Nonlinear Pricing and Misallocation *

Gideon Bornstein†
Alessandra Peter ‡

August 12, 2022.

Abstract

This paper studies the effect of nonlinear pricing on markups and misallocation. We develop a general equilibrium model of firms that are allowed to set a quantity-dependent pricing schedule—contrary to the typical assumption in macroeconomic models. Without the restriction to linear pricing, markup heterogeneity is no longer a sign of misallocation. Larger firms charge higher markups, yet 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, NYU Macro Lunch, MIT, Harvard, UT Austin, Barcelona Summer Forum, SED, and the NBER EFG for insightful comments. The paper benefited from thoughtful discussions by Michael Peters and Kieran Larkin. 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

Many goods and services feature complicated pricing schedules. Data and phone plans become cheaper with every byte and minute purchased, and a six-pack of beers typically costs less than six individual cans. Despite the apparent prevalence of such pricing schedules, which have been at the center of research in IO for decades, they have been largely absent from macroeconomic models.[1] Firms are usually modeled as choosing a single price at which to sell their output. That is, they are restricted to linear pricing. Together with consumers' first-order conditions, the linear pricing assumption implies a tight relationship between the relative price of a good and its equilibrium quantity.

In this paper, we explore the theoretical and quantitative importance of nonlinear pricing schedules for allocative efficiency. We develop a model of heterogeneous firms that can offer a menu of prices to heterogeneous consumers. We show three key results. First, when firms are not restricted to linear prices, markups are no longer a sufficient statistic for misallocation. Larger firms charge higher markups, yet the allocation of labor across firms is efficient. Second, under nonlinear pricing, a new type of misallocation arises. For each good, consumers with a high taste buy too much of it, consumers with a low taste, too little. Last, we show that nonlinear pricing also breaks the link between the aggregate markup in the economy and the distortion in labor supply.

In the last part of the paper, we quantify 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 linear prices, they would induce large welfare losses. Finally, we find that the undersupply of aggregate labor is an order of magnitude smaller with nonlinear relative to linear pricing.

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 might not be 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 allocation sold to the high-taste consumer equates marginal utility with marginal cost, while 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

[1] For a set of classic examples of markets with nonlinear pricing, see, for instance, Chapter 2 of Wilson (1993).

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 differences* (CED). Under this condition, the difference in the efficient allocation between consumer types is proportional to firm productivity. Many 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 CED, 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. All firms distort allocations across their consumers in exactly the same way. Therefore, reallocating labor across firms cannot alleviate misallocation across consumers.

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 of 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 use micro data on consumer packaged goods from the Nielsen Retail Scanner dataset. The Nielsen dataset has the key advantage that we can see prices paid for different quantities (i.e., package sizes). We first 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 only a 4% increase in price. We then use moments of the data to calibrate the baseline model as well as a model that is identical except for the restriction that firms must charge linear 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 induces 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 5%, which leads to additional welfare losses of 0.2%. Strikingly, a standard linear pricing model would conclude that labor is 34% lower than in the first best allocation. This is true even though the aggregate markup is higher in the baseline model with nonlinear pricing. When implementing the optimal linear pricing subsidies, welfare losses would now be 13% 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)), 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 τ_{ij}. A higher τ_{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 of variety j consumed 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 τ_{ij} can take one of two values: 1 or $\tau > 1$.[2] Each consumer has a high preference τ for a random subset of goods of measure π. 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., to commit 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. We assume that firms must offer a single menu to all households, rather than tailoring the price schedule $p(q)$ to each individual consumer. This assumption reflects that, 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.[3]

Given the price schedule, each consumer chooses the quantity that maximizes her utility. Firms anticipate consumers' choices and solve the following problem:

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

[2] 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.

[3] 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: (i) the individual rationality constraint of the low type (IR_1)—the consumer with a low taste must have non-negative surplus from her bundle; and (ii) the incentive compatibility constraint of the high type (IC_τ)—the high-taste consumer must weakly prefer her bundle to the one tailored toward the low-taste consumer. 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.[4]

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

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

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

For both types of consumers, firms choose a bundle that equates marginal revenue to the cost of supplying an additional unit. Marginal revenue is always equal to the marginal utility, as that is the consumer's willingness to pay for an additional unit.

Equation (2.4) pins down the optimal quantity sold to high-taste consumers. For the high-taste consumer, the marginal cost of supplying an additional unit is simply equal to the marginal production cost, c_j/P. 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

[4]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.

"no distortion at the top" result only holds *conditional* on the aggregate price index P. Below, we show that P is not equal to the price index prevailing in the efficient allocation and hence the classic result of no distortion at the top no longer holds in general equilibrium.

Equation (2.5) pins down the optimal quantity sold to low-taste consumers.[5] 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, in order for them to remain indifferent between the two bundles. The amount by which they must reduce the charges is equal to $(\tau-1)u'(q_{1j})$, 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 price charged for 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 $(1-\pi)/(1-\tau\pi) > 1$. The wedge is increasing in the share of high types π as well as the taste difference τ. The larger the fraction of high-taste consumers buying from any firm, the more they are willing to distort the allocation to the low-taste ones in order to be able to charge more at the top. Similarly, the bigger is the taste difference, the more attractive each additional unit in the low bundle becomes to high-taste consumers and the more that allocation must be distorted downwards.

Firms are able to extract the full consumer surplus from low-taste customers—they are indifferent between their bundle and not buying from the firm at all. 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 charging the entire consumer surplus for it, and reducing the price charged to the high-taste consumer who therefore gets a positive surplus.

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

[5] We assume that preferences as well as the distribution of c_j are such that all firms are active and optimally choose to serve both types. For a comprehensive analysis of market participation under linear and nonlinear pricing, see Attanasio and Pastorino (2020).

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

For high-taste consumers, the markup (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^{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:

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

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, (2.4) and (2.5), 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.[6] 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, conditional on the aggregate price index in the efficient allocation, all firms would employ less labor 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.

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

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 consumer by firm- or only firm-specific production taxes and subsidies.

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 high type outweighs the oversupply to the low type and vice versa. To this end, it is helpful to define the following property of preferences.

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

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

where mc_j is the real marginal cost of firm j, c_j/P^{FB}.

The elasticity of differences measures how the optimal consumption difference between high- and low-taste consumers of a particular good varies with the cost of producing that good.[7] 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 differences 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 differences (CED). That is,

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

[7] While we defined the elasticity in terms of *taste* differences, one could equally interpret it as an elasticity of *price* differences. 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 δ.

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

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 differences (henceforth, CED), the difference in consumption between high- and low-taste consumers is proportional to firm productivity.

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

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 differences (henceforth, CED), 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. Let $\widetilde{P_j}$ be the price index that equates firm-level production to the first best for a firm with production cost c_j. It 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 reason for that 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 differences in tastes, 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,

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

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 labor demand must be identical to the first best.

Taxes and subsidies. Consider the problem of a social planner who maximizes utilitarian social welfare.[8] We start by considering a planner who has access to fully flexible subsidies. That is, she can set firm- by consumer-specific taxes and subsidies. With access to such a flexible set of instruments, she can restore the efficient allocation. Perhaps more surprisingly, the subsidies that implement the efficient allocation are constant across firms.

PROPOSITION **3.** *A social planner can restore the first-best allocation by implementing a subsidy to sales to low-taste consumers that is constant across firms.*

Note that Proposition 3 holds irrespective of the shape of preferences, i.e., also if Assumption 1 does not hold. In the language of Baqaee and Farhi (2020), the only distortions to resource allocation are the wedges between the marginal utility of low-taste consumers and the marginal production costs (Equation 2.5). Firm-level markups do not distort the allocation of resources. Since the micro distortions do not vary across firms, the planner can restore allocative efficiency by imposing a uniform sales subsidy of $s_1 = \frac{\pi(\tau-1)}{1-\pi}$ for the low-taste consumer.

In reality, implementing a subsidy that applies only to the sales of low-taste consumers is clearly challenging. We therefore consider a restricted problem, in which the planner has access to a set of fully flexible *firm-specific* taxes and subsidies t_j. We model the taxes (or subsidies) set by the social planner as production taxes. When the planner 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:

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

The rationale for using firm-level taxes and subsidies is to shift consumption from high- to low-taste consumers by reallocating production across firms. Whenever the planner subsidizes a firm, its sales to both consumer types increase. If the share of additional production that is sold to low-taste consumer is heterogeneous across firms, then there is scope to alleviate misallocation across consumers. The planner will subsidize firms who sell a larger share of the additional production to low-taste consumers, at the expense of other firms.

We show in Proposition 4, however, that under the maintained assumption on preferences, misallocation across consumers cannot be mitigated using firm-level taxes. In fact, the optimal taxes set

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

by the social planner are identically zero.[9]

PROPOSITION **4.** *Suppose preferences satisfy Assumption 1. Then, imposing no subsidies and taxes at the firm level is optimal.*

To understand why Assumption 1 implies that the planner does not want to impose any taxes, we use the following property of CED preferences.[10]

LEMMA **1** (Implications of constant elasticity of differences.)**.** *Suppose preferences $u(.)$ satisfy Assumption 1. Then $q_{\tau j} = \alpha_{0\tau} + \alpha_{1\tau} q_j$ and $q_{1j} = -\alpha_{01} + \alpha_{11} q_j$ for some scalars $\alpha_{0\tau}, \alpha_{01} \geq 0, \alpha_{1\tau}, \alpha_{11} > 0$.*

Lemma 2 shows that firms' expansion curves are *linear*. That is, for every additional unit of firm-level production, q_j, a fraction $\pi\alpha_{1\tau}$ is sold to high-taste consumers and a fraction $(1-\pi)\alpha_{11}$ to low-taste ones. Importantly, these fractions are independent of firm cost and hence constant across firms. Whenever the social planner reallocates labor across firms, such reallocation does not affect the misallocation across consumers. Therefore, she has no incentive to reallocate production, and the optimal taxes and subsidies are zero.[11]

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:

$$\begin{aligned} \max_{\{p_j, q_{1j}, q_{\tau j}\}} \quad & (\pi q_{\tau j} + (1-\pi) q_{1j}) (p_j - c_j) \\ \text{s.t.} \quad & \tau u'(q_{\tau j}) = \frac{p_j}{P}, \qquad (3.5) \\ & u'(q_{1j}) = \frac{p_j}{P}. \qquad (3.6) \end{aligned}$$

No misallocation within firms. From the two demand curves (3.5)–(3.6), 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

[9] 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.

[10] We are grateful to Michael Peters for helping us clarify the intuition behind Proposition 4.

[11] When preferences do not feature CED, the expansion curves are not necessarily linear and there is potential to reduce misallocation by taxing and subsidizing firms. In Appendix B, we explore the properties of such firm-level subsidies quantitatively using Kimball preferences.

