consumers, there is no misallocation across firms. There are, however, losses from misallocation across consumers. Consumers are indifferent between consuming the efficient allocation or consuming an additional 0.83% of all goods on top of the market allocation. When firms must choose linear pricing schedules, there is no misallocation across consumers. Reallocating production across firms leads to welfare gains of 0.36%, in consumption equivalent units. So, the welfare gains of moving from the market allocation to the efficient one are more than twice as large in the nonlinear pricing environment relative to the linear pricing one.

The source of misallocation in the baseline model is the distortion of consumption bundles. In the

[19] In Online Appendix A.3, we show that a model with a continuum of types, when calibrated to the same data, yields a nearly identical price and quantity schedule in equilibrium.

Table 4: Welfare Gains of Fixing Misallocation

| Baseline Model | Linear Pricing | Baseline with LP subsidies |
|---|---|---|
| 0.83% | 0.36% | -0.43% |

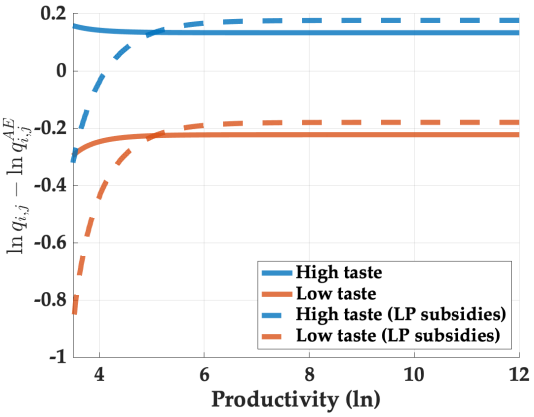
**Notes:** This table reports the welfare gains in the equilibrium with perfect allocative efficiency relative to the baseline model (column 1), the model with linear pricing (column 2), and the baseline model when the optimal linear pricing subsidy schedule is implemented (column 3). All welfare gains are measured in consumption equivalent terms—that is, the uniform increase in consumption that would make households indifferent between the two equilibria.

data, high-taste consumers are sold, on average, 65 log points more than low-taste consumers. In the efficient allocation, that difference would only be 43 log points. This means that a large share of the difference in package sizes offered by firms is not a result of differences in consumer preferences but rather a distortion to guarantee separation of types.

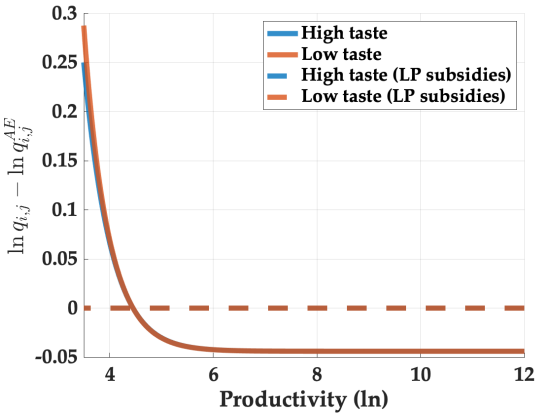
The left panel of Figure 2 illustrates the misallocation of consumption across consumers in the benchmark model (solid lines). Relative to the equilibrium with perfect allocative efficiency ($q_{ij}^{AE}$), high-taste consumers are oversupplied with each good, whereas low-taste consumers are allocated too little.[20] In terms of marginal utilities, the distortion relative to the allocatively efficient equilibrium is constant. Figure 2 shows that also in terms of quantities, the extent to which consumers are over- and undersupplied is similar across firms of different levels of productivity.

Figure 2: Allocative inefficiency by consumer type

(a) Nonlinear Pricing




(b) Linear Pricing




**Notes**: The two panels present the log difference between the decentralized allocation and the quantity under perfect allocative efficiency ($q_{ij}^{AE}$) for each consumer type. The dashed lines present the log difference between the market allocation with the linear pricing subsidies and $q_{ij}^{AE}$. Panel A presents the results for the model with nonlinear pricing, and Panel B presents the results for the linear pricing model.

If firms were restricted to linear pricing, there would be no distortion in marginal utilities across consumers. Figure 2b shows that also in terms of quantities, there is essentially no distortion between high and low types. The solid blue and red lines lie on top of each other. Instead, the amount of

[20] In the equilibrium with perfect allocative efficiency, $q_{ij}^{AE}$ is defined as the quantities chosen by a social planner who is restricted to the same aggregate supply of labor as in the market equilibrium.

distortion relative to the efficient quantity for each type is a function of firm productivity. Small, low-productivity firms are too large, and large, high-productivity firms are too small.

A social planner who has access to only firm-level taxes and subsidies cannot achieve allocative efficiency when firms charge nonlinear prices. However, in an environment in which firms are restricted to posting linear prices, the social planner could restore allocative efficiency with a set of firm-specific subsidies and taxes. This point is illustrated in Figure 2b.[21]

Panel A in Figure 2 shows the impact of implementing the subsidies that *would be optimal* if firms were restricted to posting linear prices in the nonlinear pricing environment.[22] Since the subsidies are increasing in firm size, they would induce large firms to expand at the expense of small firms. Since firm sizes were optimal to begin with, these subsidies and taxes *induce misallocation* across firms but do not alter misallocation within firms. In the aggregate, this distortion of firm-level quantities would lead to welfare losses of 0.43% (Table 4), which is larger than the welfare gains of 0.36% this policy sets out to achieve.

### 5.4 Aggregate markup and labor supply

In this section, we evaluate the impact of within-firm distortions from nonlinear pricing on aggregate labor supply. Table 5 summarizes the results. In the first-best allocation, aggregate labor is 7.3% higher than in the market equilibrium. The first-best allocation yields welfare gains of 1.1% relative to the market allocation—that is, an additional 0.25% of welfare gains relative to the efficient allocation when labor supply is fixed, which we discussed in the previous section. The social planner wants to increase overall labor by 7.3%, but that increase is not uniform across firms. The result we show in Proposition 6 also holds in an environment with positive but finite Frisch elasticity. The employment of the bottom 50% of firms is 10% higher in the first-best allocation relative to the market allocation, whereas the employment of the top 50% grows by 7%.

Table 5: First-Best and Second-Best Allocations Relative to Market Allocation

| | **Nonlinear Pricing** | | **Linear Pricing** |
|---|---|---|---|
| | FB | SB | FB & SB |
| Aggregate Labor | +7.3% | +6.9% | +88% |
| Welfare Gains | +1.1% | +0.2% | +18.8% |

**Notes:** This table presents the difference between the first- and second-best allocations relative to the market allocation. The first two columns present the results for our benchmark model. The final column presents the results for the model with a linear pricing restriction. Under linear pricing, the first- and second-best allocations coincide. Welfare gains are in consumption equivalent units.

A planner with access to only firm-level taxes and subsidies cannot achieve the first-best allocation as they cannot solve the misallocation of consumption within firms. They can, however, raise welfare by inducing more workers to join the labor force. To achieve the second-best allocation, the planner

[21] Allocative efficiency under linear pricing is achieved through a set of unique *relative* subsidies. The overall *level* of subsidies is indeterminate. We set the overall level of subsidies such that aggregate labor supply remains unchanged.

[22] When applying the optimal linear pricing subsidies in the nonlinear pricing environment, we again set their level such that the aggregate level of labor remains constant.

imposes a uniform subsidy of 7.1% across all firms. This policy increases aggregate labor by 6.9% and raises welfare by 0.2% in consumption equivalent units.

When imposing a linear pricing assumption, researchers would conclude that the optimal level of labor is 88% higher than in the market equilibrium. Under linear pricing, as shown by Edmond et al. (2021), the aggregate labor wedge is driven by the aggregate markup in the economy. Since the estimated level of markup in the linear pricing model is about 80%, the distortion in aggregate labor is large.

The tight link between the aggregate markup and the labor wedge breaks under nonlinear pricing. The estimated level of aggregate markup in the nonlinear pricing model is even larger, yet aggregate labor is only 7% below its optimal level. The labor supply decision compares the disutility of supplying an additional unit of labor to the utility gain from additional consumption. Under linear pricing, the aggregate markup directly implies that purchasing additional consumption is expensive. Under nonlinear pricing, the aggregate markup is independent of the marginal price of an additional product. In Appendix Section A.3, we show that consumers can, at the margin, purchase additional units of the goods for which they have a high taste at their marginal cost. It is therefore only the marginal cost, and not the aggregate markup, that shows up in the intratemporal first-order condition. Households still undersupply labor with nonlinear pricing because in equilibrium they consume too much of the goods for which they have a high taste–the goods they can purchase at marginal cost.[23]

Under linear pricing, a social planner with access to only firm-level taxes and subsidies can implement the first-best allocation. To do so, she would need to offer large subsidies. Not only are the required subsidies massive (85% on average), but they would also be larger for the large, high-markup firms. If firms can offer nonlinear pricing schedules, implementing these subsidies would lead to large welfare losses on the order of 19%. The welfare losses stem from two sources. First, the optimal linear pricing subsidies allocate disproportionately more workers to the larger firms, which is the exact opposite of what is optimal under nonlinear pricing. Second, the high level of subsidies leads to a large increase in aggregate labor—a level much larger than the optimal level under nonlinear pricing.

## 6 Conclusion

We develop a model of heterogeneous firms that can offer a menu of prices to consumers with different tastes for the product. Allowing firms to charge quantity-dependent prices fundamentally changes the mapping between markups, misallocation, and welfare. Under general conditions on preferences, there is no misallocation across firms, despite the fact that larger and more productive firms charge higher markups. Further, we point to a new source of misallocation, which is across consumers of the same firm. To maximize profits, high-taste consumers are allocated too much of each good and low-taste consumers too little.

