Nonlinear Pricing and Misallocation*

Gideon Bornstein†
Alessandra Peter‡

June 10, 2025

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 substantially larger than those from misallocation across firms under linear pricing.

*We thank the editor (Mikhail Golosov), three anonymous referees, Scott Baker, Hugo Hopenhayn, Pete Klenow, Erik Madsen, Virgiliu Midrigan, Alessandro Pavan, Ivan Werning, and EAGLS, as well as numerous seminar and conference participants for helpful suggestions and comments. The paper benefited from thoughtful discussions by Michael Peters, Kieran Larkin, and Joel David. We also thank Tanvi Jindal and Paige Stevenson, 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 and NBER; gideonbo@wharton.upenn.edu
‡New York University, CEPR, and NBER; 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 beer typically costs less than six individual cans. Despite the apparent prevalence of such pricing schedules, which have been at the center of research in industrial organization 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, and 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.

We then quantify the model using data from the retail sector. We find that the departure from linear pricing significantly alters the welfare losses from misallocation as well as the effect of commonly studied policy instruments. Misallocation across consumers leads to five times larger welfare losses compared to misallocation across firms with linear pricing. Further, the undersupply of aggregate labor is about three times smaller with nonlinear relative to linear pricing. From a policy perspective, implementing the taxes and subsidies that would restore efficiency in an economy in which firms are restricted to linear pricing would lead to large welfare losses.

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

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

clear the labor market. As a result, the allocation of high-taste consumers is also distorted, and there is misallocation across consumers of the same firm: high-taste consumers are allocated too much of the good, whereas those with low taste consume too little.

We next analyze the allocation of production across firms. To do so, we define a condition on preferences: *constant elasticity of 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 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, and a 10% increase in package size is, on average, associated with only a 4% increase in total 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.

The assumption of linear pricing significantly alters the welfare consequences of market power along two dimensions: misallocation and the aggregate labor wedge. Misallocation across consumers under nonlinear pricing is equivalent to a 0.7% loss in permanent consumption. This loss is substantially larger than the welfare loss from misallocation across firms one would infer when calibrating a standard model with linear pricing (0.14%) to the same data moments. While both models are calibrated to the same level of aggregate markup (30%), the distortion to aggregate labor is significantly smaller under nonlinear pricing. In the linear pricing economy, labor is distorted downwards by 24%. Under nonlinear

pricing, aggregate labor should only be 6% higher, and the welfare losses from labor distortions are correspondingly smaller.

Finally, we use the calibrated model to analyze commonly studied policy tools. If a social planner were to implement size-dependent subsidies–which would restore efficiency under linear pricing–, she would instead induce welfare losses of 0.12% under nonlinear pricing. If a policy maker were to try and correct the inferred labor wedge under linear pricing as well, she would cause welfare losses of nearly 2%.

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 (Jan De Loecker, Jan Eeckhout and Gabriel 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 (Diego Restuccia and Richard Rogerson (2007); Chang-Tai Hsieh and Peter J Klenow (2009)), arising from oligopolistic competition (Andrew Atkeson and Ariel Burstein (2008)) or limit pricing (Michael Peters (2020)), or as a result of preferences featuring variable elasticity of substitution (Chris Edmond, Virgiliu Midrigan and Daniel Yi Xu (2022); Corina Boar and Virgiliu Midrigan (2024)). While most of the models do not exactly fit into the Swati Dhingra and John 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 canonical model of second-degree price discrimination that is commonly used in theoretical IO (see Michael Spence (1977); Michael Mussa and Sherwin Rosen (1978); Eric Maskin and John Riley (1984); Jean Tirole (1988); Robert B Wilson (1993) and references therein). We show that the classic result of "no distortion at the top" (as initially discovered by Mirrlees (1971)) no longer holds and that consumption of the high-type consumer is distorted upwards. The theoretical literature has proposed several other types of changes to the classic model that would imply upward and downward distortions for the highest type.[2] Relative to this literature, our main contribution is to develop a tractable general equilibrium version of the classical model with heterogeneous firms. Our approach maintains the quasi-linear structure from the perspective of each individual firm, allowing us to leverage the classical results and maintain tractability while altering the efficiency properties of the equilibrium allocation through general equilibrium channels.

A smaller literature within macroeconomics has used and extended classical IO models to analyze the welfare consequences of firms engaging in first- or second-degree price discrimination. Attanasio and Pastorino (2020) study a large cash transfer program in rural Mexico and show that quantity discounts by sellers have significantly increased following its introduction. Their theoretical model focuses on which consumers are excluded from the market under linear vs. nonlinear pricing. Unlike

[2]Examples of modifications to the classical model that yield "distortion at the top" include: (i) departure from quasi-linearity (e.g., Kazumura, Mishra and Serizawa (2020)), (ii) departure from single crossing (e.g., Schottmueller (2015)), (iii) endogenizing market structure (Gomes and Pavan (2016)); and (iv) endogenizing the information structure (e.g., Mensch and Ravid (2022)), (v) heterogeneous outside options (Jullien (2000)).

our environment, their analysis studies the behavior of a representative firm in partial equilibrium, assuming consumers have quasi-linear preferences. Patrick Kehoe, Brad Larsen and Elena Pastorino (2020) study how information technology shapes equilibrium outcomes when firms compete dynamically and can price discriminate. They study equilibrium in a duopoly market, abstracting from general equilibrium effects.

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 David Argente, Munseob Lee and Sara Moreira (2024), Ariel Burstein, Vasco M Carvalho and Basile Grassi (2020), Gideon Bornstein (2021), Hassan Afrouzi, Andres Drenik and Ryan Kim (2020), and Liran Einav, Peter J Klenow, Jonathan D Levin and Raviv 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 market equilibrium and planner's allocation. Section 3 discusses the main misallocation results of the paper and compares them to a setup with linear pricing. In Section 4, we introduce the data and quantify the model. Finally, Section 5 concludes.

2 Model

2.1 Environment

Households. The economy is populated by a measure 1 of households $i \in [0,1]$. Households have idiosyncratic tastes over a measure 1 of consumption goods $j \in [0,1]$, and have disutility from supplying labor. The level of taste consumer i has for variety j is denoted by τ_{ij}, which can take one of two values: 1 or $\tau > 1$.[3] When $\tau_{ij} = \tau$, household i derives higher utility from consuming good j.

$$U_i = \int_0^1 \tau_{ij} u(q_{ij}) dj - \nu \frac{l_i^{1+\varphi}}{1+\varphi}, \tag{2.1}$$

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 $u(0) = 0$.

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. The realization of taste shocks τ_{ij} of each household i is private information. All firms in the economy are jointly owned by households. Household income consists of labor earnings as well as any profits rebated by firms. Labor is chosen as the numéraire. The budget constraint of household i is given by

$$\int_0^1 p_j(q_{ij}) q_{ij} dj = l_i + \Pi, \tag{2.2}$$

[3] Our results are not sensitive to the assumption of two types. In Online Appendix A, we repeat the analysis for an environment with a continuum of tastes.

where Π denotes aggregate profits in the economy. Note that the price a household pays for variety j can vary with the units purchased, q_{ij}. Households choose labor and consumption to maximize their utility, taking pricing schedules and aggregate profits as given.

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 cost c_j, which is distributed according to $F(c_j)$.

Firms choose a menu of prices $p_j(q)$ to maximize profits. The firm's problem is given by

$$\begin{aligned} \Pi_j \equiv \max_{\{p_j(\cdot), q_{1j}, q_{\tau j}\}} \quad & \pi \left(p_j(q_{\tau j}) - c_j\right) q_{\tau j} + (1 - \pi) \left(p_j(q_{1j}) - c_j\right) q_{1j} \\ \text{s.t.} \quad & q_{\tau j} \in \operatorname*{argmax}_{q \geq 0} \ \tau u(q) - \frac{p_j(q) q}{P}, \\ & q_{1j} \in \operatorname*{argmax}_{q \geq 0} \ u(q) - \frac{p_j(q) q}{P}, \end{aligned} \tag{2.3}$$

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. Firms take consumer choices—represented in the two constraints—as given, which are a function of their pricing schedule $p_j(q)$ as well as the aggregate price index P. The aggregate price index P is an equilibrium outcome that measures the cost of purchasing an additional unit of utility.[4]

Since consumer tastes are private information, firms cannot offer type-dependent pricing schedules. They must instead offer a single menu $p_j(q)$ to all households.[5] That is, firms engage in second-degree price discrimination. While we assume that tastes are private information, the single pricing schedule could also be motivated by legal or practical requirements that make it impossible to charge different consumers different prices for the same quantity purchased.

Contrary to our setup, the classical assumption in the macroeconomic literature is that firms may only charge linear pricing schedules, i.e., $p(q) = p$. This benchmark arises in equilibrium if a frictionless secondary market allows goods to be repackaged and resold at zero cost. In contrast, we assume that no such market exists, as secondary markets may fail to emerge for several practical reasons. First, such resale activity may simply be too costly, due to the expenses associated with packaging, logistics, and distribution. Second, unauthorized repackaging and resale of branded consumer goods may infringe trademark rights.[6] Third, repackaging may be entirely infeasible, as would be the case if consumers choose *quality* rather than quantity.[7]

[4] Technically, the price index P is the Lagrange multiplier on the household's budget constraint (2.2).

[5] Note that equation (2.3) places no additional restrictions on the pricing function $p_j(q)$. In doing so, we implicitly rule out that consumers could purchase multiple units of a particular quantity. In our calibration, firms optimally charge markups that are decreasing in quantity. As a result, consumers never prefer purchasing multiple units.

[6] Such actions may violate trademark protections under the Lanham Act. In some cases, especially for regulated goods, they may also contravene safety, labeling, or licensing requirements enforced by federal or state authorities.

[7] For example, mobile phones are sold in multiple configurations. A high-end model cannot be repackaged into two mid-tier models.

2.2 Equilibrium

An equilibrium is a set of firm-level pricing schedules $\{p_j(\cdot)\}_{j=0}^1$, quantities $\{q_{1j}, q_{\tau j}\}_{j=0}^1$, labor $\{l_i\}_{i=0}^1$, aggregate profits Π, as well as an aggregate price index P, such that

1. Given the pricing schedules, the equilibrium quantities and labor maximize households' utility (2.1) subject to their budget constraint (2.2).

2. Given the aggregate price index P, the pricing schedules and equilibrium quantities solve the firms' problems.

3. The aggregate price index is defined as the Lagrange multiplier on the budget constraint (2.2) in the household's problem.

4. Aggregate profits in the economy are given by

$$\Pi = \int_0^1 \Pi_j dj \tag{2.4}$$

5. The labor market clears:

$$\int_0^\infty \left[\pi q_{\tau j} + (1-\pi) q_{1j}\right] \, c_j \, dF(c_j) = \int_0^1 l_i di. \tag{2.5}$$

2.3 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 and the same information structure. The planner solves

$$\max_{\{q_{ij}, l_i\}} \quad \int_i \int_j \tau_{ij} u(q_{ij}) dj di - \int_i \nu \frac{l_i^{1+\varphi}}{1+\varphi} di \tag{2.6}$$

$$\text{s.t.} \quad \int_i \int_j q_{ij} c_j dj di = \int_i l_i di$$

$$\int_j \tau_{ij} u(q_{ij}) dj - \nu \frac{l_i^{1+\varphi}}{1+\varphi} \geq \int_j \tau_{ij} u(q_{i'j}) dj - \nu \frac{l_{i'}^{1+\varphi}}{1+\varphi} \quad \forall \{i, i'\}, \tag{2.7}$$

where the first constraint is the resource constraint while the second constraint is the incentive compatibility constraint. The latter constraint implies that each household truthfully reports its type and has no incentive to consume the bundle of a different household.

Proposition **1.** *The incentive compatibility constraints* (2.7) *in the planner's problem are not binding. The efficient allocation features constant labor supply across households,* $l_i = L^{FB}$*, and consumption allocations satisfy the following optimality conditions*

$$u'(q_{ij}^{FB}) = \frac{c_j}{\tau_{ij}} \frac{1}{P^{FB}}, \qquad \forall \{i, j\} \tag{2.8}$$

where $P^{FB} = \frac{1}{\nu(L^{FB})^{\varphi}}$.

All proofs are relegated to Appendix A.

Contrary to private information problems à la Mirrlees (1971), the constrained-efficient allocation is identical to the full-information planner's allocation in this environment, i.e., the first-best allocation. The key property of the model that delivers this result is the iid assumption on the taste shifters τ_{ij}. While each consumer has a high taste for a unique set of goods, the distribution of tastes across different firms of different productivities is identical. In fact, we show that each bundle the social planner offers (each possible deviation in (2.7)) uses the same amount of resources in production. Households therefore have no incentives to misreport their type; doing so would give them less consumption of some good for which they have a high taste in exchange for more consumption of a good for which they have a low taste. One straightforward way to implement the planner's allocation is to offer each variety at a constant price equal to marginal cost; all households would then self-select into their respective first-best bundle.

Given that the incentive compatibility constraint of the planner's problem is slack, the efficient allocation then simply equates each consumer's marginal utility to the production cost. We refer to the planner's inverse shadow cost of resources as P^{FB}, since, analogously to the market allocation, it measures the cost of producing an additional unit of utility. The level of consumption in the efficient allocation is then pinned down by the disutility of labor.

Proposition (1) implies that marginal utilities of all consumers of a given variety are equalized in the efficient allocation:

$$\frac{\tau u'(q_{\tau j}^{\text{FB}})}{u'(q_{1j}^{\text{FB}})} = 1. \tag{2.9}$$

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

$$\frac{\tau u'(q_{\tau j}^{\text{FB}})}{\tau u'(q_{\tau k}^{\text{FB}})} = \frac{u'(q_{1j}^{\text{FB}})}{u'(q_{1k}^{\text{FB}})} = \frac{c_j}{c_k}. \tag{2.10}$$

2.4 Market allocation

In this section, we characterize the decentralized equilibrium. We start by simplifying the firm's problem and deriving optimal prices and quantities as a function of the aggregate price index P. We finish by proving existence and uniqueness of the decentralized equilibrium.

The two constraints in the firm's problem (2.3) are each an infinite set of inequalities: the surplus from the quantity the household chooses must be larger than that from any other quantity. To solve (2.3), 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

$$\begin{aligned} \max_{\{q_{1j},q_{\tau j},p_{1j},p_{\tau j}\}} \quad & \pi q_{\tau j}\left(p_{\tau j}-c_j\right)+(1-\pi)q_{1j}\left(p_{1j}-c_j\right) && (2.11)\\ \text{s.t} \quad & u(q_{1j})-\frac{p_{1j}q_{1j}}{P}=0 && [IR_1]\\ & \tau u(q_{\tau j})-\frac{p_{\tau j}q_{\tau j}}{P}=\tau u(q_{1j})-\frac{p_{1j}q_{1j}}{P} && [IC_\tau] \end{aligned}$$

where $p_{1j} \equiv p_j(q_{1j})$ and $p_{\tau j} \equiv p_j(q_{\tau j})$.[8]

Optimal quantities. Conditional on the aggregate price index P, which firms take as given, the quantities offered to high- and low-taste consumers respectively solve

$$\tau\, u'(q_{\tau j}) = \frac{c_j}{P}, \tag{2.12}$$

$$u'(q_{1j}) = \frac{c_j}{P} + \frac{\pi}{1-\pi}(\tau-1)u'(q_{1j}) = \frac{1-\pi}{1-\tau\pi}\,\frac{c_j}{P}. \tag{2.13}$$

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 (2.12) 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 "no distortion at the top" result only holds *conditional* on the aggregate price index P. In the next section, 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 (2.13) pins down the optimal quantity sold to low-taste consumers.[9] 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 low-taste consumers, firms need to charge high-taste ones less, 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

[8] While the firm's problem pins down p_{1j} and $p_{\tau j}$, the prices 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.

[9] We assume that preferences as well as the distribution of c_j are such that all firms optimally choose to serve both types.

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.

Optimal prices. 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 markup 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

$$\mu_{1j} = \frac{1-\pi}{1-\tau\pi}\,\psi\left(q_{1j}\right), \tag{2.14}$$

$$\mu_{\tau j} = \left(1-(\tau-1)\frac{u(q_{1j})}{\tau u(q_{\tau j})}\right)\,\psi(q_\tau j), \tag{2.15}$$

where $\psi(q)$ is defined by

$$\psi(q) \equiv \frac{u(q)}{qu'(q)}. \tag{2.16}$$

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

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

The markup charged to low-taste consumers, (2.14), 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 (2.15) 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 (2.12), we know that the quantity sold to the high-taste consumer is identical to the case of a monopolist who can perfectly price discriminate. 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.

Labor supply. Since all households face the same aggregate price index P, they 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,

$$\nu L^{\varphi} = \frac{1}{P}. \tag{2.17}$$

Existence and uniqueness of equilibrium. We finish this section by proving that the equilibrium of the economy exists and is unique.

PROPOSITION **2.** *An equilibrium exists and is unique.*

The proof is relatively straightforward. Each firm's total production is monotonic in the aggregate price index P. When the aggregate price index equals zero, firms do not produce, and as the price goes to infinity, firms' optimal quantities also go to infinity. A single crossing condition then implies that there is a single aggregate price index P for which the labor market clears. In Appendix A.5, we show how the equilibrium aggregate price index P is supported despite preferences not featuring the standard quasi-linearity.

3 Market distortions

In this section, we analyze market distortions along several dimensions. We start by studying allocative efficiency, holding constant the aggregate level of labor. We characterize (i) misallocation of consumption within firms across consumers, and (ii) misallocation of production across firms. Next, we analyze market distortions when labor supply is endogenous. We show that the aggregate level of labor is distorted downwards and compare market and planner allocations. We conclude the section by contrasting market distortions under nonlinear pricing to a version of the model in which firms must set linear prices.

3.1 Misallocation within firms

We start by analyzing the allocation of goods across different types of consumers within each firm. To do so, we compare the market allocation to the allocation of a social planner who must employ the same aggregate amount of labor as the market equilibrium. To distinguish this allocation from the first-best—in which aggregate labor can adjust as well, we denote it using an FE superscript (fixed-labor efficient allocation).[10] Comparing the efficient allocation, (2.9) and (2.10), to the market

[10] Note that this restriction on the planner's problem (3.4) only affects the inverse shadow cost of resources, now defined as P^{FE}: instead of being related to labor supply via the disutility of labor, it is defined to be the level that ensures that aggregate labor is the same as in the market allocation. As in the first best allocation, the incentive compatibility constraints faced by the planner do not bind in the efficient allocation.

allocation, (2.12) and (2.13), we obtain the following relationship:

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

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. In general equilibrium, this distortion leads to misallocation of consumption across consumers of the same firm, as formalized in the following proposition.