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.[12]

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

$$\left(\frac{\tau_{ij}u'\left(q_{ij}\right)}{c_j}\right) \Big/ \left(\frac{\tau_{lk}u'\left(q_{lk}\right)}{c_k}\right) = \frac{\mu_j}{\mu_k}, \qquad \forall\left\{(i,l) \in \{1,\tau\}, (j,k) \in [0,1]\right\} \tag{3.7}$$

where $\mu_j \equiv \frac{p_j}{c_j}$ is the markup of firm j. We obtain equation (3.7) 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.[13] 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 ϵ_j:

$$\mu_j = \frac{\epsilon_j}{\epsilon_j - 1}. \tag{3.8}$$

The effective demand elasticity, ϵ_j, is a weighted average of the demand elasticities of the two consumers:

$$\epsilon_j = \alpha_j\, \epsilon(q_{\tau j}) + (1 - \alpha_j)\, \epsilon(q_{1j}), \tag{3.9}$$

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 following proposition.

PROPOSITION **5.** *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:*

[12] 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.

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

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 subsidies are increasing in productivity.*

Proposition 5 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 the higher a firm's productivity, the more it restricts 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 free to set pricing schedules. Note that in the baseline economy with nonlinear pricing, there is markup heterogeneity as well.[14] The higher a firm's productivity, the more it sells and the higher the markup it charges to consumers (Proposition 10 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.

The key result of no misallocation across firms despite markup heterogeneity holds in a world without consumer heterogeneity as well. If all consumers have the same preferences, or firms are allowed to perfectly price discriminate, there would also be no relationship between markups and allocative efficiency.

4 Elastic labor supply

In the previous sections, we assumed that aggregate labor supply is inelastic (i.e., the Frisch elasticity is zero). We now extend the analysis to the case of elastic labor supply, in which market power of firms also has the potential to distort the overall employment level in the economy.

Setup. Household utility is given by

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

[14] This is true as long as preferences display variable elasticity of substitution. When preferences exhibit CES, markups are constant across firms, just as in a model with linear pricing.

where l_i is the amount of labor supplied by household i, ν governs the degree of disutility of labor, and φ is the inverse Frisch elasticity. We denote by L the aggregate level of labor in the economy. Since all households face the same aggregate price index, all households supply the same level of labor and $l_i = L$. The intratemporal FOC of the household equates the marginal disutility of labor to the marginal value of an extra unit of income.

$$\nu L^{\varphi} = \frac{1}{P}, \tag{4.2}$$

As before, the aggregate price index P measures the value of an additional dollar. Given P, the firm's problem and equilibrium allocations are identical to the ones derived 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 $\{q_{1j}, q_{\tau j}, L\}_{j=0}^{1}$ as well as an aggregate price index P such that prices and quantities solve the firm's problem, the household's intratemporal FOC (4.2) holds, and the labor market clears:

$$\int_0^{\infty} \left(\pi q_{\tau j} + (1-\pi) q_{1j}\right)\, c_j\, dF(c_j) = L. \tag{4.3}$$

4.1 Efficient allocation

The social planner solves the following problem:

$$\begin{aligned} \max_{\{l_i, q_{ij}\}} \quad & \int_i \left[\left(\int_j \tau_{ij} u(q_{ij}) dj \right) - \nu \int_i \frac{l_i^{1+\varphi}}{1+\varphi} \right] di \\ \text{s.t.} \quad & \int_i \int_j q_{ij} c_j dj di = \int_i l_i di. \end{aligned}$$

The optimal allocations and aggregate labor supply are given by

$$\tau_{ij} u'(q_{ij}^{\mathrm{FB}}) = \frac{c_j}{P^{FB}} \tag{4.4}$$

$$\nu \left(l_i^{FB}\right)^{\varphi} = \frac{1}{P^{FB}} \tag{4.5}$$

for all i and j, where P^{FB} is the inverse Lagrange multiplier on the aggregate resource constraint.

As in the case of inelastic labor supply, equation (4.4) 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 their relative marginal costs of production.

4.2 Aggregate labor and misallocation

Undersupply of aggregate labor. We start by comparing the aggregate level of labor in equilibrium and in the efficient allocation.

PROPOSITION **6.** *In the market equilibrium, the aggregate level of labor L is lower than in the efficient allocation.*

Households supply labor until the disutility of a marginal hour worked equals the real wage—the inverse of the aggregate price index. Note that the price index is higher in equilibrium than in the efficient allocation. If the aggregate price in the decentralized equilibrium were equal to P^{FB}, firms would produce the efficient level for high-taste consumers while supplying too little to low-taste consumers. In that case, aggregate labor demand would be lower than aggregate labor supply. It must therefore be that P has to be greater than P^{FB}. With a higher price of consumption, households supply less labor in equilibrium.

Misallocation across firms. With inelastic labor supply, we showed 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 a positive Frisch elasticity. The fact that labor is under-supplied in the aggregate leads to a change in the distribution of market shares across firms.

PROPOSITION **7.** *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.*

To understand the intuition behind Proposition 7, consider the aggregate price index that equates labor demand of all firms to their efficient level of labor. Denote it by $\widetilde{P}$. From Proposition 2 we know such unique aggregate price index exists. Since aggregate labor is lower in the decentralized equilibrium than in the efficient allocation (Proposition 6), it must be that $P < \widetilde{P}$. That is, the aggregate price index in the decentralized equilibrium must be lower than $\widetilde{P}$ to clear the excess demand for labor.

Since the demand elasticity is lower for high-productivity firms, the reduction in the aggregate price index reduces their production by less. So the employment share of high-productivity firms is higher than in the efficient allocation.

Optimal taxes and subsidies. Consider the problem of a planner who can impose firm-level taxes and subsidies and use lump-sum transfers. Proposition 7 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 **8.** *Suppose preferences satisfy Assumption 1. Then, the optimal firm-level subsidies are positive and constant across firms.*

Similar to the case of inelastic labor supply, the planner has no incentive to impose heterogeneous subsidies (Proposition 4). This is true despite the fact that employment shares in the market equilibrium are now different from those in the efficient allocation (Proposition 7). 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 and the optimal subsidy is constant across firms.

The planner's rationale behind a implementing a positive subsidy is the following. When she evaluates whether firm j should be allocated one additional worker, she compares the marginal disutility of labor to the marginal utility of the additional consumption:

$$\nu L^{\varphi} = \gamma \frac{\tau u'(c_{\tau j})}{c_j} + (1-\gamma)\frac{u'(c_{1j})}{c_j}, \tag{4.6}$$

where $\gamma \in (0,1)$ is the share of additional production that will be sold to high-taste consumers.[15]

When a household chooses the optimal level of labor they compare the distutility of labor to the real wage, $1/P$. We can use the firm's optimality conditions, equations (2.4–2.5), to rewrite the labor supply condition as follows,

$$\nu L^{\varphi} = \gamma \frac{\tau u'(c_{\tau j})}{c_j} + (1-\gamma)\frac{1-\tau\pi}{1-\pi}\frac{u'(c_{1j})}{c_j}, \tag{4.7}$$

where we've used the fact that $\frac{1}{P} = \frac{\tau u'(c_{\tau j})}{c_j} = \frac{1-\tau\pi}{1-\pi}\frac{u'(c_{1j})}{c_j}$. The only difference between equations (4.6) and (4.7) is the second term on the RHS. The wedge between the marginal utility of low- and high-taste consumers, $\frac{1-\tau\pi}{1-\pi} < 1$, implies that the real wage is lower than the marginal utility of using an additional worker to produce output for consumption by low-taste consumers.

Since the social benefits of an additional unit of labor (RHS of equation (4.6)) are greater than the benefits perceived by households (RHS of equation (4.7)), the planner chooses a positive subsidy to raise the level of aggregate production.

The difference between equations (4.6) and (4.7) demonstrates another stark difference between linear- and nonlinear-pricing models. Under nonlinear pricing, the undersupply of aggregate labor is driven entirely by the marginal utility wedge between low- and high-taste consumers. Under linear pricing, Edmond et al. (2021) show that the undersupply of aggregate labor is driven entirely by the aggregate markup. Relaxing the linear pricing restriction breaks the link between the aggregate labor wedge and the aggregate markup in the economy.

5 Quantitative exploration

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 (package size). We start by showing that nonlinear pricing is abundant and quantitatively significant. We then use product-level data on sales and purchases to calibrate the structural parameters of the 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 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

[15] The proof of Proposition 8 shows that γ is constant across all firms.

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.

Finally, we study the inefficiency from distortions 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.[16] We focus on core grocery goods, which include the departments of dry groceries, frozen food, and dairy.[17] We use data from a single week in 2017.[18]

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 the model, we show that nonlinear pricing is abundant in the data. First, the vast majority of products are sold in 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 fixed effects. Table 2 presents the results. The first specification includes both product line and store-level fixed effects. The second specification includes product line

[16] 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.

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

[18] 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)	(3)	(4)
Size (ln)	-0.61	-0.64	-0.56	-0.39
(s.e.)	(0.0001)	(0.0001)	(0.0004)	(0.0003)
Product line & store f.e.	✔			
Product line × store f.e.		✔	✔	✔
Sample	All	All	Dairy	Expensive
Observations	88.3M	69.7M	5.0M	4.5M

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.

by store fixed effects. The estimates suggest that the price of a 10% larger package size is only about 4% higher.

Through the lens of our model, prices are optimally chosen by firms in order to cater to consumers with different tastes. In the data, some consumers may purchase large package sizes and store the goods (see, e.g., Nevo and Wong (2019) and Baker, Johnson, and Kueng (2020)). The degree of nonlinear pricing could therefore also be a reflection of storage costs, credit constraints, or higher demand elasticities of consumers who tend to buy in bulk. To alleviate this concern, we consider a specification in which we restrict the sample to dairy products, which have short shelf lives. Column (3) in Table 2 shows that also for this sample, where storage is less likely to be feasible, we see a substantial degree of nonlinear pricing.

In the theory, production costs scale linearly with size. In the data, the cost of, for example, packaging might not grow linearly with product size, implying that larger packages have a lower average production cost. This is less of a concern for expensive products, for which the indirect costs such as packaging are relatively small. Specification (4) in Table 2 restricts the sample to the 5% most expensive product lines, those with an average price above $7.99. We find substantial and significant degree of nonlinearity also for these products. The point estimate suggests that a 10% larger package size is associated with only a 6% higher price.[19]

By interpreting the quantity discounts in the data as nonlinear pricing, we implicitly assume that packages of different sizes are perceived as the same good by consumers. For some goods, such as a six-pack of beers, this is likely to be a good approximation. For others, such as large vs small bottles of carbonated beverages, differently sized packages might not be perfect substitutes. The assumption of perfect substitutability across package sizes is stronger than necessary for the main results. For instance, our model can accommodate the notion that purchasing a smaller quantity could be more desirable for all consumers.[20]

[19] Note that a fixed cost per unit sold, such as for example the time it takes to restock shelves or the cashier to scan the item cannot explain the patterns we see. In both a linear and nonlinear model, these costs are fixed per unit sold and therefore only affect the extensive margin of sizes offered, but not the price per unit.