When firms can charge nonlinear prices, the link between the aggregate markup and labor supply

[23] Because of concavity, the marginal utility of consumption of high-taste goods is too low relative to the first-best allocation.

breaks. While there is an undersupply of labor in equilibrium, its magnitude is a function of misallocation across consumers and is independent of the aggregate markup. In the first-best allocation, all firms employ more workers, but a disproportionate share of new workers go to small firms, whose employment share goes up. This result is in stark contrast to the policy prescriptions from a model that assumes firms are restricted to setting linear prices. Under the latter assumption, large, high-markup firms are too small and should be subsidized.

To illustrate the quantitative importance of the new source of misallocation, we calibrate the model to micro data from the retail sector. We show that nonlinear pricing is prevalent and that modeling quantity-dependent prices substantially changes welfare conclusions. Implementing the subsidies and taxes that are optimal under linear pricing would lead to welfare losses of about 20%.

In this paper, we studied how nonlinear pricing shapes misallocation in the goods market, assuming households are ex-ante identical. Two important questions are left for future research. First, what are the distributional consequences of nonlinear pricing in an environment with income inequality? Does nonlinear pricing lead to inefficiently low level of consumption for low-income households, and how does misallocation depend on the degree of inequality? Second, do firms with monoposony power set nonlinear wages? And if so, how does this wage setting behavior shape misallocation in the labor market?

## References


Afrouzi, H., A. Drenik, and R. Kim (2021): "Growing by the Masses: Revisiting the Link between Firm Size and Market Power," *Available at SSRN 3703244*. 1

Argente, D., M. Lee, and S. Moreira (2019): "The Life Cycle of Products: Evidence and Implications," *Available at SSRN 3163195*. 1

Atkeson, A. and A. Burstein (2008): "Pricing-to-market, trade costs, and international relative prices," *American Economic Review*, 98, 1998–2031. 1

Baqaee, D. R. and E. Farhi (2020): "Productivity and misallocation in general equilibrium," *The Quarterly Journal of Economics*, 135, 105–163. 1, 3.3

Boar, C. and V. Midrigan (2019): "Markups and inequality," Tech. rep., National Bureau of Economic Research. 1

Bornstein, G. (2021): "Entry and profits in an aging economy: The role of consumer inertia," Tech. rep., mimeo. 1

Burstein, A., V. M. Carvalho, and B. Grassi (2020): "Bottom-up markup fluctuations," Tech. rep., National Bureau of Economic Research. 1

De Loecker, J., J. Eeckhout, and G. Unger (2020): "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 135, 561–644. 1

Dhingra, S. and J. Morrow (2019): "Monopolistic competition and optimum product diversity under firm heterogeneity," *Journal of Political Economy*, 127, 196–232. 1, 2.1, 3.3, A.1

Edmond, C., V. Midrigan, and D. Y. Xu (2021): "How costly are markups?" Tech. rep., National Bureau of Economic Research. 1, 3.3, 5.4

Einav, L., P. J. Klenow, J. D. Levin, and R. Murciano-Goroff (2021): "Customers and retail growth," Tech. rep., National Bureau of Economic Research. 1

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and manufacturing TFP in China and India," *The Quarterly journal of economics*, 124, 1403–1448. 1

Maskin, E. and J. Riley (1984): "Monopoly with incomplete information," *The RAND Journal of Economics*, 15, 171–196. 1

Melitz, M. J. and G. I. Ottaviano (2008): "Market size, trade, and productivity," *The review of economic studies*, 75, 295–316. 1, 3.2

Mirrlees, J. A. (1971): "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, 38, 175–208. 1, 5

Mussa, M. and S. Rosen (1978): "Monopoly and product quality," *Journal of Economic Theory*, 18, 301–317. 1, 2.1

Myerson, R. B. (1981): "Optimal Auction Design," *Mathematics of Operations Research*, 6, 58–73. 25

Peters, M. (2020): "Heterogeneous markups, growth, and endogenous misallocation," *Econometrica*, 88, 2037–2073. 1

Restuccia, D. and R. Rogerson (2008): "Policy Distortions and Aggregate Productivity with Heterogeneous Plants," *Review of Economic Dynamics*, 11, 707–720. 1

Spence, M. (1977): "Nonlinear prices and welfare," *Journal of Public Economics*, 8, 1–18. 1

Tirole, J. (1988): *The Theory of Industrial Organization*, Cambridge, MA: MIT Press. 1, 5

Wilson, R. B. (1993): *Nonlinear pricing*, Oxford University Press on Demand. 1

# A Proofs

## A.1 Benchmark Misallocation Results

Proof of Proposition 1. From equations (2.4–2.5) and (2.11) we have that:

$$\frac{u'\left(q_{\tau j}\right)}{u'\left(q_{\tau j}^{FB}\right)} = \frac{P^{FB}}{P}, \tag{A.1}$$

$$\frac{u'\left(q_{1j}\right)}{u'\left(q_{1j}^{FB}\right)} = \frac{1-\pi}{1-\tau\pi}\frac{P^{FB}}{P}. \tag{A.2}$$

The equations above, together with the fact that $u'(q)$ is decreasing in $q$ imply that one of three cases must hold: (i) if $\frac{P}{P^{FB}} > 1$ then $q_{\tau j} > q_{\tau j}^{\text{FB}}$ and $q_{1j} > q_{1j}^{\text{FB}}$ for all $j$, (ii) if $\frac{P}{P^{FB}} \in \left(\frac{1-\tau\pi}{1-\pi}, 1\right)$ then $q_{\tau j} > q_{\tau j}^{\text{FB}}$ and $q_{1j} < q_{1j}^{\text{FB}}$ for all $j$, and (iii) if $\frac{P}{P^{FB}} < \frac{1-\tau\pi}{1-\pi}$ then $q_{\tau j} < q_{\tau j}^{\text{FB}}$ and $q_{1j} < q_{1j}^{\text{FB}}$ for all $j$.

Aggregate labor market clearing implies that

$$\int_0^1 c_j\left(\pi q_{\tau j} + (1-\pi)q_{1j}\right)dj = \int_0^1 c_j\left(\pi q_{\tau j}^{\text{FB}} + (1-\pi)q_{1j}^{\text{FB}}\right)dj,$$

so that neither option (i) nor option (iii) are consistent with equilibrium. Therefore, it must be that $\frac{P}{P^{FB}} \in \left(\frac{1-\tau\pi}{1-\pi}, 1\right)$, so that $q_{\tau j} > q_{\tau j}^{\text{FB}}$ and $q_{1j} < q_{1j}^{\text{FB}}$ for all $j$. ■

Proof of Proposition 2.

Equations (2.4–2.5), together with the concavity of $u(\cdot)$, imply that the production of all firms is increasing in the aggregate price index $P$. Therefore, there is a unique level of the aggregate price index that clears the labor market.

Let $\widetilde{P}_j$ be the aggregate price index such that the firm-level production of a firm with marginal cost $c_j$ in equilibrium is identical to its overall production in the efficient allocation: $(1-\pi)\left[q_{1j}^{\text{FB}} - q_{1j}\right] - \pi\left[q_{\tau j} - q_{\tau j}^{\text{FB}}\right] = 0$. Using (2.5–2.4), this can be written as:

$$(1-\pi)\left[(u')^{-1}\left(\frac{c_j}{P^{FB}}\right) - (u')^{-1}\left(\frac{1-\pi}{1-\tau\pi}\frac{c_j}{\widetilde{P}_j}\right)\right] - \pi\left[(u')^{-1}\left(\frac{c_j}{\tau\widetilde{P}_j}\right) - (u')^{-1}\left(\frac{c_j}{\tau P^{FB}}\right)\right] = 0. \tag{A.3}$$

Assumption 1 implies that $\partial\log(q_{\tau j} - q_{\tau j}^{FB})/\partial\log(\mathrm{c}_j) = \eta$. This follows from Equation (3.2), when relabeling $x = c_j/(\tau\tilde{P}_j)$ and $\tau = \tilde{P}_j/P^{FB}$. Similarly, $\partial\log(q_{1j}^{FB} - q_{1j})/\partial\log(\mathrm{c}_j) = \eta$.

Now consider a firm with $c_k = (1+\Delta)c_j$. Using Assumption 1, we have that

$$\begin{aligned}\pi\left(q_{\tau,k}(\tilde{P}_j) - q_{\tau,k}^{FB}\right) \qquad & - (1-\pi)\left(q_{1,k}^{FB} - q_{1,k}(\tilde{P}_j)\right) = \\ \pi(1+\Delta)^{\eta}\left(q_{\tau,j}(\tilde{P}_j) - q_{\tau,j}^{FB}\right) & - (1-\pi)(1+\Delta)^{\eta}\left(q_{1,j}^{FB} - q_{1,j}(\tilde{P}_j)\right) = 0.\end{aligned}$$

Since there is a unique level of the aggregate price index such that the labor market clears, it must

be that $P = \widetilde{P_j}$. Hence, the equilibrium firm-level production and employment for all firms is identical to the ones in the efficient allocation.

■

LEMMA **1** (Implications of constant elasticity of taste differential.)**.** *Suppose preferences* $u(.)$ *satisfy Assumption 1. Then*

1. $(u')^{-1}(x) = -\beta_0 + \beta_1 x^{-\eta}$
2. $q_{1j} = -\beta_0 + \beta_1 \left(\frac{c_j}{P}\frac{1-\pi}{1-\tau\pi}\right)^{-\eta}$
3. $q_{\tau j} = -\beta_0 + \beta_1 \left(\frac{c_j}{P}\frac{1}{\tau}\right)^{-\eta}$

*for some* $\beta_0 \geq 0, \beta_1 \geq 0$.

