Quantifying the Welfare Gains of Variety: A Sufficient Statistics Approach

Kory Kroft, Jean-William P. Laliberté, René Leal-Vizcaíno, and Matthew J. Notowidigdo*

April 2017

Abstract

This paper develops a new revealed-preference approach for valuing changes in product variety. Our key contribution is to derive a sufficient statistics formula for the "variety effect", the change in consumer surplus resulting from a change in the number of products available to consumers. We show that the variety effect can be represented graphically using a standard demand-and-supply diagram. We demonstrate that our sufficient statistics formula is valid for a wide class of models including both continuous and discrete choice. We give a microeconomic foundation for our key assumption of parallel inverse demands and provide an asymptotic result. Next, we illustrate the value of our approach by applying it to two classic questions. First, we consider whether product variety is insufficient or excessive relative to the social optimum. Second, we consider the welfare effects of taxation with endogenous product variety. Central to both applications are reduced-form estimates of the effect of taxes on variety and the effect of taxes on prices and quantities in two cases: where variety is held constant and where variety responds to a change in taxes through entry or exit of firms. Combining rich retail scanner data from grocery stores in the U.S. with detailed state and county sales tax data and using within- and between-state variation in sales tax rates and tax exemptions, we find that at current sales tax rates and entry costs, product variety is lower than socially optimal. Finally, we estimate a large effect of sales taxes on product variety, and we find that the marginal excess burden of sales taxes is significantly larger when incorporating the indirect effect of taxes on product variety.

JEL codes: H21, H71, F12, L13.

Keywords: Optimal product variety, imperfect competition, free entry, sales taxes, excess burden.

---

*Kroft: University of Toronto and NBER, kory.kroft@utoronto.ca; Laliberté: University of Toronto, jw.plaliberte@mail.utoronto.ca; Leal-Vizcaíno: Northwestern University, renelealv@u.northwestern.edu; Notowidigdo: Northwestern University and NBER; noto@northwestern.edu. We thank Simon Anderson, Raj Chetty, Amy Finkelstein, Nathan Hendren, Louis Kaplow, Henrik Kleven, Jesse Shapiro, Rob Porter, Aviv Nevo, Stephen Coate, and numerous seminar participants for helpful comments. We thank Robert French, Pinchuan Ong, Adam Miettinen, Boriana Miloucheva, Shahar Rotberg, Marc-Antoine Schmidt, Jessica Wagner and Haiyue Yu for extremely valuable research assistance. We gratefully acknowledge funding from the Social Sciences and Humanities Research Conucil (SSHRC). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the SSHRC.

1 Introduction

Quantifying the benefits to consumers from new products is important for a broad range of economic issues. In Industrial Organization (IO), it is central to whether markets provide an efficient level of product variety (Spence 1976ab; Dixit and Stiglitz 1977; Mankiw and Whinston 1986). In International Trade, it is crucial to a full accounting of the gains from trade (Feenstra 1994; Broda and Weinstein 2006).

The typical approach to measuring the gains from greater product variety in both the IO and Trade literatures has been to specify and estimate structural models of demand. In IO, the standard approach involves estimating random utility models that allow for rich observed and unobserved heterogeneity across both product characteristics and consumer tastes and examining the impact of greater variety on consumer surplus (see, e.g., Petrin 2002). This approach allows for a welfare analysis of counterfactual scenarios, but requires modeling assumptions that may be hard to assess and may have meaningful effects on the ultimate welfare estimates. Moreover, fully-specified structural models are naturally specific to the particular setting, which makes it hard to extend and compare results from these exercises to other settings.

By contrast, the standard approach in Trade to measuring the welfare gains from product variety involves estimating parametric models of demand that ignore much of the heterogeneity that is the focus in the IO literature (see, e.g., Broda and Weinstein 2006). This allows researchers to analyze the consequences of changes in product variety in a large number of markets simultaneously, but requires researchers to make strong functional assumptions on preferences such as Constant Elasticity of Substitution (CES) demand.

In this paper, we propose a new revealed preference approach to studying the welfare consequences of changes in product variety, what we label the "variety effect".[1] Rather than specify a consumer's utility function as a function of the number of offered varieties in the market, the spirit of our approach is to use the market demand curve (for a given number of products) to evaluate changes in consumer surplus due to changes in variety. We begin with a model of symmetric products and show that the variety effect is given by the area between the market demand curves before and after the change in variety. This is similar to the approach to value new goods using the area under the demand curve for the new good (Hausman 1996; Bhattacharya 2015).

To identify the variety effect, we assume that the market inverse demand curve shifts in parallel in response to exogenous changes in variety, and we show that under this assumption there are

[1] To be clear, there is also a price effect associated with changes in variety which affects consumer surplus. The variety effect is the change in consumer surplus due to the change in variety holding market prices constant.

three key "sufficient statistics" for identification: the price elasticity of market demand when variety is held constant, the price elasticity of market demand when variety can vary, and the change in market output with respect to a change in the number of varieties. We use these insights to provide a graphical representation of the variety effect in a demand-and-supply framework. As in Einav, Finkelstein, and Cullen (2014), we view these graphs as providing useful intuition and therefore as an important contribution on their own.

We demonstrate that our sufficient statistics formula is valid for a wide class of models with symmetric preferences, including both continuous and discrete choice, the key assumption being that inverse aggregate demands are parallel for different variety levels. To give a microeconomic foundation for the parallel demands assumption, first we restrict ourselves to the random utility model and show that the symmetric nested logit model gives rise to parallel inverse demands. Next, we show that for a wide class of distributions of the random utility shock, the inverse aggregate demands are asymptotically parallel as the variety level increases. This result comes from extreme value theory: when the random utility shocks are independent and identically distributed, the distribution of the maximum order statistic converges to a Gumbel distribution (the same as the logit shocks) for a wide range of distributions. We show numerically that this convergence happens very quickly for many standard shock distributions (Normal, Gamma and Exponential).[2] Second, we come back to the continuous choice model and we characterize the utility functions that give rise to parallel demands.

Departing from the symmetric model, we show that our sufficient statistics formula is robust to allowing for asymmetric products and prices. First we show that we can still obtain the variety effect as the area between the market demand curves before and after the change in variety. Then, under the assumption that all prices are increased simultaneously and in the same proportion in response to a change in the number of varieties, we show how to summarize the area between the market demand curves in terms of our sufficient statistics. Importantly, we show that we can state the sufficient statistics in terms of the elasticity of market demand with respect to the *uniform price response* and therefore we do not need to estimate how total output changes with each individual price separately. We also consider an extension to probabilistic entry and provide an analysis of the LeChatelier Principle in the context of product differentiation.

We next illustrate the value of our approach by applying it to two important economic questions. First, we revisit the classic IO question of whether there is too little or too much product variety in the free-entry equilibrium. To shed light on this, we consider a general model of imperfect competition

[2]Extreme value theory has been used in economics in a random utility context by Gabaix et. al. (2016) to show that there might exist robustly high equilibrium markups in large markets that are insensitive to the degree of competition as the number of firms increases.

with symmetric firms (nesting Cournot, Bertrand and Perfect Collusion, all of which are captured in a reduced-form by a conjectural variations parameter) and derive the social marginal welfare gain or loss from a small change in the number of varieties. We show that whether variety is insufficient or excessive depends on the relative magnitude of the variety effect and the "business-stealing effect", which is a negative externality arising because the marginal entrant does not account for the harm it imposes on its competitors (see Mankiw and Whinston 1986). The business-stealing effect jointly depends on the firm's markup and the response of firm-level output to a change in the number of varieties. Determining whether variety is socially optimal at the current equilibrium therefore requires identifying both the variety effect and the business-stealing effect. A key insight of the paper is to show, somewhat surprisingly, that one may recover the business-stealing effect locally using a purely reduced-form approach, which we describe in more detail below.