[20] Suppose for example that that there is a penalty v for purchasing the larger package (e.g., the large bottle of Coca-

5.2 Calibration

Recall that our model consists of two types of households and, therefore, two different sizes offered to consumers. To map the data to the model, we split products in each product line into two categories: small and large. All products smaller than the median 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.[21]

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%.

We assume that preferences satisfy Assumption 1. In Lemma 2 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 β_0, β_1, and η are structural parameters.

Firm productivity is assumed to follow a Pareto distribution with shape parameter θ.[22] Following the macro literature, we set the Frisch elasticity to 1 ($\varphi = 1$). We normalize $\beta_0 = \beta_1 = 1$. This normalization is without loss of generality.[23] The disutility of labor, ν, is calibrated such that in the market equilibrium, aggregate labor supply is equal to 1.[24]

We calibrate the four structural parameters—the elasticity of differences η, the taste shifter τ, the share of high-taste consumers π, and the Pareto shape θ—to match six key moments in the data. We set the fraction of high-taste consumers, π, to match the share of purchases of the large size in the data. The mapping between this data moment and π 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.

Identification of η, θ, and τ. The size difference between the large and small bundles is informative of both the taste shifter τ, and the elasticity of differences η. A large size difference can be rationalized

Cola loses its carbonation). This penalty applies irrespective of the taste shifter for the good. The utility of consuming the small bundle is still $\tau_{ij}u(q_{1j})$, while consuming the large bundle gives utility $\tau_{ij}(1-\upsilon)u(q_{\tau j})$. One can show that the allocations in such alternative model are identical to our benchmark specification but with $\tau_{\text{new}} \equiv \tau(1-\upsilon)$.

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

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

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

[24] 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 ν adjusts so that aggregate labor is equal to 1 regardless of the level of φ.

Table 3: Calibrated Moments and Parameters

A. Moments

Moment	Data	Model	
		Benchm.	Linear pricing
Fraction buying large q	51%	51%	51%
$\mathbb{E}\left[\ln q_{j\tau} - \ln q_{j1}\right]$	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%	96.7%	95.4%

B. Parameters

Parameter		Model	
		Benchm.	Linear pricing
π	Share of high-taste consumers	0.51	0.51
τ	High-taste demand shifter	1.17	1.35
η	Elasticity of differences	1.86	2.13
θ	Pareto shape	0.86	1.13
Externally Set			
φ	Inverse Frisch elasticity	1	1

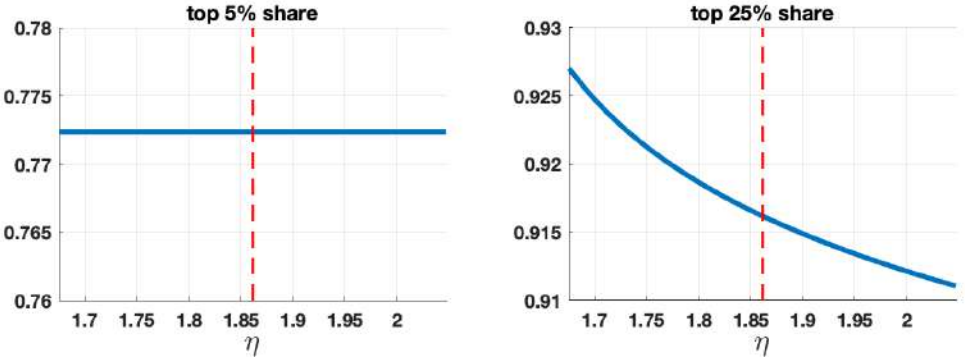
Notes: Panel A presents the model fit. The data column presents the moments we target in our estimation procedure. The second column presents the model moments of our benchmark specification in which firms can offer nonlinear pricing schedules. The third column presents the model moments for a specification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

by either a large taste shifter or a higher elasticity of differences.

The distribution of sales across firms is informative of both the elasticity of differences η, and the Pareto shape θ. Recall that our model features variable elasticity of substitution. For the largest firms, the elasticity is equal to η. At the bottom of the productivity distribution, the elasticity depends less and less on η and goes to ∞. Therefore, the market share of the top 5% of firms is governed by both η and θ, while the market share of the top 25% depends primarily on θ.

To illustrate the joint identification of η and θ, we consider combinations of both parameters such that the market share of the top 5% of firms remains fixed. Figure 1 plots the market share of the top 25% for these different combinations of η and θ.[25] When η is higher, the same productivity difference translates into a larger quantity difference. To keep the market share of the top 5% of firms fixed, the tail of the productivity distribution must be thinner. That is, the value of θ must rise with η. Since the elasticity of differences η is relatively more important for high productivity firms, when both η and θ go up, the market share of the top 25% goes down. Therefore, by matching market shares at the top as well as at the middle of the productivity distribution, we can separately identify η and θ.

Figure 1: Identification of the Pareto shape and elasticity of differences



Notes: This figure plots the market shares of the top 5% and 25% when varying the elasticity of differences and the Pareto shape parameter so that the top 5% market share remains unchanged. The vertical dashed line indicates the point estimate for the elasticity of differences, η.

[25] In the quantification, we also target the market shares of the top 10% and top 50%.

Linear pricing 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 this 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 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.[26]

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 2, 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.

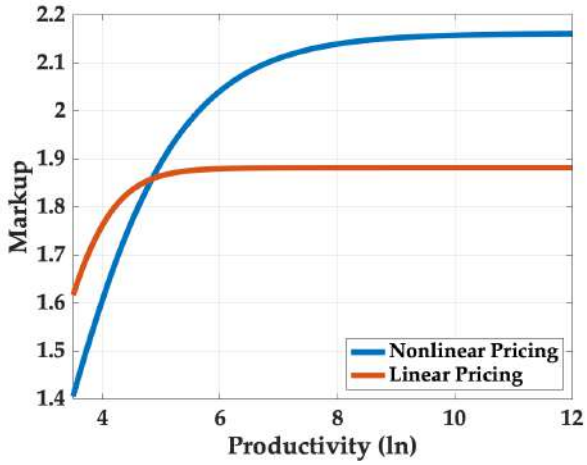
Table 4 reports the welfare implications of misallocation for both models. When firms are not restricted to linear pricing and optimally charge different markups to different consumers, there is no misallocation across firms. There are, however, losses from misallocation across consumers. Consumers are indifferent between consuming the efficient allocation or consuming an additional 0.82% 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 0.27%, 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 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 39 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 3 illustrates the misallocation of consumption across consumers in the

[26] 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.

Figure 2: 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 cost-weighted average markups, which is identical to the ratio of total sales to total costs of the firm.

Table 4: Welfare Gains of Fixing Misallocation

Baseline Model	Linear Pricing	Baseline with LP subsidies
0.82%	0.27%	-0.36%

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.

benchmark model (solid lines). Relative to the equilibrium with perfect allocative efficiency (q_{ij}^{AE}), high-taste consumers are oversupplied each good, whereas low-taste consumers are allocated too little.[27] In terms of marginal utilities, the distortion relative to the allocatively efficient equilibrium is constant. Figure 3 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.

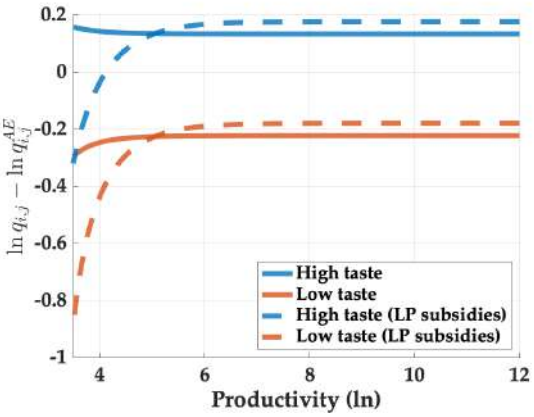
If firms were restricted to linear pricing, there would be no distortion in marginal utilities across consumers. Figure 3b shows that also in terms of quantities, there is essentially no distortion between high and low types. The solid blue and red lines lie on top of each other. Instead, the amount of 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

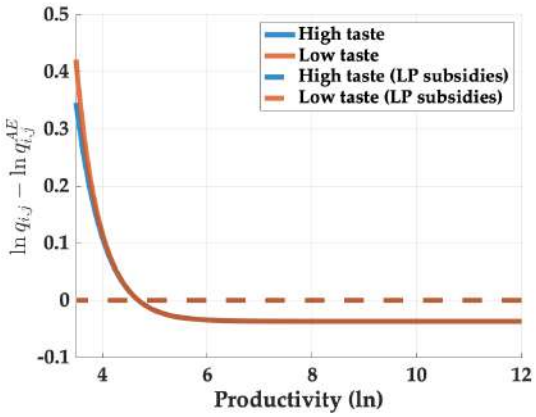
[27] In the equilibrium with perfect allocative efficiency, q_{ij}^{AE} is defined as the quantities chosen by a social planner who is restricted to the same aggregate supply of labor as in the market equilibrium.

Figure 3: 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.

subsidies and taxes. This point is illustrated in Figure 3b.[28]

Panel A in Figure 3 shows the impact of implementing the subsidies that *would be optimal* if firms were restricted to posting linear prices in the nonlinear pricing environment.[29] 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 0.36% (Table 4), which is larger than the welfare gains of 0.27% this policy sets out to achieve.

5.4 Aggregate markup and labor supply

In this section, we evaluate the impact of across-consumer distortions on aggregate labor supply. Table 5 summarizes the results. In the first best allocation, aggregate labor is 4.8% higher than in the market equilibrium. The first-best allocation yields welfare gains of 0.99% relative to the market allocation—that is, an additional 0.17% 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 4.8%, but that increase is not uniform across firms. The employment of the bottom 50% of firms is 6.3% higher in the first-best allocation relative to the market allocation, whereas the employment of the top 50% grows by 4.7%.

A planner with access to only firm-level taxes and subsidies cannot achieve the first-best allocation

[28] 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.

[29] 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.

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

	Nonlinear Pricing		**Linear Pricing**
	FB	SB	FB & SB
Aggregate Labor	+4.8%	+4.6%	+52%
Welfare Gains	+0.99%	+0.15%	+12.1%

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.

as they cannot solve the misallocation of consumption across consumers. They can, however, raise welfare by inducing more workers to join the labor force. To achieve the second-best allocation, the planner imposes a uniform subsidy of 7% across all firms. This policy increases aggregate labor by 4.6% and raises welfare by 0.15% in consumption equivalent units.