PROOF OF LEMMA **1.** Let $g(x) \equiv (u')^{-1}(x)$ and $\gamma \equiv \frac{1}{\tau}$. From the definition of the elasticity of taste differential (3.2), we have

$$-\eta = \frac{\partial \log\left(g(x\gamma) - g(x)\right)}{\partial \log(x)}. \tag{A.4}$$

We can rearrange to obtain:

$$-\eta\left[g(x\gamma) - g(x)\right] = \frac{\partial g(x\gamma)}{\partial \log(x)} - \frac{\partial g(x)}{\partial \log(x)}.$$

Taking derivatives and rearranging, we get

$$-\eta\left(g(x\gamma) - g(x)\right) = x\left[g'(\gamma x)\gamma - g'(x)\right].$$

Differentiating w.r.t. $\log(\gamma)$:

$$-\eta g'(x\gamma)x\gamma = x\left(g''(x\gamma)x\gamma\gamma + g'(x\gamma)\gamma\right),$$

which simplifies to

$$\frac{g''(x\gamma)\gamma x}{g'(x\gamma)} = -\eta - 1. \tag{A.5}$$

Equation (A.5) implies that $g'(x)$ is iso-elastic and can be written as

$$g'(x) = -\eta x^{-\eta-1},$$

or

$$g(x) = x^{-\eta} + c_1. \tag{A.6}$$

This proves part 1 of Lemma 1. Part 2 and 3 then directly follow from the firm's optimality conditions (2.4) and (2.5).

Finally, $\beta_1 \geq 0$ and $\beta_0 \geq 0$ follow from the fact that we assumed well-behaved preferences, i.e. that $u(q)$ satisfies $u'(q) \geq 0 \, \forall q \geq 0$ and $u''(q) \leq 0 \, \forall q \geq 0$ ■

Proof of Proposition **3.**

Let's first set up the planner's problem using the primal approach. The planner chooses taxes and subsidies to all firms, $\{t_j\}$, such that its budget is balanced. By choosing taxes and subsidies the planner has control over the firm-level employment of all firms in the economy. We take the primal approach and write the planner's problem as follows:

$$\begin{aligned} \max_{\{l_j, q_{1j}, q_{\tau j}\}_{j=0}^{1}} \quad & \int_0^1 \left[\pi\tau u\left(q_{\tau j}\right) + (1-\pi) u\left(q_{1j}\right)\right] dj, && \text{(A.7)} \\ \text{s.t.} \quad & u'\left(q_{1j}\right) = \frac{1-\pi}{1-\tau\pi}\, \tau\, u'\left(q_{\tau j}\right), && \text{for all } j \\ & \pi q_{\tau j} + (1-\pi) q_{1j} = \frac{l_j}{c_j}, && \text{for all } j \\ & \int l_j dj = 1. \end{aligned}$$

Taking first order conditions, we obtain

$$[q_{\tau j}]: \qquad \pi\tau u'\left(q_{\tau j}\right) + \frac{1-\pi}{1-\tau\pi}\, \tau\, u''\left(q_{\tau j}\right) \mu_j = \pi\theta_j, \tag{A.8}$$

$$[q_{1j}]: \qquad (1-\pi) u'\left(q_{1j}\right) - u''\left(q_{1j}\right) \mu_j = (1-\pi)\theta_j, \tag{A.9}$$

$$[l_j]: \qquad \frac{\theta_j}{c_j} = \lambda. \tag{A.10}$$

where $\mu_j$, $\theta_j$, and $\lambda$ are the Lagrange multipliers on the three constraints, respectively. Multiplying equation (A.8) by $\frac{1-\tau\pi}{\tau(1-\pi)} \frac{u''(q_{1j})}{u''(q_{\tau j})}$ and adding to equation (A.9), we obtain:

$$\pi \frac{1-\tau\pi}{1-\pi} u'\left(q_{\tau j}\right) \frac{u''\left(q_{1j}\right)}{u''\left(q_{\tau j}\right)} + (1-\pi) u'\left(q_{1j}\right) = \left[\pi \frac{1-\tau\pi}{\tau(1-\pi)} \frac{u''\left(q_{1j}\right)}{u''\left(q_{\tau j}\right)} + (1-\pi)\right] \theta_j.$$

Rearranging we have

$$\theta_j = \gamma_j \tau u'\left(q_{\tau j}\right) + \left(1-\gamma_j\right) u'\left(q_{1j}\right), \tag{A.11}$$

where

$$\gamma_j = 1 - \frac{1-\pi}{(1-\pi) + \pi \frac{1-\tau\pi}{1-\pi} \frac{u''(q_{1j})}{u''(q_{\tau j})}}.$$

Note that $\gamma_j$ represents the share of additional production allocated to high-taste consumers when

$l_j$ increases. The third optimality condition (A.10) inmplies that

$$\frac{\gamma_j \tau u'(q_{\tau j}) + (1-\gamma_j) u'(q_{1j})}{c_j} = \lambda, \quad \text{for all } j. \tag{A.12}$$

Equation (A.12) indicates that in the optimal allocation, the planner is indifferent between reallocation a unit of labor from one firm to another firm.

We will now show that the nonlinear pricing equilibrium allocations satisfy equation (A.12). First, using equations (2.6–2.7), the LHS of equation (A.12) becomes

$$\frac{\gamma_j + \frac{1-\pi}{1-\tau\pi}(1-\gamma_j)}{P}, \quad \text{for all } j.$$

Using Lemma 1, we have that

$$\frac{u''(q_{1j})}{u''(q_{\tau j})} = \left(\frac{u'(q_{1j})}{u'(q_{\tau j})}\right)^{1+\eta} = \left(\frac{1-\pi}{1-\tau\pi}\right)^{1+\eta},$$

where the last equality follows from equations (2.4–2.5). From the definition of $\gamma_j$, we see that $\gamma_j$ is constant across firms in the equilibrium allocation. Denote its value by $\gamma$. Therefore, the LHS of equation (A.12) becomes

$$\frac{\gamma + \frac{1-\pi}{1-\tau\pi}(1-\gamma)}{P}, \quad \text{for all} j.$$

Setting $\lambda = \frac{\gamma + \frac{1-\pi}{1-\tau\pi}(1-\gamma)}{P}$, we have that equation (A.12) holds for all $j$. The first order conditions of the planner then pin down the values of $\mu_j$ and $\theta_j$, for all $j$. We conclude that the equilibrium allocations coincide with the constrained efficient allocation. Therefore, the optimal firm-level taxes and subsidies are all zero.

■

## A.2 Endogenous Labor Supply

Proof of Proposition 5. Since $P = P^{FB} = \frac{1}{\nu}$, we have that $q_{\tau j} = q_{\tau j}^{FB}$ and $q_{1j} < q_{1j}^{FB}$ and

$$L = \int_0^1 c_j((1-\pi)q_{\tau j} + \pi q_{1j})dj \tag{A.13}$$

$$< \int_0^1 c_j((1-\pi)q_{\tau j}^{FB} + \pi q_{1j}^{FB})dj = L^{FB}. \tag{A.14}$$

■

Proof of Proposition 6. The excess employment ($\omega_j$) is given by

$$\omega_j = \frac{\frac{(\pi q_{\tau j} + (1-\pi)q_{1j})c_j}{L}}{\frac{(\pi q_{\tau j}^{\mathrm{FB}} + (1-\pi)q_{1j}^{\mathrm{FB}})c_j}{L^{\mathrm{FB}}}} = \frac{\pi q_{\tau j} + (1-\pi)q_{1j}}{\pi q_{\tau j}^{\mathrm{FB}} + (1-\pi)q_{1j}^{\mathrm{FB}}} \frac{L^{\mathrm{FB}}}{L}$$

where $L$ and $L^{\text{FB}}$ are aggregate labor. Using Lemma 1 as well as the fact that with endogenous labor supply $P = P^{\text{FB}}$,

$$\omega_j = \frac{\beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \left[\pi(\tau)^{-\eta} + (1-\pi)\left(\frac{1-\pi}{1-\tau\pi}\right)^{-\eta}\right] - \beta_0}{\beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \left[\pi(\tau)^{-\eta} + (1-\pi)\right] - \beta_0} \frac{L^{\text{FB}}}{L}$$

Taking derivatives wrt $c_j$

$$\frac{\partial \omega_j}{\partial c_j} = \frac{\left(\eta\beta_1 \left(\frac{c_j}{P}\right)^{-\eta-1}\right)\left(\left[\pi(\tau)^{-\eta} + (1-\pi)\left(\frac{1-\pi}{1-\tau\pi}\right)^{-\eta}\right] - \left[\pi(\tau)^{-\eta} + (1-\pi)\right]\right)}{\left(\beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \left[\pi(\tau)^{-\eta} + (1-\pi)\right] - \beta_0\right)^2} \frac{L^{\text{FB}}}{L}$$

Since $\eta \geq 0$ and $\beta_1 \geq 0$, it follows that

$$\begin{aligned}
&\frac{\partial \omega_j}{\partial c_j} < 0 \\
\iff &\left(\left[\pi(\tau)^{-\eta} + (1-\pi)\left(\frac{1-\pi}{1-\tau\pi}\right)^{-\eta}\right] - \left[\pi(\tau)^{-\eta} + (1-\pi)\right]\right) \leq 0 \\
\iff &\left(\frac{1-\pi}{1-\tau\pi}\right)^{-\eta} \leq 1 \\
\iff &\frac{1-\pi}{1-\tau\pi} \geq 1 \\
\iff &\tau \geq 1.
\end{aligned}$$

and higher productivity firms have higher excess market shares.

■

Proof of Proposition 7.