The second application we consider is the welfare effect of taxation, which is of longstanding interest in Public Economics. Standard formulas for measuring the welfare effects of taxes typically assume markets are competitive (Harberger 1964) or markets are characterized by imperfect competition but the number of firms and products are fixed (Auerbach and Hines 2001, Weyl and Fabinger 2013). In differentiated product markets, taxes can distort product variety, since they reduce firm profitability, thereby leading to exit for those firms at the margin. Our contribution is to derive a new formula for the marginal welfare gain from increasing commodity taxes in a general model of imperfect competition covering a wide range of market conduct when variety is endogenous. While the standard formula emphasizes the response of total output to the tax due to a fiscal externality, our formula shows additionally that one needs to account for the variety effect which arises since taxes distort product variety. Interestingly, the impact on welfare coming through a change in variety is ambiguous and depends on whether variety is excessive or insufficient at the prevailing equilibrium. We show how our new formula connects to existing formulas for the welfare effects of taxes under imperfect competition (e.g., Besley 1989).[3]

Next, we show how *both* the marginal welfare gain of varieties and the marginal welfare gain of taxation can be identified using exogenous variation in taxes. The key insight is that taxes shift variety in the market. As a result, it is straightforward to apply our sufficient statistics formula for the variety effect. To identify the business-stealing effect using a sufficient statistics approach, we exploit the firm's free-entry condition and show that, under a symmetry assumption, it's possible to recover a firm's markup and the response of output to a change in varieties using the following

[3]Although we focus on taxes in this paper, our results are valid for any type of cost shock; for example, a tariff. Thus, we view the methods developed in this paper as being portable across the different fields of economics, such as Public Economics, Trade and IO.

reduced-form parameters: the change in market output with respect to the tax when variety can vary and when variety is held fixed, the change in market price with respect to the tax when variety can vary, and the change in product variety with respect to the tax. Interestingly, our sufficient statistics formula does not require observing a firm's cost or the market conduct parameter, which may be difficult to estimate. Comparing the variety effect against the business-stealing effect reveals whether product variety is too high or low relative to the social optimum. The variety effect and the impact of taxes on market-level output and price also allows us to implement the marginal welfare gain of taxation.

Finally, as a last step in our analysis, we combine rich retail scanner data from grocery stores in the U.S. with detailed state and county sales tax data. Our data allows us to estimate both the "short-run" and "long-run" effects of taxes. To operationalize our framework, we assume that the long-run effects of taxes correspond to situations where variety may adjust to taxes and the short-run effets of taxes correspond to situations where variety is fixed, an assumption we test below. Following the literature, the long-run effects of taxes are estimated using *cross-sectional* variation in sales tax rates and exemptions between and within counties in a difference-in-differences research design. We validate these cross-sectional results in a "border pair" subsample that compares similar counties in different states that share a state border, following the empirical strategy laid out in Holmes (1998), Dube et al. (2010) and Hagedorn, Manovskii and Mitman (2016). The short-run effects of taxes are estimated by exploiting *time-series* variation in tax rates, and using fixed effects panel models to focus on within-store changes over time in prices and expenditures in response to changes in sales taxes.

Our preferred estimates suggest a large effect of sales taxes on product variety in the long run. The long-run effects of taxes on output are larger (in magnitude) than the short-run effects and we find similar effects on after-tax prices in the short run and long run. When we implement our sufficient statistics formulas, we find that at current sales tax rates and entry costs, there is insufficient entry. In other words, the variety effect dominates the business stealing effect. We also find that the marginal excess burden of sales taxes is significantly larger than estimates ignoring the indirect effect of taxes on product variety, reflecting the fact that sales taxes have an additional social cost of moving product variety even further away from the social optimum. These results are of course specific to the retail sector and may not generalize, but they illustrate the potential for our approach to be usefully applied in other settings to learn more about the welfare effects of increasing product variety.

A key advantage of our framework is that it illustrates in a transparent way the reduced-form

estimates that are needed to conduct a welfare analysis. In this sense, our framework is broadly related to the recent work on sufficient statistics (e.g., Chetty 2009), and we show how the welfare analysis of product variety can be implemented using a key set of estimable reduced-form parameters. Conditional on these statistics, the researcher does not need to estimate additional structural parameters governing consumer tastes or firm costs, which may give our approach an advantage of robustness to misspecifying these aspects of the problem; additionally, cost data are often not readily available and not straightforward to estimate. As a result, it is relatively straightforward to implement, which makes it applicable to the settings studied in IO, Public Economics and Trade.

These advantages come with important limitations. One is that our approach based on reduced-form estimates is inherently "local"; we can estimate whether variety at the current equilibrium is "too high" or "too low" relative to the social optimum and we can estimate the marginal welfare gain from small increases in variety. However, we cannot readily solve for the socially optimal level of variety as in Berry and Waldfogel (1999) and Gentzkow, Shapiro and Sinkinson (2014) without imposing more structure on the problem. This is analogous to the optimal unemployment insurance (UI) literature that uses sufficient statistics to estimate the marginal welfare gain from small changes in UI benefit levels as opposed to the globally optimal UI benefit level (Baily 1978, Chetty 2006, Chetty 2008, Kroft and Notowidigdo 2016). Another limitation of our modeling approach is that it is not well-suited to studying endogenous changes in product characteristics, such as the strategic product re-positioning that is examined in Wollmann (2016). Lastly, estimating the cost function without cost data would require specifying a form of competition. Even then, given the inherently local nature of our approach, we would only be able to estimate a linear approximation of the supply function. Therefore, by not fully specifying the form of competition, we cannot easily analyze the counterfactual welfare effects of alternative policies.

Finally our paper builds on and contributes to several literatures. First, our paper relates to the vast literature on optimal product variety. This old literature dates back to Spence (1976), Dixit and Stiglitz (1977) and Mankiw and Whinston (1986) with more recent contributions by Anderson, de Palma and Nesterov (1995), Berry and Waldfogel (1999), Gentzkow, Shapiro and Sinkinson (2014), Berry, Eizenberg and Waldfogel (2015). Similarly, our paper relates to the welfare economics of new goods (Trajtenberg 1989, Hausman 1996, Hausman and Leonard 2002, Petrin 2002, Bhattacharya 2015). Much of this literature is purely theoretical; this paper is the first to adopt a sufficient statistics approach.

Second, our paper directly relates to the literature on commodity taxation with imperfect competition. Key papers in this area include Seade (1987), Stern (1987), Myles (1989), Besley (1989),

Delipalla and Keen (1992), Anderson, de Palma and Kreider (2001a, 2001b), Auerbach and Hines (2001), Weyl and Fabinger (2013) and Gillitzer, Kleven and Slemrod (2015). These papers typically assume a specific form of firm competition and impose specific structure on consumer preferences, and some of these papers focus purely on the short-run equilibrium holding the number of varieties fixed. Our paper is distinguished by allowing for both differentiated products and free entry in deriving our welfare expressions, without having to specify the form of imperfect competition (e.g. Bertrand vs Cournot).

Lastly, our paper relates to the literature on the gains from trade and product variety. Key papers in this area are Feenstra (1994), Broda and Weinstein (2006), Arkolakis, Costinot and Rodriguez-Clare (2012), Melitz and Redding (2015) and Atkin, Faber and Gonzales (2016). These papers typically specify a utility function and estimate the variety effect under that restriction; our framework does not impose strong functional forms on preferences, but rather uses demand curves to identify a consumer's love for variety.