When imposing a linear pricing assumption, researchers would conclude that the optimal level of labor is 52% 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 aggregate markup in the linear pricing model is 85%, the distortion in aggregate labor is large.

As explained in Section 4, under nonlinear pricing, the tight link between the aggregate markup and the labor wedge breaks. The estimated level of aggregate markup in the nonlinear pricing model is even larger, yet aggregate labor is only 5% below its optimal level.

Under linear pricing, a social planner with access to 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 (87% on average), but they would also be larger for the large, high-markup firms. If firms are not restricted to linear pricing schedules, implementing these subsidies would lead to large welfare losses on the order of 13%. 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.

5.5 Alternative specifications

In this section, we consider three alternative specifications to the baseline model. First, we quantify a model in which idiosyncratic tastes τ_{ij} are drawn from a continuous distribution and show that equilibrium prices and quantities are similar to the two-type model. Second, we consider a model with a continuum of types, but restrict firms to offer only two bundles. We find that the degree of misallocation is similar to the baseline model. Last, we allow for preferences that do not feature CED and find a still very small role for misallocation across firms.

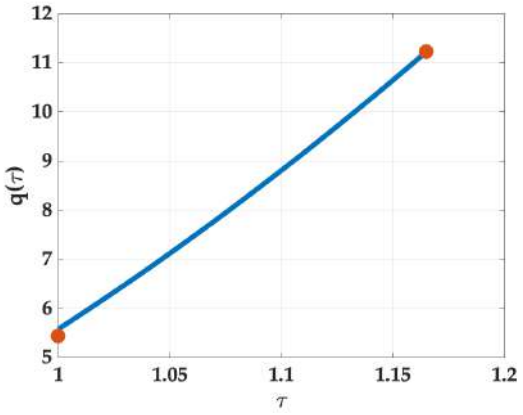
5.5.1 Continuum of tastes with continuum of bundles

In the baseline model, there are two types of consumers facing each firm: $\tau_{ij} \in \{1, \tau\}$. To the extent that there is much more unobserved heterogeneity in the world, approximating it with two types risks understating the degree of distortions across consumers. To quantify this, we re-calibrate the model assuming that tastes are distributed uniformly between 1 and τ. With a continuum of types, instead of a single incentive compatibility constraint, the firm is facing a continuum of local incentive compatibility constraints. We set up the problem and characterize its solution in Online Appendix A.

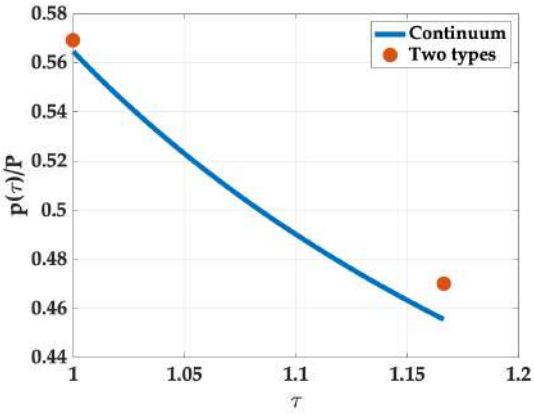
When a firm is facing a continuum of types, under the standard regularity conditions in the mechanism design literature, it offers a smooth nonlinear pricing schedule. The equilibrium price and quantity of the bundles sold to the lowest and highest taste consumers are very similar to those of the benchmark model. Figure 4 plots the quantities and relative prices as a function of τ for the median firm and compares them to the benchmark specification.

Figure 4: Continuum vs. Two Tastes: Median Firm

(a) Quantities



(b) Relative prices



Notes: The solid line represents the allocation as a function of τ of the median firm. The two red markers represent the allocations of the median firm in the benchmark specification.

5.5.2 Continuum of tastes with two bundles

Products are usually offered in just a few sizes. To match this feature of the data, we analyze and estimate a model in which taste shifters τ_{ij} are drawn from a continuous distribution, yet firms may only offer two bundles. We describe the setup in detail in Online Appendix A.4. In this model, a new dimension of misallocation arises: the set of consumers buying the small bundle might be different between planner and market allocations. We show theoretically that with CED preferences, the set of consumers purchasing each bundle is independent of firm productivity. That is, also this additional dimension of misallocation is constant across firms.

Quantitatively, we find that the welfare costs of misallocation under this specification is similar to our benchmark specification. The welfare gain from removing misallocation is 0.92% in consumption

equivalent units, compared to 0.82% in our benchmark specification.

5.5.3 Kimball preferences

To quantify the importance of CED for the main misallocation results, we estimate the baseline model assuming that preferences feature a Kimball (1995) aggregator using the specification of Klenow and Willis (2016).[30] Relative to an environment with CED, there is scope for improving on the market allocation by using firm-level taxes and subsidies. However, the extent of across-firm misallocation is small. While the overall welfare losses from misallocation are 0.53%, welfare only increases by 0.01% with the optimal firm-level taxes and subsidies.

6 Conclusion

Many goods and services feature complicated, nonlinear pricing schedules. We embed this feature of pricing into a macro model by developing a general equilibrium framework of heterogeneous firms that can offer a menu of prices to consumers with different tastes. 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 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 13%.

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. i.e., how does income inequality shape consumption inequality? Does nonlinear pricing lead to inefficiently low levels of consumption for low-income households, and how does misallocation depend on the degree of inequality? Second, do firms with monopsony power set nonlinear wages? And if so, how does this wage setting behavior shape misallocation in the labor market?

[30] For details on model and calibration, see Online Appendix B.

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

Attanasio, O. and E. Pastorino (2020): "Nonlinear Pricing in Village Economies," *Econometrica*, 88, 207–263. 5

Baker, S. R., S. G. Johnson, and L. Kueng (2020): "Financial returns to household inventory management," Tech. rep., National Bureau of Economic Research. 5.1

Baqaee, D. R. and E. Farhi (2020): "Productivity and misallocation in general equilibrium," *The Quarterly Journal of Economics*, 135, 105–163. 3.2, 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, 4.2, 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

Fudenberg, D. and J. Tirole (1991): *Game theory*, MIT press. A.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

Kimball, M. (1995): "The Quantitative Analytics of the Basic NeomonetaristModel," *Journal of Money, Banking, and Credit*, 27. 5.5.3

Klenow, P. J. and J. L. Willis (2016): "Real Rigidities and Nominal Price Changes," *Economica*, 83, 443–472. 5.5.3

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, 3.2

Mirrlees, J. A. (1971): "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, 38, 175–208. 1, 6

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. 32

Nevo, A. and A. Wong (2019): "The elasticity of substitution between time and market goods: Evidence from the Great Recession," *International Economic Review*, 60, 25–51. 5.1

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, 6

Wilson, R. B. (1993): *Nonlinear pricing*, Oxford University Press on Demand. 1, 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 **2** (Further Implications of constant elasticity of differences.)**.** *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 **2.** Let $g(x) \equiv (u')^{-1}(x)$ and $\gamma \equiv \frac{1}{\tau}$. From the definition of the elasticity of differences (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 2. 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 Lemma 1.

Using Lemma 2,

$$q_j \equiv \pi q_{\tau j} + (1-\pi) q_{1j} = -\beta_0 + \beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \hat{\tau} \tag{A.7}$$

where $\hat{\tau} \equiv \pi\tau^{\eta} + (1-\pi)^{1-\eta}(1-\tau\pi)^{\eta}$. Given the linear affine form of q_{1j} and $q_{\tau j}$, these can then be rewritten as

$$q_{1j} = \beta_0 \left(\left(\frac{1-\tau\pi}{1-\pi}\right)^{\eta} \frac{1}{\hat{\tau}} - 1 \right) + \left(\frac{1-\tau\pi}{1-\pi}\right)^{\eta} \frac{1}{\hat{\tau}} q_j \tag{A.8}$$

$$q_{\tau j} = \beta_0 \left(\frac{\tau^{\eta}}{\hat{\tau}} - 1 \right) + \frac{\tau^{\eta}}{\hat{\tau}} q_j \tag{A.9}$$

and Lemma 1 follows with

$$\alpha_{0\tau} = \beta_0 \left(\frac{\tau^{\eta}}{\hat{\tau}} - 1 \right) \tag{A.10}$$

$$\alpha_{1\tau} = \left(\frac{1-\tau\pi}{1-\pi}\right)^{\eta} \frac{1}{\hat{\tau}} \tag{A.11}$$

$$\alpha_{01} = \beta_0 \left(\frac{\tau^{\eta}}{\hat{\tau}} - 1 \right) \tag{A.12}$$

$$\alpha_{01} = \left(\frac{1-\tau\pi}{1-\pi}\right)^{\eta} \frac{1}{\hat{\tau}} \tag{A.13}$$

The sign restrictions are directly implied by Lemma 2 as well as the fact that $\tau \geq 1$. ■

Proof of Proposition 3. Consider a uniform subsidy to the sales of low-taste consumption goods, denoted by s. The firm's problem can then be written as follows

$$\begin{aligned} \max_{\{q_{1j}, q_{\tau j}, p_{1j}, p_{\tau j}\}} \quad & \pi q_{\tau j} \left(p_{\tau j} - c_j\right) + (1-\pi) q_{1j} \left((1+s) p_{1j} - c_j\right) && \text{(A.14)} \\ \text{s.t} \quad & u(q_{1j}) - \frac{p_{1j} q_{1j}}{P} = 0 && [IR_1] \\ & \tau u(q_{\tau j}) - \frac{p_{\tau j} q_{\tau j}}{P} = (\tau - 1) u(q_{j1}) && [IC_{\tau}] \end{aligned}$$

Taking first order conditions, we implicitly obtain the equilibrium quantities from the following

optimality conditions

$$\tau\, u'(q_{\tau j}) = \frac{c_j}{P}, \tag{A.15}$$

$$u'(q_{1j}) \left(s + \frac{1-\pi\tau}{1-\pi} \right) = \frac{c_j}{P} \tag{A.16}$$

If $s = \frac{\pi(\tau-1)}{1-\pi}$, the latter optimality condition becomes

$$u'(q_{1j}) = \frac{c_j}{P}. \tag{A.17}$$

Note that equations (A.15–A.17) are identical to the optimality conditions in the efficient allocation. Therefore, a uniform sales subsidy of $s = \frac{\pi(\tau-1)}{1-\pi}$ to the sales of low-taste consumption goods implements the efficient allocations.

■

Proof of Proposition 4.