PROPOSITION **3.** *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.*

The result that the allocation of low-taste consumers is distorted downwards is familiar from the micro theory literature.[11] However, the standard *no distortion at the top* result no longer holds in general equilibrium. Instead, high-taste consumers are allocated too much of each good.

To build intuition for the *distortion at the top* result, consider the market allocation under the fixed-labor efficient allocation price index P^{FE}. Under such price index, the consumption of high-taste consumers would be identical to the efficient allocation and the consumption of low-taste consumers would be distorted downwards. However, such an allocation cannot clear the labor market.

In general equilibrium, the aggregate price index P must therefore be higher than in the FE 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.

3.2 Misallocation across firms

Next we turn to the question of misallocation across firms. We compare the *overall* production of firm j, and hence its employment, to the efficient allocation. Firm-level production is equal to a weighted average of quantities sold to each type of consumer, $q_j = \pi q_{1j} + (1-\pi)q_{\tau j}$. To analyze misallocation across firms, it is helpful to define the following property of preferences.

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

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

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

The elasticity of differences measures how the difference in optimal consumption between high- and low-taste consumers of a particular good varies with the cost of producing that good. If the elasticity

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

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 $q_{ij}^{\text{FE}} = (u')^{-1}\left(\frac{1}{\tau_{ij}}\frac{c_j}{P^{\text{FE}}}\right)$, the elasticity of differences is ultimately a function of the inverse marginal utility:[12]

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

In general, the combination of oversupply to high types and undersupply to low types could lead to arbitrary patterns of firm-level output relative to the efficient allocation. In Proposition 4, we show one of the key results of the paper: for a large class of preferences, defined formally in Assumption 1, the over-supply to the high type exactly offsets the under-supply to the low type for all firms. 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,

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

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 Marc J Melitz and Gianmarco IP 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.[13]

Proposition **4.** *Suppose preferences satisfy Assumption 1. Then, the equilibrium levels of firm-level production and employment are identical to the efficient allocation.*

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 efficient allocation for a firm with production cost c_j. It is implicitly defined as

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

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 $\partial \log(q_{\tau j} - q_{\tau j}^{FE})/\partial \log(\text{c}_j) = \eta$. 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 $\tau = \tilde{P}_j/P^{FE}$, the result follows.

[12] Alternatively, one can define the elasticity of differences using the first best allocations, $q_{\tau j}^{\text{FB}} - q_{1j}^{\text{FB}}$, rather than the ones in the fixed-labor efficient allocation, $q_{\tau j}^{\text{FE}} - q_{1j}^{\text{FE}}$. Both definitions boil down to equation (3.2).

[13] In Appendix A, Lemma 2, we show that a corollary of Assumption 1 is that the inverse marginal utility takes the following form: $(u')^{-1}(x) = -\beta_0 + \beta_1 x^{\eta}$.

Relative to the planner allocation, the market behaves *as if* preferences of the high type were shifted by $\tilde{P}_j/P^{AE}$. Since there is constant elasticity of differences in tastes, the elasticity of the difference between the efficient and market allocations is also constant. Similarly, $\partial \log(q_{1j}^{AE} - q_{1j})/\partial \log(\mathrm{c}_j) = \eta$.

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

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

When the difference in quantities sold to the two types of consumers scales proportionately with costs, under- and oversupply relative to the efficient allocation are also proportional to cost. Therefore, in order for the labor market to clear, the employment of each firm must be identical to the efficient allocation.

In summary, all firms, irrespective of their productivity, employ the efficient amount of labor and produce the efficient amount of output. However, they allocate that output to consumers in a distorted fashion—too much is sold to high-taste consumers, too little to low–taste consumers.

3.3 Aggregate labor distortion

We now turn to analyze market distortions relative to the first best allocation. We start by comparing the aggregate level of labor in the two equilibria and then study the distribution of employment across firms.

In the first best allocation, the planner chooses the amount of labor that equates marginal disutility of labor supply with the marginal product of labor (see Proposition (1)). In the decentralized equilibrium, labor supply is chosen to equate marginal disutility to the real wage.

There are two forces that make the aggregate level of labor in the decentralized equilibrium differ from the first best one. First, the real wage may not equal the marginal product of labor, and second, the marginal product of labor may differ between the two equilibria. The latter effect is straightforward: since there is misallocation of consumption in the decentralized equilibrium, any unit of labor produces less utility—the marginal product of labor is lower. The former effect is more subtle and is unique to our environment with nonlinear pricing.

In the decentralized equilibrium, the real wage is lower than the marginal product of labor. This wedge is also stemming from the presence of misallocation across consumers. From equation (2.12), we know that the inverse of the aggregate price index, i.e., the real wage, is equal to the marginal utility of high-taste consumption divided by the marginal cost. That is, the real wage is equal to the marginal product of labor if that were employed for producing high-taste consumption only. However, when the aggregate level of labor goes up, the real wage falls, and production of both low- and high-taste consumption rises, not only that of high-taste. When households choose their labor supply, they do not take into account this general equilibrium effect. Because the marginal utility of consumption is strictly higher for low-taste consumers, the real wage is lower than the marginal product of labor in the decentralized equilibrium.

Both of these forces reduce the aggregate level of labor in the decentralized equilibrium, as formalized by the following proposition.

PROPOSITION **5.** *In the market equilibrium, the aggregate level of labor L is lower than in the first best allocation.*

Let's now turn to study the distribution of workers across firms. In the previous section, we showed that when the aggregate level of labor is fixed, the decentralized equilibrium market share of all firms is identical to their market share in the efficient allocation (see Proposition 7). This is no longer the case with variable aggregate labor. The fact that labor is under-supplied in the aggregate leads to a change in the distribution of employment shares across firms.

PROPOSITION **6.** *Let a firm's* excess employment share *be the ratio between its equilibrium employment share and the employment share in the first best allocation. Suppose preferences satisfy Assumption 1. Then, the following hold:*

1. *If preferences feature decreasing demand elasticities, excess employment shares are increasing in firm productivity.*
2. *If preferences feature increasing demand elasticities, excess employment shares are decreasing in firm productivity.*
3. *If preferences feature constant demand elasticities, excess employment shares are constant across all firms.*

To understand the intuition behind Proposition 6, consider the aggregate price index that equates labor demand of all firms to the first-best level of labor. Denote it by $\widetilde{P}$. From Proposition 4 we know such unique aggregate price index exists. Since aggregate labor is lower in the decentralized equilibrium than in the first best allocation (Proposition 5), 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.

The effect of the change in the aggregate price index on the distribution of employment across firms depends on the demand elasticity. If, for example, demand elasticities are decreasing in quantity, then high-productivity firms face lower demand elasticities. This implies that the change in the aggregate price index has a smaller effect on their production, and hence on their employment. As a result, under decreasing demand elasticities, the employment share of high-productivity firms is higher than in the first best allocation.

In summary, the misallocation across consumers leads to an inefficiently low level of labor in the decentralized equilibrium. Depending on the demand elasticity, relative firm sizes are also distorted.

3.4 Taxes and subsidies

So far, we have compared the decentralized allocation to the efficient as well as the first-best ones. We now turn to study the problem of a social planner who has access to a limited set of taxes and subsidies. In particular, we restrict the planner to have access to only *firm-level* taxes and subsidies.

Since the planner can achieve the first best as well as the efficient allocation without observing types, she can implement these allocations if she has access to unrestricted tools. This can be achieved, for example, by mandating that firms price linearly at marginal cost. In reality, such mandate is likely unfeasible. 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 subject to a set of implementability constraints:

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

The first two constraints ensure that, given a set of taxes t_j and an aggregate price index P, firms optimally supply q_{1j} units to low taste consumers and $q_{\tau j}$ units to high taste ones. The third equation is the aggregate resource constraint while the last one ensures that households are on their labor supply curve.

Optimal taxes and subsidies with fixed labor supply. In a similar vein to sections 3.1–3.2, we start by restricting the planner to keep the aggregate level of labor constant. That is, the planner solves problem (3.4), subject to an additional constraint that L is equal to its level in the decentralized equilibrium. The proposition below shows that in this case, the planner chooses not to impose any taxes or subsidies—i.e., $t_j = 0\,\forall j$.

PROPOSITION **7.** *Suppose preferences satisfy Assumption 1. Then, if the planner cannot change the aggregate level of labor, imposing no subsidies and taxes at the firm level is optimal.*

When the aggregate level of labor is fixed, the only 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.

Proposition 7 shows, however, that under Assumption 1, misallocation across consumers cannot be mitigated using firm-level taxes. Indeed, the optimal taxes set by the social planner are identically

zero. To understand why, we use the following property of CED preferences.[14]

LEMMA **1** (Implications of constant elasticity of differences.)**.** *Suppose preferences $u(.)$ satisfy Assumption 1. Then, $q_{1j} = \alpha_1 + \frac{\delta}{1-\pi} q_j$ and $q_{\tau j} = \alpha_\tau + \frac{1-\delta}{\pi} q_j$ for some scalars $\alpha_1, \alpha_\tau \in \mathbb{R}$, and $\delta \in (0,1)$.*

Lemma 1 shows that firms' expansion curves are *linear*. That is, for every additional unit of firm-level production, q_j, a fraction δ to low-taste ones and a fraction $(1-\delta)$ is sold to high-taste consumers. 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.[15]

Optimal taxes and subsidies with endogenous labor supply. We next turn to study optimal taxes and subsidies when the planner is not restricted to keep the level of aggregate labor fixed. We showed in Propositions 5 and 6 that in the first best, aggregate labor is higher and, in general, the allocation of labor across firms is different. The following proposition shows that the planner optimally imposes a *uniform* subsidy, irrespective of the shape of demand elasticities.

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 7). Despite the fact that employment shares in the market equilibrium are now different from those in the efficient allocation (Proposition 6), Assumption 1 implies that firms have linear expansion curves. That is, 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 reason the planner imposes a subsidy at all is that, in the decentralized equilibrium, the real wage is lower than the marginal product of labor. As discussed in the previous section, this difference stems from the presence of misallocation across consumers. We provide a more detailed intuition below. The marginal product of labor is equal to the value of allocating an additional worker to some firm j. Under Assumption 1, marginal products are equalized across firms; hence the identity of firm j is irrelevant. As shown in the case of fixed labor supply, adding a worker to firm j results in δ/c_j more units allocated to low-taste consumers and $(1-\delta)/c_j$ more units to high-taste consumers. Thus, the marginal product of labor is equal to

$$MPL = \delta \frac{\tau u'(q_{\tau j})}{c_j} + (1-\delta) \frac{u'(q_{1j})}{c_j}. \tag{3.5}$$

Given that we normalized the nominal wage to 1, the real wage in the economy is $1/P$. We can use the firm's optimality conditions, equations (2.12)–(2.13), to rewrite the real wage as

[14] We are grateful to Michael Peters for helping us clarify the intuition behind Proposition 7.

[15] 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 Online Appendix B, we explore the properties of such firm-level subsidies quantitatively using Kimball preferences.

$$\frac{1}{P} = \delta \frac{\tau u'(q_{\tau j})}{c_j} + (1-\delta)\frac{1-\tau\pi}{1-\pi}\frac{u'(q_{1j})}{c_j}, \tag{3.6}$$

where we 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 difference between the marginal product of labor (3.5) and the real wage (3.6) is equal to

$$MPL - \frac{1}{P} = (1-\delta)\frac{\pi}{1-\pi}(\tau - 1)\frac{u'(q_{1j})}{c_j} > 0. \tag{3.7}$$

Since the labor wedge is positive, the planner optimally sets a positive level of subsidy for all firms to raise the level of aggregate production.

3.5 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.[16] Firms therefore solve the following problem:[17]

$$\begin{aligned} \max_{\{p_j, q_{1j}, q_{\tau j}\}} \quad & (\pi q_{\tau j} + (1-\pi) q_{1j})(p_j - c_j) \\ \text{s.t.} \quad & \tau u'(q_{\tau j}) = \frac{p_j}{P}, \qquad (3.8) \\ & u'(q_{1j}) = \frac{p_j}{P}. \qquad (3.9) \end{aligned}$$

No misallocation within firms. From the two demand curves (3.8)–(3.9), it follows directly that marginal utilities are equated 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 3 states that under nonlinear pricing, reallocating a firm's production from high-taste to low-taste consumers raises welfare.

The intuition for the difference is straightforward. With linear pricing, both types of consumers equate their marginal utility with the real price of the good. Since both types of consumers face the same price, their marginal utilities are equal. With nonlinear pricing, firms ensure separation between the two types of consumers by restricting the quantity sold to the low type, increasing its marginal utility relative to the high type.[18]

[16] The linear pricing restriction naturally arises if there is a secondary market in which products can be repackaged and sold at no cost.

[17] All formal derivations are relegated to Appendix B.

[18] 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.

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

$$\left(\frac{\tau_{ij}u'\left(q_{ij}\right)}{c_j}\right) \Big/ \left(\frac{\tau_{i'j'}u'\left(q_{i'j'}\right)}{c_{j'}}\right) = \frac{\mu_j}{\mu_{j'}}, \qquad \forall\left\{(i,i') \in [0,1], (j,j') \in [0,1]\right\}, \tag{3.10}$$

where $\mu_j \equiv \frac{p_j}{c_j}$ is the markup of firm j. We obtain equation (3.10) 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. When markups are heterogeneous, there is misallocation across firms. Firms that charge higher markups are underproducing, whereas firms with relatively lower markups are overproducing.

The equilibrium markups charged by 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:

$$\mu_j = \frac{\epsilon_j}{\epsilon_j - 1}. \tag{3.11}$$

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

$$\epsilon_j = \alpha_j\, \epsilon(q_{\tau j}) + (1-\alpha_j)\, \epsilon(q_{1j}), \tag{3.12}$$

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 share of sales going to high-taste consumers. 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).

If preferences feature decreasing demand elasticities, then high-productivity firms charge a higher markup. Because of the higher relative markup, these large, high-productivity firms are *too small* relative to the efficient allocation.[19] To raise welfare, a 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 can set nonlinear pricing schedules. Note that in the baseline economy with nonlinear pricing, there is markup heterogeneity across firms as well.[20] If preferences feature decreasing demand elasticities, the higher a firm's productivity, the more it sells and the higher the markup it charges to consumers.[21] There is, however, no misallocation across firms. As a result, observing large firms that charge high

[19] Proposition 11 in Appendix A confirms that the classic result of the macro literature on markups and misallocation also holds in our setup with consumer heterogeneity.

[20] 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.

[21] Proposition 9 in Appendix A formalizes this.

markups does not imply that these firms should be subsidized. As long as pricing schedules can be set nonlinearly, subsidizing large, high-markup firms increases misallocation and leads to welfare losses. The key result of no misallocation across firms despite markup heterogeneity also holds in a world without consumer heterogeneity. If all consumers have the same preferences, or firms are allowed to perfectly price discriminate, there would be no relationship between markups and allocative efficiency either. Such economy would be efficient while displaying heterogeneous markups across firms.

Aggregate labor distortion. As pointed out by Edmond, Midrigan and Xu (2022), in the economy with linear pricing, the labor wedge—that is, the difference between the real wage and the marginal product of labor—is driven entirely by the aggregate cost-weighted markup. A higher aggregate markup reduces the real wage while not changing the marginal product of labor. The source of the labor wedge is entirely different than in the benchmark economy, (3.7). With nonlinear pricing, the labor wedge is independent of the aggregate level of markup; it only depends on the marginal utility wedge between low- and high-taste consumers. Thus, relaxing the linear pricing assumption breaks the link between the labor wedge and the aggregate markup in the economy. In section, 4, we explore the quantitative importance of this difference.

3.6 Market structure vs. preferences

The source of heterogeneous market power in our environment is consumer preferences. We follow Kimball (1995) and Dhingra and Morrow (2019) in assuming that preferences feature variable elasticity of substitution. An alternative source used in the literature to obtain heterogeneous markups is market structure; the seminal paper using this approach is Atkeson and Burstein (2008). In this subsection, we show that our main results hold also when heterogeneity in market power is driven by oligopolistic competition rather than preferences. First, there is misallocation across consumers of the same firm. Second, we show that relative markups are not indicative of whether a social planner would optimally choose to subsidize or tax a firm. In fact, we find that under a large set of parameters, the social planner optimally chooses to *tax* the large, high-productivity firm despite its relative markup being higher.

Setup. Consumers have nested CES preferences over a continuum of product types $m \in (0,1)$ with two firms, j and k, producing one variety of the product each.[22] Consumers have idiosyncratic tastes towards different product types, $\tau_{im} \in \{1, \tau\}$ with $\tau > 1$. The utility of household i is given by

$$U_i = \int_0^1 \tau_{im} q_{im}^{\frac{\sigma-1}{\sigma}} dm, \tag{3.13}$$

where

$$q_{im} = \left(q_{ijm}^{\frac{\epsilon-1}{\epsilon}} + q_{ikm}^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}}. \tag{3.14}$$

[22] See Online Appendix D for the detailed setup.

We assume that all product types are symmetric and omit the m subscript in the equations below for brevity. The two firms have the same linear production function as in the main text and differ along their cost of production. Firms choose a bundle of quantity and price designed respectively for the high- and low-taste consumer to maximize profits. The main difference from our benchmark model is that the marginal utility of consumption from one firm depends not only on the aggregate price index, but also directly on the quantity consumed from the other firm in the product type. We solve for a Nash Equilibrium.

Market allocation. The market allocation features a similar structure to the case with monopolistic competition. Conditional on the aggregate price index P, firms choose a quantity sold to high-taste consumers that equates marginal utility to marginal cost, while the quantity sold to the low-taste consumers is distorted downwards:[23]

$$\frac{\partial U_i}{\partial q_{j\tau}} = \frac{c_j}{P}, \tag{3.15}$$

$$\frac{\partial U_i}{\partial q_{j1}} = \frac{1-\pi}{1-\pi\tau\psi_j}\frac{c_j}{P}, \tag{3.16}$$

where $\psi_j \in (\frac{1}{\tau}, 1)$ is given by

$$\psi_j = \left(\frac{q_{k\tau}^{\frac{\epsilon-1}{\epsilon}} + q_{j1}^{\frac{\epsilon-1}{\epsilon}}}{q_{k1}^{\frac{\epsilon-1}{\epsilon}} + q_{j1}^{\frac{\epsilon-1}{\epsilon}}} \right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1}.$$

The allocation for firm k is defined symmetrically.