Let's first set up the planner's problem using the primal approach. The planner chooses taxes and subsidies to all firms, $\{t_j\}$, such that its budget is balanced. By choosing taxes and subsidies the planner has control over the firm-level employment of all firms in the economy, as well as the aggregate quantity of labor. We take the primal approach and write the planner's problem as follows:

$$\begin{aligned}
\max_{\{l_j, q_{1j}, q_{\tau j}, L\}_{j=0}^{1}} \quad & -\nu L + \int_0^1 \left[\pi\tau u(q_{\tau j}) + (1-\pi)u(q_{1j})\right] dj, \\
\text{s.t.} \quad & u'(q_{1j}) = \frac{1-\pi}{1-\tau\pi}\,\tau\, u'(q_{\tau j}), && \text{for all } j \\
& \pi q_{\tau j} + (1-\pi) q_{1j} = \frac{l_j}{c_j}, && \text{for all } j \\
& \int l_j dj = L.
\end{aligned}$$

Taking first order conditions, we obtain

$$\begin{aligned}
[q_{\tau j}]: &\qquad \pi\tau u'\left(q_{\tau j}\right)+\frac{1-\pi}{1-\tau\pi}\,\tau\, u''\left(q_{\tau j}\right)\mu_j=\pi\theta_j,\\
[q_{1j}]: &\qquad (1-\pi)u'\left(q_{1j}\right)-u''\left(q_{1j}\right)\mu_j=(1-\pi)\theta_j,\\
[l_j]: &\qquad \frac{\theta_j}{c_j}=\lambda,\\
[L]: &\qquad \nu=\lambda,
\end{aligned}$$

where $\mu_j$, $\theta_j$, and $\lambda$ are the Lagrange multipliers and the three constraints, respectively. As in the case of fixed labor supply, we can combine the first two conditions to obtain:

$$\theta_j=\gamma_j\tau u'\left(q_{\tau j}\right)+\left(1-\gamma_j\right)u'\left(q_{1j}\right), \tag{A.15}$$

where

$$\gamma_j=1-\frac{1-\pi}{(1-\pi)+\pi\frac{1-\tau\pi}{1-\pi}\frac{u''(q_{1j})}{u''(q_{\tau j})}}.$$

Let $t_j$ denote the tax levied on production by firm $j$, such that the marginal cost it faces is $c_j(1+t_j)$. From the firm's quantity choices in equilibrium, we then have that

$$\begin{aligned}
\tau u'(q_{\tau j})&=c_j\nu(1+t_j)\\
u'(q_{1j})&=c_j\nu(1+t_j)\frac{1-\pi}{1-\tau\pi}
\end{aligned}$$

which uses the fact that in equilibrium, $P=\frac{1}{\nu}$. Plugging this back into (A.15), we see that the optimal level of taxes depend only on $\gamma_j$.

$$1=(1+t_j)\left(\gamma_j+(1-\gamma_j)\frac{1-\pi}{1-\tau\pi}\right). \tag{A.16}$$

Using Lemma 1, we can write $\gamma_j$ as

$$\gamma_j=\frac{\pi\left(\frac{1-\tau\pi}{1-\pi}\right)^{\eta}}{(1-\pi)+\pi\left(\frac{1-\tau\pi}{1-\pi}\right)^{\eta}}.$$

Note first that $\gamma_j$ is independent of $c_j$. Hence, (A.16) implies that firm-level taxes or subsidies are constant across firms $j$. Further, since $\gamma_j<1$ and $\frac{1-\pi}{1-\tau\pi}<1$, (A.16) also implies that $1+t_j<1$. Taken together, we have that $t_j=t<0\,\forall j$.

■

## A.3 Additional Propositions and Proofs

PROPOSITION **8.** *Under Assumption 1, an equilibrium exists and is unique.*

PROOF OF PROPOSITION **8.** Using the optimality conditions of the firm, (2.4) and (2.5), write labor market clearing directly as a function of $P$:

$$\int_0^1 c_j \left[\pi(u')^{-1}\left(\frac{c_j}{P}\frac{1}{\tau}\right) + (1-\pi)(u')^{-1}\left(\frac{c_j}{P}\frac{1-\pi}{1-\tau\pi}\right)\right] dj = 1. \tag{A.17}$$

Using Lemma (1), $\lim_{x\to\infty}(u')^{-1}(x) = -\beta_0 \leq 0$ and $\lim_{x\to 0}(u')^{-1}(x) = \infty$. So when $P \to 0$ , no firm wants to produce positive quantities, and when $P \to \infty$, production goes to infinity. Since $u(.)$ is continuously differentiable, there exists a $P > 0$ such that (A.17) holds. Since marginal utility $u'(.)$ is decreasing everywhere, $P$ is unique.[24] ■

**Supporting the aggregate price index in equilibrium.** Recall that the aggregate price index $P$ measures the price of obtaining an additional unit of utility. In Proposition 8, we show that there exists a unique $P$ that clears the labor market. Below, we show how this aggregate price index can be supported by the pricing decision of firms.

Recall that while the price each firm charges for the high- and low-type bundles is unique, the prices firms charge for quantities that are not purchased in equilibrium are indeterminate. Firms can charge arbitrary prices for $q_j \notin \{q_{1j}, q_{\tau j}\}$ as long as neither of the two consumer types wants to deviate and purchase that quantity. To rationalize the aggregate price index, we assume that firms offer any quantity $q > q_{\tau j}$ for the overall price $p_{\tau j}q_{\tau j} + \tilde{p}_j(\tilde{q} - q_{\tau j})$. That is, firms offer units over and above the high-type bundle for $\tilde{p}_j$.

We first derive the value of $\tilde{p}_j$ that supports the equilibrium level of the aggregate price index $P$. The following equation pins down $\tilde{p}_j$,

$$\frac{1}{P} = \frac{\tau u'(q_{\tau j})}{\tilde{p}_j}, \tag{A.18}$$

where the LHS is the utility gain from an extra unit of expenditure in equilibrium, and the RHS is the additional utility of spending an extra dollar on $q_{\tau j}$. Using the firm's optimality condition for $q_{\tau j}$:

$$\tilde{p}_j = c_j. \tag{A.19}$$

Equation (A.19) implies that in order to support the aggregate price index $P$ in equilibrium, firms need to offer additional units above the high-type bundle for marginal cost.

Finally, note that individual rationality constraint of high-type consumers and incentive compatibility of low-type consumers imply that no consumer wants to deviate and purchase a quantity greater than $q_{\tau j}$ for all $j$.

[24] In the proposition and proof, we maintained the assumption that primitives are such that all firms choose to serve all consumers in equilibrium, i.e. the solution to (2.4) and (2.5) is weakly positive even for the highest cost firm. A similar continuity argument proves existence and uniqueness of equilibrium in the absence of this restriction.

PROPOSITION **9.** *Suppose preferences satisfy Assumption 1. Then, firms with higher productivity (i.e., low production costs) charge higher markups at the firm-level* $\left(\mu_j \equiv \frac{\pi\,(p_{1j}q_{1j})+(1-\pi)\,(p_{\tau j}q_{\tau j})}{c_j(\pi q_{1j}+(1-\pi)q_{\tau j})}\right)$.

PROOF OF PROPOSITION **9.** Using the firm's optimality conditions to substitute out prices, the inverse of the markup is given by

$$\frac{1}{\mu_j} = \frac{(1-\pi\tau)u(q_{j1})/\psi(q_t)}{(1-\pi\tau)u(q_{j1})+\pi\tau u(q_{j\tau})} + \frac{\pi\tau u(q_{j\tau})/\psi(q_{j\tau})}{(1-\pi\tau)u(q_{j1})+\pi\tau u(q_{j\tau})} \tag{A.20}$$

where $\psi(q) \equiv \frac{u(q)}{qu'(q)}$. Using Lemma 1, we have that

$$\mu_j = \frac{\eta}{\eta-1}\beta_1^{\frac{1}{\eta}} \frac{\pi\tau\left(q_{\tau j}+\beta_0\right)^{\frac{\eta-1}{\eta}} + (1-\pi\tau)\left(q_{1j}+\beta_0\right)^{\frac{\eta-1}{\eta}} - \left(\beta_0\right)^{\frac{\eta-1}{\eta}}}{\pi\tau q_{\tau j}\left(q_{\tau j}+\beta_0\right)^{-\frac{1}{\eta}} + (1-\pi\tau)q_{1j}\left(q_{1j}+\beta_0\right)^{-\frac{1}{\eta}}}. \tag{A.21}$$

Let

$$x_1 \equiv \frac{c_j}{P}\frac{1-\pi}{1-\tau\pi},$$
$$x_\tau \equiv \frac{c_j}{P}\frac{1}{1-\tau}.$$

Using the expressions for quantities, we get

$$\begin{aligned}\mu_j &= \frac{\eta}{\eta-1}\frac{\pi\tau\beta_1^{\frac{\eta-1}{\eta}}x_\tau^{1-\eta} + (1-\pi\tau)\beta_1^{\frac{\eta-1}{\eta}}x_1^{1-\eta} - \left(\beta_0\right)^{\frac{\eta-1}{\eta}}}{\pi\tau\left(-\beta_0+\beta_1 x_\tau^{-\eta}\right)\beta_1^{-\frac{1}{\eta}}x_\tau + (1-\pi\tau)\left(-\beta_0+\beta_1 x_1^{-\eta}\right)\beta_1^{-\frac{1}{\eta}}x_1} \\ &= \frac{\eta}{\eta-1}\frac{\beta_1^{\frac{\eta-1}{\eta}}\left(c_j/P\right)^{1-\eta}\widetilde{\tau} - \beta_0^{\frac{\eta-1}{\eta}}}{\beta_1^{\frac{\eta-1}{\eta}}\left(c_j/P\right)^{1-\eta}\widetilde{\tau} - \left(c_j/P\right)\beta_0\beta_1^{-\frac{1}{\eta}}},\end{aligned} \tag{A.22}$$

where

$$\widetilde{\tau} \equiv (\pi\tau)^\eta\,\pi^{1-\eta} + (1-\tau\pi)^\eta(1-\pi)^{1-\eta}. \tag{A.23}$$