The rest of the paper proceeds as follows. Section 2 considers the symmetric model, and section 3 describes how to extend this model to incorporate asymmetries, multi-product firms, probabilistic entry and long-run outside market adjustments. Section 4 explores the applications to the socially optimal product variety and the welfare effects of taxation using a version of the simple model described in section 2. Section 5 describes our data. Second 6 outlines our econometric framework. Second 7 contains our empirical results. Section 8 reports our welfare calibrations and Section 9 concludes.

2 Symmetric Model of Demand

We begin with two classes of models with symmetric preferences. In section 2.1.1, we consider a class of discrete choice (random utility) models and in section 2.1.2, we consider continuous choice models. We show that for each model consumer surplus is equal to the integral of aggregate demand (see proposition 1 and 1'). Next, we proceed to the welfare analysis of product variety, where we define and provide graphical representation of the price effect and variety effect. We then derive a sufficient statistics formula for the variety effect under a "parallel demands" assumption and consider several microfoundations that satisfy parallel demands.

2.1 Preferences, Demand, and Consumer Surplus

2.1.1 Symmetric Discrete Choice Model

Individuals indexed by i choose to purchase a single product $j \in \{1, ..., J\}$ or the outside option $j = 0$. The number of products in the market, J, is our measure of product variety.

Preferences. We consider a population of statistically identical and independent consumers of mass unity. The utility of individual i who purchases product j is given by:

$$u_{ij}(y_i, p_j) = \alpha(y_i - p_j) + \delta_j + (1 - \sigma)\nu_i + \sigma\varepsilon_{ij} \tag{1}$$

where y_i is the consumer's income, p_j is the price of good j and $(1 - \sigma)\nu_i + \sigma\varepsilon_{ij}$ is an idiosyncratic match value between consumer i and product j, which captures heterogeneity in tastes across consumers and products. The utility of individual i who chooses the outside option is given by $u_{i0} = \alpha y_i + \varepsilon_{i0}$. In general, we assume that for $j \neq 0$, the random utility shocks (ε_{ij}) are identical and independently (continuously) distributed (i.i.d.) and independent of (ν_i), but we allow ε_{i0} to be correlated with ν_i.

Demand. Given the preferences in equation (1), we may define the demand for product j as

$$q_j(p_1, \ldots, p_J, J) = \mathbb{P}\left(u_{ij}(y_i, p_j) = \max_{j' \in \{0,\ldots,J\}} u_{ij'}(y_i, p_{j'})\right) \tag{2}$$

Aggregate demand (for all products excluding the outside good) when $p_j = p$ for all $j = 1 \ldots J$ for a fixed number of varieties J is defined as

$$Q(p, J) = \sum_{j=1}^{J} q_j(p, J) \tag{3}$$

Similarly, define $P(Q, J)$ to be the inverse aggregate demand corresponding to $Q(p, J)$.

Consumer Surplus. In the discrete choice model we define consumer surplus as the expected maximum utility normalized by the marginal utility of income

$$CS(p_1, \ldots, p_J, J) = \frac{1}{\alpha}\mathbb{E}\left[\max_{j \in \{0,\ldots,J\}} u_{ij}(y_i, p_j)\right]$$

Proposition 1. *Under the assumption of no income effects and symmetric prices $p_j = p$, we can represent consumer surplus as*

$$CS(p, J) = \int_p^\infty Q(s, J)ds \tag{4}$$

Thus, consumer surplus is equal to the integral of aggregate demand.

Proof. This is a consequence of the Williams-Daly-Zachary Theorem (See McFadden 1981) which states that $\frac{\partial CS(p_1,\ldots,p_J,J)}{\partial p_j} = -q_j(p_1,\ldots,p_J,J)$. Therefore, $\frac{dCS(p,J)}{dp} = \sum_{j=1}^{J} \frac{\partial CS(p,\ldots,p,J)}{\partial p_j} = -Q(p,J)$. If follows that $CS(p,J) = \int_p^\infty Q(s,J)ds + k$. The integration constant does not depend on (p,J) and can be disregarded for comparative statics. □

2.1.2 Symmetric Continuous Choice Model

We now introduce the continuous choice model and show the analogous result to proposition 1.

Preferences. Let the representative consumer's utility function given by

$$u(q_1,\ldots,q_J,m) = h_J(q_1,\ldots,q_J) + m$$

for any $h_J : \{1,\ldots,J\} \to \mathbb{R}$ which is symmetric in all its arguments, continuously differentiable, strictly quasi-concave and $h(0,\ldots,0) = 0$ and where the linear good m is interpreted as money.

Demand. Imagine the consumer is facing symmetric prices $p_j = p$ for all j, and define $H_J(Q) = h_J\left(\frac{Q}{J},\ldots,\frac{Q}{J}\right)$.

The solution to the consumer problem then is given by

$$u^*(p,J,y) = \max_Q H_J(Q) + y - pQ$$

From the first-order condition, we obtain the family of inverse demands $P(Q,J) = H_J'(Q)$. Furthermore, it is easy to see that given the optimal aggregate quantity $Q(p,J)$ for price p, the strict quasi-concavity of h implies the consumer chooses symmetric quantities $q_j = \frac{Q}{J}$ for all j.

Consumer Surplus. Finally, note that

$$u^*(p,J,y) = H_J(Q(p,J)) + y - p * Q(p,J) = \int_0^{Q(p,J)} P(s,J) + y - p * Q(p,J) = \int_p^\infty Q(s,J)ds + y$$

proving proposition 1 for the continuous choice model.

Proposition 1'. *If the consumer has preferences as described above, namely $u(q_1,\ldots,q_J,m) = h_J(q_1,\ldots,q_J) + m$ for any h_J symmetric in all its arguments, continuously differentiable, strictly quasi-concave and $h(\mathbf{0}) = 0$, and if the consumer faces symmetric prices $p_j = p$, we can represent consumer surplus as*

$$CS(p,J) \equiv u^*(p,J,y) = \int_p^\infty Q(s,J)ds + y \tag{5}$$

Thus, consumer surplus is equal to the integral of aggregate demand up to a constant.

Furthermore, none of the assumptions on utility stated in the proposition are too restrictive. We show that for any family of downward sloping aggregate demands there exists a utility function $u_J : \mathbb{R}^{J+1} \to \mathbb{R}$ satisfying the conditions of the proposition that rationalizes them; in particular, this preferences are symmetric across different varieties and are strictly quasi-concave. Let $P(Q, J)$ be continuously differentiable and striclty decreasing in Q. Let H be any antiderivative $\int P(Q, J)dQ$, which exists because $P(Q, J)$ is differentiable. Then the following is a strictly quasi-concave direct utility function that rationalizes $P(Q, J)$ for integer J when all prices p_j in the market are equal:

$$u(q_1, \ldots, q_J, m) = H\left(\left(J^{\rho-1}\sum_{j=1}^{J} q_j^{\rho}\right)^{\frac{1}{\rho}}\right) + m$$

for some $\rho \in (0, 1)$. Furthermore, we can make sense of J as a continuous variable if we permit a continuum of varieties $q : [0, J] \to \mathbb{R}$ and let

$$u_J(q, m) = H\left(\left(\int_0^J J^{\rho-1} q^{\rho}(j)dj\right)^{\frac{1}{\rho}}\right) + m$$

Finally, in this case $CS(p, J) = \int_p^{\infty} Q(s, J)ds$ is differentiable in both p and J.