The market allocation (3.15)–(3.16) take the same form as in our benchmark model (2.12)–(2.13) with the only difference being the presence of $\psi_j \leq 1$. Intuitively, the incentive compatibility constraint in this setup is less binding relative to our benchmark setup and hence the distortion to low-taste consumers is not as large. To understand why the IC constraint is less binding, note that the utility of consuming a given quantity from firm j depends not only on the taste shifter τ_{ij} but also on the quantity consumed from the competitor firm, q_{ki}. The high-taste consumer purchases more of the other variety k, which implies that she values the bundle offered by firm j to the low type *less* than in the baseline model. All else equal, this makes deviating to the small bundle less attractive and leads to a less binding IC constraint.

Misallocation. The market allocations feature the same types of distortions as in our benchmark model. Conditional on the price index and allocations of the competitor, firms distort the allocation to the low-taste consumer downwards. And due to general equilibrium effects, the high-taste consumer allocation is larger than its first-best allocation. That is, there is misallocation within firms between

[23] In the equations below, we assume that parameters are such that the incentive compatibility constraint binds for the high-taste consumer—as it does in the benchmark model. We show in Online Appendix D that there are parametrizations under which the incentive compatibility for the high taste consumer is slack and the individual rationality is binding instead. In those cases, the optimality conditions are still given by (3.15)–(3.16), but with $\psi_j = 1/\tau$.

high- and low-taste consumers as the one studied in section 3.1.

Unlike our benchmark model, the downward distortion to low-taste consumer varies across firms. Assume without loss of generality that firm j is more productive so that $c_j \leq c_k$. In that case, $\psi_j > \psi_k$ so that the productive firm distorts the low-taste consumer allocations relatively more. Because the productive firm distort more across their consumers, the social planner has an incentive to implement firm-level taxes and subsidies. We confirm numerically that the planner may choose to *tax* the high-productivity firm in order to reduce the average within-firm misallocation.[24] As the high productivity firm charges higher markups from both consumers, this result is in stark contrast to the standard result obtained under linear pricing in which the high-markup firm should be subsidized.

3.7 Imperfect substitutability within firms

In our benchmark model, we assume that utility only depends on the total quantity consumed. That is, purchasing multiple small-sized packages is *perfectly substitutable* with consuming a single large-sized package. Expanding our environment to incorporate imperfectly substitutable goods is not straightforward. Typical definitions of imperfect substitutability apply to pre-specified goods (e.g., apples vs. oranges or Coca-Cola vs. Pepsi), rather than different quantities of the same good, where the quantities sold in each package size are chosen by firms. Nonetheless, we show that the assumption of perfect substitutability does not drive our two main results: (i) nonlinear pricing can lead to misallocation across consumers within a firm, (ii) firm-level markups may not be informative of the degree of misallocation across firms. The key assumption in our setup that underlies the difference relative to standard models of linear pricing is that goods are *indivisible*, that is, firms can mandate a minimum quantity sold (package size) together with price. Consumers cannot purchase fractions of large units.

To show that our results are robust to imperfect substitutability, we consider three variants of our baseline model, which we present in detail in Online Appendix C. The first model we consider preserves the choice of firms and the social planner of what a "large" and "small" package is. In this specification, we assume that consumers' utility is not only a function of the total quantity consumed across all possible sizes s, $\sum_s n_{ij}^s \times q_{ij}^s$, but also depends on *how* the quantities are purchased. The latter is summarized by a utility penalty $\nu(n_s, q_s)$, that may depend on package size as well as the number of units purchased:

$$\tau_{ij} u \left(\sum_s n_{ij}^s \times q_{ij}^s \right) - \sum_s \nu(n_{ij}^s, q_{ij}^s). \tag{3.17}$$

Purchasing two small bundles now no longer gives the same utility as purchasing one large one, even if the physical quantity were to be the same. As a result, the incentive compatibility constraint of the high-taste consumer depends not only on her taste shifter τ_{ij}, but also on the relative utility penalty of purchasing a large package. Despite this difference, our main results continue to hold. Conditional on an aggregate price index, there is still no incentive to distort the quantity sold to high-

[24] For example, such a tax arises with: $\sigma = 3$, $\epsilon = 4$, $c_k/c_j = 1.5$, $\tau = 1.5$, $\pi = 0.5$. In that case, the optimal tax on the productive firm is 1.95%.

taste consumers, while the one to low-taste consumers is distorted downwards. In general equilibrium, there is too much consumption for high-taste consumers and too little for low-taste consumers.

The allocation of production across firms now depends on the properties of $u'(.)-v'(.)$. If we extend the definition of CED to the combined utility function, we can show that Proposition 7 continues to hold.[25] The allocation of labor to firms is efficient, despite the fact that markups vary across firms.

In the other two alternative models of imperfect substitutability we consider, package sizes are chosen by nature. This allows us to connect to more standard models of imperfect substitutability. In one model, we maintain the discrete choice nature of the problem, but add an idiosyncratic taste towards every specific package size.[26] In the other one, we allow consumers to purchase multiple bundles and assume there is a constant elasticity of substitution across different-sized packages. Also in these two models we show that there is misallocation across consumers and that firms' relative markup is not indicative of whether they should be taxed or subsidized.

3.8 Discussion

In this section, we discuss the main assumptions of the model and their importance for the theoretical results.

First, taste shifters can take one of two values only, $\tau_{ij} \in \{1, \tau\}$. This assumption simplifies the exposition, but is not essential for the results. In Online Appendix A, we prove Propositions 3–7 for the case where the support of taste shifters is continuous and firms can offer a continuum of bundles. We also consider the case where taste shifters are continuously distributed, but firms can only offer two bundles. We prove that our theoretical results hold in this case as well.

Second, firms use a constant-return-to-scale (CRS) technology. The assumption of a CRS technology implies that the marginal cost of a firm is independent of its total production and that there are no fixed costs. While simplifying our analysis, this assumption is not driving our results. In Online Appendix E, we analyze a version of the model in which firms have general cost function $C_j(q_j)$. We show that the main results continue to hold: there is misallocation across consumers—too much is sold to those with high taste, too little to those with low taste—, but not across firms—the firm-level of employment and output are identical to the efficient allocation.[27]

Third, preferences satisfy CED. Outside of this class of preferences, Propositions 4–7 no longer hold. However, the fundamental source of misallocation is still across consumers: firms distort bundles in order to extract maximal surplus. As a result, the social planner could implement the efficient allocation with a subsidy to sales to low-taste consumers, and this subsidy would be constant across firms. If the planner is additionally restricted to firm-level taxes and subsidies, as we analyze in the baseline model, she would now optimally use these. In Online Appendix B, we lay out the model with

[25] If the combined utility function features CED, consumers lose v% of utility from purchasing a larger size.

[26] This definition of imperfect substitutability is reminiscent of Anderson, De Palma and Thisse (1988), who show that firm-level demand in the multinomial logit model takes the CES form. Also in our environment, the price elasticity of demand for an individual package of a firm is finite.

[27] Our model can also accommodate a "fixed cost" at the product level. Suppose there is a cost f_j to being in the store, regardless of size. Then, the firm's objective function becomes $\pi\left[q_{\tau j}\left(p_{\tau j}-c_j\right)-f_j\right]+(1-\pi)\left[q_{1j}\left(p_{1j}-c_j\right)-f_j\right]$. Such a fixed cost may affect which types of consumers the firm chooses to serve, but would not affect quantity sold or price charged to any given type.

one commonly used example of non-CED preferences, Kimball (1995). Quantitatively, we find that the welfare gains from reallocation of production across firms are minimal.

Fourth, we assume that taste shifters (τ) as well as the share of customers with a high taste towards a good (π) do not vary across firms. This assumption allows us to cleanly study the new channel we propose in this paper—misallocation across consumers. If the distribution of tastes were to vary across firms, there would be an additional source of misallocation, as the level of distortion introduced by nonlinear pricing would no longer be constant across firms. In particular, the planner would have an incentive to impose a relative tax on firms with larger distortions, where the size of the distortion is equal to $(1-\pi_j)/(1-\tau_j\pi_j)$.[28]

Finally, we assume that consumers may only purchase a single bundle—either the small or the large one. That is, consumers are not allowed to purchase multiple units of the same size, nor fractions of a bundle. In the calibration, we find that no consumer prefers purchasing multiple units. So the key restriction that allows firms to price discriminate is the indivisibility of bundles. In order to pay the lower unit price $p_{\tau j}$, consumers must purchase the large bundle, $q_{\tau j}$.

4 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 five times as large as those under linear pricing. Moreover, implementing a tax system that would eliminate misallocation under linear pricing significantly worsens misallocation under nonlinear pricing.

Finally, we study the inefficiency from distortions to aggregate labor supply in the nonlinear and linear pricing environments. Although both environments feature a large average markup, the distortion to aggregate labor is substantially larger under linear pricing. Nonlinear pricing breaks the link between aggregate markups and the aggregate labor supply.

4.1 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.[29] We focus on core grocery goods, which include the departments of dry groceries,

[28] Note that we have assumed taste shifters are log-additive with the utility from consumption ($\tau_{ij}u(q_{ij})$). An alternative assumption would be that taste shifters enter inside the utility function ($u(\tau_{ij}q_{ij})$). Under the latter assumption, the level of distortion varies across firms so that the planner would have an incentive to impose firm-level taxes.

[29] 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.

Table 1: Summary Statistics

Number of products	165,053
Number of product modules	556
Number of product lines	41,949
Share of sales in multi-size product lines	90.4%
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.

frozen food, and dairy.[30] We use data from a single week in 2017.[31]

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: 90.4% 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:

$$\ln p_{ujs} = \beta \ln q_{ujs} + \Gamma \mathbf{X}_{ujs} + \epsilon_{ujs}, \tag{4.1}$$

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 by store fixed effects. The estimates suggest that the total 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 Scott R Baker, Stephanie G Johnson and Lorenz 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 the concern that all per-unit price variation is driven by this channel, 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,

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

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

Table 2: Nonlinear Pricing in the Retail Sector

Dependent variable:	price per unit (ln)			
	(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	3.3M

Notes: This table reports the results of regression (4.1). 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.

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 total price.[32]

The purpose of the regression analysis presented in this section is not to claim that all per-unit price variation is due to heterogeneous consumer tastes. Instead, the results indicate that firms are not restricted to set constant per-unit prices.

4.2 Calibration

Recall that our baseline theoretical analysis considers an environment with two types of households and, therefore, two different sizes offered to consumers. In the data, firms offer many different sizes. We therefore extend the model for the quantitative analysis to allow for a continuum of household types.[33] In particular, the taste of household i towards product j, τ_{ij}, is distributed uniformly between 1 and $\bar{\tau}$. We further assume that preferences satisfy Assumption 1, so that the inverse marginal utility takes the following form,

$$\left(u'\right)^{-1}(x) = -\beta_0 + \beta_1 x^{-\eta}, \tag{4.2}$$

where β_0, β_1, and η are structural parameters.

We treat each product line as a firm in the model.[34] Firm productivity is assumed to follow a Pareto distribution with shape parameter θ. Following the literature, we consider two values for the

[32] 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.

[33] In Online Appendix A, we set up the full model with a continuum of types and show that our theoretical results carry over.

[34] In doing so, we abstract from multi-product firms as well the division of surplus between the manufacturer, wholesaler, and retailer. Absent contracting frictions, the latter abstraction is without loss of generality.

Frisch elasticity: 1 ($\varphi = 1$), as is standard in macro models of misallocation (see Edmond, Midrigan and Xu (2022)) as well as a lower Frisch elasticity of 0.25 ($\varphi = 4$) as suggested by Raj Chetty, Adam Guren, Day Manoli and Andrea Weber (2011). We normalize $\beta_1 = 1$. This normalization is without loss of generality. [35] The disutility of labor, ν, is calibrated such that in the market equilibrium, aggregate labor supply is equal to 1.[36]

We calibrate the four structural parameters—the elasticity of differences η, the taste dispersion $\bar{\tau}$, the Pareto shape θ, and the degree to which the demand elasticity varies with quantity sold β_0—to match four key moments in the data. These parameters are calibrated to match the dispersion of package sizes within firms, the degree of market concentration, the aggregate markup, and the elasticity of firm-level markups with respect to size. The first two moments are computed from the Nielsen data—the dispersion of package sizes within firms is measured as the average sales-weighted standard deviation of log package size within a product line, while for market concentration we match the average sales share of the top 5% of product lines within a product module.

The last two moments are chosen to be consistent with the literature on markups and misallocation across firms. We target a cost-weighted aggregate markup of 1.3, which is within the range of estimates for the U.S.[37] For the elasticity of firm-level markup to size, we target 0.03 as estimated by Edmond, Midrigan and Xu (2022) using Compustat data.[38] Parameters are chosen to minimize the sum of squared deviations of model to data moments.

The first model column of Table 3 presents the results of the benchmark calibration. The upper bound of the taste distribution ($\bar{\tau}$) is estimated to be 1.19. This estimate implies that, for any given quantity consumed, the marginal utility of the highest taste consumer is 19% larger than that of the lowest taste. This taste dispersion allows our model to match the dispersion of package size within firms. The Pareto shape, governing the dispersion of productivities across firms, is estimated to be 3.14. This value allows our model to match the high degree of concentration observed in the data—the top 5% of firms account for 73% of sales.

Next, the elasticity of differences is estimated to be 4.15. This value also corresponds to the demand elasticity of the most productive firms. This elasticity helps our model to match the aggregate 30% markup we target.[39]. Finally, we estimate a value of β_0 of 0.02 which generates that, on average, firms that are 10% larger charge a 0.3% higher markup.

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 3.5. 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

[35] See Appendix A.6 for a formal argument.

[36] Note that the estimation of the structural parameters $\{\bar{\tau}, \theta, \eta, \beta_0\}$ 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.

[37] Edmond, Midrigan and Xu (2022) consider aggregate markups between 1.1–1.4. Sangani (2024) estimates an average cost-weighted retail markup of 1.32. We choose to match the cost-weighted rather than sales-weighted aggregate markup, since the former is equal to the aggregate labor wedge under linear pricing.

[38] The firm-level markup is defined as the cost-weighted markup across all consumers. Since production features constant returns to scale, this is equivalent to revenues over costs.

[39] If preferences were CES ($\beta_0 = 0$), then η would map one for one into the aggregate markup.

Table 3: Calibrated Moments and Parameters

Parameter		Model		Moment		Model	
		Benchm.	LP		Data	Benchm.	LP
$\bar{\tau}$	Highest taste	1.19	1.47	Package size disp.	0.44	0.44	0.44
θ	Pareto shape	3.14	3.57	Sales share top 5%	0.73	0.73	0.73
η	Elasticity of diffs	4.15	4.20	Aggregate markup	1.3	1.3	1.3
β_0	Departure from CES	0.02	0.06	Markup-size elast.	0.03	0.03	0.03

Notes: This table reports the calibrated parameters for the benchmark model as well as the linear pricing (LP) model. The LP model is identical except for the fact that firms are restricted to charging a constant unit price. The right panel reports the data moments used for calibration as well as their model counterparts.

allows us to compare the magnitude of misallocation to what researchers would conclude if they used a standard linear pricing model calibrated to match the same moments. 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 targeted moments. However, only our benchmark model, in which firms are allowed to price nonlinearly, is able to generate significant dispersion in unit prices within the same product line. Estimating regression (4.1) in the benchmark model, we obtain an estimate of -0.2, while the coefficient is equal to 0, by construction, in the linear pricing model. That is, the benchmark model accounts for about one-third of the amount of nonlinearity of prices that we observe in the data.

4.3 Misallocation

We first study the welfare costs of misallocation, that is, keeping aggregate labor constant. We consider both misallocation of production across firms and of consumption across households. Table 4 reports the results.

Table 4: Welfare Costs of Misallocation (Fixed Labor Supply)

	Benchmark	Linear Pricing	Benchmark w/LP subsidies
Welfare Gains	0.68%	0.14%	-0.12%

Notes: This table reports the welfare gains in the fixed-labor efficient allocation relative to the benchmark model (column 1), the model with linear pricing (column 2), and the benchmark 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. These welfare gains are computed while keeping the aggregate level of labor unchanged.

Under nonlinear pricing, the costs of misallocation are entirely due to misallocation across consumers. In the market allocation, the standard deviation of package sizes (ln) is 0.44. This dispersion is larger than the efficient allocation, as high-taste consumers are allocated too much of the good. The dispersion of package sizes in the fixed-labor efficient allocation is half as large (0.21 vs. 0.44). This distortion leads to sizable welfare costs. Consumers are indifferent between consuming the FE allocation or consuming an additional 0.68% of all goods on top of the market allocation.

When firms must choose linear pricing schedules, there is no misallocation across consumers, since

markups are, by assumption, equalized within firms. There are, however, losses from misallocation across firms. Larger firms charge higher markups and, as a consequence, are too small relative to the efficient allocation. Optimally reallocating production across firms leads to welfare gains of 0.14% in consumption equivalent units.[40]

A social planner who has access to firm-level taxes and subsidies can restore the FE allocation in the model with linear prices. This is achieved by a size-dependent subsidy scheme, in which the most productive firms, that charge the highest markups, are subsidized, while the small, low-markup firms are taxed. Column 3 of Table 4 reports the welfare costs of implementing this subsidy if firms are not restricted to linear pricing.[41] That is, had we interpreted the same data moments through the lens of a standard linear pricing model and implemented the seemingly optimal subsidy scheme, this would have induced welfare losses of 0.12%. Such a subsidy scheme cannot alleviate welfare losses from misallocation across consumers, but instead distorts the allocation of labor across firms and thereby generates additional losses.

This result emphasizes a key take-away from this paper, namely that the mapping from markup dispersion to misallocation relies on the assumption of linear pricing. Both models were calibrated to match the same degree of markup dispersion across firms: an elasticity of markups with respect to firm size of 0.03. However, only if firms charge a linear pricing schedule is markup dispersion a sign of misallocation. To build further intuition, we now show how the equilibrium allocation under nonlinear pricing can be mapped to "wedges" in the sense of Restuccia and Rogerson (2007), Hsieh and Klenow (2009), and Baqaee and Farhi (2020).

Wedge accounting. Baqaee and Farhi (2020) show that, to a first order, the losses from misallocation are equal to

$$\mathcal{L} \approx \frac{1}{2} \times \varepsilon \times \text{var}_{\lambda} \left(\ln \omega_{ij} \right), \tag{4.3}$$

where ε is the elasticity of substitution across firms, $\text{var}_{\lambda} \left(\ln \omega_{ij} \right)$ is the sales-weighted variance of the log of the wedges, and ω_{ij} denotes the wedge of consumer-firm pair ij. This wedge is defined as the ratio between the marginal utility of that consumer and the real marginal cost. That is, $\omega_{ij} = \left(\tau_{ij} u'(q_{ij}) \right) / \left(\frac{c_j}{P} \right)$.