Rewrite this as

$$\mu_j = \frac{\eta}{\eta-1}\frac{(c_j/P)^{1-\eta}\alpha-\gamma}{(c_j/P)^{1-\eta}\alpha-(c_j/P)\delta}, \tag{A.24}$$

where

$$\alpha = \beta_1^{\frac{\eta-1}{\eta}}\widetilde{\tau} \qquad > 0, \tag{A.25}$$
$$\gamma = \beta_0^{\frac{\eta-1}{\eta}} \qquad > 0, \tag{A.26}$$
$$\delta = -\beta_0\beta_1^{-\frac{1}{\eta}} \qquad > 0. \tag{A.27}$$

Since $\frac{\eta}{\eta-1} < 0$, the sign of the derivative is

$$\text{sign}\left(\frac{\partial \mu_j}{\partial(c_j/P)}\right) = -\text{sign}\underbrace{\left(\eta\alpha\delta\left(\frac{c_j}{P}\right)^{1-\eta} + \alpha\gamma(1-\eta)\left(\frac{c_j}{P}\right)^{-\eta} - \gamma\delta\right)}_{\equiv Z(c_j)}. \tag{A.28}$$

We need to show that markups are higher for more productive firms (those with lower costs). That is, $\left(\frac{\partial \mu_j}{\partial(c_j/P)}\right) < 0$, or $Z(c_j) \geq 0$ everywhere. If $Z(c_j) \geq 0$ at its minimum, then it's positive everywhere.

$$\underset{c_j}{\text{argmin}} \quad Z(c_j) = \frac{\gamma}{\delta}. \tag{A.29}$$

Plugging back in we get that the derivative is positive if and only if

$$\alpha\delta^{\eta-1} > \gamma^{\eta}. \tag{A.30}$$

Which simplifies to

$$\widetilde{\tau} \geq 1. \tag{A.31}$$

Write $\widetilde{\tau}$ as a function of $\tau$. For any $(\pi, \eta)$, $\widetilde{\tau}(1) = 1$. Then, as long as $\widetilde{\tau}(\tau)' \geq 0$, we have that $\widetilde{\tau} \geq 1 \quad \forall \tau \geq 1$.

$$\begin{aligned}\widetilde{\tau}'(\tau) &= \eta\pi\tau^{\eta-1} - \eta\pi(1-\tau\pi)^{\eta-1}(1-\pi)^{1-\eta} \\ &= \eta\pi\left[\tau^{\eta-1} - (1-\tau\pi)^{\eta-1}(1-\pi)^{1-\eta}\right],\end{aligned}$$

which is positive if and only if $\tau^{\eta-1} \geq (1-\tau\pi)^{\eta-1}(1-\pi)^{1-\eta}$. Since $\eta - 1 \geq 0$:

$$\begin{aligned}\widetilde{\tau}'(\tau) \geq 0 &\iff \tau \geq \frac{1-\tau\pi}{1-\pi} \\ &\iff \tau \geq 1.\end{aligned}$$

■

## A.4 Identification

Proposition **10** (Normalization of $\beta_1$)**.** *Holding fixed the set of structural parameters other than $\beta_1$, $\{\beta_0, \eta, \tau, \pi, \theta\}$, the markups and allocations in the market equilibrium as well as allocations in the first-best allocation are identical for all $\beta_1 > 0$.*

Proof of Proposition **10.** Let $\widetilde{\beta_1} \equiv \beta_1 P^{\eta}$. Using Lemma 1, we can re-write the optimal quantities

sold on the market equilibrium as

$$q_{1j} = -\beta_0 + \widetilde{\beta_1} c_j^{-\eta} \left( \frac{1-\pi}{1-\tau\pi} \right)^{\eta} \tag{A.32}$$

$$q_{\tau j} = -\beta_0 + \widetilde{\beta_1} c_j^{-\eta} \left( \frac{1}{\tau} \right)^{\eta} \tag{A.33}$$

So, for any ${\beta_1}'$ there is a $P'$ such that $\widetilde{\beta_1}' = \widetilde{\beta_1}$ and hence allocations are unchanged. Note that $P'$ ($P^{FB'}$) is the unique price index that clears the labor market, and hence the equilibrium level of the price index.

We have that market allocations are independent of the level of $\beta_1$. We now turn to show that also equilibrium markups do not depend on $\beta_1$. From Lemma 1, we obtain

$$\psi(q) = \frac{u(q)}{qu'(q)} = \frac{\eta}{\eta - 1} \left[ \left( 1 + \frac{\beta_0}{q} \right) - \beta_0^{\frac{\eta-1}{\eta}} \frac{(q+\beta_0)^{\frac{1}{\eta}}}{q} \right]. \tag{A.34}$$

Note that $\psi(\cdot)$ does not depend on $\beta_1$. Using this fact together with the fact that allocations are unchanged, we have that markups are also unchanged from equations (2.6) and (2.7).

Similarly, we can show that first-best allocations are independent of $\beta_1$. Let $\widetilde{\beta_1}^{FB} \equiv \beta_1 \left( P^{FB} \right)^{\eta}$. Using Lemma 1, we can re-write the first-best quantities ((2.11)) as

$$q_{1j}^{FB} = -\beta_0 + \widetilde{\beta_1}^{FB} c_j^{-\eta} \tag{A.35}$$

$$q_{\tau j}^{FB} = -\beta_0 + \widetilde{\beta_1}^{FB} c_j^{-\eta} \left( \frac{1}{\tau} \right)^{\eta} \tag{A.36}$$

So, for any ${\beta_1}^{FB'}$ there is a $P^{FB'}$ such that $\widetilde{\beta_1}^{FB'} = \widetilde{\beta_1}^{FB}$ and hence allocations are unchanged. Note that $P^{FB'}$ is indeed the inverse Lagrange multiplier on the planner's problem, as it clears the labor market.

■

Proposition **11** (Normalization of $\beta_0$)**.** *Consider a set of structural parameters $\{\beta_0, \beta_1, \eta, \tau, \pi, \theta\}$. If we multiply $\beta_0$ by a constant $\alpha > 0$ and divide $c_j$ for all $j$ by the same constant, then:*

1. *Markups in the market equilibrium are identical.*
2. *Allocations in both the market equilibrium and the first-best are scaled by the constant $\alpha$.*

Proof of Proposition **11.** Using Lemma 1, we have that the quantities sold in the market equilibrium are given by

$$q_{1j} = -\beta_0 + \beta_1 P^{\eta} c_j^{-\eta} \left( \frac{1-\pi}{1-\tau\pi} \right)^{\eta}, \tag{A.37}$$

$$q_{\tau j} = -\beta_0 + \beta_1 P^{\eta} c_j^{-\eta} \left( \frac{1}{\tau} \right)^{\eta}. \tag{A.38}$$

Consider $\beta_0' = \alpha\beta_0$, $c_j' = \alpha c_j$ and $P' = \alpha^{\frac{1-\eta}{\eta}} P$. Then from the equations above we obtain

$$q_{1j}' = \alpha q_{1j}, \tag{A.39}$$

$$q_{\tau j}' = \alpha q_{\tau j}. \tag{A.40}$$

Since all costs are divided by $\alpha$, the labor ($l_j' = q_j' c_j' = \alpha q_j \, c_j/\alpha = l_j$) needed to produce the allocations for the scaled $\beta_0$ is unchanged. Therefore $P' = \alpha^{\frac{\eta-1}{\eta}} P$ is indeed the equilibrium level of the price index.

Turning to the markups, we will start by showing that $\psi(q_{ij})$ remain unchanged. From equation (A.34) we have

$$\psi(q) = \frac{\eta}{\eta - 1}\left[\left(1 + \frac{\beta_0}{q}\right) - \left(\frac{\beta_0}{q}\right)^{\frac{\eta-1}{\eta}}\left(1 + \frac{\beta_0}{q}\right)^{\frac{1}{\eta}}\right] \tag{A.41}$$

Since both quantities and $\beta_0$ are scaled by $\alpha$, we have that $\psi(q_{ij})$ are unchanged for all $i$ and $j$. Using the equilibrium markup levels from equations (2.6–2.7) we have that markups are unchanged in the new equilibrium.

Finally, let's show that also first-best allocations are all scaled by $\alpha$. From Lemma 1 and equation (2.11) we have

$$q_{1j}^{FB} = -\beta_0 + \beta_1 \left(P^{FB}\right)^{\eta} c_j^{-\eta} \tag{A.42}$$

$$q_{\tau j}^{FB} = -\beta_0 + \beta_1 \left(P^{FB}\right)^{\eta} c_j^{-\eta} \left(\frac{1}{\tau}\right)^{\eta} \tag{A.43}$$

Similarly, for $\beta_0' = \alpha\beta_0$ we can choose $P^{FB'} = \alpha^{\frac{1-\eta}{\eta}} P^{FB}$. All first-best allocations are then scaled by $\alpha$. With the scaled down production costs, the labor market clears and we confirm that $P^{FB'}$ is indeed the inverse Lagrange multiplier on the planner's problem.

■

# B Linear Pricing: Setup and Proofs

## B.1 Linear Pricing Equilibrium

When firm are restricted to linear pricing, the household's problem is given by

$$\begin{aligned} \max_{\{c_{ij}\}} \quad & \int_0^1 \tau_{ij} u(q_{ij}) dj \\ \text{s.t.} \quad & \int_0^1 p_j c_{ij} = I, \end{aligned} \tag{B.1}$$

where $I$ is the income of households. Taking first order conditions, we obtain

$$\tau_{ij} u'(q_{ij}) = \frac{p_j}{P},$$

where $P$ is the inverse Lagrange multiplier.