2.2 Welfare Effects of Variety

In this section, we consider the impact of a small change in the number of varieties J on consumer surplus. The starting point of this section is the expression for consumer surplus as the integral of aggregate demand $CS(p, J) = \int_p^{\infty} Q(s, J)ds$ which we have proved for both the discrete choice model and continuous choice model. In what follows we we will assume that J is a continuous variable and $Q(p, J)$ is defined for any $J \in \mathbb{R}$ and continuously differentiable. We first introduce some definitions.

Definition 1. The "price effect" is defined as

$$-Q\frac{dp}{dJ} \tag{6}$$

It arises since market prices may change when firms enter or exit the market. Due to the envelope theorem, there is no first-order effect on welfare due to re-optimization when prices change, so only the mechanical effect of a price change affects consumer welfare.

Definition 2. The "variety effect" is defined as

$$\Lambda(Q,J) \equiv \int_{P(Q,J)}^{\infty} \frac{\partial Q}{\partial J}(s,J)ds \tag{7}$$

Holding the effect on prices constant, a new variety increases welfare since consumers exhibit a "love of variety". The variety effect depends on how aggregate demand responds to a change in variety.

The total effect on consumer welfare is the sum of the price effect and the variety effect:

$$\frac{dCS}{dJ} = -Q\frac{dp}{dJ} + \Lambda \tag{8}$$

Each term in equation (8) is illustrated in Figure 1 where we consider a reduction in product variety ($\Delta J < 0$). The price effect is the rectangular area where the base is the pre-existing output level and the height is the change in prices. The variety effect is given by the area between the demand curves. To see the intuition for the variety effect, consider a market where consumers choose among a set of products. The aggregate demand curve for these products is given by equation (3). The inverse of this curve gives the maximum willingness to pay for products in the market across individuals. Now consider a reduction in the number of products available. In this case, some consumers will no longer be able to purchase their most preferred option. Thus, the maximum possible utility attainable will be lower for these consumers. We can represent this a shift down in the inverse aggregate demand curve. The area between the inverse aggregate demand curves before and after the change in variety (above the market price) corresponds exactly to the variety effect.

Another way to see this is to define the average change in willingness to pay for inframarginal units as variety J changes as

$$\overline{\frac{\partial P}{\partial J}}(Q,J) = \frac{1}{Q}\int_0^Q \frac{\partial P}{\partial J}(s,J)ds.$$

We can then show the variety effect is the total change in willingness to pay for the units that are actually exchanged:

$$\Lambda = \int_p^{\infty} \frac{\partial Q}{\partial J}(s,J)ds = \int_0^Q \frac{\partial P}{\partial J}(s,J)ds = Q\overline{\frac{\partial P}{\partial J}} \tag{9}$$

A key objective of this paper is to try and establish a method to identify the variety effect using reduced-form methods. The next theorem states that one may recover the variety effect in the general model above using a "sufficient statistics" approach (Chetty 2009).

Theorem 1. *The variety effect Λ is equal to the average change in willingness to pay times the*

quantity demanded:

$$\Lambda(Q,J) = Q\overline{\frac{\partial P}{\partial J}}(Q,J) \tag{10}$$

Furthermore, if the inverse aggregate demands are parallel, $\frac{\partial P}{\partial Q}(Q,J) = \frac{\partial P}{\partial Q}(Q,J')$ *for all* J, J' *and* Q*, then the average change in willingness to pay is equal to the marginal change in willingness to pay,* $\overline{\frac{\partial P}{\partial J}}(Q,J) = \frac{\partial P}{\partial J}(Q,J)$*. Therefore we obtain:*

$$\overline{\frac{\partial P}{\partial J}} = \left(\frac{dP}{dQ} - \frac{dP}{dQ}\Big|_J\right)\frac{dQ}{dJ} \tag{11}$$

where $\frac{dP}{dQ}\Big|_J = \frac{\partial P}{\partial Q}$ *denotes the slope of inverse demand when variety* J *is held fixed and* $\frac{dP}{dQ} = \frac{dP(Q(J),J)}{dJ}/\frac{dQ}{dJ}$ *denotes the slope of inverse demand when* J *is variable.*

Proof: See Appendix.

The expression for the variety effect in equations (10) and (11) can be most easily understood geometrically. Figure 2 considers a reduction in variety and shows that the variety effect is the area of the rectangle with base Q_0 and height $d \equiv P^* - P_1$. The base Q_0 is simply the aggregate output prior to the change in variety and is observable. The height d captures the change in willingness to pay as J changes, therefore $d = \overline{\frac{\partial P}{\partial J}}$. However, d is not directly observable, it depends on the price induced by the change in variety, P_1, which is observable and the market price that would prevail at the same level of output but on the original demand curve, P^*, which is unobservable. To see how to recover an expression for d, consider a change in variety from J_0 to J. Note from Figure 2 that d must satisfy the following relationship $Q(P(J),J) = Q(P(J)+d, J_0)$.[4] To solve for d, note that for a small change in J:

$$\begin{aligned} dQ &\approx Q(P(J),J) - Q(P_0,J_0) \\ &= Q(P(J)+d,J_0) - Q(P_0,J_0) \\ &\approx \frac{dQ}{dP}|_{J_0}(d + P(J) - P_0) \\ &\approx \frac{dQ}{dP}|_{J_0}(d + \frac{dP}{dQ}dQ) \end{aligned}$$

Thus, rearranging and solving for d yields:

$$d \approx \left(\frac{dP}{dQ}|_J - \frac{dP}{dQ}\right)dQ \tag{12}$$

[4] Another way to characterize would be $Q(P_0,J_0) = Q(P_0 - d, J_1)$.

In economic terms, d gives the reduction in the willingness to pay for the marginal unit. The upshot is that it's possible to identify d using the price elasticities of demand when J is fixed and when J is variable. Moving forward, we interpret these as the "short-run" and the "long-run" price elasticities of demand, respectively. As long as one can estimate the short-run and the long-run price elasticity of demand, then one can recover the variety effect using purely reduced-form methods. In Section 4, we will show how to recover the variety effect associated with a commodity tax change.

2.3 Microfoundations for Parallel Demands

In this section, we consider the conditions under which aggregate demands satisfy the parallel demands assumption. We start with the discrete choice model and then move to the continuous choice model.

2.3.1 Discrete Choice Model

We first describe the Nested Logit model which features parallel aggregate demands. Next we prove an asymptotic result that a wide class of random utility models feature parallel demands.

Example 1. If the random utility shocks (ε_{ij}) are drawn from the Gumbel distribution, and $(1-\sigma)\nu_i$ has the distribution derived in Cardell (1997), then this model corresponds to the nested logit model in which there are only two nests: one which includes $j = 1, ..., J$ and the other which includes only the outside option $j = 0$. Then

$$q_i(p_1, \ldots, p_J, J) = \frac{\left(\sum_{j=1}^{J} e^{\frac{\delta_j - \alpha p_j}{\sigma}}\right)^{\sigma}}{1 + \left(\sum_{j=1}^{J} e^{\frac{\delta_j - \alpha p_j}{\sigma}}\right)^{\sigma}} \frac{e^{\frac{\delta_i - \alpha p_i}{\sigma}}}{\sum_{j=1}^{J} e^{\frac{\delta_j - \alpha p_j}{\sigma}}}$$

As the parameter σ goes to 0, the only random term in (1) is ν_i which is constant across all $j \neq 0$. When $\sigma = 1$, we retrieve the logit model. The parameter σ/α characterizes a consumer's "love of variety". When σ/α is large, consumers value greater variety (higher J). This parameter is inversely related to the elasticity of substitution considered in trade models with a representative consumer with a CES utility function. In fact, it has been shown that the logit model aggregates to the CES model (Anderson, de Palma and Thisse 1987).[5] If we focus on a symmetric equilibrium where prices

[5]The formal connection requires one to introduce a second stage where individuals choose a continuous quantity of the good.