Under linear pricing, these wedges are equal to firm-level markups. That is, $\omega_{ij} = \mu_j$. Under nonlinear pricing, equation (4.3) continues to hold but the wedges are no longer driven by markups. Instead, the wedges depend only on consumer tastes, $\omega_{ij} = \omega(\tau_i)$.[42] The dispersion in these wedges, which is the source of misallocation, is due to within-firm dispersion rather than across-firm dispersion. The sales-weighted variance of the wedges that arise under nonlinear pricing is about 6 times larger than those under linear pricing.[43] This ratio is similar to the ratio in welfare losses due to misallocation

[40] The closest paper in terms of model and calibration strategy is Edmond, Midrigan and Xu (2022). Welfare losses of 0.14% are considerably smaller than their estimated welfare gains from optimal firm-level taxes. This is largely due to the fact that their setup features both endogenous capital accumulation and a roundabout production structure, both of which significantly amplify any welfare effects of distortions.

[41] 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.

[42] In Appendix A, we show that $\omega(\tau) = \frac{\tau}{\tau - h(\tau)^{-1}}$, where $h(\tau)$ is the hazard rate of taste-shifter τ.

[43] The first-order approximation for the loss due to misallocation (4.3) is quite accurate. Under nonlinear pricing, this

across the nonlinear and linear pricing calibrations, which is about 5.

The size of welfare losses with linear and nonlinear pricing. The welfare gains of eliminating misallocation are about five times larger in the nonlinear pricing environment relative to the linear pricing one.[44] Ultimately, the relative size of misallocation across the two models depends on the severity of misallocation across firms vs. across consumers.

The degree of misallocation across consumers in our benchmark model is mostly identified by matching the dispersion of package sizes within firms. Recalibrating the two models to match a 10% lower package size dispersion results in a 19% decline in the welfare costs of misallocation under nonlinear pricing and almost no change in the welfare costs under linear pricing.

Similarly, the degree of misallocation across firms under linear pricing is mostly governed by the elasticity of firm-level markups to their size. Recalibrating the two models to match a 10% higher elasticity raises the welfare costs of misallocation under linear pricing by 22% and leads to almost no change in the welfare costs under nonlinear pricing.

4.4 Aggregate markup and labor supply

In this section, we quantify distortions in aggregate labor supply. Table 5 summarizes the results. We first consider the calibration with a Frisch elasticity of 1. In the first best allocation, aggregate labor is 6.4% higher than in the market equilibrium. The first-best allocation yields welfare gains of 0.92% relative to the market allocation—that is, an additional 0.24% 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 6.4%, but that increase is not uniform across firms. The employment of the bottom 50% of firms is 7.5% higher in the first-best allocation relative to the market allocation, whereas the employment of the top 50% grows only by 6.3%.[45]

A planner with access to only firm-level taxes and subsidies cannot achieve the first-best allocation as they cannot solve the misallocation of consumption 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 7.4% across all firms. This policy increases aggregate labor by 5.9% and raises welfare by 0.2% in consumption equivalent units.

When imposing a linear pricing assumption, researchers would conclude that the optimal level of labor is 23.5% higher than in the market equilibrium. Under linear pricing, as shown by Edmond, Midrigan and Xu (2022), the aggregate labor wedge is driven by the aggregate markup in the economy, which is targeted at 30%.[46]

formula delivers a welfare loss of 0.66% relative to the actual 0.68%.

[44] Instead of recalibrating the parameters of the model to match the same data moments, we could have imposed the linear pricing assumption using the previously calibrated parameters and use this for comparison. The welfare losses from misallocation would have been 0.14% in this calibration. However, when we impose the linear pricing assumption in this way, the model predicts a relationship between markups and market size that is far from the data, making any comparison between the models hard to interpret.

[45] Since the top 50% of firms employ 94% of workers in the decentralized equilibrium, the overall labor increase is much closer to the increase in employment for the top 50% of firms.

[46] In fact, Edmond, Midrigan and Xu (2022) find comparable degrees of undersupply of labor, ranging from 30.1% when targeting an aggregate markup of 1.25 to 42.1% with an aggregate markup of 1.35.

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

		Nonlinear Pricing		**Linear Pricing**
		FB	SB	FB & SB
Frisch elas. $= 1$	Aggregate Labor	+6.4%	+5.9%	+23.5%
	Welfare Gains	+0.92%	+0.20%	+2.81%
Frisch elas. $= 0.25$	Aggregate Labor	+1.8%	+1.7%	+6.3%
	Welfare Gains	+0.75%	+0.06%	+0.88%

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. The first two rows use $\varphi = 1$, the last two rows use $\varphi = 4$. Under linear pricing, the first- and second-best allocations coincide. Welfare gains are in consumption equivalent units. In contrast to Table 4, the welfare gains in the table take into account changes in aggregate labor.

As explained in Section 3, the tight link between the aggregate markup and the labor wedge breaks under nonlinear pricing. With the same aggregate markup, the optimal level of labor is only 6.4% higher than its market level. One can understand this result using the wedges, ω_{ij}, we derived in the previous subsection. To a first order, the aggregate labor wedge is equal to the cost-weighted average of these wedges. Under nonlinear pricing, these wedges are not driven by markups but by the distortion across consumers due to price discrimination. Their average is equal to 7.3%, much lower than the 30% aggregate markup.

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 large (31% 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 1.67%. 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.

When labor supply is less elastic, the same aggregate markup leads to a smaller distortion in aggregate labor. With a Frisch elasticity of 0.25, the aggregate labor distortion is 1.8% under nonlinear pricing and is equal to 6.3% under the assumption of linear pricing.

4.5 Alternative specifications

In this section, we consider two alternative specifications to the benchmark model. First, we consider an alternative environment in which firms are restricted to offer only two bundles. We find that the degree of misallocation is similar to the benchmark model. Second, we allow for preferences that do not feature CED. Also in this environment, the role for misallocation across firms is negligible.

4.5.1 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.3. 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.

We calibrate this version of the model to the same targets, resulting in small variation in the structural parameters.[47] We find that the welfare costs of misallocation under this specification is very close to our benchmark specification. The welfare gain from removing misallocation is 0.71% in consumption equivalent units, compared to 0.68% in our benchmark specification. This result suggests that the welfare costs of misallocation are robust to considering a finite number of package sizes.

It is worth noting that when firms are restricted to offering fewer package sizes than there are consumer types, the equilibrium may involve a mixture of linear and nonlinear pricing. In order to cater more closely to each type, firms might sell multiple units of a single package at a linear price; effectively bypassing the restriction on the number of package sizes. In the calibrated model, such mixed pricing strategy does not arise in equilibrium: each consumer chooses to purchase only one unit. Analyzing environments in which linear and nonlinear pricing co-exist is a promising avenue for future research.

4.5.2 Kimball preferences

To quantify the importance of CED for the main misallocation results, we estimate the benchmark model assuming that preferences feature a Kimball (1995) aggregator using the specification of Klenow and Willis (2016).[48] The overall welfare losses from misallocation are similar to our benchmark model (0.79% vs. 0.68% with CED preferences).

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 negligible: Welfare only increases by 0.02% with the optimal firm-level taxes and subsidies. That is, even when allowing for preferences which give rise to across-firm misallocation, total misallocation is driven almost entirely by the distortion of quantities across consumers, as in our benchmark model.

4.6 Market Concentration, Markups, and Misallocation

A large body of evidence has documented that the aggregate markup as well as the degree of market concentration in the economy have been rising over the past decades.[49] In this section, we evaluate

[47] See Table A.2 in Online Appendix A.

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

[49] For the rise in markups see De Loecker, Eeckhout and Unger (2020), Edmond, Midrigan and Xu (2022), and Hall (2018). For the rise in concentration see Gutierrez and Philippon (2017), and Autor et al. (2020).

how the welfare consequences of these secular changes depend on whether firms can offer nonlinear pricing schedules.

To evaluate the change in misallocation over time, we recalibrate our benchmark model as well as the linear pricing model to match the U.S. economy in the 1980s. We do so by estimating the elasticity of taste differences (η) and productivity Pareto shape (θ) to match the lower level of aggregate markup and lower degree of concentration in the 1980s. We keep all other parameters—the degree of taste heterogeneity ($\bar{\tau}$), the departure from CES (β_0), and the Frisch elasticity (φ)—fixed.[50] We target an aggregate markup of 1.15, 15 basis points below our benchmark estimate, consistent with the change in the cost-weighted markup estimated in De Loecker, Eeckhout and Unger (2020). We target a decline of 17% in the market share of the top 5% of firms consistent with evidence in Smith and Ocampo (Forthcoming).[51]

Both our benchmark model and the linear pricing model interpret the lower aggregate markup and lower degree of market concentration as a result of a higher elasticity of demand η and a thinner Pareto tail $1/\theta$.[52] The implications for welfare, however, are opposite.

If firms are restricted to price linearly, our model's welfare conclusions echo the large body of literature studying changes in market power over time. Misallocation across firms was significantly lower in the 1980s (welfare costs of 0.05% compared to 0.14%). With a lower aggregate markup, there were also much lower welfare losses compared to the first-best allocation, in which labor is chosen optimally (0.89% compared to 2.81%).

Under nonlinear pricing, welfare losses entirely stem from misallocation across consumers. With a higher elasticity of taste differences (η) and, by assumption, the same amount of taste dispersion, there is more scope for quantity distortions across consumers. In fact, the welfare costs of misallocation are 0.98% compared to 0.68% in the 2010's. Increased misallocation across consumers also translates into a higher labor wedge: in the first-best allocation, welfare would be 1.17% higher compared to 0.92% in the 2010's.

We caveat these findings by noting that the values of the other structural parameters ($\bar{\tau}$, β_0, φ) could have also changed over this time period. The analysis in this subsection should not be viewed as arguing that misallocation has necessarily gone down over the past decades. Instead, we emphasize that assuming firms use linear pricing schedules may alter not only quantitative welfare predictions, but also their qualitative direction. We leave for future research the analysis of how misallocation has changed over time as our micro data only goes back to the late 2000's.

[50] The analysis in this sub-section assumes $\varphi = 1$.

[51] Smith and Ocampo (Forthcoming) find that the market share of the top four firms selling a product category in a commuting zone was 17% lower in 1982 compared to 2012.

[52] The elasticity of demand rises from 4.15 in our benchmark calibration (4.2 under linear pricing) to 7.21 (7.37 under linear pricing). The Pareto shape rises from 3.14 in our benchmark calibration (3.57 under linear pricing) to 6.87 (7.64 under linear pricing).

5 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 increases. 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 1.7%.

In this paper, we studied how nonlinear pricing shapes misallocation in the goods market, assuming households are ex-ante identical. An important topic we leave open for future research is the distributional consequences of nonlinear pricing. 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?

References

Afrouzi, Hassan, Andres Drenik, and Ryan Kim. 2020. "Growing by the Masses: Revisiting the Link between Firm Size and Market Power." Working paper. 1

Anderson, S. P., A. De Palma, and J. F. Thisse. 1988. "The CES and the logit: Two related models of heterogeneity." *Regional Science and Urban Economics*, 18. 26

Argente, David, Munseob Lee, and Sara Moreira. 2024. "The Life Cycle of Products: Evidence and Implications." *Journal of Political Economy*, 132(2): 337–390. 1

Atkeson, Andrew, and Ariel Burstein. 2008. "Pricing-to-market, trade costs, and international relative prices." *American Economic Review*, 98(5): 1998–2031. 1, 3.6, D

Attanasio, Orazio, and Elena Pastorino. 2020. "Nonlinear Pricing in Village Economies." *Econometrica*, 88(1): 207–263. 1

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms*." *The Quarterly Journal of Economics*, 135(2): 645–709. 49

Baker, Scott R, Stephanie G Johnson, and Lorenz Kueng. 2020. "Financial returns to household inventory management." National Bureau of Economic Research. 4.1

Baqaee, David Rezza, and Emmanuel Farhi. 2020. "Productivity and misallocation in general equilibrium." *The Quarterly Journal of Economics*, 135(1): 105–163. 4.3, 4.3

Boar, Corina, and Virgiliu Midrigan. 2024. "Markups and Inequality." *Review of Economic Studies*. Advance online publication. 1

Bornstein, Gideon. 2021. "Entry and Profits in an Aging Economy: The Role of Consumer Inertia." Working paper. 1

Broda, Christian, and David E. Weinstein. 2010. "Product Creation and Destruction: Evidence and Price Implications." *American Economic Review*, 100(3): 691–723. C.3

Burstein, Ariel, Vasco M Carvalho, and Basile Grassi. 2020. "Bottom-up markup fluctuations." National Bureau of Economic Research. 1

Chetty, Raj, Adam Guren, Day Manoli, and Andrea Weber. 2011. "Are Micro and Macro Labor Supply Elasticities Consistent? A Review of Evidence on the Intensive and Extensive Margins." *American Economic Review*, 101(3): 471–75. 4.2

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. "The rise of market power and the macroeconomic implications." *The Quarterly Journal of Economics*, 135(2): 561–644. 1, 49, 4.6

Dhingra, Swati, and John Morrow. 2019. "Monopolistic competition and optimum product diversity under firm heterogeneity." *Journal of Political Economy*, 127(1): 196–232. 1, 2.4, 3.5, 3.6, A.1

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu. 2022. "How costly are markups?" *Journal of Political Economy*. 1, 3.5, 4.2, 37, 40, 4.4, 46, 49

Einav, Liran, Peter J Klenow, Jonathan D Levin, and Raviv Murciano-Goroff. 2021. "Customers and retail growth." National Bureau of Economic Research. 1

Fudenberg, Drew, and Jean Tirole. 1991. *Game theory*. MIT press. A.1

Gomes, Renato, and Alessandro Pavan. 2016. "Many-to-many matching and price discrimination." *Theoretical Economics*, 11(3): 1005–1052. 2

Gutierrez, German, and Thomas Philippon. 2017. "Declining Competition and Investment in the U.S." *NBER Working Paper*. 49

Hall, Robert E. 2018. "New evidence on the markup of prices over marginal costs and the role of mega-firms in the us economy." National Bureau of Economic Research. 49

Hsieh, Chang-Tai, and Peter J Klenow. 2009. "Misallocation and manufacturing TFP in China and India." *The Quarterly journal of economics*, 124(4): 1403–1448. 1, 4.3

Jullien, Bruno. 2000. "Participation Constraints in Adverse Selection Models." *Journal of Economic Theory*. 2

Kazumura, Tomoya, Debasis Mishra, and Shigehiro Serizawa. 2020. "Mechanism design without quasilinearity." *Theoretical Economics*, 15(2): 511–544. 2

Kehoe, Patrick, Brad Larsen, and Elena Pastorino. 2020. "Dynamic Competition in the Era of Big Data." *Working Paper*. 1

Kimball, M.S. 1995. "The Quantitative Analytics of the Basic NeomonetaristModel." *Journal of Money, Banking, and Credit*, 27(4). 3.6, 3.8, 4.5.2

Klenow, Peter J., and Jonathan L. Willis. 2016. "Real Rigidities and Nominal Price Changes." *Economica*, 83: 443–472. 4.5.2

Maskin, Eric, and John Riley. 1984. "Monopoly with incomplete information." *The RAND Journal of Economics*, 15(2): 171–196. 1

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

Mensch, Jeffrey, and Doron Ravid. 2022. "Monopoly, product quality, and flexible learning." *arXiv preprint arXiv:2202.09985*. 2

Mirrlees, James A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies*, 38: 175–208. 1, 2.3, 11

Mussa, Michael, and Sherwin Rosen. 1978. "Monopoly and product quality." *Journal of Economic Theory*, 18(2): 301–317. 1, 2.4

Myerson, Roger B. 1981. "Optimal Auction Design." *Mathematics of Operations Research*, 6(1): 58–73. 54

Nevo, Aviv, and Arlene Wong. 2019. "The elasticity of substitution between time and market goods: Evidence from the Great Recession." *International Economic Review*, 60(1): 25–51. 4.1

Peters, Michael. 2020. "Heterogeneous markups, growth, and endogenous misallocation." *Econometrica*, 88(5): 2037–2073. 1

Restuccia, Diego, and Richard Rogerson. 2007. "Policy Distortions and Aggregate Productivity with Heterogeneous Plants." NBER Working Paper 13018. 1, 4.3

Sangani, Kunal. 2024. "Markups across the income distribution: Measurement and implications." *Available at SSRN 4092068*. 37

Schottmueller, Christoph. 2015. "Adverse selection without single crossing: Monotone solutions." *Journal of Economic Theory*. 2

Smith, Dominic A, and Sergio Ocampo. Forthcoming. "The Evolution of US Retail Concentration." *American Economic Journal: Macroeconomics*. 4.6, 51

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

Tirole, Jean. 1988. *The Theory of Industrial Organization*. Cambridge, MA:MIT Press. 1, 11

Wilson, Robert B. 1993. *Nonlinear pricing*. Oxford University Press on Demand. 1, 1

A Proofs

A.1 Efficient allocation

Proof of Proposition 1.

Let's conjecture that the incentive compatibility constraints are not binding. The planner's problem can then be written as

$$\max_{\{q_{ij},l_i\}} \quad \int_i \int_j \tau_{ij} u(q_{ij}) dj di - \int_i \nu \frac{l_i^{1+\varphi}}{1+\varphi} di$$
$$\text{s.t.} \quad \int_i \int_j q_{ij} c_j dj di = \int_i l_i di$$

We will first show that the labor allocation as well as the resources devoted to the production of consumption for each household in the economy is identical. Then, we will show that given a resource constraint for the consumption of an individual household, the optimal consumption bundle that would be chosen by the household is identical to the one they are given by the planner. Thus, they have no incentive to misreport their type.

Taking first order conditions of the planner's problem we obtain

$$\tau_{ij} u'(q_{ij}) = c_j \lambda, \qquad \forall \{i,j\} \tag{A.1}$$
$$\nu l_i^{\varphi} = \lambda, \tag{A.2}$$

where λ is the Lagrange multiplier on the planner's resource constraint. The second optimality condition implies that labor is constant across households. Denote the optimal level of labor by L^{FB}.

The first optimality condition implies that q_{ij} varies across households only through τ_{ij} and c_j. That is, for any variety with cost c_j, the consumption of all households with a low taste towards that variety is identical, and the consumption of all households with a high taste towards that variety is identical. Let $q_1^{FB}(c)$ and $q_\tau^{FB}(c)$ denote the efficient level of consumption of a variety whose marginal cost of production is c for consumers with a low and high taste towards that variety, respectively.