The firm's problem is then given by

$$\begin{aligned} \max_{\{p_j, q_{1j}, q_{\tau j}\}} \quad & (\pi q_{\tau j} + (1-\pi) q_{1j}) (p_j - c_j) \\ \text{s.t.} \quad & \tau u'(q_{\tau j}) = \frac{p_j}{P}, \\ & u'(q_{1j}) = \frac{p_j}{P}. \end{aligned} \tag{B.2}$$

Taking first order conditions, we have

$$\begin{aligned} [p_j]: \quad & (\pi q_{\tau j} + (1-\pi) q_{1j}) = \frac{\nu_{1j} + \nu_{\tau j}}{P}, \\ [q_{\tau j}]: \quad & \pi (p_j - c_j) = -\tau u''(q_{\tau j}) \nu_{\tau j}, \\ [q_{1j}]: \quad & (1-\pi) (p_j - c_j) = -u''(q_{1j}) \nu_{1j}, \end{aligned}$$

where $\nu_{1j}$ and $\nu_{\tau j}$ are the Lagrange multipliers on the demand functions for low- and high-taste consumers, respectively. Define $\epsilon(q)$ to be the inverse elasticity of marginal utility:

$$\epsilon(q) \equiv -\frac{u'(q)}{q u''(q)}.$$

We can use the demand function to rewrite the last two first order conditions as follows:

$$\pi (p_j - c_j) q_{\tau j} = \frac{p_j}{P} \frac{1}{\epsilon (q_{\tau j})} \nu_{\tau j}, \tag{B.3}$$

$$(1-\pi)(p_j - c_j) q_{1j} = \frac{p_j}{P} \frac{1}{\epsilon (q_{1j})} \nu_{1j}. \tag{B.4}$$

Multiplying each equation by $\epsilon(q_{ij})/p_j$ and summing the two conditions, we have

$$\frac{p_j - c_j}{p_j} \left( \pi q_{\tau j} \epsilon(q_{\tau j}) + (1-\pi) q_{1j} \epsilon(q_{1j}) \right) = \frac{1}{P} \left( \nu_{\tau j} + \nu_{1j} \right)$$

Using the first order condition with respect to $p_j$ we finally obtain

$$\frac{p_j}{p_j - c_j} = \frac{\pi q_{\tau j} \epsilon(q_{\tau j}) + (1-\pi) q_{1j} \epsilon(q_{1j})}{\pi q_{\tau j} + (1-\pi) q_{1j}} \tag{B.5}$$

Defining the firm-level markup as $\mu_j \equiv \frac{p_j}{c_j}$, this equation becomes

$$\frac{\mu_j}{\mu_j - 1} = \alpha_j \epsilon(q_{\tau j}) + (1 - \alpha_j) \epsilon(q_{1j}), \tag{B.6}$$

where $\alpha_j$ is the production share sold to high-taste consumers:

$$\alpha_j = \frac{\pi q_{\tau j}}{\pi q_{\tau j} + (1-\pi) q_{1j}}.$$

PROPOSITION **12.** *Suppose preferences satisfy Assumption 1 and exhibit variable elasticity of substitution, there is misallocation across firms in the linear pricing equilibrium. In particular,*

1. *Firms with higher productivity ($1/c_j$) charge higher markups.*
2. *Firms that charge high markups sell too little relative to the efficient allocation.*
3. *The optimal firm-level taxes are positive for low-markup firms, and negative (subsidies) for high-markup firms. That is, $\exists\ \bar{c}_j$ s.t. $t(c_j) \leq 0$ if $c_j \leq \bar{c}_j$ and $t(c_j) \geq 0$ if $c_j \geq \bar{c}_j$.*

PROOF OF PROPOSITION **12.**

1. Using Lemma 1,

$$\epsilon(q_{ij}) = \eta\left(\frac{\beta_0}{q_{ij}} + 1\right)$$

Using this expression, (B.6) simplifies to

$$\left(1 - \frac{1}{\mu_j}\right)^{-1} = \eta\left(1 + \frac{\beta_0}{q_j^2}\right) \tag{B.7}$$

Derivative wrt to $q_j$:

$$\begin{aligned} &\frac{\partial\left(1 - \frac{1}{\mu_j}\right)^{-1}}{\partial q_j} = -\eta\frac{\beta_0}{q_j} < 0 \\ \Rightarrow\quad &\frac{\partial \mu_j}{\partial q_j} > 0 \end{aligned} \tag{B.8}$$

And firms that sell higher $q_j$ charge higher markups. Using Lemma 1 together with the consumers' FOCs, we get that

$$q_j = \pi q_{\tau j} + (1-\pi)q_{1j} = -\beta_0 + \beta_1\left(\mu_j\frac{c_j}{P}\right)^{-\eta}(\pi\tau^{\eta} + (1-\pi)) \tag{B.9}$$

Since $\partial\mu_j/\partial q_j > 0$, (B.9) implies that $\partial q_j/\partial c_j < 0$ and therefore $\partial mu_j/\partial c_j < 0$: more productive firms charge higher markups.

2. The demand function with linear pricing implies

$$q_{ij} = (u')^{-1}\left(\frac{\mu_j c_j}{\tau_{ij}P}\right), \tag{B.10}$$

while from equation (2.11), we have that in the efficient allocation,

$$q_{ij}^{FB} = (u')^{-1}\left(\frac{1}{\tau_{ij}P^{FB}}\right). \tag{B.11}$$

Using Lemma 1 and summing over the two consumer types, we have

$$q_j = -\beta_0 + \beta_1 \left(\pi\tau^\eta + (1-\pi)\right) \left(\frac{c_j \mu_j}{P}\right)^{-\eta}, \tag{B.12}$$

$$q_j^{FB} = -\beta_0 + \beta_1 \left(\pi\tau^\eta + (1-\pi)\right) \left(\frac{c_j}{P^{FB}}\right)^{-\eta} \tag{B.13}$$

Let $\bar{\mu}$ be such that $\frac{\bar{\mu}}{P} = \frac{1}{P}^{FB}$. Since $\beta_1 > 0$ and $\eta > 0$, equations (B.12-B.13) imply that

$$\begin{aligned} q_j &< q_j^{FB} \qquad \text{if} \quad \mu_j > \bar{\mu}, \\ q_j &> q_j^{FB} \qquad \text{if} \quad \mu_j < \bar{\mu}. \end{aligned}$$

That is, high-markup firms sell too little relative to the efficient allocation while low-markup firms sell too much. Note that there is a strictly positive mass of firms with markups both below and above the threshold. Otherwise, the labor market doesn't clear.

3. Consider a planner who can tax and subsidize firm-level production. We will show how the planner can implement the efficient allocation. The firm's problem becomes

$$\begin{aligned} \max_{\{p_j, q_{1j}, q_{\tau j}\}} \quad & \left(\pi q_{\tau j} + (1-\pi) q_{1j}\right) \left(p_j - c_j(1+t_j)\right) \\ \text{s.t.} \quad & \tau u'(q_{\tau j}) = \frac{p_j}{P}, \\ & u'(q_{1j}) = \frac{p_j}{P}. \end{aligned}$$

Following the same steps as in the problem without taxes, we obtain

$$\frac{\mu_j}{\mu_j - 1} = \alpha_j \epsilon(q_{\tau j}) + (1-\alpha_j)\, \epsilon(q_{1j}), \tag{B.14}$$

where $\alpha_j$ is the production share sold to high-taste consumers:

$$\alpha_j = \frac{\pi q_{\tau j}}{\pi q_{\tau j} + (1-\pi) q_{1j}}.$$

The demand function can be written as

$$\tau_{ij} u'(q_{ij}) = \frac{\mu_j (1+t_j) c_j}{P}, \tag{B.15}$$

Let $\tilde{\mu}_j$ be defined explicitly as follows:

$$\frac{\tilde{\mu}_j}{\tilde{\mu}_j - 1} = \alpha_j \epsilon(q_{\tau j}^{FB}) + (1-\alpha_j)\, \epsilon(q_{1j}^{FB}), \tag{B.16}$$

so that $\tilde{\mu}_j$ is the markup the firm would like to set when production is equal to the efficient

allocation. Now, let the planner's tax be such that

$$1 + t_j = \frac{1}{\tilde{\mu}_j} S, \tag{B.17}$$

for some positive scalar $S$. From equations (2.11) and (B.15) we have that if $P = SP^{FB}$, equilibrium and efficient allocation coincide and the labor market clears. Since labor demand of all firms is increasing in $P$, $P = SP^{FB}$ is the unique equilibrium and the planner successfully implements the efficient allocations by setting taxes according to equation (B.17). The scalar $S$ is set so that total taxes are equal to total subsidies.

Finally, we want to show that $\tilde{\mu}_j$ is decreasing in $c_j$. From equation (B.16), we have that $\tilde{\mu}_j$ is decreasing in $c_j$ if and only if $\alpha_j \epsilon(q_{\tau j}^{FB}) + (1-\alpha_j)\,\epsilon(q_{1j}^{FB})$ is increasing in $c_j$. Define

$$\tilde{\epsilon}_j = \frac{\pi q_{\tau j}^{FB} \epsilon(q_{\tau j}^{FB}) + (1-\pi) q_{1j}^{FB} \epsilon(q_{\tau j}^{FB})}{\pi q_{\tau j}^{FB} + (1-\pi) q_{1j}^{FB}}. \tag{B.18}$$

From Lemma 1, $\epsilon(q) = \eta\left(\frac{\beta_0}{q} - 1\right)$. Plugging the expression into equation (B.18) we obtain

$$\tilde{\epsilon}_j = -\eta + \frac{\eta \beta_0}{\pi q_{\tau j}^{FB} + (1-\pi) q_{1j}^{FB}} \tag{B.19}$$

Since both $q_{1j}^{FB}$ and $q_{\tau j}^{FB}$ are decreasing in $c_j$, we have that $\tilde{\epsilon}_j$ is increasing in $c_j$. Hence, $\tilde{\mu}_j$ is decreasing in $c_j$. Let $\bar{c}_j$ be the cost of a firm for which the planner's optimal tax is equal to zero. Denote by $\bar{\mu}_j$ the markup of that firm. For all $c_j > \bar{c}_j$, we have that $\mu_j < \bar{\mu}_j$ and that $t_j < 0$. Similarly, for all $c_j < \bar{c}_j$, we have that $\mu_j > \bar{\mu}_j$ and that $t_j > 0$.