$p_j = p$ and attributes $\delta_j = \delta$ for all j, aggregate demand is equal to

$$Q(p, J) = \frac{J^\sigma e^{\delta - \alpha p}}{1 + J^\sigma e^{\delta - \alpha p}}$$

Observe then the inverse aggregate demand curve is given by

$$P(Q, J) = \frac{\delta}{\alpha} + \frac{\sigma}{\alpha} \log J - \frac{1}{\alpha} \log \left(\frac{Q}{1 - Q} \right)$$

which is is separable in J and Q. This implies that exogenous shifts in variety move the inverse aggregate demand curve in parallel. Additionally, the coefficient on log variety is different than the coefficient on output. Thus, in this model, the value of additional variety is not pinned down fully by the price elasticity of demand, which is in contrast to the standard models considered in the trade literature (see for example Feenstra 1994, Broda and Weinstein 2006). In the nested logit model, consumer surplus is given by the familiar "log sum" expression $CS = \frac{1}{\alpha} \log \left(1 + J^\sigma e^{\delta - \alpha p} \right)$, also referred to as the "inclusive value."[6] Thus, we get a closed-form solution for the variety effect. In particular, one can show that $\Lambda = \frac{\sigma}{\alpha} q$. The variety effect in this parametric model is equal to the willingness to pay for additional variety ($\frac{\sigma}{\alpha}$) multiplied by the market share of a (symmetric) product (q).[7] Intuitively, if product variety increases by one, then evaluating the effect on consumer surplus requires the post-entry market share of the new good. If this is zero, then consumers don't value the new good and hence the welfare gains are nil.

Next, we show that there is a large class of models that admit a nested logit approximation. The random utility models in this class have in common that the distribution of the maxima of the shocks is asymptotically Gumbel, which implies that the aggregate inverse demands are asymptotically parallel. Therefore, for any model in this class, the structural estimates of the price effect and the variety effect will asymptotically converge to the structural estimates of the nested logit, which in turn can be estimated with reduced-form parameters, namely the price elasticities of demand when J is fixed and when J is variable.

Summarizing, a sufficient condition to get inverse parallel demands is that the random utility shocks (ε_{ij}) are iid Gumbel, independent of the size of σ, the distribution of ν_i and the distribution of ε_{i0}.[8] Furthermore, if the shocks (ε_{ij}) are assumed to be iid, then they have to be Gumbel in

[6] See Train (2003) for derivation.

[7] In this simple nested logit model, the marginal willingness to pay (WTP) for additional variety is the same as the average WTP.

[8] Remember the starting point is the random utility specification $u_{ij} = \alpha(y_i - p_j) + \delta_j + (1 - \sigma)\nu_i + \sigma\varepsilon_{ij}$, and $u_{i0} = \alpha y_i + \varepsilon_{i0}$.

order to satisfy the inverse parallel demands condition.[9] Nonetheless, using results from extreme value theory we can show that for a large class of iid shocks, the inverse demands are asymptotically parallel. We now define the class of models that admit the asymptotic approximation, and provide a useful sufficient condition to show that a given model is in the class.

Definition 3. Let (ε_j) be iid distributed according to a continuous cdf F. We say that F is in the domain of attraction of the Gumbel distribution if

$$\max_{j\in\{1,\ldots,J\}} \varepsilon_j \overset{a}{\sim} Gumbel(\mu(J),\sigma(J)),$$

as $J\to\infty$ for some location and dispersion parameters $(\mu(J),\sigma(J))$.

Lemma 1. *Let x_0 be the supremum of the support of a cdf F that is twice continuously differentiable. If F satisfies that $\lim_{x\to x_0}\frac{F''(x)(1-F(x))}{F'^2}=-1$ then F is in the domain of attraction of the Gumbel distribution.*

See Resnick (1987) for a proof of the lemma and a full characterization of the domain of attraction of the Gumbel distribution. The characterization is outside the scope of the paper and the lemma is enough for our purposes. For example, if (ε_j) are iid $N(0,\sigma^2)$ or exponential the above lemma applies.

The next theorem is the micro foundation for our key assumption of parallel inverse demands and states that inverse demands become parallel as variety increases for any random utility model with shocks in the Gumbel domain of attraction.

Theorem 2. *Let the random utility shocks (ε_j) be iid and distributed according to F in the domain of attraction of the Gumbel distribution. Then, for any large enough J and K there exists d such that for all $p\in\mathbb{R}$ we have $Q(p,J)\approx Q(p+d,K)$. Specifically, for all p we have*

$$Q(p,J)=\mathbb{P}\left(\max_{j\in\{1,\ldots,J\}} u_{ij}(p)>u_{i0}\right)\overset{a}{\sim}\mathbb{P}\left(\max_{j\in\{1,\ldots,K\}} u_{ij}(p+d)>u_{i0}\right)=Q(p+d,K)$$

Therefore the inverse demands are approximately parallel $P(Q,K)\approx P(Q,J)+d$ for all Q, for large enough J and K.

Proof: See Appendix.

[9] Without the independence assumption, there are other distributions of shocks that also give rise to parallel demands. If we do not assume independence of the shocks then the Gumbel distribution is not necessary. In this case, for any family of (downward sloping, increasing in J) parallel inverse demands there is a joint density of shocks that rationalizes it. In general, the shocks are not Gumbel and are correlated in this construction.

In Figure 3, we assess this approximation theorem by numerically simulating different random utility models and calculating the bias that arises from assuming demands are parallel. Specifically, we simulate a model of a large number of consumers choosing with utility over products given by equation (1). We choose $\alpha = 1$ and $y = 1$ in the simulation and we consider a range of different shock distributions (Gumbel, Normal, Gamma, and Pareto). We then repeat this procedure for a range of different values of J to assess how the bias from assuming parallel demand varies with J when consider a hypothetical 20 percent increase in the number of products (from initial value of J). We compute the welfare gains exactly using numerical methods and compare the exact welfare gain to the approximate gains implied by assuming parallel demands based on the formula in equation (10). The results in Figure 3 show that the bias that arises from assuming parallel demands is a function of the number of varieties in the market, where bias is measured as the difference between the estimated (approximate) variety effect and the exact variety effect. The benchmark distribution is Gumbel where we know from theory that the demand curves are exactly parallel and therefore the bias is zero. For both the Normal and Exponential distributions, we find that the bias is small in magnitude and converges to 0 fairly quickly as the number of varieties increase. On the other hand, with a Pareto distribution, there is a bias of roughly 20 percent, which does not vanish as varieties increase. In this case, the variety effect computed using our sufficient statistics formula is a lower bound on the true variety effect.

2.3.2 Continuous Choice Models

Inverse aggregate demands are parallel, $\frac{\partial P}{\partial Q}(Q, J) = \frac{\partial P}{\partial Q}(Q, J')$ for all J, J' and Q, if and only if they are linearly separable in Q and J. Therefore any family of parallel inverse demands can be written as $P(Q, J) = a(J) - f(Q)$ for increasing and differentiable functions $a(J)$ and $f(Q)$. Letting F be any antiderivative of f. Then

$$u_J(q, m) = \left(\int_0^J a(J)^\rho J^{\rho-1} q(j)^\rho dj \right)^{\frac{1}{\rho}} - F\left(\int_0^J q(j) dj \right) + m$$

for some $\rho \in (0, 1)$ is a direct utility function that rationalizes $P(Q, J)$ given that firms are playing a symmetric price equilibrium.