Denote by R_i the resources devoted to the production of consumption by household i

$$R_i \equiv \int_j q_{ij} c_j dj.$$

Using the optimal quantity function we can rewrite the resources devoted to the production of consumption by household i as

$$R_i = \int_{\mathcal{J}_1^i} q_1^{FB}(c_j) c_j dj + \int_{\mathcal{J}_\tau^i} q_\tau^{FB}(c_j) c_j dj,$$

where $\mathcal{J}_1^i$ denotes the varieties for which household i has a low taste ($j : \tau_{ij} = 1$), and $\mathcal{J}_\tau^i$ denote the varieties for which household i has a high taste ($j : \tau_{ij} = \tau$). The iid assumption on taste shocks

implies that

$$R_i = (1-\pi)\int q_1^{FB}(c)cdF(c) + \pi\int q_\tau^{FB}(c)cdF(c). \tag{A.3}$$

The RHS of (A.3) has no household subscript, thus proving that the resources devoted to the production of consumption by individual households do not vary across households. The aggregate resource constraint then implies $R_i = L^{FB}$ for all i.

Consider now the problem of a household which maximizes their utility from consumption subject to an individual resource constraint:

$$\begin{aligned} \max_{q_j} \quad & \int_j \tau_{ij} u(q_{ij}) dj \\ \text{s.t.} \quad & \int_j q_{ij} c_j dj di = L \end{aligned} \tag{A.4}$$

Solving the problem we obtain the same optimality conditions of the planner:

$$\tau_{ij} u'(q_{ij}) = c_j \lambda.$$

Since also the resources devoted to the household consumption are identical across the individual and the planner's problems, we obtain that the optimal allocation of the household is identical to the planner's one. Since the production resources to devoted to all households, this implies that the household cannot gain utility from consumption from misreporting their type. Moreover, since all labor quantities are identical, we confirm that no household has an incentive to misreport their type. That is, the incentive compatibility in the planner's problem are not binding.

We have proved that the first-best allocation is identical to the constrained efficient allocation. Rearranging the optimality conditions (A.1)–(A.2), we obtain

$$\begin{aligned} u'(q_{ij}) &= \frac{c_j}{\tau_{ij}} \frac{1}{P^{FB}}, \qquad \forall\{i,j\} \\ P^{FB} &= \frac{1}{\nu (L^{FB})^{\varphi}}, \end{aligned}$$

where P^{FB} is the inverse of the Largrange multiplier on the planner's problem.

■

A.2 Existence and uniqueness of equilibrium

Proof of Proposition **2.** Using the optimality conditions of the firm, (2.12) and (2.13) and the labor supply optimal condition, we write labor market clearing directly as a function of P:

$$\int_0^1 c_j \left[\pi (u')^{-1}\left(\frac{c_j}{P}\frac{1}{\tau}\right) + (1-\pi)(u')^{-1}\left(\frac{c_j}{P}\frac{1-\pi}{1-\tau\pi}\right)\right] dj = \nu^{-\frac{1}{\varphi}} P^{-\frac{1}{\varphi}}. \tag{A.5}$$

When $P \to 0$, the RHS goes to ∞ while the LHS is a finite number. When $P \to \infty$, the RHS

goes to 0 while the LHS goes to ∞. Due to concavity of $u(\cdot)$, the LHS is increasing in P while the RHS is decreasing in P. Moreover the LHS is continuous in P (as the utility function is continuously differentiable) and also the RHS is continuous in P. A single crossing argument then implies that there exists a unique $P > 0$ such that (A.5) holds with equality.[53]

■

A.3 Benchmark misallocation results

Proof of Proposition **3.** From equations (2.12–2.13) and (2.8) we have that:

$$\frac{u'(q_{\tau j})}{u'\left(q_{\tau j}^{AE}\right)} = \frac{P^{AE}}{P}, \tag{A.6}$$

$$\frac{u'(q_{1j})}{u'\left(q_{1j}^{AE}\right)} = \frac{1-\pi}{1-\tau\pi}\frac{P^{AE}}{P}. \tag{A.7}$$

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

Aggregate labor market clearing implies that

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

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

Proof of Proposition **4.**

Equations (2.12–2.13), 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: $(1-\pi)\left[q_{1j}^{\text{AE}} - q_{1j}\right] - \pi\left[q_{\tau j} - q_{\tau j}^{\text{AE}}\right] = 0$. Using (2.13–2.12), this can be written as:

$$(1-\pi)\left[(u')^{-1}\left(\frac{c_j}{P^{AE}}\right) - (u')^{-1}\left(\frac{1-\pi}{1-\tau\pi}\frac{c_j}{\widetilde{P}_j}\right)\right] - \pi\left[(u')^{-1}\left(\frac{c_j}{\tau\widetilde{P}_j}\right) - (u')^{-1}\left(\frac{c_j}{\tau P^{AE}}\right)\right] = 0. \tag{A.8}$$

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

[53] 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 (2.12) and (2.13) 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.

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

$$\pi\left(q_{\tau,k}(\tilde{P}_j) - q^{AE}_{\tau,k}\right) \qquad - (1-\pi)\left(q^{AE}_{1,k} - q_{1,k}(\tilde{P}_j)\right) =$$
$$\pi(1+\Delta)^{\eta}\left(q_{\tau,j}(\tilde{P}_j) - q^{AE}_{\tau,j}\right) - (1-\pi)(1+\Delta)^{\eta}\left(q^{AE}_{1,j} - q_{1,j}(\tilde{P}_j)\right) = 0.$$

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 β_0, and $\beta_1 \geq 0$.

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

$$-\eta = \frac{\partial \log\left(g(x\gamma) - g(x)\right)}{\partial \log(x)}. \tag{A.9}$$

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

$$\frac{g''(x\gamma)\gamma x}{g'(x\gamma)} = -\eta - 1. \tag{A.10}$$

Equation (A.10) implies that $g'(x)$ is iso-elastic and can be written as

$$g'(x) = -\eta\beta_1 x^{-\eta-1},$$

or

$$g(x) = \beta_1 x^{-\eta} - \beta_0. \tag{A.11}$$

This proves part 1 of Lemma 2. Part 2 and 3 then directly follow from the firm's optimality conditions (2.12) and (2.13).

Finally, $\beta_1 \geq 0$ must hold in order for utility to be concave. To see this, notice that

$$u''(q) = -\frac{1}{\eta}\frac{1}{\beta_1}u'(q)^{1+\eta}, \tag{A.12}$$

which is negative iff $\frac{1}{\eta}\frac{1}{\beta_1} > 0$. Since $\eta > 1$ by assumption, this implies that $u(.)$ is concave only if $\beta_1 > 0$.

■

Proof of Lemma 1.

Using Lemma 2,

$$q_j \equiv \pi q_{\tau j} + (1-\pi)q_{1j} = -\beta_0 + \beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \hat{\tau} \tag{A.13}$$

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

$$q_{1j} = \beta_0 \left(\left(\frac{1-\tau\pi}{1-\pi}\right)^\eta \frac{1}{\hat{\tau}} - 1\right) + \left(\frac{1-\tau\pi}{1-\pi}\right)^\eta \frac{1}{\hat{\tau}} q_j \tag{A.14}$$

$$q_{\tau j} = \beta_0 \left(\frac{\tau^\eta}{\hat{\tau}} - 1\right) + \frac{\tau^\eta}{\hat{\tau}} q_j \tag{A.15}$$

and Lemma 1 follows with

$$\alpha_\tau = \beta_0 \left(\frac{\tau^\eta}{\hat{\tau}} - 1\right) \tag{A.16}$$

$$\delta_\tau = \frac{\tau^\eta}{\hat{\tau}} \tag{A.17}$$

$$\alpha_1 = \beta_0 \left(\left(\frac{1-\tau\pi}{1-\pi}\right)^\eta \frac{1}{\hat{\tau}} - 1\right) \tag{A.18}$$

$$\delta_1 = \left(\frac{1-\tau\pi}{1-\pi}\right)^\eta \frac{1}{\hat{\tau}} \tag{A.19}$$

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

PROOF OF PROPOSITION **5.** 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

$$\nu L^D(P)^{\varphi} = \frac{1}{P}. \tag{A.20}$$

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

$$\nu \left(L^{FB}\right)^{\varphi} = \frac{1}{P^{FB}} \tag{A.21}$$

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}:

$$L = \left(\frac{1}{\nu P}\right)^{\frac{1}{\varphi}} < \left(\frac{1}{\nu P^{FB}}\right)^{\frac{1}{\varphi}} = L^{FB}. \tag{A.22}$$

■

PROOF OF PROPOSITION **6.** The excess employment (ω_j) is given by

$$\omega_j = \frac{\frac{(\pi q_{\tau j} + (1-\pi) q_{1j}) c_j}{L}}{\frac{\left(\pi q_{\tau j}^{\mathrm{FB}} + (1-\pi) q_{1j}^{\mathrm{FB}}\right) c_j}{L^{\mathrm{FB}}}} = \frac{\pi q_{\tau j} + (1-\pi) q_{1j}}{\pi q_{\tau j}^{\mathrm{FB}} + (1-\pi) q_{1j}^{\mathrm{FB}}} \frac{L^{\mathrm{FB}}}{L} \tag{A.23}$$

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

$$\omega_j = \frac{\beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \hat{\tau} - \beta_0}{\beta_1 \left(\frac{c_j}{P^{FB}}\right)^{-\eta} \hat{\tau}^{FB} - \beta_0} \frac{L^{\mathrm{FB}}}{L} \tag{A.24}$$

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

$$\frac{\partial \omega_j}{\partial c_j} = \frac{-\beta_1 c_j^{-\eta-1} P^{\eta} \hat{\tau} \left[\beta_1 \left(\frac{c_j}{P^{FB}}\right)^{-\eta} \hat{\tau}^{FB} - \beta_0\right] + \beta_1 c_j^{-\eta-1} \left(P^{FB}\right)^{\eta} \hat{\tau}^{FB} \left[\beta_1 \left(\frac{c_j}{P}\right)^{-\eta} \hat{\tau} - \beta_0\right]}{\left(\beta_1 \left(\frac{c_j}{P^{FB}}\right)^{-\eta} \hat{\tau}^{FB} - \beta_0\right)^2} \frac{L^{\text{FB}}}{L} \tag{A.25}$$

Using the fact that the demand elasticity is decreasing in quantity consumed, we obtain $\beta_0 > 0$. Rearranging (A.25) and using the fact that we also have $\beta_1 > 0$ and $c_j > 0$, we obtain:

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

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 4, 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

$$\frac{P^{FB}}{\tilde{P}} = \left(\frac{\hat{\tau}}{\hat{\tau}^{FB}}.\right)^{\frac{1}{\eta}} \tag{A.26}$$

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.20) and (A.21). Therefore, to clear the labor market it must be that $P < \tilde{P}$ and we have

$$\frac{P^{FB}}{P} > \frac{P^{FB}}{\tilde{P}} = \left(\frac{\hat{\tau}}{\hat{\tau}^{FB}}.\right)^{\frac{1}{\eta}} \tag{A.27}$$

■

A.4 Taxes and subsidies

Proof of Proposition 7.

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

approach and write the planner's problem as follows:

$$\begin{aligned}
\max_{\{l_j, q_{1j}, q_{\tau j}\}_{j=0}^{1}} \quad & \int_0^1 \left[\pi\tau u\left(q_{\tau j}\right) + (1-\pi)u\left(q_{1j}\right)\right] dj, && \text{(A.28)}\\
\text{s.t.} \quad & u'\left(q_{1j}\right) = \frac{1-\pi}{1-\tau\pi}\,\tau\, u'\left(q_{\tau j}\right), && \text{for all } j\\
& \pi q_{\tau j} + (1-\pi) q_{1j} = \frac{l_j}{c_j}, && \text{for all } j\\
& \int l_j dj = 1.
\end{aligned}$$

Taking first order conditions, we obtain

$$[q_{\tau j}]: \quad \pi\tau u'\left(q_{\tau j}\right) + \frac{1-\pi}{1-\tau\pi}\,\tau\, u''\left(q_{\tau j}\right)\mu_j = \pi\theta_j, \tag{A.29}$$

$$[q_{1j}]: \quad (1-\pi)u'\left(q_{1j}\right) - u''\left(q_{1j}\right)\mu_j = (1-\pi)\theta_j, \tag{A.30}$$

$$[l_j]: \quad \frac{\theta_j}{c_j} = \lambda. \tag{A.31}$$

where μ_j, θ_j, and λ are the Lagrange multipliers on the three constraints, respectively. Multiplying equation (A.29) by $\frac{1-\tau\pi}{\tau(1-\pi)}\frac{u''(q_{1j})}{u''(q_{\tau j}}$ and adding to equation (A.30), we obtain:

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

Rearranging we have

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

where

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

Note that γ_j represents the share of additional production allocated to high-taste consumers when l_j increases. The third optimality condition (A.31) iwmplies that

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

Equation (A.33) 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.33). First, using equations (2.14–2.15), the LHS of equation (A.33) becomes

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

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 (2.12–2.13). 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.33) 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.33) 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.

■

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:

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

Taking first order conditions, we obtain

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

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:

$$\theta_j = \gamma_j\tau u'(q_{\tau j}) + (1-\gamma_j)\,u'(q_{1j}), \tag{A.34}$$

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

$$\lambda = \gamma_j \frac{1}{c_j} \tau u'(q_{\tau j}) + (1-\gamma_j) \frac{1}{c_j} u'(q_{1j}), \tag{A.35}$$

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.35), we see that the optimal level of taxes are given by

$$(1+t_j) = \frac{\lambda P}{\gamma_j + (1-\gamma_j)\frac{1-\pi}{1-\tau\pi}}. \tag{A.36}$$

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

$$(1+t_j) = \frac{1}{\gamma_j + (1-\gamma_j)\frac{1-\pi}{1-\tau\pi}}. \tag{A.37}$$

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.37) 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.37) also implies that $1 + t_j < 1$. Taken together, we have that $t_j = t < 0 \, \forall j$.

■

A.5 Additional propositions and proofs

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 2, 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 $\tilde{p}_j$,

$$\frac{1}{P} = \frac{\tau u'(q_{\tau j})}{\tilde{p}_j}, \tag{A.38}$$

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 $q_{\tau j}$:

$$\tilde{p}_j = c_j. \tag{A.39}$$

Equation (A.39) 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 **9.** *Suppose preferences satisfy Assumption 1 and the elasticity of demand is decreasing in the quantity consumed. Then, firms with higher productivity (i.e., low production costs) charge higher markups at the firm-level* $\left(\mu_j \equiv \frac{\pi (p_{1j} q_{1j}) + (1-\pi) (p_{\tau j} q_{\tau j})}{c_j (\pi q_{1j} + (1-\pi) q_{\tau j})}\right)$.

PROOF OF PROPOSITION **9.** Using Lemma 1,

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

Since η is positive, β_0 must be positive for the demand elasticity to decline with quantity. Using the firm's optimality conditions to substitute out prices, the inverse of the markup is given by

$$\frac{1}{\mu_j} = \frac{(1-\pi\tau) u(q_{j1}) / \psi(q_t)}{(1-\pi\tau) u(q_{j1}) + \pi\tau u(q_{j\tau})} + \frac{\pi\tau u(q_{j\tau}) / \psi(q_{j\tau})}{(1-\pi\tau) u(q_{j1}) + \pi\tau u(q_{j\tau})} \tag{A.40}$$

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

$$\mu_j = \frac{\eta}{\eta-1}\beta_1^{\frac{1}{\eta}}\frac{\pi\tau\left(q_{\tau j}+\beta_0\right)^{\frac{\eta-1}{\eta}}+(1-\pi\tau)\left(q_{1j}+\beta_0\right)^{\frac{\eta-1}{\eta}}-\left(\beta_0\right)^{\frac{\eta-1}{\eta}}}{\pi\tau q_{\tau j}\left(q_{\tau j}+\beta_0\right)^{-\frac{1}{\eta}}+(1-\pi\tau)q_{1j}\left(q_{1j}+\beta_0\right)^{-\frac{1}{\eta}}}. \tag{A.41}$$

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

$$\begin{aligned}\mu_j &= \frac{\eta}{\eta-1}\frac{\pi\tau\beta_1^{\frac{\eta-1}{\eta}}x_\tau^{1-\eta}+(1-\pi\tau)\beta_1^{\frac{\eta-1}{\eta}}x_1^{1-\eta}-\left(\beta_0\right)^{\frac{\eta-1}{\eta}}}{\pi\tau\left(-\beta_0+\beta_1x_\tau^{-\eta}\right)\beta_1^{-\frac{1}{\eta}}x_\tau+(1-\pi\tau)\left(-\beta_0+\beta_1x_1^{-\eta}\right)\beta_1^{-\frac{1}{\eta}}x_1} \\ &= \frac{\eta}{\eta-1}\frac{\beta_1^{\frac{\eta-1}{\eta}}\left(c_j/P\right)^{1-\eta}\widetilde{\tau}-\beta_0^{\frac{\eta-1}{\eta}}}{\beta_1^{\frac{\eta-1}{\eta}}\left(c_j/P\right)^{1-\eta}\widetilde{\tau}-\left(c_j/P\right)\beta_0\beta_1^{-\frac{1}{\eta}}},\end{aligned} \tag{A.42}$$

where

$$\widetilde{\tau} \equiv (\pi\tau)^\eta\,\pi^{1-\eta}+(1-\tau\pi)^\eta(1-\pi)^{1-\eta}. \tag{A.43}$$

Rewrite this as

$$\mu_j = \frac{\eta}{\eta-1}\frac{(c_j/P)^{1-\eta}\alpha-\gamma}{(c_j/P)^{1-\eta}\alpha-(c_j/P)\delta}, \tag{A.44}$$

where

$$\alpha = \beta_1^{\frac{\eta-1}{\eta}}\widetilde{\tau} \qquad > 0, \tag{A.45}$$
$$\gamma = \beta_0^{\frac{\eta-1}{\eta}} \qquad > 0, \tag{A.46}$$
$$\delta = -\beta_0\beta_1^{-\frac{1}{\eta}} \qquad > 0. \tag{A.47}$$

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

$$\text{sign}\left(\frac{\partial\mu_j}{\partial(c_j/P)}\right) = -\text{sign}\underbrace{\left(\eta\alpha\delta\left(\frac{c_j}{P}\right)^{1-\eta}+\alpha\gamma(1-\eta)\left(\frac{c_j}{P}\right)^{-\eta}-\gamma\delta\right)}_{\equiv Z(c_j)}. \tag{A.48}$$

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.

$$\underset{c_j}{\text{argmin}} \quad Z(c_j) = \frac{\gamma}{\delta}. \tag{A.49}$$

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

$$\alpha\delta^{\eta-1} > \gamma^{\eta}. \tag{A.50}$$

Which simplifies to

$$\widetilde{\tau} \geq 1. \tag{A.51}$$

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.6 Identification

Proposition **10** (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 **10.** 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

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

$$q_{\tau j} = -\beta_0 + \widetilde{\beta_1} c_j^{-\eta}\left(\frac{1}{\tau}\right)^{\eta} \tag{A.53}$$

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

$$\psi(q) = \frac{u(q)}{qu'(q)} = \frac{\eta}{\eta - 1}\left[\left(1 + \frac{\beta_0}{q}\right) - \beta_0^{\frac{\eta-1}{\eta}}\frac{(q+\beta_0)^{\frac{1}{\eta}}}{q}\right]. \tag{A.54}$$

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 (2.14) and (2.15).