■

Proposition **13** (Normalization of $\beta_1$)**.** *Holding fixed the set of structural parameters other than $\beta_1$, $\{\beta_0, \eta, \tau, \pi, \theta\}$, the markups and allocations in the market equilibrium with linear pricing as well as allocations in the first-best allocation are identical for all $\beta_1 > 0$.*

Proof of Proposition **13.**

The price $p_j$ and the two quantities $q_{1j}$ and $q_{\tau j}$ are given by equation (B.5) and the two constraints in the firm problem (B.2). Let $\widetilde{\beta_1} \equiv \beta_1 P^{\eta}$. Using Assumption 1, we can rewrite the three equilibrium conditions as

$$p_j = \widetilde{\beta_1}^{\frac{1}{\eta}} \left(q_{1j} + \beta_0\right)^{-\frac{1}{\eta}} \tag{B.20}$$

$$p_j = \widetilde{\beta_1}^{\frac{1}{\eta}} \tau \left(q_{\tau j} + \beta_0\right)^{-\frac{1}{\eta}} \tag{B.21}$$

$$\frac{p_j}{p_j - c_j} = \frac{\pi q_{\tau j} \eta \left(\frac{\beta_0}{q_{\tau j}} + \beta_0\right) + (1-\pi) q_{1j} \eta \left(\frac{\beta_0}{q_{1j}} + \beta_0\right)}{\pi q_{\tau j} + (1-\pi) q_{1j}} \tag{B.22}$$

The first two equations only depend on $\widetilde{\beta_1}$ and the third is entirely independent of $\beta_1$. So, for any ${\beta_1}'$ there is a $P'$ such that $\widetilde{\beta_1}' = \widetilde{\beta_1}$ and hence allocations and prices are unchanged. Note that $P'$ ($P^{FB'}$) is the unique price index that clears the labor market, and hence the equilibrium level of the price index.

The first-best allocations also solve Equations (B.20) and (B.21) with $P^{FB}$ instead. With $\widetilde{\beta^F B_1} \equiv \left(\beta_1^F B\right)^\eta$, the allocations are independent of $\beta_1$ by the same argument.

■

Proposition **14** (Normalization of $\beta_0$)**.** *Consider a set of structural parameters $\{\beta_0, \beta_1, \eta, \tau, \pi, \theta\}$. If we multiply $\beta_0$ by a constant $\alpha > 0$ and divide $c_j$ for all $j$ by the same constant, then:*

1. *Markups in the market equilibrium with linear pricing are identical.*
2. *Allocations in both the market equilibrium and the first-best with linear pricing are scaled by the constant $\alpha$.*

Proof of Proposition **14.**

The price $p_j$ and the two quantities $q_{1j}$ and $q_{\tau j}$ are again given by equation (B.5) and the two constraints in the firm problem (B.2). Let $\beta_0' = \alpha\beta_0$, $c_j' = c_j/\alpha$. Conjecture that $q_{ij}' = \alpha q_{ij}$ and $p_j' = p_j/\alpha$. Using Lemma 1, we can again show that the three optimality conditions hold for all $\alpha > 0$.

$$p_j' = \beta_1^{\frac{1}{\eta}} P' \tau \left(q_{1j}' + \beta_0'\right)^{-\frac{1}{\eta}} \tag{B.23}$$

$$p_j' = \beta_1^{\frac{1}{\eta}} P' \left(q_{1j}' + \beta_0'\right)^{-\frac{1}{\eta}} \tag{B.24}$$

$$\frac{p_j}{p_j - c_j} = \frac{\pi q_{\tau j}' \epsilon(q_{\tau j}') + (1-\pi) q_{1j}' \epsilon(q_{1j}')}{\pi q_{\tau j}' + (1-\pi) q_{1j}'} \tag{B.25}$$

Setting $P' = \alpha^{\frac{1-\eta}{\eta}}$, all three conditions hold. For the third optimality condition, we used the fact that $\epsilon(q') = \eta\left(\frac{\beta_0'}{q'} + 1\right)$ and hence independent of $\alpha$. Since all costs are divided by $\alpha$, the labor ($l_j' = q_j' c_j' = \alpha q_j \, c_j/\alpha = l_j$) needed to produce the allocations for the scaled $\beta_0$ is unchanged. Therefore $P' = \alpha^{\frac{\eta-1}{\eta}} P$ is indeed the equilibrium level of the price index.

The first-best allocations also solve Equations (B.23) & (B.24), with $c_j'$ instead of $p_j'$ and the price index replaced by $P^{FB}$, the Lagrange multiplier on the aggregate resource constraint. For $\beta_0' = \alpha\beta_0$ we can choose $P^{FB'} = \alpha^{\frac{1-\eta}{\eta}} P^{FB}$. All first-best allocations are then scaled by $\alpha$. With the scaled down production costs, the labor market clears and we confirm that $P^{FB'}$ is indeed the inverse Lagrange multiplier on the planner's problem.

■

# Online Appendix

## A Continuum of Types

In this appendix, we set up the baseline model from Section 2 for an environment in which consumer tastes are drawn from a continuous distribution. We show that the propositions and proofs remain the same. We then compare the quantitative results to the baseline calibration from Section 5. The implied price dispersion is somewhat smaller, but the allocation of goods closely resembles the two types model.

### A.1 Theory: Model Setup

Household preferences are as before, with the only difference that taste shifter $\tau_{ij}$ are drawn from a cumulative distribution function $G(\tau)$ with support on $[1, \bar{\tau}]$. The CDF $G$ is continuously differentiable, and has non-decreasing hazard rate, $h(\tau) \equiv \frac{g(\tau)}{1-G(\tau)}$.[25]

**Firms.** Each firm $j$ chooses a pricing schedule $p(q)$ that maximizes expected profits. This pricing schedule also implies a mapping of consumer taste $\tau$ to a quantity purchased $q(\tau)$. Since firms cannot condition on type, they must ensure that consumers self-select into their type's bundle.

$$\max_{\{q_j(\tau), p_j(q)\}} \quad \int_{\tau} q_j(\tau) \left(p_j(q_j(\tau)) - c_j\right) dG(\tau) \tag{A.1}$$
$$q_j(\tau) \in \operatorname*{argmax}_{q \geq 0} \left[\tau u(q) - \frac{p_j(q)q}{P}\right], \qquad \forall \tau$$

The set of constraints in Problem (A.1) states that each consumer type $\tau$ must prefer their allocation to not buying the good ($q = 0$, the IR constraint) and to buying any other positive quantity (the set of IC constraints).[26] We solve the problem of the firm using standard tools from the mechanism design literature (see Fudenberg and Tirole, 1991). In the solution to this problem, the individual rationality constraint binds for the lowest types ($\tau_{ij} = 1$), while the set of incentive compatibility constraints for these consumers are slack. For all other consumers, the only binding constraint is the downward local incentive compatibility constraint.

**Firm-level optimal prices and quantities.** The quantity sold to consumers of a particular taste $\tau$ is implicitly given by

$$\tau u'(q_j(\tau)) = \frac{c_j}{P} \frac{\tau}{\tau - [h(\tau)]^{-1}} \tag{A.2}$$

[25] This assumption is common and necessary in order to use the standard mechanism design tools, see Myerson (1981)

[26] As before, we assume that the distribution of tastes $G(\tau)$, the distribution of firm productivities $F(c)$ and preference parameters are such that all firms optimally choose to serve all types of consumers.

Firms choose a quantity $q_j(\tau)$ that equates the marginal utility of each consumer, $\tau u'(q_j(\tau))$, to the *effective cost* of the good. The effective cost consists of two components. First, the real marginal cost of producing the good is $c_j/P$. Second, selling an additional unit entails a *shadow* cost. In order to ensure that consumers with higher taste are still willing to purchase their designated quantity, the prices these consumers pay must go down.

In choosing the optimal quantity offered to consumers with taste $\tau$, the firm takes into account the measure of consumers with that given taste, $g(\tau)$, who will now purchase an additional unit, relative to the measure of consumers with a higher taste for the good, $1 - G(\tau)$, who must now be charged a marginally lower price. This is the hazard rate $h(\tau)$. The higher is the hazard rate, the higher is the measure of consumers with taste $\tau$ relative to consumers with higher tastes, and the lower is the shadow cost of selling an additional unit to consumers with taste $\tau$.

Markups charged by the firm are given by

$$\mu_{ij} = \psi\left(q_{ij}\right) \frac{\tau_{ij}}{\tau_{ij} - h^{-1}(\tau_{ij})} \left[1 - \frac{\int_0^i \tau_{kj} u(q_{kj}) dk}{\tau_{ij} u(q_{ij})}\right] \tag{A.3}$$

The term $\psi(q)$ is the *social markup*, a term coined by Dhingra and Morrow (2019). If firms could perfectly price discriminate, they would extract the full consumer surplus from each of their consumers. The markup charged from each consumer would be equal to the social markup $\psi(q_{ij})$. With nonlinear pricing, firms are able to extract the full consumer surplus only of the consumers with the lowest taste. Consumers with a high taste on the other hand have a positive consumer surplus, which is necessary to achieve separation.