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:

$$\begin{aligned} \max_{\{l_j, q_{1j}, q_{\tau j}\}_{j=0}^{1}} \quad & \int_0^1 \left[\pi\tau u\left(q_{\tau j}\right) + (1-\pi) u\left(q_{1j}\right) \right] dj, && \text{(A.18)} \\ \text{s.t.} \quad & u'\left(q_{1j}\right) = \frac{1-\pi}{1-\tau\pi}\, \tau\, u'\left(q_{\tau j}\right), && \text{for all } j \\ & \pi q_{\tau j} + (1-\pi) q_{1j} = \frac{l_j}{c_j}, && \text{for all } j \\ & \int l_j dj = 1. \end{aligned}$$

Taking first order conditions, we obtain

$$[q_{\tau j}]: \qquad \pi\tau u'\left(q_{\tau j}\right) + \frac{1-\pi}{1-\tau\pi}\, \tau\, u''\left(q_{\tau j}\right) \mu_j = \pi\theta_j, \tag{A.19}$$

$$[q_{1j}]: \qquad (1-\pi) u'\left(q_{1j}\right) - u''\left(q_{1j}\right) \mu_j = (1-\pi)\theta_j, \tag{A.20}$$

$$[l_j]: \qquad \frac{\theta_j}{c_j} = \lambda. \tag{A.21}$$

where μ_j, θ_j, and λ are the Lagrange multipliers on the three constraints, respectively. Multiplying equation (A.19) by $\frac{1-\tau\pi}{\tau(1-\pi)} \frac{u''(q_{1j})}{u''(q_{\tau j})}$ and adding to equation (A.20), 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

$$\theta_j = \gamma_j \tau u'(q_{\tau j}) + (1-\gamma_j) u'(q_{1j}), \tag{A.22}$$

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 γ_j represents the share of additional production allocated to high-taste consumers when l_j increases. The third optimality condition (A.21) iwmplies that

$$\frac{\gamma_j \tau u'(q_{\tau j}) + (1-\gamma_j) u'(q_{1j})}{c_j} = \lambda, \quad \text{for all } j. \tag{A.23}$$

Equation (A.23) 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.23). First, using equations (2.6–2.7), the LHS of equation (A.23) becomes

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

Using Lemma 2, 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 γ_j, we see that γ_j is constant across firms in the equilibrium allocation. Denote its value by γ. Therefore, the LHS of equation (A.23) 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.23) holds for all j. The first order conditions of the planner then pin down the values of μ_j and θ_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 **6.** Let $L^D(P)$ denote total labor demand by firms given an aggregate price index P. Since all firm-level quantities are increasing in P, $L^D(P)$ is an increasing function of P. In general equilibrium, labor supply equals labor demand, so that

$$\nu L^D(P)^{\varphi} = \frac{1}{P}. \tag{A.24}$$

Note that the LHS of the equation above is increasing in P, while the RHS is decreasing in P. So there is a unique level of P that clears the labor market.

Denote by P^{FB} the inverse Lagrange multiplier on the aggregate resource constraint of the planner's problem. From the planner's optimality conditions we have that

$$\nu \left(L^{FB}\right)^{\varphi} = \frac{1}{P^{FB}} \tag{A.25}$$

First, we show that $P > P^{FB}$. Suppose by contradiction that $P = P^{FB}$. This directly implies that for all j $q_{\tau j} = q_{\tau j}^{FB}$ and $q_{1j} < q_{1j}^{FB}$, so that $L^D(P^{FB}) < L^{FB}$. Labor supply on the other hand is equal between the market allocation and the first-best When $P = P^{FB}$. Hence the labor market cannot clear if $P = P^{FB}$. Since the LHS is increasing in P and the RHS is decreasing in P, we can similarly rule out any $P < P^{FB}$. It must therefore be that $P > P^{FB}$.

Finally, we use the optimal labor supply condition together with $P > P^{FB}$ to show that aggregate labor supply is lower than L^{FB}:

$$L = \left(\frac{1}{\nu P}\right)^{\frac{1}{\varphi}} < \left(\frac{1}{\nu P^{FB}}\right)^{\frac{1}{\varphi}} = L^{FB}. \tag{A.26}$$

■

Proof of Proposition 7. The excess employment (ω_j) is given by

$$\omega_j = \frac{\frac{(\pi q_{\tau j} + (1-\pi) q_{1j}) c_j}{L}}{\frac{\left(\pi q_{\tau j}^{\mathrm{FB}} + (1-\pi) q_{1j}^{\mathrm{FB}}\right) c_j}{L^{\mathrm{FB}}}} = \frac{\pi q_{\tau j} + (1-\pi) q_{1j}}{\pi q_{\tau j}^{\mathrm{FB}} + (1-\pi) q_{1j}^{\mathrm{FB}}} \frac{L^{\mathrm{FB}}}{L} \tag{A.27}$$

where L and L^{FB} are aggregate labor. Using Lemma 2,

$$\omega_j = \frac{\beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \hat{\tau} - \beta_0}{\beta_1 \left(\frac{c_j}{P^{FB}}\right)^{-\eta} \hat{\tau}^{FB} - \beta_0} \frac{L^{\mathrm{FB}}}{L} \tag{A.28}$$

where $\hat{\tau} \equiv \pi (\tau)^{-\eta} + (1-\pi) \left(\frac{1-\pi}{1-\tau\pi}\right)^{-\eta}$ and $\hat{\tau}^{FB} \equiv \pi (\tau)^{-\eta} + (1-\pi)$.

Taking derivatives wrt c_j

$$\frac{\partial \omega_j}{\partial c_j} = \frac{-\beta_1 c_j^{-\eta-1} P^{\eta} \hat{\tau} \left[\beta_1 \left(\frac{c_j}{P^{FB}}\right)^{-\eta} \hat{\tau}^{FB} - \beta_0\right] + \beta_1 c_j^{-\eta-1} \left(P^{FB}\right)^{\eta} \hat{\tau}^{FB} \left[\beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \hat{\tau} - \beta_0\right]}{\left(\beta_1 \left(\frac{c_j}{P^{FB}}\right)^{-\eta} \hat{\tau}^{FB} - \beta_0\right)^2} \frac{L^{\mathrm{FB}}}{L} \tag{A.29}$$

Rearranging (A.29) and using the fact that $\beta > 0, \beta_1 > 0$ and $c_j > 0$

$$\frac{\partial \omega_j}{\partial c_j} \leq 0$$
$$\iff P^{\eta}\hat{\tau} \leq \left(P^{FB}\right)^{\eta}\hat{\tau}^{FB}$$
$$\text{or } \frac{P^{FB}}{P} \geq \left(\frac{\hat{\tau}}{\hat{\tau}^{FB}}\right)^{\frac{1}{\eta}}$$

We now prove that this condition holds. Let $\tilde{P}$ be the price index that equates labor demand of firm j to its labor demand in the efficient allocation. Using the CED assumption, and following exactly the same steps as Proposition 2, we know that such price index also equates the labor demand of all other firms to their labor demand in the efficient allocation. In such case, $\omega_j = 1$ for all j and $\frac{\partial \omega_j}{\partial c_j} = 0$. So that

$$\frac{P^{FB}}{\tilde{P}} = \left(\frac{\hat{\tau}}{\hat{\tau}^{FB}}.\right)^{\frac{1}{\eta}} \tag{A.30}$$

Note that $\tilde{P} > P^{FB}$ since as long as $\tau > 1$, $\hat{\tau} < \hat{\tau}^{FB}$. But with $\tilde{P} > P^{FB}$, labor supply by households is lower than in first-best. This follows directly from the consumers' intratemporal FOCs (A.24) and (A.25). Therefore, to clear the labor market it must be that $P < \tilde{P}$ and we have

$$\frac{P^{FB}}{P} > \frac{P^{FB}}{\tilde{P}} = \left(\frac{\hat{\tau}}{\hat{\tau}^{FB}}.\right)^{\frac{1}{\eta}} \tag{A.31}$$

■

Proof of Proposition 8.

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:

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

Taking first order conditions, we obtain

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

where μ_j, θ_j, and λ 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:

$$\theta_j=\gamma_j\tau u'\left(q_{\tau j}\right)+\left(1-\gamma_j\right)u'\left(q_{1j}\right), \tag{A.32}$$

where

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

Using the third optimality condition, we obtain

$$\lambda=\gamma_j\frac{1}{c_j}\tau u'\left(q_{\tau j}\right)+\left(1-\gamma_j\right)\frac{1}{c_j}u'\left(q_{1j}\right), \tag{A.33}$$

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

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

where P is the resulting aggregate price index in the economy with firm-level taxes. Plugging this back into (A.33), we see that the optimal level of taxes are given by

$$(1+t_j)=\frac{\lambda P}{\gamma_j+(1-\gamma_j)\frac{1-\pi}{1-\tau\pi}}. \tag{A.34}$$

Using the final optimality condition of the planner and the labor supply condition in the market equilibrium we have that $\lambda=\frac{1}{P}$, so that

$$(1+t_j)=\frac{1}{\gamma_j+(1-\gamma_j)\frac{1-\pi}{1-\tau\pi}}. \tag{A.35}$$

Using Lemma 2, we can write γ_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 γ_j is independent of c_j. Hence, (A.35) implies that firm-level taxes or subsidies are constant across firms j. Further, since $\gamma_j < 1$ and $\frac{1-\pi}{1-\tau\pi} < 1$. Therefore, (A.35) 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 **9.** *Under Assumption 1, an equilibrium exists and is unique.*

PROOF OF PROPOSITION **9.** Using the optimality conditions of the firm, (2.4) and (2.5), write labor market clearing directly as a function of P:

$$\int_0^1 c_j \left[\pi (u')^{-1} \left(\frac{c_j}{P}\frac{1}{\tau}\right) + (1-\pi)(u')^{-1} \left(\frac{c_j}{P}\frac{1-\pi}{1-\tau\pi}\right)\right] dj = 1. \tag{A.36}$$

Using Lemma (2), $\lim_{x\to\infty}(u')^{-1}(x) = -\beta_0 \leq 0$ 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.36) holds. Since marginal utility $u'(.)$ is decreasing everywhere, P is unique.[31] ■

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 9, 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.

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 $\tilde{p}_j$,

$$\frac{1}{P} = \frac{\tau u'(q_{\tau j})}{\tilde{p}_j}, \tag{A.37}$$

[31] 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.

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 $q_{\tau j}$:

$$\tilde{p}_j = c_j. \tag{A.38}$$

Equation (A.38) 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.