We now describe examples of demand functions generated from continuous choice models that feature parallel demands.

Example 2. Bulow and Pfleiderer (1981) obtain the following three classes of tractable inverse

demands as the unique curves with the property that a constant fraction of cost is always passed on to the consumer as marginal cost shifts:

1. $P(Q,J) = \alpha_J - \beta_J Q^{\delta}$, for $\delta > 0$,
2. $P(Q,J) = \alpha_J - \beta_J log(Q)$,
3. $P(Q,J) = \alpha_J + \beta_J Q^{1/\eta}$, for $\eta < 0$, which is the constant elasticity inverse demand shifted by the intercept α_J.

As before, solving the integrability problem we can obtain direct utility functions that generate this inverse aggregte demands when firms are setting symmetric prices. For example, for the first class we can see that the following utility works

$$u(q_1,\ldots,q_J,m) = \alpha_J \left(J^{\rho-1} \sum_{i=1}^{J} q_i^{\rho} \right)^{\frac{1}{\rho}} - \beta_J \frac{\left(\sum_{i=1}^{J} q_i\right)^{\delta+1}}{\delta+1} + m$$

An important case is when $\beta_J = \beta$ for all J, then the inverse aggregate demands are linearly separable in J and Q and we say that they shift in parallel as J moves. The fact that these are the only classes of curves for which marginal costs are passed-on in a constant fraction makes them a tractable benchmark and therefore they have been popular in applied work. Fabinger and Weyl (2016) generalize Bulow and Pfleiderer's (1981) analysis and characterize a bigger class of "tractable equilibrium forms" of the form $P(Q,J) = \alpha_J + \beta Q^t + \gamma Q^u$ which allow for greater modelling flexibility. Again, as long as β and γ are independent of J then we say that the inverse demands shift in parallel

3 Extensions

3.1 Asymmetric Products and Prices

So far we have assumed a symmetric price equilibrium. We now relax this assumption and allow for asymmetric prices. Consider the case where the number of products goes from J to M (with $M > J$). In this case, the variety effect is equal to the following line integral:[10]

$$\Lambda = \int_0^{\infty} \sum_{j=J+1}^{M} q_j(\mathbf{p_J}, p_{J+1}+s, p_{J+2}+s, \ldots, p_M+s)ds$$

where $\mathbf{p_J} = (p_1, p_2, ..., p_J)$ are the prices for the existing J products before the introduction, and $(p_{J+1},\ldots,p_M)$ are the prices at which the new $M-J$ products are introduced. This can be restated

[10] This holds for both continuous and discrete choice, see Bhattacharya (2016).

in terms of aggregate demands as:

$$\Lambda = \int_0^\infty Q_M(\mathbf{p_M} + s\mathbf{1_M})ds - \int_0^\infty Q_J(\mathbf{p_J} + s\mathbf{1_J})ds$$

where $\mathbf{1_K}$ is a $K-$dimensional vector of ones, and $\mathbf{p_M} = (\mathbf{p_J}, p_{J+1}, \ldots, p_M)$.

Next, we assume that there exists some d such that $Q_M(\mathbf{p_M} + (s+d)\mathbf{1_M}) = Q_J(\mathbf{p_J} + s\mathbf{1_J})$ for all $s \in \mathbb{R}$.[11] In other words, increase prices starting from $\mathbf{p_M}$ by some constant amount d until total quantity demanded equals quantity demanded when there are J products in the market. Under this assumption, it follows that:

$$\Lambda = \int_0^d Q_M(\mathbf{p_M} + s\mathbf{1_M})ds \tag{13}$$

By the mean value theorem for integrals, there exists $d' \in [0, d]$ such that

$$\Lambda = \int_0^d Q_M(\mathbf{p_M} + s\mathbf{1_M})ds = d * Q_M(\mathbf{p_M} + d'\mathbf{1_M}) \tag{14}$$

Therefore, as $M \to J$ we have $d' \to d$ and we obtain a first order approximation for the variety effect:

$$\Lambda \approx d * Q_J(\mathbf{p_J})$$

We now have an expression for the variety effect that is similar to expression in the case of symmetric prices, equation (10). All that remains is to characterize d. Let $\mathbf{p}_M^1$ be the *new* prices after the introduction of the $M - J$ products. Then,

$$dQ = Q_M(\mathbf{p_M^1}) - Q_J(\mathbf{p_J})$$

Using the definition of d,

$$dQ = Q_M(\mathbf{p_M^1}) - Q_M(\mathbf{p_M} + d\mathbf{1_M})$$

Assume, furthermore, that prices all change by the same amount, $s = dp = (\mathbf{p_M^1} - \mathbf{p_M})_j$ for all

[11] Note that for the symmetric price case, we considered a decrease in variety. Here, we consider an increase in variety so the sign of d changes.

j, after the introduction of the new products.[12] Then,

$$\begin{aligned} dQ &= Q_M(\mathbf{p_M} + s\mathbf{1_M}) - Q_M(\mathbf{p_M} + d\mathbf{1_M}) \\ &= Q_M(\mathbf{p_M} + s\mathbf{1_M}) - Q_M(\mathbf{p_M}) + Q_M(\mathbf{p_M}) - Q_M(\mathbf{p_M} + d\mathbf{1_M}) \\ &= \sum_{j=1}^{M} \frac{\partial Q_M}{\partial p_j} s - \sum_{j=1}^{M} \frac{\partial Q_M}{\partial p_j} d \\ &= (s-d) \sum_{j=1}^{M} \frac{\partial Q_M}{\partial p_j} \\ &= (s-d) \left. \frac{dQ_M}{d\mathbf{p}} \right|_M \end{aligned}$$

Where $\left. \frac{dQ_M}{d\mathbf{p}} \right|_M = \frac{dQ(\mathbf{p_M} + t\mathbf{1_M})}{dt} = \sum_{j=1}^{M} \frac{\partial Q_M}{\partial p_j}$. Thus, if we observe the change in market output when all prices are increased simultaneously, we do not need to estimate each partial derivative separately and we can directly estimate $\left. \frac{dQ_M}{d\mathbf{p}} \right|_M$. When we take the model to the data, we assume a tax change is passed on to the consumers symmetrically across products, therefore the tax inducing the kind of uniform price change that allows to estimate the total derivative $\left. \frac{dQ_M}{d\mathbf{p}} \right|_M$ without having to estimate the partial derivatives (which would require a different kind of independent variation in the price data).

Finally, as $M \to J$ and using that $s = dp$ we can rearrange the last equation to get the analogous expression to equations (11) and (12):

$$d = \left(\frac{dp}{dQ} - \left. \frac{d\mathbf{p}}{dQ_J} \right|_J \right) dQ$$

Similarly, under symmetric pass-through a first order approximation to the price effect is $-Qdp$.[13]

3.2 Probabilistic Entry

In this section, we extend the symmetric firm model to allow for probabilistic entry.[14] We assume that nature draws a fixed cost. Let the equilibrium price and quantity functions be given respectively by $p(J)$ and $q(J)$. Assume that for every draw of the fixed cost, there is a uniquely determined number J of firms that enter the market. Then the distribution of fixed costs determines an equilibrium

[12] A class of models for which we obtain symmetric pass-through is that of linear-quadratic revenues with constant marginal costs.