**Efficient Allocation** The first-best allocation solves the planner's problem as in Equation (2.10). The optimal allocations are given by

$$u'(q_{ij}^{\text{FB}}) = \frac{c_j}{\tau_{ij}} \frac{1}{P^{\text{FB}}}, \tag{A.4}$$

where $P^{\text{FB}}$ is the inverse Lagrange multiplier on the aggregate resource constraint.

## A.2 Theory: Propositions and Proofs

Proposition **15.** *In equilibrium, there is a cut-off taste $\hat{\tau}$ for each good $j$ such that all consumers with $\tau > \hat{\tau}$ are allocated too much, and all consumers with $\tau < \hat{\tau}$ are allocated too little of the good.*

Proof of Proposition **15.** From equations (A.2) and (A.4) we have that:

$$\frac{u'\left(q_{\tau j}\right)}{u'\left(q_{\tau j}^{FB}\right)} = \frac{P^{FB}}{P} \omega(\tau) \tag{A.5}$$

where $\omega(\tau) \equiv \frac{\tau}{\tau - [h(\tau)]^{-1}}$ Given that the hazard rate is non-decreasing, $\omega(\tau)$ in decreasing in $\tau$. Further, $\omega(\bar{\tau}) = 1$ and hence $\omega(\tau) \geq 1 \, \forall \tau$.

As in the model with two types, one of three cases must hold: (i) $P^{FB}/P > 1$ and therefore $q_{\tau j} < q_{\tau j}^{\text{FB}}$ $\forall\{\tau, j\}$, (ii) $P^{FB}/P \leq \omega(1)$ and therefore $q_{\tau j} \geq q_{\tau j}^{\text{FB}}$ $\forall\{\tau, j\}$, or (iii) $P^{FB}/P \in (\omega(1), 1)$ and therefore, for each $j$, $q_{\tau j} > q_{\tau j}^{\text{FB}}$ for some $\tau$ and $q_{\tau j} < q_{\tau j}^{\text{FB}}$ for others.

Only (iii) is consistent with labor market clearing. Let $\hat{\tau}$ be given by $\omega(\hat{\tau}) = P^{FB}$. Given we are in case (iii), $\hat{\tau} \in (1, \bar{\tau})$. It follows that first, for all $j$ $q_{\hat{\tau} j} = q_{\hat{\tau} j}^{FB}$. Second, since $\omega'(\tau) \leq 0$, $q_{\hat{\tau} j} > q_{\hat{\tau} j}^{FB}$ $\forall\, \tau > \hat{\tau}$ and $q_{\hat{\tau} j} < q_{\hat{\tau} j}^{FB}$ $\forall\, \tau < \hat{\tau}$.

■

PROPOSITION **16.** *Suppose preferences satisfy Assumption 1. Then, the equilibrium levels of firm-level production and employment are identical to the efficient allocation.*

PROOF OF PROPOSITION **16.**

From equation (A.2), it follows again that there is a unique level of the aggregate price index such that the labor market clears.

Let $\widetilde{P}_j$ be the aggregate price index such that the firm-level production of a firm with marginal cost $c_j$ in equilibrium is identical to its overall production in the efficient allocation.

$$\int_1^{\hat{\tau}} \left[ q_j^{FB}(\tau) - q_j(\tau, \widetilde{P}_j) \right] dG(\tau) - \int_{\hat{\tau}}^{\bar{\tau}} \left[ q_j(\tau, \widetilde{P}_j) - q_j^{FB}(\tau) \right] dG(\tau) = 0 \tag{A.6}$$

By the same argument as in the Proof of Proposition 3, Assumption 1 implies that $\widetilde{P}_j$ is independent of firm cost hence total production is equal to first-best for all firms.

■

PROPOSITION **17.** *Suppose preferences satisfy Assumption 1. Then, the optimal firm-level subsidies and taxes are zero.*

PROOF OF PROPOSITION **17.**

Let's first set up the planner's problem using the primal approach. The planner chooses taxes and subsidies to all firms, $\{t_j\}$, such that its budget is balanced. By choosing taxes and subsidies the planner has control over the firm-level employment of all firms in the economy. We take the primal approach and write the planner's problem as follows:

$$\begin{aligned} \max_{\{l_j, q_j(\tau)\}_{j=0}^1} \quad & \int_0^1 \int_\tau \tau u\left(q_j(\tau)\right) g(\tau) d\tau dj, && \\ \text{s.t.} \quad & \frac{\tau}{\omega(\tau)} u'(q_j(\tau)) = \bar{\tau} u'(q_j(\bar{\tau})) && \forall\, (\tau, j) \\ & \int_\tau q_j(\tau) g(\tau) d\tau = \frac{l_j}{c_j}, && \forall\, j \\ & \int_0^1 l_j dj = 1. && \end{aligned} \tag{A.7}$$

Taking first order conditions, we obtain

$$[q_j(\tau)]: \qquad \tau u'\left(q_j(\tau)\right) g(\tau) - \mu_j(\tau)\frac{\tau u''(q_j(\tau))}{\omega(\tau)} g(\tau) = \theta_j g(\tau), \tag{A.8}$$

$$[q_j(\bar{\tau})]: \qquad \bar{\tau} u'\left(q_j(\bar{\tau})\right) g(\bar{\tau}) + \int_\tau \mu_j(\tau)\, \bar{\tau}\, u''(q_j(\bar{\tau}))\, g(\tau)\, d\tau = \theta_j g(\bar{\tau}), \tag{A.9}$$

$$[l_j]: \qquad \frac{\theta_j}{c_j} = \lambda, \tag{A.10}$$

where $\mu_j(\tau)$, $\theta_j$, and $\lambda$ are the (sets of)s Lagrange multipliers on the three constraints, respectively. Combining conditions (A.8) and (A.9), we get

$$\bar{\tau} u'\left(q_j(\bar{\tau})\right) g(\bar{\tau}) + \bar{\tau}\int_\tau \omega(\tau) u'(q_j(\tau)) \frac{u''(q_j(\bar{\tau}))}{u''(q_j(\tau))} g(\tau) d\tau = \left[ g(\bar{\tau}) + \bar{\tau}\int_\tau \frac{\omega(\tau)}{\tau}\frac{u''(q_j(\bar{\tau}))}{u''(q_j(\tau))} g(\tau) d\tau \right] \theta_j \tag{A.11}$$

Substituting out $\theta_j$ using (A.10) and using the fact that, under Assumption 1 $u''\left(q_j(\tau)\right)/u''\left(q_j\bar{\tau}\right) = \left(u'\left(q_j(\tau)\right)/u'\left(q_j(\bar{\tau})\right)\right)^{1+\eta}$, it follows that the optimality condition of the planner (A.11) holds at the market allocations characterized by (A.2). The resulting Lagrange multiplier $\lambda$ on the aggregate resource constraint is given by

$$\lambda = \frac{1}{P}\frac{g(\bar{\tau}) + \bar{\tau}^{-\eta}\int_\tau \omega(\tau)^{1-\eta}\tau^\eta g(\tau) d\tau}{g(\bar{\tau}) + \bar{\tau}^{-\eta}\int_\tau \omega(\tau)^{-\eta}\tau^\eta g(\tau) d\tau} \tag{A.12}$$

which is indeed independent of firm $j$. We conclude that the equilibrium allocations coincide with the constrained efficient allocation. Therefore, the optimal firm-level taxes and subsidies are all zero.

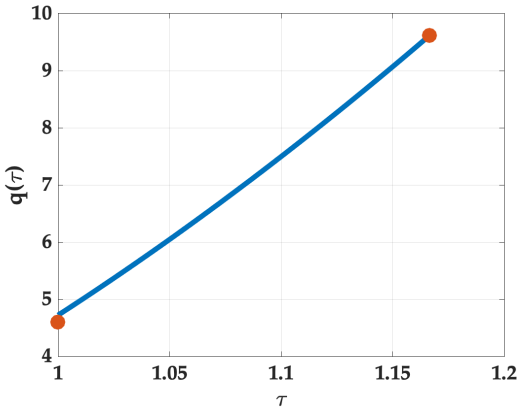
■

### A.3 Quantitative Model

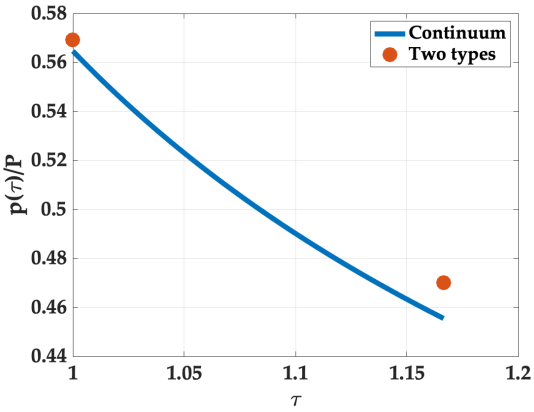
We keep all parameters at the values calibrated for the model with 2 types and set $\bar{\tau} = \tau$. The distribution of types $G(\tau)$ is assumed to follow a uniform distribution. Figure A.1 compares the market allocation under the continuum of types environment to our benchmark environment with two types. We plot the pricing and allocation of the firm with median productivity, but the results are similar for firms of all productivity levels. The left panel presents the quantity produced for each taste and the right panel the relative price charged as a function of consumer tastes. The allocations in the two models are very similar.

Figure A.1: Continuum and Two Types Comparison (Median Firm)

(a) Quantity




(b) Relative price




**Notes**: The two figures present the quantity and relative price of the median firm as a function of consumer taste. The solid blue line refers to the continuum of types market equilibrium, and the two red markers represent the market equilibrium of our benchmark model with two types.