PROPOSITION **10.** *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 **10.** Using the firm's optimality conditions to substitute out prices, the inverse of the markup is given by

$$\frac{1}{\mu_j} = \frac{(1-\pi\tau)u(q_{j1})/\psi(q_t)}{(1-\pi\tau)u(q_{j1})+\pi\tau u(q_{j\tau})} + \frac{\pi\tau u(q_{j\tau})/\psi(q_{j\tau})}{(1-\pi\tau)u(q_{j1})+\pi\tau u(q_{j\tau})} \tag{A.39}$$

where $\psi(q) \equiv \frac{u(q)}{qu'(q)}$. Using Lemma 2, we have that

$$\mu_j = \frac{\eta}{\eta-1}\beta_1^{\frac{1}{\eta}} \frac{\pi\tau\left(q_{\tau j}+\beta_0\right)^{\frac{\eta-1}{\eta}} + (1-\pi\tau)\left(q_{1j}+\beta_0\right)^{\frac{\eta-1}{\eta}} - \left(\beta_0\right)^{\frac{\eta-1}{\eta}}}{\pi\tau q_{\tau j}\left(q_{\tau j}+\beta_0\right)^{-\frac{1}{\eta}} + (1-\pi\tau)q_{1j}\left(q_{1j}+\beta_0\right)^{-\frac{1}{\eta}}}. \tag{A.40}$$

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

$$\begin{aligned}\mu_j &= \frac{\eta}{\eta-1}\frac{\pi\tau\beta_1^{\frac{\eta-1}{\eta}}x_\tau^{1-\eta} + (1-\pi\tau)\beta_1^{\frac{\eta-1}{\eta}}x_1^{1-\eta} - \left(\beta_0\right)^{\frac{\eta-1}{\eta}}}{\pi\tau\left(-\beta_0+\beta_1 x_\tau^{-\eta}\right)\beta_1^{-\frac{1}{\eta}}x_\tau + (1-\pi\tau)\left(-\beta_0+\beta_1 x_1^{-\eta}\right)\beta_1^{-\frac{1}{\eta}}x_1} \\ &= \frac{\eta}{\eta-1}\frac{\beta_1^{\frac{\eta-1}{\eta}}\left(c_j/P\right)^{1-\eta}\widetilde{\tau} - \beta_0^{\frac{\eta-1}{\eta}}}{\beta_1^{\frac{\eta-1}{\eta}}\left(c_j/P\right)^{1-\eta}\widetilde{\tau} - \left(c_j/P\right)\beta_0\beta_1^{-\frac{1}{\eta}}},\end{aligned} \tag{A.41}$$

where

$$\widetilde{\tau} \equiv (\pi\tau)^\eta\,\pi^{1-\eta} + (1-\tau\pi)^\eta(1-\pi)^{1-\eta}. \tag{A.42}$$

Rewrite this as

$$\mu_j = \frac{\eta}{\eta - 1} \frac{(c_j/P)^{1-\eta}\alpha - \gamma}{(c_j/P)^{1-\eta}\alpha - (c_j/P)\delta}, \tag{A.43}$$

where

$$\alpha = \beta_1^{\frac{\eta-1}{\eta}} \widetilde{\tau} \qquad > 0, \tag{A.44}$$

$$\gamma = \beta_0^{\frac{\eta-1}{\eta}} \qquad > 0, \tag{A.45}$$

$$\delta = -\beta_0 \beta_1^{-\frac{1}{\eta}} \qquad > 0. \tag{A.46}$$

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

$$\text{sign}\left(\frac{\partial \mu_j}{\partial (c_j/P)}\right) = -\text{sign} \underbrace{\left(\eta\alpha\delta\left(\frac{c_j}{P}\right)^{1-\eta} + \alpha\gamma(1-\eta)\left(\frac{c_j}{P}\right)^{-\eta} - \gamma\delta\right)}_{\equiv Z(c_j)}. \tag{A.47}$$

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.

$$\underset{c_j}{\text{argmin}} \quad Z(c_j) = \frac{\gamma}{\delta}. \tag{A.48}$$

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

$$\alpha\delta^{\eta-1} > \gamma^{\eta}. \tag{A.49}$$

Which simplifies to

$$\widetilde{\tau} \geq 1. \tag{A.50}$$

Write $\widetilde{\tau}$ as a function of τ. For any (π, η), $\widetilde{\tau}(1) = 1$. 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 **11** (Normalization of β_1)**.** *Holding fixed the set of structural parameters other than β_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 **11.** Let $\widetilde{\beta_1} \equiv \beta_1 P^\eta$. Using Lemma 2, we can re-write the optimal quantities sold on the market equilibrium as

$$q_{1j} = -\beta_0 + \widetilde{\beta_1} c_j^{-\eta} \left(\frac{1-\pi}{1-\tau\pi}\right)^\eta \tag{A.51}$$

$$q_{\tau j} = -\beta_0 + \widetilde{\beta_1} c_j^{-\eta} \left(\frac{1}{\tau}\right)^\eta \tag{A.52}$$

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 β_1. We now turn to show that also equilibrium markups do not depend on β_1. From Lemma 2, we obtain

$$\psi(q) = \frac{u(q)}{qu'(q)} = \frac{\eta}{\eta-1}\left[\left(1+\frac{\beta_0}{q}\right) - \beta_0^{\frac{\eta-1}{\eta}} \frac{(q+\beta_0)^{\frac{1}{\eta}}}{q}\right]. \tag{A.53}$$

Note that $\psi(\cdot)$ does not depend on β_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 β_1. Let $\widetilde{\beta_1}^{FB} \equiv \beta_1 \left(P^{FB}\right)^\eta$. Using Lemma 2, we can re-write the first-best quantities ((2.11)) as

$$q_{1j}^{FB} = -\beta_0 + \widetilde{\beta_1}^{FB} c_j^{-\eta} \tag{A.54}$$

$$q_{\tau j}^{FB} = -\beta_0 + \widetilde{\beta_1}^{FB} c_j^{-\eta} \left(\frac{1}{\tau}\right)^\eta \tag{A.55}$$

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 **12** (Normalization of β_0)**.** *Consider a set of structural parameters $\{\beta_0, \beta_1, \eta, \tau, \pi, \theta\}$. If we multiply β_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 α.*

Proof of Proposition **12.** Using Lemma 2, we have that the quantities sold in the market equilibrium are given by

$$q_{1j} = -\beta_0 + \beta_1 P^\eta c_j^{-\eta} \left(\frac{1-\pi}{1-\tau\pi}\right)^\eta, \tag{A.56}$$

$$q_{\tau j} = -\beta_0 + \beta_1 P^\eta c_j^{-\eta} \left(\frac{1}{\tau}\right)^\eta. \tag{A.57}$$

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

$$q_{1j}' = \alpha q_{1j}, \tag{A.58}$$

$$q_{\tau j}' = \alpha q_{\tau j}. \tag{A.59}$$

Since all costs are divided by α, the labor ($l_j' = q_j' c_j' = \alpha q_j \, c_j/\alpha = l_j$) needed to produce the allocations for the scaled β_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.53) we have

$$\psi(q) = \frac{\eta}{\eta-1}\left[\left(1+\frac{\beta_0}{q}\right) - \left(\frac{\beta_0}{q}\right)^{\frac{\eta-1}{\eta}}\left(1+\frac{\beta_0}{q}\right)^{\frac{1}{\eta}}\right] \tag{A.60}$$

Since both quantities and β_0 are scaled by α, 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 α. From Lemma 2 and equation (2.11) we have

$$q_{1j}^{FB} = -\beta_0 + \beta_1 \left(P^{FB}\right)^\eta c_j^{-\eta} \tag{A.61}$$

$$q_{\tau j}^{FB} = -\beta_0 + \beta_1 \left(P^{FB}\right)^\eta c_j^{-\eta} \left(\frac{1}{\tau}\right)^\eta \tag{A.62}$$

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 α. 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 ν_{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

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

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 α_j is the production share sold to high-taste consumers:

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

Proof of Proposition **5.**

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:

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

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

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

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 2 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 α_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 2, $\epsilon(q) = \eta \left(\frac{\beta_0}{q} - 1 \right)$. 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 β_1)**.** *Holding fixed the set of structural parameters other than β_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 β_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 β_1 by the same argument.

■

Proposition **14** (Normalization of β_0)**.** *Consider a set of structural parameters $\{\beta_0, \beta_1, \eta, \tau, \pi, \theta\}$. If we multiply β_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 α.*

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 α. Since all costs are divided by α, the labor ($l_j' = q_j' c_j' = \alpha q_j \, c_j/\alpha = l_j$) needed to produce the allocations for the scaled β_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 α. 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 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. In the last part, we restrict firms to offering 2 bundles only. We show that with CED preferences, the allocation of consumers to the two bundles is independent of firm productivity and quantify the model.

A.1 Theory: Model setup

Household preferences are as before, with the only difference that taste shifter τ_{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, $h(\tau) \equiv \frac{g(\tau)}{1-G(\tau)}$.[32]

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

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

The set of constraints in Problem (A.1) states that each consumer type τ 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).[33] 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 τ is implicitly given by

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

[33] 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.

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

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 τ, 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 τ relative to consumers with higher tastes, and the lower is the shadow cost of selling an additional unit to consumers with taste τ.

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^{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 τ. 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 τ 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 $q_{\hat{\tau} j} < q_{\hat{\tau} j}^{FB} \quad \forall \, \tau < \hat{\tau}$.

■

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 4, 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:

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

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 $\mu_j(\tau)$, θ_j, and λ 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 θ_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 λ 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 4 in the main text 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.

A.4 Two bundles: Theory

Household preferences and production technology are as before. However, firms may only offer two bundles (q_1, p_1) and (q_τ, p_τ). Let $\hat{\tau}_j$ be the threshold below which consumers buy q_1 and above which q_τ.

A.4.1 Market allocation

The firm's problem is given by

$$\begin{aligned} \max_{\{q_{1j}, q_{\tau j}, p_{1j}, p_{\tau j}, \hat{\tau}_j\}} \quad & (p_{1j} - c_j) q_{1j} G(\hat{\tau}) + (p_{\tau j} - c_j) q_{\tau j} (1 - G(\hat{\tau})) \\ \text{s.t.} \quad & u(q_{1j}) = \frac{p_{1j} q_{1j}}{P}, \\ & \hat{\tau}_j u(q_{\tau j}) - \frac{p_{\tau j} q_{\tau j}}{P} = \hat{\tau}_j u(q_{1j}) - \frac{p_{1j} q_{1j}}{P}. \end{aligned}$$

which uses the usual result that the IR constraint only binds for the lowest type and the IC only for the threshold consumer $\hat{\tau}$.

The optimal quantities q_{1j} and $q_{\tau j}$ solve

$$\hat{\tau}_j u'(q_{\tau j}) = \frac{c_j}{P} \tag{A.13}$$

$$u'(q_{1j}) = \frac{G(\hat{\tau}_j)}{1 - \hat{\tau}_j (1 - G(\hat{\tau}_j))} \frac{c_j}{P} \tag{A.14}$$

Similar to the baseline model with two types, conditional on the aggregate price index P, the threshold type $\hat{\tau}_j$ is sold the optimal quantity and there is a wedge $\frac{G(\hat{\tau}_j)}{1 - \hat{\tau}_j (1 - G(\hat{\tau}_j))} > 1$ that distorts the allocation the lowest type downwards.