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 ((2.8)) as

$$q_{1j}^{FB} = -\beta_0 + \widetilde{\beta_1}^{FB} c_j^{-\eta} \tag{A.55}$$

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

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.

■

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

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

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:

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

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

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

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

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

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

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

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

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

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

Proposition **11.** *If preferences exhibit variable elasticity of substitution, there is misallocation across firms in the linear pricing equilibrium. In particular, if the elasticity of demand is decreasing in the quantity consumed:*

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.*

Proof of Proposition **11.**

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 (2.8), 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

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

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

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

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 (2.8) 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 **12** (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 **12.**

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.

■

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 4. 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, $h(\tau) \equiv \frac{g(\tau)}{1-G(\tau)}$.[54]

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).[55] 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

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

[55] 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 (2.6). 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 **13.** *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 **13.** 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 **14.** *Suppose preferences satisfy Assumption 1. Then, the equilibrium levels of firm-level production and employment are identical to the efficient allocation.*

Proof of Proposition **14.**

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 7, Assumption 1 implies that $\widetilde{P_j}$ is independent of firm cost hence total production is equal to first-best for all firms.

■

Proposition **15.** *Suppose preferences satisfy Assumption 1. Then, the optimal firm-level subsidies and taxes are zero.*

Proof of Proposition **15.**

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 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.[56]

[56] We maintain the assumption that firms choose to serve all customers. Note that in theory, a firm may choose its bundles in a way that excludes some low-taste consumers. We confirm in our quantitative analysis that no firm do not

A.3.1 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 **16.** *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.3.2 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

have incentive to exclude any customer.

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

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

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

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

A.4 Two bundles: Proofs

Proof of Proposition **16.**

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 $\hat{\tau}_j$

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

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

■

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

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

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 **18.**

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 4. 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

Table A.1: Calibrated moments and parameters when firms offer 2 bundles

Parameter		Model		Moment		Model	
		Benchm.	2 sizes		Data	Benchm.	2 sizes
$\bar{\tau}$	Highest taste	1.19	1.22	Package size dispersion	0.44	0.44	0.44
θ	Pareto shape	3.14	3.12	Sales share top 5%	0.73	0.73	0.73
η	Elasticity of differences	4.15	4.11	Aggregate markup	1.3	1.3	1.3
β_0	Departure from CES	0.02	0.02	Markup elast. w.r.t. size	0.03	0.03	0.03

of any firm to the first-best.

■

A.5 Two bundles: Quantification

The model moments and calibrated parameters are displayed in Table A.1.

The misallocation costs are summarized in Table A.2. The welfare costs from misallocation with two package sizes are similar to our benchmark results: 0.71% relative to 0.68%.

Table A.2: Welfare Costs of Misallocation

Baseline Model	2 sizes
0.68%	0.71%

Notes: This table reports the welfare costs in the decentralized equilibrium relative to the efficient allocation for our benchmark model (column 1), and the model in which firms are restricted to choosing 2 sizes only. All welfare costs are measured in consumption equivalent terms—that is, the uniform decline in consumption that would make households indifferent between the two equilibria.

B Kimball preferences

In this section, we describe the model and its equilibrium with Kimball preferences. We map the Kimball aggregator to the canonical second-degree price discrimination problem with continuum of types.

B.1 Canonical second-degree price discrimination results

There is a continuum of types $\tau \in [1, \bar{\tau}]$ which are drawn according to a pdf $g(\tau)$. We solve for the problem of a firm facing a constant marginal cost which is equal to c. Let the utility of consumer with type τ be given by

$$U(\tau) = b(q, \tau) - T,$$

where q is the amount consumed and T is the amount paid to the firm. We assume that the cross derivative is positive, $b_{q\tau}(\cdot) > 0$.

Following the standard derivations, we obtain the following:[57]

$$\frac{\partial b(q,\tau)}{\partial q} = c + h(\tau)^{-1} \frac{\partial^2 b(q,\tau)}{\partial \tau \partial q} \tag{B.1}$$

where $h(\tau) \equiv \frac{g(\tau)}{1-G(\tau)}$. And

$$T(\tau) = b(q(\tau), \tau) - \int_1^\tau \frac{\partial b(q(t), t)}{\partial t} dt \tag{B.2}$$

B.2 Mapping Kimball preferences to the canonical problem

With Kimball preferences, the aggregate consumption is implicitly defined as follows

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

Instead of j notation, let's work directly with productivity of the firm (c). Then, the equation becomes

$$\int \int \tau \Upsilon \left(\frac{q(c,\tau)}{Q_i} \right) dG(\tau) dF(c) = 1 \tag{B.4}$$

This shows that Q_i doesn't vary by i, where we used the fact that taste draws are iid there is a continuum of firms. So we will use Q from here on:

$$\int \int \tau \Upsilon \left(\frac{q(c,\tau)}{Q} \right) dG(\tau) dF(c) = 1 \tag{B.5}$$

We want to map this into the general formulation of section B.1. We shall use the following Lemma.

[57] For detailed derivations, see, for example, https://faculty.haas.berkeley.edu/hermalin/continuous_2nd_degree.pdf.

LEMMA **3.** *The utility from consuming q instead of 0 for a consumer with taste τ is*

$$\tau Q D\left[\Upsilon\left(\frac{q}{Q}\right)-\Upsilon\left(0\right)\right]. \tag{B.6}$$

where

$$D=\frac{1}{\int\int \tau\Upsilon'\left(\frac{q(c,\tau)}{Q}\right)\frac{q(c,\tau)}{Q}dG(\tau)dF(c)}. \tag{B.7}$$

Proof. Aggregate consumption is defined by

$$\int \tau_{ij}\Upsilon\left(\frac{q_{ij}}{Q_i}\right)dj=1, \tag{B.8}$$

We start by deriving an expression for $\frac{\partial Q}{\partial q_{ij}}$.Totally differentiating, we get

$$\int \tau_{iz}\Upsilon'\left(\frac{q_{iz}}{Q}\right)q_{iz}dz\frac{1}{Q^2}dQ=\frac{1}{Q}\Upsilon'\left(\frac{q_{ij}}{Q}\right)\tau_{ij}dq_i, \tag{B.9}$$

which yields

$$\frac{\partial Q}{\partial q_{ij}}=Q\frac{\tau_{ij}\Upsilon'\left(\frac{q_{ij}}{Q}\right)}{\int \tau_{iz}\Upsilon'\left(\frac{q_{iz}}{Q}\right)q_{iz}dz} \tag{B.10}$$

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

$$\int_{\underline{q}}^{\bar{q}}\frac{\partial Q}{\partial q_{ij}}dq_{ij}. \tag{B.11}$$

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

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

where

$$D=\frac{Q}{\int \tau_{iz}\Upsilon'\left(\frac{q_{iz}}{Q}\right)q_{iz}dz}. \tag{B.12}$$

And for $\underline{q} = 0$, we obtain the desired expression. ■

We can then set $U(\tau)$ as follows,

$$U(\tau) = \tau QDP\left[\Upsilon\left(\frac{q}{Q}\right) - \Upsilon(0)\right] - T,$$

so that

$$b(q,\tau) = \tau QDP\left[\Upsilon\left(\frac{q}{Q}\right) - \Upsilon(0)\right].$$

This implies that

$$\frac{\partial b(q,\tau)}{\partial q} = \tau DP\Upsilon'\left(\frac{q}{Q}\right),$$
$$\frac{\partial^2 b(q,\tau)}{\partial \tau \partial q} = DP\Upsilon'\left(\frac{q}{Q}\right)$$

The optimality condition (B.1) becomes

$$\tau D\Upsilon'\left(\frac{q}{Q}\right) = \frac{c}{P} + h(\tau)^{-1} D\Upsilon'\left(\frac{q}{Q}\right). \tag{B.13}$$

so that

$$\tau D\Upsilon'\left(\frac{q}{Q}\right) = \frac{c}{P}\frac{\tau}{\tau - h(\tau)^{-1}} \tag{B.14}$$

and

$$\frac{T(\tau)}{P} = \tau QD\left[\Upsilon\left(\frac{q(\tau)}{Q}\right) - \Upsilon(0)\right] - QD\int_1^\tau \left[\Upsilon\left(\frac{q(t)}{Q}\right) - \Upsilon(0)\right] dt \tag{B.15}$$

B.3 Klenow-Willis specification

We use the Klenow-Willis formulation. That is,

$$\Upsilon(q) = 1 + (\sigma - 1)\exp\left(\frac{1}{\varepsilon}\right)\varepsilon^{\sigma/\varepsilon - 1}\left(\Gamma\left(\frac{\sigma}{\varepsilon}, \frac{1}{\varepsilon}\right) - \Gamma\left(\frac{\sigma}{\varepsilon}, \frac{q^{\varepsilon/\sigma}}{\varepsilon}\right)\right), \tag{B.16}$$

$$\Upsilon'(q) = \frac{\sigma - 1}{\sigma} e^{\frac{1 - q^{\epsilon/\sigma}}{\epsilon}}, \tag{B.17}$$

$$\Upsilon''(q) = -\frac{\sigma - 1}{\sigma} e^{\frac{1 - q^{\epsilon/\sigma}}{\epsilon}}\left(\frac{1}{\sigma} q^{\epsilon/\sigma - 1}\right), \tag{B.18}$$

We can invert $\Upsilon'(q)$ to

$$\left(\Upsilon'\right)^{-1}(x) = \left(1 - \epsilon \ln\left(\frac{\sigma}{\sigma - 1} x\right)\right)^{\frac{\sigma}{\epsilon}} \tag{B.19}$$

so that

$$\left(\frac{q}{Q}\right)^{\frac{\epsilon}{\sigma}} = 1 - \epsilon \ln\left(\frac{\sigma}{\sigma - 1}\right) - \epsilon \ln c + \epsilon \ln\left(\tau - h(\tau)^{-1}\right) + \epsilon \ln(PD) \tag{B.20}$$

Plugging in the Klenow-Willis formulation to the definition of Q equation, we get

$$\int\int \tau\Gamma\left(\frac{\sigma}{\epsilon}, \frac{\left(\frac{q(c,\tau)}{Q}\right)^{\epsilon/\sigma}}{\epsilon}\right) dG(\tau)dF(c) = \int\int \tau\Gamma\left(\frac{\sigma}{\epsilon}, \frac{1}{\epsilon}\right) dG(\tau)dF(c) \tag{B.21}$$

B.3.1 Second-best allocation

Consider a social planner that can set firm-level taxes and subsidies. Let's write the planner's problem of minimizing labor subject to providing a unit of aggregate consumption.

$$\min_{\{q(c,\tau),s(c)\}} \int\int cq(c,\tau)dG(\tau)dF(c) \tag{B.22}$$

$$\text{s.t.} \qquad \tau\Upsilon'\left(q(c,\tau)\right) = \frac{\tau}{\tau - h(\tau)^{-1}}c(1-s(c)), \qquad [\gamma(c,\tau)] \tag{B.23}$$

$$\int\int \tau\Upsilon\left(q(c,\tau)\right) dG(\tau)dF(c) = 1, \qquad [\mu] \tag{B.24}$$

Take FOC:

$$[s(c)]: \qquad \int \frac{\tau}{\tau - h(\tau)^{-1}}\gamma(c,\tau)dG(\tau) = 0, \tag{B.25}$$

$$[q(c,\tau)]: \qquad c = \tau\Upsilon''(q(c,\tau))\gamma(c,\tau) + \tau\Upsilon'(q(c,\tau))\mu \tag{B.26}$$

Numerical algorithm to compute the second-best allocation. Below we outline a sketch of the algorithm we use to compute the second-best allocation:

1. Guess Lagrange multiplier μ.
2. For each cost level c:
   (a) Guess the subsidy of firms with such cost level, $s(c)$
   (b) Use (B.23) to to obtain $q(c,\tau)$
   (c) Use (B.26) to obtain $\gamma(c,\tau)$
   (d) Check if (B.25) holds
   (e) Iterate until finding $s(c)$
3. Use the constraint (B.24) as a residual equation.
4. Iterate until correct μ is found and residual equation is approximately zero.

B.4 Quantitative analysis

The model moments and calibrated parameters are displayed in Table B.1.

The misallocation costs are summarized in Table B.2. The welfare costs from misallocation with Kimball preferences are similar to our benchmark results: 0.79% relative to 0.68%.

Table B.1: Calibrated Moments and Parameters under Kimball

Parameter		Model		Moment		Model	
		Benchm.	Kimball		Data	Benchm.	Kimball
$\bar{\tau}$	Highest taste	1.19	1.21	Package size dispersion	0.44	0.44	0.44
θ	Pareto shape	3.14	3.18	Sales share top 5%	0.73	0.73	0.63
η	Elasticity of differences	4.15	–	Aggregate markup	1.3	1.3	1.32
β_0	Departure from CES	0.02	–	Markup elast. w.r.t. size	0.03	0.03	0.03
σ	Degree of demand elasticity	–	5.38				
ϵ	Degree of superelasticity	–	0.87				

Table B.2: Welfare Costs of Misallocation

Baseline Model	Kimball Specification
0.68%	0.79%

Notes: This table reports the welfare costs in the decentralized equilibrium relative to the efficient allocation for our benchmark model (column 1), and the model with Kimball preferences (column 2). All welfare costs are measured in consumption equivalent terms—that is, the uniform decline in consumption that would make households indifferent between the two equilibria.

Proposition 7 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 0.02% in consumption equivalent units. That is, it eliminates a very small portion of the overall welfare costs due to misallocation (0.79%).

C Imperfect substitution within firms

In the benchmark model, firms offer a quantity to be sold to low-taste consumers and a quantity to high-taste consumers. Recall that the utility of household i from purchasing quantity q_{ij} is given by

$$U_{ij} = \tau_{ij} u(q_{ij}),$$

where τ_{ij} is the taste of household i for the product of firm j. Because q_{ij} specifies the *total* amount consumed by household i, purchasing multiple small-sized packages is perfectly substitutable with consuming a single large-sized package.

As we explain in detail below, the *perfect substitutes assumption* does not drive our two main results: (i) nonlinear pricing can lead to misallocation across consumers within a firm, (ii) firm-level markups may not be informative of the degree of misallocation across firms. The key assumption in our setup that underlies the difference relative to standard models of linear pricing is that goods are *indivisible*, that is, firms can mandate a minimum quantity sold (package size) together with price. Consumers cannot purchase fractions of large units.

Expanding our environment to incorporate imperfectly substitutable goods is not straightforward. Typical definitions of imperfect substitutability apply to pre-specified goods (e.g., apples vs. oranges or Coca-Cola vs. Pepsi), rather than different quantities of the same good, where the quantities are chosen by firms.

In this section, we consider three alternative ways of incorporating imperfect substitution across the same product sold in a variety of sizes. The first preserves the choice of firms and the social planner of what a "large" and "small" package is, but allows consumers to purchase multiple bundles of the same firm. In this specification, we assume consumers' utility is not only a function of the total quantity consumed, but also depends on how the quantities are purchased, i.e., in which package sizes. The second and third models consider environments where units are chosen by nature, allowing us to connect to more standard models of imperfect substitutability. In the second model, we maintain the discrete choice nature of the problem, but add an idiosyncratic taste towards every specific sized bundle. In the third specification, we allow consumers to purchase multiple bundles and assume there is a constant elasticity of substitution across the different sized bundles. In all three, we show that our two main results continue to hold.

C.1 Disutility of package size

Suppose that consumer utility not only depends on the total quantity of the good purchased, but also on *how* it is purchased, for example, the way in which goods are packaged. Let q_s, $s \in \{1, ..., S\}$ be the physical quantity of the good contained in each available size s, and let n_s be the number of packages of each size purchased.

Utility now has three components.

$$\tau u\left(\sum_s n_s \times q_s\right) - \sum_s n_s \times v(q_s) - f\left(\sum_s n_s\right) \tag{C.1}$$

The first component is the same as in the main text and represents the utility flow from physical quantity; say, fluid ounces of Coca Cola. The second component, $v(q_s)$, is an increasing function of package size that captures any disutility from the *way in which* the fluid ounces are purchased, e.g., disutility of carrying a large container, or more rapid depreciation once the package has been opened.[58] The third component captures a fixed cost of purchasing or opening a package.[59]

Taking into account the disutility associated with buying large vs small units, the two are no longer perfect substitutes, as long as $2v(q_l) - f(2) \neq v(2q_l) - f(1)$.

Social planner's problem. Given an aggregate price index P^{FB}, the planner chooses a set of sizes for each consumer type τ to maximize (C.1). For simplicity, normalize $f(1) = 0$ and assume that $f(2)$ is large enough such that the planner chooses to deliver the first-best quantities to each type in a single package.

The first-best allocations then solve

$$\tau_{ij} u'(q_{ij}^{FB}) - v'(q_{ij}^{FB}) = \frac{c_j}{P^{FB}} \tag{C.2}$$

Market allocation. The firm's problem has the same objective as in the main text, but the IR and the IC constraints are now different:

$$\begin{aligned} \max_{\{q_{1j}, q_{\tau j}, p_{1j}, p_{\tau j}\}} \quad & \pi q_{\tau j} \left(p_{\tau j} - c_j\right) + (1-\pi) q_{1j} \left(p_{1j} - c_j\right) && (C.3) \\ \text{s.t} \quad & u(q_{1j}) - v(q_{1j}) - \frac{p_{1j} q_{1j}}{P} = 0 && [IR_1] \\ & \tau u(q_{\tau j}) - v(q_{\tau j}) - \frac{p_{\tau j} q_{\tau j}}{P} = \tau u(q_{1j}) - v(q_{1j}) - \frac{p_{1j} q_{1j}}{P} && [IC_\tau] \end{aligned}$$

Note that we omitted the possibility that the consumer may want to purchase two small bundles; i.e. $f(2)$ is large enough.[60]

The quantities sold the high and low type respectively solve

$$\tau u'(q_{\tau j}) - v'(q_{\tau j}) = \frac{c_j}{P} \tag{C.4}$$

$$\frac{1 - \tau\pi}{1 - \pi} u'(q_{1j}) - v'(q_{1j}) = \frac{c_j}{P} \tag{C.5}$$

Similarly to the baseline model, any distortion at the top only comes from the difference in price indices between the planner and the market. The allocation to the low type features the same wedge as before, but the wedge only affect the first part of utility, since we assumes that the preference shifter does not affect utility of package size.[61].