[13] When there is one large incumbent with several products and several small incumbents with one product each, it might be unreasonable to assume that pass-through is symmetric.

[14] In light of propositions 1 and 1', everything that follows goes through for both the continuous and discrete choice models.

distribution F of variety J and consumer surplus from an ex ante perspective is given by:

$$CS = \int \int_p^{\infty} Q(s, J) ds dF(J) \tag{15}$$

Moreover, when there is an exogenous change in variety from the distribution F_1 to F_2 we may calculate, for the discrete case, that the variety effect is:

$$\Lambda = \int \int_p^{\infty} Q(s, J) ds dF_2(J) - \int \int_p^{\infty} Q(s, J) ds dF_1(J) \tag{16}$$

Suppose there exists $d(J_2, J_1)$ such that $Q_{J_2}(p + (s + d(J_2, J_1)) = Q_{J_1}(p)$ for all s. Let the conditional distribution of new variety be given by $F_{2|1}(J_2|J_1)$. Then,

$$\begin{aligned}
\Lambda &= \int \int \left[\int_p^{\infty} Q(s, J_2) ds - \int_p^{\infty} Q(s, J_1) ds \right] dF_{2|1}(J_2|J_1) dF_1(J_1) \\
&= \int \int \left[\int_0^{d(J_2, J_1)} Q(p + s, J_2) ds \right] dF_{2|1}(J_2|J_1) dF_1(J_1) \\
&= \int \int \left[\int_0^{d(J_2, J_1)} Q(p - d(J_2, J_1) + s, J_1) ds \right] dF_{2|1}(J_2|J_1) dF_1(J_1) \\
&\approx \int \int d(J_2, J_1) dF_{2|1}(J_2|J_1) Q(p, J_1) dF_1(J_1) \\
&= E[d(J_2, J_1) * Q(p, J_1)]
\end{aligned}$$

Thus, we obtain the familiar formula for the variety effect in terms of the product of the aggregate demand and the expected vertical shift of the inverse demand, the second of which is the average change in willingness to pay. Similarly, letting the bars denote expectations:

$$\begin{aligned}
d\bar{Q} &= E_{F_2}(Q(p(J), J) - E_{F_1}(Q(p(J), J)) \\
&= E_{F_2}(Q(p(J) + d(J, J_0), J_0) - E_{F_1}(Q(p(J) + d(J, J_0), J_0)) \\
&\approx E_{F_2 - F_1}\left[\frac{\partial Q}{\partial p}(p_0, J_0) * (p(J) + d(J, J_0) - p_0) \right] \\
&= \frac{\partial Q}{\partial p}(p_0, J_0) E_{F_2 - F_1}(p(J) + d(J, J_0)) \\
&= \frac{\partial Q}{\partial p}(p_0, J_0)(d\bar{p} + E(d))
\end{aligned}$$

Therefore $E(d) = \left(\frac{dp}{d\bar{Q}}|_J - \frac{d\bar{p}}{d\bar{Q}} \right) d\bar{Q}$ corresponds to the probabilistic version of expression (12).

3.3 The Principle of Le Chatelier and Parallel Inverse Demands

In this section, we extend the model by incorporating an outside market represented by the variable y and we assume the consumer can only adjust y in the long run. We start from a continuous choice model where all firms in the inside market are symmetric, we denote p the symmetric equilibrium price of the inside market, and Q the aggregate quantity.

Let $u(Q, y, J) - pQ$ be the utility function of the consumer and assume u is supermodular and quasiconcave. Let

$$Q^*(y, p, J) = \underset{Q}{\operatorname{argmax}}\, u(Q, y, J) - pQ$$

be the aggregate demand of the inside good conditional on (p, y, J), and let

$$y^*(p, J) = \underset{y}{\operatorname{argmax}}\, u(Q^*(y, p, J), y, J) - pQ^*(y, p, J)$$

be the optimal choice of y given (p, J). Finally, define the long-run aggregate demand $Q(J) = Q^*(y^*(p(J), J), p(J), J)$.

Observe the long-run change in aggregate demand for the inside market given an exogenous change in variety J has three components:

$$\frac{dQ(J)}{dJ} = \frac{\partial Q^*}{\partial p}\frac{dp(J)}{dJ} + \frac{\partial Q^*}{\partial y}\frac{dy(p(J), J)}{dJ} + \frac{\partial Q^*}{\partial J} \tag{17}$$

the indirect effect of variety through equilibrium price p, the indirect effect of variety through the outside variable y, and the direct effect of variety J.

Assume the following parallel inverse demands condition:

Assumption. *(Parallel Inverse demands) For all J and all y there exists d such that for all p then $Q(y, p, J) = Q(y_0, p + d, J_0)$.*

In particular, for all J there exists $d(J)$ such that $Q(J) = Q(y_0, p(J) + d(J), J_0)$. Then

$$\begin{aligned} dQ &= Q(J_1) - Q(J_0) \\ &= Q(y_0, p_1 + d, J_0) - Q(y_0, p_0, J_0) \\ &\approx \frac{\partial Q^*}{\partial p} * (dp + d) \end{aligned}$$

And so we can calculate the vertical shift

$$d \approx \left(\frac{dp}{dQ^*} - \frac{dp}{dQ} \right) * dQ \tag{18}$$

Define the indirect utility function

$$w(y, p, J) = u(Q^*(y, p, J), y, J) - pQ^*(y, p, J)$$

and note from the consumer perspective in a long-run equilibrium welfare is $v(J) = w(y^*(p(J), J), p(J), J)$. Taking the first order conditions:

$$\begin{aligned} \frac{dv(J)}{dJ} &= \frac{\partial w}{\partial p}\frac{dp}{dJ} + \frac{\partial w}{\partial J} + \frac{\partial w}{\partial y}\frac{dy}{dJ} \\ &= -Q\frac{dp}{dJ} + \Lambda + \frac{\partial w}{\partial y}\frac{dy}{dJ} \end{aligned}$$

Furthermore, the parallel inverse demands condition implies $-Q * d \approx \left(\Lambda + \frac{\partial w}{\partial y}\frac{dy}{dJ} \right) dJ$ and so

$$dv(J) \approx -Q * (dp + d) \tag{19}$$

In other words, we can estimate the welfare effect in (19) by estimating pass-through and the vertical shift parameter through equation (18). To estimate the second, we need the short-run slope of demand (keeping both variety J and the outside market demand y fixed) and the long-run slope of demand when both y and J are adjusted. However, estimating the vertical shift parameter is not enough to estimate the variety effect, Λ, since the vertical shift includes indirect effects of variety through the outside market y. An application of the Le Chatelier Principle (Milgrom and Roberts 1996) shows the slope of demand in the very long run (when both J and y are adjusted) is steeper than when only variety J adjusts, therefore $-Q * d$ would be overestimating Λ. In summary, the love for variety assumption and the Le Chatelier Principle together imply the following bounds:

$$0 \leq \Lambda \leq -Q * d'(J)$$

We have shown how to apply the parallel demands assumption in a model with an outside market y to calculate the welfare effect $\frac{dv}{dJ}$ with the reduced form estimates that are analogous to those used in the baseline model. If we are interested in calculating the variety effect Λ we need one more estimate: the long-run slope of demand where J is variable but y is kept constant $\left.\frac{dQ}{dJ}\right|_y = \frac{dQ^*(y,p(J),J)}{dJ}$. Then

$$\Lambda = Q * \left(\frac{\frac{dp}{dJ}}{\frac{dQ}{dJ}\Big|_y} - \frac{1}{\frac{\partial Q^*}{\partial p}} \right) * \frac{dQ}{dJ}\Big|_y$$

4 Applications

We now consider several applications of our model. First, we revisit the classic question of whether free-entry is efficient and show how one can shed light on this question using reduced-form empirical methods. Second, we consider the marginal welfare gain or loss of a small tax change in the context of product variety and show to empirically implement it. We begin by describing firms, market structure and the government.