The threshold type who is indifferent between the two bundles solves:

$$\frac{c_j}{P} = \left(\hat{\tau}_j - \frac{1 - G(\hat{\tau}_j)}{g(\hat{\tau}_j)} \right) \frac{u(q_{\tau j}) - u(q_{1j})}{q_{\tau j} - q_{1j}} \tag{A.15}$$

The threshold type is a function of the hazard ratio associated with the distribution of consumer tastes $G(.)$. In general, it depends on the productivity of the firm. However, we show that, as long as preferences feature CED, the threshold type in independent of firm productivity. While all firms might choose to offer bundles that induce an inefficient allocation of consumers to quantities purchased, this distortion is constant across firms.

PROPOSITION **18.** *Suppose preferences satisfy Assumption 1. Then the allocation of consumer tastes to the low and high bundles in the market equilibrium is identical for all firms. That is,* $\hat{\tau}_j = \hat{\tau} \; \forall j$.

A.4.2 Efficient allocation

Suppose the social planner can also only choose two quantities for each firm, q_{1j}^{FB} and $q_{\tau j}^{FB}$. The two quantities imply a cut-off type $\hat{\tau}_j^{FB}$. They are the solution to

$$\max_{\{q_{1j},q_{\tau j},\hat{\tau}_j\}} \int_j \left[\int_1^{\hat{\tau}_j} \tau u(q_{1j})dG(\tau) + \int_{\hat{\tau}_j}^{\infty} \tau u(q_{\tau j})dG(\tau)\right] dj$$
$$\text{s.t.} \quad \int_j c_j \left(q_{1j}G(\hat{\tau}_j) + q_{\tau j}(1 - G(\hat{\tau}_j))\right) = 1$$

As before,let P^{FB} be the inverse Lagrange multiplier on the aggregate resource constraint. The optimal allocations and the cut-off types solve

$$\mathbb{E}\left[\tau|\tau \geq \hat{\tau}_j^{FB}\right] u'(q_{1j}^{FB}) = \frac{c_j}{P^{FB}} \tag{A.16}$$
$$\mathbb{E}\left[\tau|\tau \leq \hat{\tau}_j^{FB}\right] u'(q_{\tau j}^{FB}) = \frac{c_j}{P^{FB}} \tag{A.17}$$
$$\hat{\tau}_j^{FB} \frac{\left[u(q_{\tau j}^{FB}) - u(q_{1j}^{FB})\right]}{q_{\tau j}^{FB} - q_{1j}^{FB}} = \frac{c_j}{P^{FB}} \tag{A.18}$$

Conditional on a cut-off type $\hat{\tau}_j^{FB}$, the planner chooses the quantities that equate *expected* marginal utility to marginal cost for the set of households who purchase that bundle. The optimality condition for the cut-off type $\hat{\tau}_j^{FB}$ is similar to the market allocation with the exception that only the taste shifter enters.

From (A.18) it follows that, under CED, the cut-off type is the same for all firms also in the market allocation.

Proposition **19.** *Suppose preferences satisfy Assumption 1. Then the allocation of consumer tastes to the low and high bundles in the first-best is identical for all firms. That is,* $\hat{\tau}_j^{FB} = \hat{\tau}^{FB}\ \forall j$.

In this environment, there are two dimensions of misallocation across firms. Relative to the social planner, the market allocation induces a different cut-off type. The set of consumers that purchase the high vs low-taste bundle are different. In addition, the two quantities offered are different. Consider for example the large bundle. The social planner chooses it to maximizes the average utility—net of costs—of households purchasing that bundles. The firm chooses it to maximize utility of the cut-off type.

Importantly, however, our main result of no misallocation across firms remains. Both types of misallocation across consumers do not depend on firm productivity. As long as preferences are CED, total production of each firm in the market allocation is identical to first-best.

Proposition **20.** *Suppose preferences satisfy Assumption 1. Then, the equilibrium levels of firm-level production and employment are identical to the efficient allocation.*

A.5 Two bundles: Proofs

Proof of Proposition 18.

Using Lemma 2, we can write the differences in utility and quantities as

$$q_{\tau j} - q_{1j} = \beta_1 \left(c_j/P\right)^{-\eta} \left[\hat{\tau}_j^{\eta} - \tilde{\tau}_j^{\eta}\right] \tag{A.19}$$

$$u(q_{\tau j}) - u(q_{1j}) = \frac{\eta}{\eta - 1} \beta_1^{\frac{1}{\eta}} \left(c_j/P\right)^{1-\eta} \left[\hat{\tau}_j^{\eta-1} - \tilde{\tau}_j^{\eta-1}\right] \tag{A.20}$$

$$\tag{A.21}$$

where $\tilde{\tau}_j \equiv \frac{1-\hat{\tau}_j(1-G(\hat{\tau}_j))}{G(\hat{\tau}_j)}$. Taking ratios

$$\frac{u(q_{\tau j}) - u(q_j)}{q_{\tau j} - q_{1j}} = \frac{c_j}{P} \frac{\frac{\eta}{\eta-1}\beta_1^{\frac{1}{\eta}} \left[\hat{\tau}_j^{\eta-1} - \tilde{\tau}_j^{\eta-1}\right]}{\beta_1 \left[\hat{\tau}_j^{\eta} - \tilde{\tau}_j^{\eta}\right]} \tag{A.22}$$

Plugging into (A.15), the optimality condition for $\hat{\tau}_j$

$$\frac{1 - G(\hat{\tau}_j)}{g(\hat{\tau}_j)} + \hat{\tau}_j = \frac{\beta_1 \left[\hat{\tau}_j^{\eta} - \tilde{\tau}_j^{\eta}\right]}{\frac{\eta}{\eta-1}\beta_1^{\frac{1}{\eta}} \left[\hat{\tau}_j^{\eta-1} - \tilde{\tau}_j^{\eta-1}\right]} \tag{A.23}$$

which is independent of c_j and hence $\hat{\tau}_j = \hat{\tau}\ \forall j$

■

Proof of Proposition 19. Using Lemma (2) and the optimality conditions (A.16) and (A.17), rewrite (A.18) as

$$\hat{\tau}_j^{FB} \frac{c_j}{P^{FB}} = \frac{c_j}{P^{FB}} \frac{\frac{\eta}{\eta-1}\beta_1^{\frac{1}{\eta}} \left[\mathbb{E}\left[\tau|\tau \geq \hat{\tau}_j^{FB}\right]^{\eta-1} - \mathbb{E}\left[\tau|\tau \leq \hat{\tau}_j^{FB}\right]^{\eta-1}\right]}{\beta_1 \left[\mathbb{E}\left[\tau|\tau \geq \hat{\tau}_j^{FB}\right]^{\eta} - \mathbb{E}\left[\tau|\tau \leq \hat{\tau}_j^{FB}\right]^{\eta}\right]} \tag{A.24}$$

As before, the $\frac{c_j}{P^{FB}}$ cancel and the resulting cut-off type $\hat{\tau}_j^{FB}$ is constant across firms.

■

Proof of Proposition 20.

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:

$$G(\hat{\tau})\, q_{1j}(\widetilde{P}_j) + (1 - G(\hat{\tau}))\, q_{\tau j}(\widetilde{P}_j) = G(\hat{\tau}^{FB})\, q_{1j}^{FB} + (1 - G(\hat{\tau}^{FB}))\, q_{\tau j}^{FB} \tag{A.25}$$

$$G(\hat{\tau})\, \left[q_{1j}(\widetilde{P}_j) - q_{1j}^{FB}\right] + (1 - G(\hat{\tau}))\, \left[q_{\tau j}(\widetilde{P}_j) - q_{\tau j}^{FB}\right] = \left[G(\hat{\tau}) - G(\hat{\tau}^{FB})\right]\, \left[q_{\tau j}^{FB} - q_{1j}^{FB}\right] \tag{A.26}$$

Note that here we already used the fact that $\hat{\tau}$ and $\hat{\tau}^{FB}$ are both independent of c_j.

For the remainder of the proof, we follow the exact same steps as in the proof of Proposition 2. The only difference in the expressions for excess labor is the last term, $\left[G(\hat{\tau}) - G(\hat{\tau}^{FB})\right]\, \left[q_{\tau j}^{FB} - q_{1j}^{FB}\right]$.

Under CED, this term is proportional to the firm's cost, with the same factor of proportionality as the difference between market and first-best, $\left[q_{\tau j}(\widetilde{P}_j) - q_{\tau j}^{FB}\right]$. Hence, $\widetilde{P_j}$ equates total labor demand of any firm to the first-best.

■

A.6 Two bundles: Quantification

In order to maximize comparability to the baseline results, we choose a distribution of tastes $G(.)$ such that, in the market equilibrium, (i) a fraction π purchase the large quantity, (ii) the lower bound of the taste distribution is normalized to 1, and (iii) the cut-off type to which the high bundle is tailored is equal to the taste shifter τ.

We use a generalized Pareto distribution with location parameter equal to 1, shape ξ, and scale σ.

$$1 - G(x) = \left(1 + \frac{\xi}{\sigma}(x-1)\right)^{-\frac{1}{\xi}} \tag{A.27}$$

We choose ξ and σ such that the decentralized equilibrium is equivalent to the benchmark estimation:

$$1 - G(\tau) = \pi, \tag{A.28}$$

$$\frac{1 - G(\tau)}{g(\tau)} = \tau - \frac{\eta - 1}{\eta} \frac{\left[\tau^{\eta} - \left(\frac{1-\tau(1-G(\tau))}{G(\tau)}\right)^{\eta}\right]}{\left[\tau^{\eta-1} - \left(\frac{1-\tau(1-G(\tau))}{G(\tau)}\right)^{\eta-1}\right]}, \tag{A.29}$$

where (A.29) uses the equilibrium condition for $\hat{\tau}$, (A.15), evaluated at τ.

B Kimball preferences

In this section, we describe the model and its equilibrium with Kimball preferences. The Kimball preferences do not fall into our additive separable utility specification of Section 2. We therefore start from the household's utility before writing down the firm's problem.

Households. Households have idiosyncratic tastes over a measure 1 of varieties of consumption goods $j \in [0,1]$. The taste shifter of consumer i towards variety j is denoted by τ_{ij}. The aggregate consumption is implicitly defined as follows

$$\int \tau_{ij} \Upsilon\left(\frac{q_{ij}}{Q_i}\right) dj = 1, \tag{B.1}$$

where q_{ij} denotes the quantity consumed of variety j by household i. And Q_i denotes aggregate consumption of household i.

As in the benchmark model, the taste shifters τ_{ij} can take one of two values: 1 or $\tau > 1$. Each consumer has a high preference τ for a random subset of goods of measure π. 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.

Household utility is given by

$$U_i = \frac{1}{1-\kappa} Q_i^{1-\kappa} - \frac{\varphi}{1+\nu} L^{1+\nu}. \tag{B.2}$$

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.