[58] Here, the taste shifter only applies to the first component of utility. If we assumed that it also applies to $v(.)$, the problem becomes identical to the one we studied in the main text, with a utility function given by $u(q) - v(q)$.

[59] This is necessary to ensure that the planning problem has an interior solution for the set of q_s. Otherwise, it may be optimal to sell infinitely many packages of infinitesimal size.

[60] Allowing for a lower value of $f(2)$ such that consumers may want to purchase two small bundles would not change the tow main propositions below, but would considerably complicate notation.

[61] If the taste shifter also applied to $v(.)$, the problem would fully collapse to the original formulation

Misallocation. The following propositions show that the two main results of our baseline model with perfectly substitutable sizes continue to hold. The first mirrors Proposition 3. In general equilibrium, the allocation sold to both the high taste and the low taste consumer are distorted. The second mirrors Propositions 7 and 9: firms may charge size-dependent markups despite the fact that the allocation of production across firms is efficient.

PROPOSITION **19.** *Given a level of labor l_j, 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.*

PROOF OF PROPOSITION **19.** Given that l_j is assumed to be constant, there are three possibilities: (1) $q_{1j} = q_{1j}^{FB}$ and $q_{\tau j} = q_{\tau j}^{FB}$, (2) $q_{1j} > q_{1j}^{FB}$ and $q_{\tau j} < q_{\tau j}^{FB}$, or (3) $q_{1j} < q_{1j}^{FB}$ and $q_{\tau j} > q_{\tau j}^{FB}$.

Suppose $q_{1j} = q_{1j}^{FB}$ and $q_{\tau j} = q_{\tau j}^{FB}$. From $q_{\tau j} = q_{\tau j}^{FB}$, it then follows that $P = P^{FB}$. Using (C.2) and (C.5), $q_{1j} = q_{1j}^{FB}$ would imply that $\frac{1-\tau\pi}{1-\pi} = 1$.

Suppose that $q_{1j} > q_{1j}^{FB}$ and $q_{\tau j} < q_{\tau j}^{FB}$. From $q_{\tau j} = q_{\tau j}^{FB}$, it then follows that $P < P^{FB}$. Using (C.2) and (C.5), we would have $q_{1j} < q_{1j}^{FB}$, contradicting this case as well. ■

The intuition behind Proposition 19 is similar to the baseline model. The additional utility cost of purchasing in larger packages applies to both types of consumers, in the planner's problem as well as in the market allocation. This cost in and of itself does not create any additional distortion. As in the baseline model, firms distort the quantity sold to low-taste consumers downward in order to be able to extract more rents from the top. For the labor market to clear, the aggregate price index must therefore increase, introducing a distortion also at the top.

In general, markup variation across firms as well as any patterns of misallocation of production depend on the shapes of $u'()$ and $v'()$. However, we can still show that our main result of no misallocation with CED preferences holds; as long as we extend the definition of CED to also include the disutility of package size.

PROPOSITION **20.** *Markup variation across firms is not necessarily a sign of misallocation across firms. That is, there exist utility functions $u(.)$ and disutility functions $v(.)$ such that there is markup variation across firms, but no misallocation.*

PROOF OF PROPOSITION **20.** Suppose that the disutility of package size v is proportional to u, that is, $v(q_{ij}) = \nu u(q_{ij})$ for some $\nu < 1$. Intuitively, this would mean that one looses $\nu\%$ of utility from, say, Coke getting flat after opening a large bottle. Then, if $\tau u(q)$ features CED, so does $\tau u(q) - v(q) = (\tau - \nu)v(q)$ and there is no misallocation of labor across firms in equilibrium. ■

Intuitively, when the entire utility of purchasing and consuming a certain package size features CED, the disutility of buying a big bottle of Coca Cola acts as a proportional downward shift that affects utility of all consumers proportionally. The model with imperfect substitutes then is analogous to a model in which different sizes are perfect substitutes, but there is *less* taste dispersion across households: low-taste households have effective taste shifters of $(1-\nu)$, while the taste shifter is $\tau - \nu$ at the top.

C.2 Continuum of tastes and fixed sizes

Now consider an environment where Households consume a variety of product types, $j \in (0,1)$, but where each product type is offered in two *exogenous* sizes, q_{jl} and q_{jh}. The preferences of household i are given by

$$U_i = \int_0^1 \left[\mathbb{I}(s_{ij} = l)\left(\tau_{ij} u(q_{jl})\right) + \mathbb{I}(s_{ij} = h)\left(\tau_{ij} u(q_{jh})\right)\right] dj, \tag{C.6}$$

where $s_j \in \{n, l, h\}$ indicates whether the consumer chooses to purchase no good (n), the low-quantity bundle (l) or the high-quantity bundle (h). τ_{ij} is the idiosyncratic taste of household i for good j and is uniformly distributed between $[1, \tau]$ with $\tau < 2$.

Firm's problem. Let $\underline{\tau}$ be the threshold levels below which consumers do not purchase either bundle and $\bar{\tau}$ be the threshold above which consumers buy the high bundle. The firm's problem (multiplied by $(\tau - 1)$) is given by

$$\begin{aligned} \max_{\{p_l, p_h, \underline{\tau}, \bar{\tau}\}} \quad & (p_l - c_j) q_l (\bar{\tau} - \underline{\tau}) + (p_h - c_j) q_h (\tau - \bar{\tau}) \\ \text{s.t.} \quad & \underline{\tau} u(q_l) = \frac{p_l q_l}{P}, \\ & \bar{\tau} u(q_h) - \frac{p_h q_h}{P} = \bar{\tau} u(q_l) - \frac{p_l q_l}{P}. \end{aligned}$$

Taking FOCs,

$$[\underline{\tau}]: \quad (p_l - c_j) q_l = \lambda_1 u(q_l) \tag{C.7}$$

$$[\bar{\tau}]: \quad (p_h - c_j) q_h - (p_l - c_j) q_l = \lambda_2 \left(u(q_h) - u(q_l)\right) \tag{C.8}$$

$$[p_l]: \quad q_l(\bar{\tau} - \underline{\tau}) = \lambda_1 \frac{q_l}{P} - \lambda_2 \frac{q_l}{P} \tag{C.9}$$

$$[p_h]: \quad q_h(\tau - \bar{\tau}) = \lambda_2 \frac{q_h}{P} \tag{C.10}$$

From the FOC wrt p_h and p_l we get

$$\lambda_2 = P(\tau - \bar{\tau}) \tag{C.11}$$

$$\lambda_1 = P(\tau - \underline{\tau}) \tag{C.12}$$

Combining the FOC w.r.t. $\underline{\tau}$ with the expression for λ_1, we get

$$\frac{(p_l - c_j) q_l}{P} = (\tau - \underline{\tau})\, u(q_l).$$

Using the IR constraint,

$$\underline{\tau} u(q_l) - \frac{c_j q_l}{P} = (\tau - \underline{\tau})\, u(q_l)$$

so that

$$\underline{\tau} = \frac{\tau}{2} + \frac{1}{2} \frac{c_j}{P} \frac{q_l}{u(q_l)}. \tag{C.13}$$

Now using the FOC w.r.t. $\bar{\tau}$, combining with the expression for λ_2, we get

$$\frac{p_h q_h - p_l q_l}{P} - (q_h - q_l)\frac{c_j}{P} = (\tau - \bar{\tau})\left(u(q_h) - u(q_l)\right),$$

Using the IC constraint, we obtain

$$\bar{\tau}(u(q_h) - u(q_l)) - (q_h - q_l)\frac{c_j}{P} = (\tau - \bar{\tau})\left(u(q_h) - u(q_l)\right),$$

So that

$$\bar{\tau} = \frac{\tau}{2} + \frac{1}{2}\frac{q_h - q_l}{u(q_h) - u(q_l)}\frac{c_j}{P} \tag{C.14}$$

We assume that $q_h, q_l, u(q_h), u(q_l)$ are such that $1 < \underline{\tau}$ and $\bar{\tau} < \tau$ for all firms. We can then obtain the prices from the IR and IC constraints:

$$p_l = \underline{\tau}\frac{u(q_l)}{q_l}P,$$
$$p_h = \bar{\tau}\frac{u(q_h) - u(q_l)}{q_h}P + p_l\frac{q_l}{q_h}.$$

Misallocation. In this setup, the source of misallocation across consumers is governed by the two thresholds $\underline{\tau}$ and $\bar{\tau}$. Holding constant the overall production of a firm, a social planner may choose to give low taste consumers who fall below $\underline{\tau}$ the small packaged product at the expense of higher taste consumers who will obtain the small instead of the large packaged good. The next proposition formalizes this form of misallocation.

PROPOSITION **21.** *If a social planner chooses how to allocate the production of a firm to different consumers, they would choose a lower $\underline{\tau}$ and a higher $\bar{\tau}$. That is, they would serve more low-taste customers at the expense of downsizing the bundle of some high-taste consumers.*

Proof. The planner's problem is

$$\begin{aligned} \max_{\{\underline{\tau},\bar{\tau}\}} \quad & \frac{1}{2}(\bar{\tau}^2 - \underline{\tau}^2)u(q_l) + \frac{1}{2}(\tau^2 - \bar{\tau}^2)u(q_h) \\ \text{s.t.} \quad & (\bar{\tau} - \underline{\tau})q_l + (\tau - \bar{\tau})q_h = \frac{L_j}{c_j}. \end{aligned}$$

Taking FOC:

$$\begin{aligned} [\underline{\tau}]: \quad & \underline{\tau}u(q_l) = q_l\lambda, \\ [\bar{\tau}]: \quad & \bar{\tau}\left(u(q_h) - u(q_l)\right) = (q_h - q_l)\lambda, \end{aligned}$$

So that

$$\frac{\bar{\tau}^*}{\underline{\tau}^*} = \frac{q_h - q_l}{q_l}\frac{u(q_l)}{u(q_h) - u(q_l)} \tag{C.15}$$

While the decentralized one has

$$\frac{\bar{\tau}-\frac{\tau}{2}}{\underline{\tau}-\frac{\tau}{2}}=\frac{q_h-q_l}{q_l}\frac{u(q_l)}{u(q_h)-u(q_l)} \tag{C.16}$$

That is, for the same level of $\underline{\tau}$, the planner would choose $\bar{\tau}^*>\bar{\tau}$. Since that wouldn't exhaust the resources, it must be that $\underline{\tau}^*<\underline{\tau}$. Similarly, for the same level of $\bar{\tau}$, the planner would choose a $\underline{\tau}^*<\underline{\tau}$. Since that allocation is not feasible with the firm's employment, it must be that $\bar{\tau}^*>\bar{\tau}$. ■

Now suppose the economy is populated with two types of firms—a low productivity (c_1) and a high productivity one ($c_2<c_1$). The small and large packages sizes are also firm specific. Consider a planner that can impose taxes and subsidies at the firm level. The planner's problem is

$$\begin{aligned}
\max_{\{\underline{\tau}_1,\bar{\tau}_1,\underline{\tau}_2,\bar{\tau}_2\}} \quad & \frac{1}{2}(\bar{\tau}_1^2-\underline{\tau}_1^2)u(q_{l1})+\frac{1}{2}(\tau^2-\bar{\tau}_1^2)u(q_{h1})+\frac{1}{2}(\bar{\tau}_2^2-\underline{\tau}_2^2)u(q_{l2})+\frac{1}{2}(\tau^2-\bar{\tau}_2^2)u(q_{h2}) \\
\text{s.t.} \quad & \left[(\bar{\tau}_1-\underline{\tau}_1)q_{l1}+(\tau-\bar{\tau}_1)q_{h1}\right]c_1+\left[(\bar{\tau}_2-\underline{\tau}_2)q_{l2}+(\tau-\bar{\tau}_2)q_{h2}\right]c_2=L, \\
& \bar{\tau}_1=-\left[\frac{q_{h1}u(q_{l1})-q_{l1}u(q_{h1})}{q_{l1}\left(u(q_{h1})-u(q_{l1})\right)}\right]\frac{\tau}{2}+\frac{u(q_{l1})}{q_{l1}}\frac{q_{h1}-q_{l1}}{u(q_{h1})-u(q_{l1})}\underline{\tau}_1 \\
& \bar{\tau}_2=-\left[\frac{q_{h2}u(q_{l2})-q_{l2}u(q_{h2})}{q_{l2}\left(u(q_{h2})-u(q_{l2})\right)}\right]\frac{\tau}{2}+\frac{u(q_{l2})}{q_{l2}}\frac{q_{h2}-q_{l2}}{u(q_{h2})-u(q_{l2})}\underline{\tau}_2
\end{aligned}$$

We now show that the firm-level markup is not indicative of whether such firm is too small or too big. That is, the relative markup of a firm cannot be used to determined if such firm should be subsidized or taxed.

PROPOSITION **22.** *Markup variation across firms is not a sufficient statistic for misallocation across firms. That is, there exist parameters such that the planner would optimally subsidize some firms at the expense of others, even though such firms charge a lower markup. And, similarly, there exist parameters such that the planner would optimally subsidize some firms at the expense of others, when such firms charge a higher markup.*

Proof.

Taking first order conditions

$$\begin{aligned}
[\underline{\tau}_1]: \quad & \underline{\tau}_1u(q_{l1})=q_{l1}c_1\lambda+\frac{u(q_{l1})}{q_{l1}}\frac{q_{h1}-q_{l1}}{u(q_{h1})-u(q_{l1})}\nu_1, \\
[\underline{\tau}_2]: \quad & \underline{\tau}_2u(q_{l2})=q_{l2}c_2\lambda+\frac{u(q_{l2})}{q_{l2}}\frac{q_{h2}-q_{l2}}{u(q_{h2})-u(q_{l2})}\nu_2, \\
[\bar{\tau}_1]: \quad & \bar{\tau}_1\left(u(q_{h1})-u(q_{l1})\right)=(q_{h1}-q_{l1})c_1\lambda-\nu_1, \\
[\bar{\tau}_2]: \quad & \bar{\tau}_2\left(u(q_{h2})-u(q_{l2})\right)=(q_{h2}-q_{l2})c_2\lambda-\nu_2,
\end{aligned}$$

Combining the first order conditions we obtain

$$\underline{\tau}_iu(q_{li})=q_{li}c_i\lambda+\frac{u(q_{li})}{q_{li}}\frac{q_{hi}-q_{li}}{u(q_{hi})-u(q_{li})}\left[(q_{hi}-q_{li})c_i\lambda-\bar{\tau}_i\left(u(q_{hi})-u(q_{li})\right)\right] \tag{C.17}$$

for $i \in \{1,2\}$. Rearranging and plugging the value for $\bar{\tau}_i$:

$$\underline{\tau}_i u(q_{li}) = \left[q_{li} + \frac{u(q_{li})}{q_{li}} \frac{(q_{hi} - q_{li})^2}{u(q_{hi}) - u(q_{li})} \right] c_i \lambda - \frac{u(q_{li})}{q_{li}} (q_{hi} - q_{li}) \left[- \left[\frac{q_{hi} u(q_{li}) - q_{li} u(q_{hi})}{q_{li} \left(u(q_{hi}) - u(q_{li}) \right)} \right] \frac{\tau}{2} + \frac{u(q_{li})}{q_{li}} \frac{q_{hi} - q_{li}}{u(q_{hi}) - u(q_{li})} \underline{\tau}_i \right]. \tag{C.18}$$

We can rearrange this equation to be

$$D_i \underline{\tau}_i = A_i + B_i c_i, \tag{C.19}$$

with

$$A_i = \frac{u(q_{li})}{q_{li}} \frac{q_{hi} - q_{li}}{u(q_{hi}) - u(q_{li})} \frac{q_{hi} u(q_{li}) - q_{li} u(q_{hi})}{q_{li}} \frac{\tau}{2} \tag{C.20}$$

$$D_i = u(q_{li}) \left[1 + \frac{u(q_{li})}{q_{li}} \frac{q_{hi} - q_{li}}{u(q_{hi}) - u(q_{li})} \frac{q_{hi} - q_{li}}{q_{li}} \right] \tag{C.21}$$

where B_i's formula is irrelevant for the rest of the proof so we omit it. We have that

$$\frac{\underline{\tau}_2 - A_2/D_2}{\underline{\tau}_1 - A_1/D_1} = \frac{c_2}{c_1} \tag{C.22}$$

Recall that in the market equilibrium we have

$$\frac{\underline{\tau}_2 - \frac{\tau}{2}}{\underline{\tau}_1 - \frac{\tau}{2}} = \frac{c_2}{c_1}. \tag{C.23}$$

We confirm numerically that different sets of $\{q_{l1}, q_{l2}, q_{h1}, q_{h2}, \tau, c_1, c_2, u(\cdot)\}$ deliver different implications for the direction of subsiddies relative to markups.

■

C.3 Nested CES preferences

The last model we consider is one with nested CES preferences as in Christian Broda and David E. Weinstein (2010). Let σ denote the elasticity of substitution between different package sizes sold by the same firm, while γ is the elasticity of substitution across firms.

As in the example above, nature chooses two package sizes: q_h and q_l, and n_h and n_l denote the number of packages consumed. There is a unit continuum of identical firms in the economy that produce with a linear technology and unit cost normalized to one.

The utility of purchasing n_h units of the large package q_h and n_l units of the small package q_l from firm j is given by

$$U_i = \int_0^1 \tau_{ij} \left(\left((n_{h,ij} q_h)^{\frac{\sigma-1}{\sigma}} + (n_{l,ij} q_l)^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \right)^{\frac{\gamma-1}{\gamma}} dj \tag{C.24}$$

For simplicity, suppose that there are two types, $\tau_{ij} = 1$ and $\tau_{ij} = \tau > 1$ and there are equal shares of both types in the population.