Lemma **3.** *The additional utility from consuming x of firm j instead of consuming $0 \leq x' \leq x$ from it is given by:*

$$\frac{Q^{2-\kappa}}{\int \Upsilon'\left(\frac{q_z}{Q}\right) q_z dz} \left[\Upsilon\left(\frac{x}{Q}\right) - \Upsilon\left(\frac{x'}{Q}\right)\right]$$

Proof. Aggregate consumption is defined by

$$\int \tau_{ij} \Upsilon\left(\frac{q_{ij}}{Q_i}\right) dj = 1, \tag{B.3}$$

Let's find what's $\frac{\partial Q}{\partial q_{ij}}$. If we totally differentiate we get

$$\int \tau_{iz} \Upsilon'\left(\frac{q_{iz}}{Q}\right) q_{iz} dz \frac{1}{Q^2} dQ = \frac{1}{Q} \Upsilon'\left(\frac{q_{ij}}{Q}\right) \tau_{ij} dq_i, \tag{B.4}$$

which yields

$$\frac{\partial Q}{\partial q_{ij}} = Q \frac{\tau_{ij} \Upsilon'\left(\frac{q_{ij}}{Q}\right)}{\int \tau_{iz} \Upsilon'\left(\frac{q_{iz}}{Q}\right) q_{iz} dz} \tag{B.5}$$

We can now calculate how Q changes when we consume $\bar{q}$ of q_{ij} instead of $\underline{q}$. That is, the gain from consuming q_{ij}. This gain is given by

$$\int_{\underline{q}}^{\bar{q}} \frac{\partial Q}{\partial q_{ij}} dq_{ij}. \tag{B.6}$$

We can use the expression for $\frac{\partial Q}{\partial q_{ij}}$ to obtain:

$$\begin{aligned}
\int_{\underline{q}}^{\bar{q}} \frac{\partial Q}{\partial q_{ij}} dq_{ij} &= \int_{\underline{q}}^{\bar{q}} Q \frac{\tau_{ij} \Upsilon' \left(\frac{q_{ij}}{Q}\right)}{\int \tau_{iz} \Upsilon' \left(\frac{q_{iz}}{Q}\right) q_{iz} dz} dq_{ij} \\
&= Q \frac{1}{\int \tau_{iz} \Upsilon' \left(\frac{q_{iz}}{Q}\right) q_{iz} dz} \int_{\underline{q}}^{\bar{q}} \tau_{ij} \Upsilon' \left(\frac{q_{ij}}{Q}\right) dq_{ij} \\
&= Q \frac{1}{\int \tau_{iz} \Upsilon' \left(\frac{q_{iz}}{Q}\right) q_{iz} dz} \tau_{ij} \left. \Upsilon \left(\frac{q_{ij}}{Q}\right) \right|_{\underline{q}}^{\bar{q}} \\
&= Q^2 \frac{1}{\int \tau_{iz} \Upsilon' \left(\frac{q_{iz}}{Q}\right) q_{iz} dz} \tau_{ij} \left[\Upsilon \left(\frac{\bar{q}}{Q}\right) - \Upsilon \left(\frac{\underline{q}}{Q}\right) \right] \\
&= \tau_{ij} Q D \left[\Upsilon \left(\frac{\bar{q}}{Q}\right) - \Upsilon \left(\frac{\underline{q}}{Q}\right) \right],
\end{aligned}$$

where

$$D = \frac{Q}{\int \tau_{iz} \Upsilon' \left(\frac{q_{iz}}{Q}\right) q_{iz} dz}. \tag{B.7}$$

■

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

$$\begin{aligned}
\max_{\{q_{1j}, q_{\tau j}, p_{1j}, p_{\tau j}\}} \quad & \pi q_{\tau j} \left(p_{\tau j} - c_j\right) + (1-\pi) q_{1j} \left(p_{1j} - c_j\right) && \text{(B.8)} \\
\text{s.t} \quad & Q D \left[\Upsilon \left(\frac{q_{1j}}{Q}\right) - \Upsilon \left(0\right) \right] - \frac{p_{1j} q_{1j}}{P} = 0, && [IR_1] \\
& \tau Q D \left[\Upsilon \left(\frac{q_{\tau j}}{Q}\right) - \Upsilon \left(\frac{q_{1j}}{Q}\right) \right] = \frac{p_{\tau j} q_{\tau j}}{P} - \frac{p_{1j} q_{1j}}{P}. && [IC_\tau]
\end{aligned}$$

The second constraint uses the fact that the individual rationality constraint of the low-taste consumer holds with equality.

Firm-level optimal prices and quantities. Taking first order conditions of the firm's problem we obtain:

$$[q_{\tau j}]: \quad \pi\left(p_{\tau j}-c_j\right)+\lambda_\tau\left(\tau D\Upsilon'\left(\frac{q_{\tau j}}{Q}\right)-\frac{p_{\tau j}}{P}\right)=0 \tag{B.9}$$

$$[q_{1j}]: \quad (1-\pi)\left(p_{1j}-c_j\right)-\lambda_\tau\left(\tau D\Upsilon'\left(\frac{q_{1j}}{Q}\right)-\frac{p_{1j}}{P}\right)+\lambda_1\left(D\Upsilon'\left(\frac{q_{1j}}{Q}\right)-\frac{p_{1j}}{P}\right)=0 \tag{B.10}$$

$$[p_{\tau j}]: \quad \pi q_{\tau j}-\lambda_\tau\left(\frac{q_{\tau j}}{P}\right)=0 \tag{B.11}$$

$$[p_{1j}]: \quad (1-\pi)q_{1j}+\lambda_\tau\left(\frac{q_{1j}}{P}\right)-\lambda_1\left(\frac{q_{1j}}{P}\right)=0 \tag{B.12}$$

Using the two FOCs wrt prices, we can solve for the Lagrange multipliers:

$$\lambda_\tau=\pi P \tag{B.13}$$

$$\lambda_1=P \tag{B.14}$$

$$\tag{B.15}$$

Plugging into the first two FOC we obtain:

$$\tau\, D\Upsilon'\left(\frac{q_{\tau j}}{Q}\right)=\frac{c_j}{P}, \tag{B.16}$$

$$D\Upsilon'\left(\frac{q_{1j}}{Q}\right)=\frac{1-\pi}{1-\tau\pi}\frac{c_j}{P} \tag{B.17}$$

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

$$\mu_{1j}=\frac{1-\pi}{1-\tau\pi}\,\psi(q_{1j}), \tag{B.18}$$

$$\mu_{\tau j}=\left(1-(\tau-1)\frac{\Upsilon\left(\frac{q_{1j}}{Q}\right)-\Upsilon\left(0\right)}{\tau\left(\Upsilon\left(\frac{q_{\tau j}}{Q}\right)-\Upsilon\left(0\right)\right)}\right)\,\psi(q_{\tau j}), \tag{B.19}$$

where $\psi(q)$ is defined by

$$\psi(q)\equiv Q\frac{\Upsilon\left(\frac{q}{Q}\right)-\Upsilon\left(0\right)}{q\Upsilon'\left(\frac{q}{Q}\right)}. \tag{B.20}$$

B.1 Second-best allocation

We use the homotheticity of the Kimball preferences, and set the planner's problem as minimizing labor subject to providing a unit of aggregate consumption.

Table B.1: Calibrated Moments and Parameters under Kimball

A. Moments

Moment	Data	Model	
		Benchm.	Kimball
Fraction buying large q	51%	51%	51%
$\mathbb{E}\left[\ln q_{j\tau} - \ln q_{j1}\right]$	0.65	0.65	0.65
Sales share top 5%	73%	77%	73%
Sales share top 10%	86%	84%	86%
Sales share top 25%	97%	92%	97%
Sales share top 50%	99.6%	96.7%	99.5%

B. Parameters

Parameter		Model	
		Benchm.	Kimball
π	Share of high-taste consumers	0.51	0.51
τ	High-taste demand shifter	1.17	1.11
η	Elasticity of differences	1.86	–
ϵ	Degree of superelasticity	–	0.73
σ	Degree of demand elasticity	–	4.37
θ	Pareto shape	0.86	1.97
Externally Set			
φ	Inverse Frisch elasticity	1	1

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. The third column presents the model moments for a specification with Kimball preferences. Panel B presents the set of calibrated parameters for the two model specifications.

$$\begin{aligned} \min_{\{l_j, q_{1j}, q_{\tau j}\}_{j=0}^{1}} \quad & \int l_j dj, && \text{(B.21)} \\ \text{s.t.} \quad & \Upsilon'\left(q_{1j}\right) = \frac{1-\pi}{1-\tau\pi}\,\tau\,\Upsilon'\left(q_{\tau j}\right), && \text{for all } j \\ & \pi q_{\tau j} + (1-\pi) q_{1j} = \frac{l_j}{c_j}, && \text{for all } j \\ & \int \left[(\pi\tau\Upsilon\left(q_{\tau j}\right) + (1-\pi)\Upsilon\left(q_{1j}\right)\right] dj = 1. \end{aligned}$$

where the first constraint is the implementability constraint from the firm's problem, the second constraint is the production costs, and the third constraint is the requirement to provide one unit of aggregate consumption.

B.2 Quantitative analysis

We use the Klenow-Willis specification for the Kimball aggregator:

$$\Upsilon(x) = 1 + (\sigma - 1)\exp\left(\frac{1}{\varepsilon}\right)\varepsilon^{\sigma/\varepsilon - 1}\left(\Gamma\left(\frac{\sigma}{\varepsilon}, \frac{1}{\varepsilon}\right) - \Gamma\left(\frac{\sigma}{\varepsilon}, \frac{x^{\varepsilon/\sigma}}{\varepsilon}\right)\right), \tag{B.22}$$

where ε governs the degree of superelasticity, and σ governs the level of demand elasticity.

The model moments and calibrated parameters are displayed in Table B.1. We see that the model with Kimball preferences is able to match the data moments accurately. Notice that the calibration with Kimball preferences has one additional parameter relative to our benchmark calibration.

The misallocation costs are summarized in Table B.2. Under the Kimball calibration, the welfare gains from fixing misallocation are smaller: 0.53% relative to 0.82%.

Proposition 3 shows that under CED, the planner has no incentive to use firm-level taxes and subsidies. Since Kimball preferences do not satisfy CED, the planner can potentially reduce misallocation by imposing firm-level taxes and subsidies. We find that quantitatively, the planner can improve welfare by a very modest amount. The optimal allocation with firm-level taxes and subsidies only improves the decentralized equilibrium by 0.01% in consumption equivalent units. That is, it achieves

Table B.2: Welfare Gains of Fixing Misallocation

Baseline Model	Kimball Specification
0.82%	0.53%

Notes: This table reports the welfare gains in the equilibrium with perfect allocative efficiency relative to the baseline model (column 1), and the model with Kimball preferences (column 2). 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.

very little of the potential welfare gains of allocative efficiency (0.53%).