Social planner allocation. Suppose labor is scarce so that the planner can produce one unit of q_h and one unit of q_l per firm. Suppose further that the optimal allocation features $\{(1,0),(0,1)\}$, meaning that consumers receive 1 unit of q_l for all the goods for which they have a low taste and 1 unit of q_h for all the goods for which they have a low taste.[62]

Market allocation. Each firm chooses a unit price p_h and p_l as well as an allocation—units sold to high and low type—to maximize profits. We focus on symmetric equilibria. That is, each firm uses the same amount of labor and hence decides between the same allocations as the planner: $\{(1,0),(0,1)\}$ or $\{(0,0),(1,1)\}$. We omit the j subscript from now on.

If the firm chooses $\{(1,0),(0,1)\}$, it solves the following problem[63]

$$\begin{aligned} \max_{\{p_l,p_h\}} \quad & q_h\left(p_h - c\right) + q_l\left(p_l - c\right) && \text{(C.25)} \\ \text{s.t} \quad & q_l^{\frac{\gamma-1}{\gamma}} - \frac{p_l q_l}{P} = 0, && [IR_1] \\ & \tau q_h^{\frac{\gamma-1}{\gamma}} - \frac{p_h q_h}{P} = \tau q_l^{\frac{\gamma-1}{\gamma}} - \frac{p_l q_l}{P}. && [IC_\tau] \end{aligned}$$

If the firm chooses $\{(0,0),(1,1)\}$, it excludes the low taste consumer and charges the high taste consumer a transfer T in exchange for the full bundle.

$$\begin{aligned} \max_{\{T\}} \quad & T - c(q_h + q_l) && \text{(C.26)} \\ \text{s.t.} \quad & \tau\left(\left(q_h^{\frac{\sigma-1}{\sigma}} + q_l^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}\right)^{\frac{\gamma-1}{\gamma}} \leq \frac{T}{P}. \end{aligned}$$

The profits of the two options are given by

$$\Pi(\{(1,0),(0,1)\}) = P\left[\tau q_h^{\frac{\gamma-1}{\gamma}} + (2-\tau)\, q_l^{\frac{\gamma-1}{\gamma}}\right], \tag{C.27}$$

$$\Pi(\{(0,0),(1,1)\}) = P\tau\left[\left(q_h^{\frac{\sigma-1}{\sigma}} + q_l^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}\right]^{\frac{\gamma-1}{\gamma}}. \tag{C.28}$$

When choosing to exclude the low type, the firm can extract the full consumer surplus from the

[62] It is easy to show that the opposite allocation – q_h to low type, q_l to high type – is strictly worse and that giving both q_h and q_l to the high type is always worse than giving that same allocation to the high type. As long as $\tau q_h^{\frac{\gamma-1}{\gamma}} + q_l^{\frac{\gamma-1}{\gamma}} \geq \tau\left(\left(q_h^{\frac{\sigma-1}{\sigma}} + q_l^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}\right)^{\frac{\gamma-1}{\gamma}}$, it is optimal for the planner to choose $\{(1,0),(0,1)\}$ over $\{(0,0),(1,1)\}$.

[63] Profits are multiplied by 2 for ease of readability.

high type, but when serving both, the IC constraints ensures that the high type receives an information rent.

Misallocation. We start by showing that the market allocation may feature misallocation of consumption across consumers, where high-taste consumers consume too much of the good and low-taste consumers too little.

PROPOSITION **23.** *There exist parameters such that 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.*

PROOF OF PROPOSITION **23.** This case happens whenever the planner chooses to serve both types, but firms exclude low taste consumers and sell both packages to the high taste consumer in order to extract the full surplus.

We construct one such example below. Let $\varepsilon \geq 0$ be the difference between the social value of serving both types and the one of providing both q_h and q_l to the high type.

$$\varepsilon \equiv \tau q_h^{\frac{\gamma-1}{\gamma}} + q_l^{\frac{\gamma-1}{\gamma}} - \tau \left(\left(q_h^{\frac{\sigma-1}{\sigma}} + q_l^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \right)^{\frac{\gamma-1}{\gamma}} \tag{C.29}$$

The difference between firm profits from serving both types and only the high taste consumer can be written as

$$\frac{\Pi(\{(1,0),(0,1)\}) - \Pi(\{(0,0),(1,1)\})}{P} = \tau q_h^{\frac{\gamma-1}{\gamma}} + q_l^{\frac{\gamma-1}{\gamma}} - (\tau - 1)\, q_l^{\frac{\gamma-1}{\gamma}} - \tau \left[\left(q_h^{\frac{\sigma-1}{\sigma}} + q_l^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \right]^{\frac{\gamma-1}{\gamma}} \tag{C.30}$$

$$= \varepsilon - (\tau - 1)\, q_l^{\frac{\gamma-1}{\gamma}}, \tag{C.31}$$

which is negative for small enough ε. That, is, there are parameters such that the social planner chooses to allocate some of all goods to all people $\varepsilon \geq 0$, but the market allocates too much (q_l in addition to q_h) to the high type and too little (nothing) to the low type. ■

Finally, we show that also in this model, the relative firm-level markup is not indicative of whether a social planner would prefer to subsidize or tax the firm.

PROPOSITION **24.** *Markup variation across firms is not a sufficient statistic for misallocation across firms. That is, there exist parameters such that the planner would optimally subsidize some firms at the expense of others, even though both firms charge the same markup.*

PROOF OF PROPOSITION **24.** In the symmetric equilibrium, all firms charge the same markup. It remains to be shown that the social planner may want to re-allocate labor across firms. Suppose the social planner moves q_l workers from one firm to another. The firm that now only employs q_h workers sells that package to high types, which results in a welfare loss of

$$\text{Welfare loss} = \tau \left(\left(q_h^{\frac{\sigma-1}{\sigma}} + q_l^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \right)^{\frac{\gamma-1}{\gamma}} - \tau q_h^{\frac{\gamma-1}{\gamma}} \tag{C.32}$$

The firm that now employs $q_h + 2q_l$ workers can either sell all three units to the high type and extract the full surplus, or sell the additional unit to the low type, extract the full surplus there and reduce the transfer charged to the high type. We verify numerically that there are parameter values such that the firm chooses to sell the extra unit to the low type, while preferring $\{(0,0)(1,1)\}$ to $\{(1,0)(0,1)\}$ as before.

The welfare gains from the firm employing an additional q_l workers are given by

$$\text{Welfare gain} = q_l^{\frac{\gamma-1}{\gamma}} \tag{C.33}$$

Welfare gains exceeds losses whenever

$$q_l^{\frac{\gamma-1}{\gamma}} + \tau q_h^{\frac{\gamma-1}{\gamma}} \geq \tau \left(\left(q_h^{\frac{\sigma-1}{\sigma}} + q_l^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \right)^{\frac{\gamma-1}{\gamma}}, \tag{C.34}$$

which is precisely the maintained assumption under which serving both types is socially preferable to allocation all the good to the high taste consumer.

■

D Oligopolistic competition

In this section, we consider a modification to the baseline model whereby differences in market power of firms come not from consumer preferences, but from the market structure. In particular, we depart from monopolistic competition and set up an environment in which firms compete a la Atkeson and Burstein (2008).

Setup. There is a unit mass of product types $m \in (0,1)$. In each type, there are two firms, 1 and 2, that have a linear production technology with unit costs $c_1 < c_2$ respectively. Consumers have nested CES preferences over firms and sectors: within product type, the goods produced by each firm have an elasticity of substitution ϵ, while across sectors, good are more substitutable, with an elasticity $\sigma > \epsilon > 1$.

As in the baseline model, we allow for taste heterogeneity across consumers. Consumers have i.i.d. tastes τ_{im} towards each of the product types, where τ_{im} can take on two values, τ and 1. Consumers supply one unit of labor inelastically and own the firms.

Preferences of consumer i are given by

$$U_i = \int_0^1 \tau_{im} q_{im}^{\frac{\sigma-1}{\sigma}} dm, \tag{D.1}$$

$$q_{im} = \left(q_{ijm}^{\frac{\epsilon-1}{\epsilon}} + q_{ikm}^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}} \tag{D.2}$$

Firm's problem. We will consider the problem of firm $j \in \{1,2\}$ who competes with the other firm, $-j$. Similar to our benchmark environment, one can easily show that the IC constraint does not bind for the low-taste consumer. For now, let us assume that the IC constraint of the high-taste consumer is binding. In this case, the firm's problem is given by

$$\max_{\{p_{jl}, p_{jh}, q_{jl}, q_{jh}\}} \pi(p_{jh} - c_j) q_{jh} + (1-\pi)(p_{jl} - c_j) q_{jl}, \tag{D.3}$$

$$\text{s.t.} \quad \left(q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}} = (q_{-jl})^{\frac{\sigma-1}{\sigma}} + \frac{p_{jl} q_{jl}}{P}, \tag{D.4}$$

$$\tau \left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jh}^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}} = \tau \left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}} + \frac{p_{jh} q_{jh} - p_{jl} q_{jl}}{P}. \tag{D.5}$$

Taking first order conditions:

$$
\begin{aligned}
[p_{jh}]: \quad & \pi = \frac{1}{P}\lambda_{jh}, \\
[p_{jl}]: \quad & (1-\pi) = \frac{1}{P}\left(\lambda_{jl} - \lambda_{jh}\right), \\
[q_{jh}]: \quad & \pi(p_{jh} - c_{jh}) = \left[-\frac{\sigma-1}{\sigma}\tau\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jh}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jh}^{-\frac{1}{\epsilon}} + \frac{p_{jh}}{P}\right]\lambda_{jh}, \\
[q_{jl}]: \quad & (1-\pi)(p_{jl} - c_{jl}) = \left[-\frac{\sigma-1}{\sigma}\left(q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jl}^{-\frac{1}{\epsilon}} + \frac{p_{jl}}{P}\right]\lambda_{jl} \\
& \qquad - \left[-\frac{\sigma-1}{\sigma}\tau\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jl}^{-\frac{1}{\epsilon}} + \frac{p_{jl}}{P}\right]\lambda_{jh},
\end{aligned}
$$

From the first two equations, we obtain:

$$
\begin{aligned}
\lambda_{jh} &= \pi P, \\
\lambda_{jl} &= P.
\end{aligned}
$$

The high quantity optimality condition then becomes

$$
\frac{c_{jh}}{P} = \frac{\sigma-1}{\sigma}\tau\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jh}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jh}^{-\frac{1}{\epsilon}}. \tag{D.6}
$$

The low quantity optimality condition can be rearranged to

$$
(1-\pi)\frac{c_{jl}}{P} = \frac{\sigma-1}{\sigma}\left(q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jl}^{-\frac{1}{\epsilon}} - \pi\frac{\sigma-1}{\sigma}\tau\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jl}^{-\frac{1}{\epsilon}}
$$

or

$$
\frac{\sigma-1}{\sigma}\left(q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jl}^{-\frac{1}{\epsilon}} = \frac{1-\pi}{1-\pi\tau\psi_j}\frac{c_j}{P} \tag{D.7}
$$

where $\psi_j \in (0,1)$ is given by

$$
\psi_j = \left(\frac{q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}}{q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1}. \tag{D.8}
$$

Recall that we have assumed that the IC constraint rather than the IR constraint is binding. Note

that when both IR constraints are binding, the allocations and prices are given by

$$\frac{c_j}{P} = \frac{\sigma-1}{\sigma}\tau\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jh}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jh}^{-\frac{1}{\epsilon}}, \tag{D.9}$$

$$\frac{c_j}{P} = \frac{\sigma-1}{\sigma}\tau\left(q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jl}^{-\frac{1}{\epsilon}}, \tag{D.10}$$

$$\frac{p_{jh}q_{jh}}{P} = \tau\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jh}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}} - \tau q_{-jh}^{\frac{\sigma-1}{\sigma}}, \tag{D.11}$$

$$\frac{p_{jl}q_{jl}}{P} = \left(q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}} - q_{-jl}^{\frac{\sigma-1}{\sigma}}, \tag{D.12}$$

To check whether the IR allocations satisfies the IC constraint, we need to check the following condition:

$$\tau\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jh}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}} \geq \tau\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}} + \frac{p_{jh}q_{jh} - p_{jl}q_{jl}}{P}$$

Rearranging with the IR pricing:

$$\tau\left[q_{-jh}^{\frac{\sigma-1}{\sigma}} - \left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}}\right] + \left(q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}} - q_{-jl}^{\frac{\sigma-1}{\sigma}} \geq 0 \tag{D.13}$$

If this condition holds, the equilibrium allocation is given by the IR allocation. Otherwise, it is given by the IC allocation (D.7).

LEMMA **4.** *If the in the market equilibrium the incentive compatibility constraint is binding, it must be that* $\psi_j > \frac{1}{\tau}$.

Proof. Suppose by contradiction that $\psi_j \leq \frac{1}{\tau}$ in the IC equilibrium. Let q_{jl}^{IC} denote the IC equilibrium allocation from (D.7). Because $\psi_j \leq \frac{1}{\tau}$, we have that $q_{jl}^{IC} \geq q_{jl}^{IR}$, where the latter is defined in (D.10). From the definition of ψ_j, (D.8), this implies that $\psi_j < \frac{1}{\tau}$ for all $q_{jl} \leq q_{jl}^{IR}$. Now let's consider whether the IR allocation satisfies the IC constraint (D.13). At $q_{jl} = 0$, that equation holds with equality (trivially, $0 = 0$). The derivative of the LHS of (D.13) with respect to q_{jl} is

$$\frac{\partial LHS(D.13)}{\partial q_{jl}} = \frac{\sigma-1}{\sigma}\left(q_{-jl}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jl}^{-\frac{1}{\epsilon}} - \tau\frac{\sigma-1}{\sigma}\left(q_{-jh}^{\frac{\epsilon-1}{\epsilon}} + q_{jl}^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}\frac{\sigma-1}{\sigma}-1} q_{jl}^{-\frac{1}{\epsilon}}.$$

For this to be positive, we need

$$\psi_j(q_{jl}) \leq \frac{1}{\tau}.$$

As we showed above, this is true for all $q_{jl} \leq q_{jl}^{IR}$. Hence, the IC is satisfied in the IR allocation. Contradiction.

■

E Nonlinear costs

In this section, we consider a modification to the baseline model where firms use flexible cost functions. Instead of assuming firms face a linear production cost, we only assume that total production cost is a function of total production.

Setup. Let the production cost of firm j be denoted by $C_j(q_j)$ where $q_j = \pi q_{\tau j} + (1-\pi)q_{1j}$ is the total production of firm j. Then, the firm's problem is given by

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

Denote by $c_j(q_j) \equiv C_j'(q_j)$ the marginal cost of production of firm j. Then, the optimal quantity of firm j is given by

$$\tau\, u'(q_{\tau j}) = \frac{c_j(q_j)}{P}, \tag{E.2}$$

$$u'(q_{1j}) = \frac{1-\pi}{1-\tau\pi}\, \frac{c_j(q_j)}{P}. \tag{E.3}$$

These optimality conditions are very similar to our benchmark model equations (2.12)–(2.13)

Efficient allocation. The planner's problem, who is constrained to use the same level of aggregate labor, is given by[64]

$$\max_{\{q_{1j}, q_{\tau j}\}} \quad \int_j \left[\pi u(q_{\tau j}) + (1-\pi) u(q_{1j})\right] dj \tag{E.4}$$

$$\text{s.t.} \quad \int_j C_j\left(\pi q_{\tau j} + (1-\pi) q_{1j}\right) = \bar{L}. \tag{E.5}$$

Let P^{AE} denote the Lagrange multiplier on the planner's resource constraint. Then, the efficient allocation quantities satisfy

$$u'(q_{\tau j}^{\text{AE}}) = \frac{c_j(q_j^{AE})}{\tau}\, \frac{1}{P^{\text{AE}}}, \tag{E.6}$$

$$u'(q_{1j}^{\text{AE}}) = c_j(q_j^{AE})\, \frac{1}{P^{\text{AE}}}, \tag{E.7}$$

for all j, where $q_j^{AE} \equiv \pi q_{\tau j}^{AE} + (1-\pi) q_{1j}^{AE}$ is the quantity produced by firm j in the efficient allocation.

[64] In the planner's problem we omit the IC constraints as those will not bind in equilibrium. The proof of this argument mirrors the proof of Proposition 1

E.1 Propositions and Proofs

In this section, we prove that the two main results of our benchmark model still hold: firm-level output and employment are identical to the efficient allocation, but high-taste consumers are sold too much, and low-taste consumers too little of each good.

Proof of Proposition 4 with nonlinear production costs.

Equations (E.2–E.3), 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 production cost function C_j in equilibrium is identical to its overall production in the efficient allocation: $(1-\pi)\left[q_{1j}^{\text{AE}} - q_{1j}\right] - \pi\left[q_{\tau j} - q_{\tau j}^{\text{AE}}\right] = 0$. Using (E.3–E.2), this can be written as:

$$(1-\pi)\left[(u')^{-1}\left(\frac{c_j(q_j^{AE})}{P^{AE}}\right) - (u')^{-1}\left(\frac{1-\pi}{1-\tau\pi}\frac{c_j(q_j^{AE})}{\widetilde{P}_j}\right)\right] - \pi\left[(u')^{-1}\left(\frac{c_j(q_j^{AE})}{\tau\widetilde{P}_j}\right) - (u')^{-1}\left(\frac{c_j(q_j^{AE})}{\tau P^{AE}}\right)\right] = 0. \tag{E.8}$$

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

Now consider a firm with $c_k(q_k^{AE}) = (1+\Delta)c_j(q_j^{AE})$. Using Assumption 1, we have that

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

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.

■

Proof of Proposition 3 with nonlinear production costs and CED preferences. From the Proposition above, we know that $q_j^{AE} = q_j$. That is, the firm-level quantity sold in the decentralized equilibrium is equal to the efficient allocation level. From equations (E.2–E.3) and (E.6), together with noting that the marginal cost of firm j is the same across the two equilibria, we have that:

$$\frac{u'(q_{\tau j})}{u'\left(q_{\tau j}^{AE}\right)} = \frac{P^{AE}}{P}, \tag{E.9}$$

$$\frac{u'(q_{1j})}{u'\left(q_{1j}^{AE}\right)} = \frac{1-\pi}{1-\tau\pi}\frac{P^{AE}}{P}. \tag{E.10}$$

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

$q_{\tau j} > q_{\tau j}^{\text{AE}}$ and $q_{1j} < q_{1j}^{\text{AE}}$ for all j, and (iii) if $\frac{P}{P^{AE}} < \frac{1-\tau\pi}{1-\pi}$ then $q_{\tau j} < q_{\tau j}^{\text{AE}}$ and $q_{1j} < q_{1j}^{\text{AE}}$ for all j.

Aggregate labor market clearing implies that

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

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