4.1 Firms and Government

We start by assuming each firm j produces a single product according to the cost function $c_j(q_j) = c(q_j) = cq_j + f$ which is identical for all firms.[15] Each firm faces an valorem tax on its output τ. A given firm makes two decisions. First, each firm decides whether to enter the market given a fixed cost of entry. Second, each firm chooses p_j to maximize profits:

$$\begin{aligned} \max_{p_j} \pi_j &= p_j(1-\tau)q_j(p_1 \ldots, p_J) - cq_j(p_1 \ldots, p_J) - f \\ \text{s.t. } & \frac{\partial p_k}{\partial p_j} = \vartheta \text{ for } k \neq j \end{aligned}$$

The first-order condition for p_j is given by

$$(1-\tau)q_j + (p_j(1-\tau) - c)\left(\frac{\partial q_j}{\partial p_j} + \sum_{k \neq j} \frac{\partial q_j}{\partial p_k}\frac{\partial p_k}{\partial p_j} \right) = 0$$

We allow for different forms of behavior by letting $\frac{\partial p_k}{\partial p_j} = \vartheta$, for $k \neq j$, parametrize the degree of competition. For example, by setting $\vartheta = 0$ we obtain Bertrand competition and setting $\vartheta = 1$ we obtain perfect collusion. This is related to the way Weyl and Fabinger (2013) model competition, although they focus on tax incidence and pass-through with a fixed number of firms.[16] The conjectural variation terms only make sense when they correspond to static solution concepts or are

[15] One can define a more general cost function $c(q) + f$ for a convex variable cost function $c(q)$ and all the results go through.

[16] In the homogeneous good conjectural variations model, the first order condition is given by

$$p(1-\tau) + \frac{dp(1-\tau)}{dQ}(1+\theta) - c = 0$$

reduced forms of truly dynamic models (see Vives 2001, Riordan 1985) or supply function equilibria (Hart 1982). We do not take a stance on which is the dynamic model that ϑ captures in reduced form, instead proving that our evaluation of welfare is robust to any of the specifications that can be modeled this way.

In a symmetric equilibrium, $p_1 = p$ solves:

$$(1-\tau)q_1(p_1,p,\ldots,p) + (p_1(1-\tau)-c)\left(\frac{\partial q_1(p_1,p,\ldots,p)}{\partial p_1} + (J-1)\vartheta\frac{\partial q_1(p_1,p,\ldots,p)}{\partial p_2}\right) = 0$$

We assume the left hand side $\frac{\partial \pi_1}{\partial p_1}(p_1,p)$ is strict single crossing (from above) in p_1 and decreasing in p so that a unique symmetric equilibrium $p(J,\tau)$ exists.[17] Furthermore, we require that $\pi_j(p(J,\tau),J,\tau)$ be decreasing in J. Then, the "long run" number of firms J^* is determined by the free-entry condition $\pi_j(p(J^*,\tau),J^*,\tau) = 0$:

$$p(J,\tau)(1-\tau)q(p(J^*,\tau)) - cq(p(J^*,\tau)) - f = 0 \tag{20}$$

Finally, government revenue is given by $R = \tau * p * Q$ and social welfare W is defined as the sum of consumers' surplus (CS), producers' surplus (PS) and government revenue (R).

4.2 Socially Optimal Product Variety

It is a well-known result that the number of firms in a free-entry equilibrium, may diverge from the socially optimal number of firms (Spence 1976, Dixit and Stiglitz 1977, Mankiw and Whinston 1986, Anderson, de Palma and Nesterov 1995). Observe the marginal welfare gain of variety is given by:

$$\frac{\partial W}{\partial J}(J(\tau),\tau) = \Lambda + \pi + \tau pq + (p-c)J\frac{\partial q}{\partial J} \tag{21}$$

The private optimum is determined by equation (20), where free entry drives profits to zero; thus, we see that the private and social optimum diverge whenever $\Lambda+\tau pq+(p-c)J\frac{\partial q}{\partial J} \neq 0$. The first term is the variety effect and reflects the fact that firms create consumer surplus when they enter, a value which they may not completely internalize if they cannot extract all surplus. The second effect is the gain in government revenue which is a second externality not internalized by firms. Finally, the last

where $1+\theta \equiv \frac{dQ}{dq}$. The model nests various forms of competition such as Cournot ($\theta = 0$), Bertrand ($\theta = -1$), and perfect collusion ($\theta = J-1$) which, of course, gives the monopoly outcome.

[17] The case of strategic complementarities, where $\frac{\partial \pi_1}{\partial p_1}(p_1,p)$ is increasing in p allows for the existence of multiple symmetric equilibria, in that case assume there is a continuous and symmetric equilibrium selection $p(J,\tau)$.

term is the business-stealing effect which arises because entry affects output per firm, if q increases then entry is business enhancing otherwise entrants are stealing business from incumbent firms, in any case there is an externality imposed to the other firms (Mankiw and Whinston 1986). If the number of firms in the free-entry equilibrium diverges from the social optimum it depends on the relative size of the positive and negative externalities, starting from a benchmark without taxes ($\tau = 0$), socially optimal variety is determined by balancing the variety effect with the business-stealing effect.

Empirical Implementation. In general, theory cannot determine where $\frac{dW}{dJ} \gtreqless 0$. Thus, whether there is excessive or insufficient entry is an empirical question. There have been some attempts to tackle this question in the literature (Berry and Waldfogel 1999, Berry, Eizenberg and Waldfogel 2015). These papers mostly consider a structural approach by specifying the utility function and the nature of firm competition. In this paper, we pursue a complementary approach. We show that one may use *exogenous* variation in variety entry to identify whether there is too little or too much variety. Ignoring government revenue effects, the logic is straightforward: with an exogenous change in J we can use Theorem 1 to identify Λ. The key challenge is identifying the business-stealing term $(p-c)J\frac{\partial q}{\partial J}$. Although $\frac{\partial q}{\partial J}$ is estimable, one requires a measure of the social value of this output, $p - c$.[18] It turns out that this can be pinned down by a price effect using a free-entry envelope condition.

To fix ideas, consider a cost shifter, τ. In general, any cost shifter is valid, but we focus on taxes since this is the empirical application we consider below. The reduced-form objects we focus on are the short-run and long-run price effects, output effects and variety effects. First, we note that the variety effect can be pinned down as follows:

$$\Lambda = -Q \left[\frac{dp}{d\tau}\Big|_J \frac{\frac{dQ}{d\tau}}{\frac{dQ}{d\tau}\Big|_J} - \frac{dp}{d\tau} \right] \frac{1}{\frac{dJ}{d\tau}} \tag{22}$$

The intuition for the variety effect is illustrated in Figure 4. Here we consider a small increase in taxes. As discussed above, the base of the rectangular area is given by pre-existing output before the tax change. The height of the rectangle is given by the difference between "long-run" change in price as a result of the tax change and the "short-run" change in price re-scaled by the ratio of the long-run output effect to the short-run output effect of the tax. The re-scaling serves to extend the price effect up the demand curve so that it's measured at the long-run output level. The identification of the variety effect thus comes from a policy instrument that shifts marginal costs (such sales taxes), and is observed in a setting where variety is held constant and in a setting where

[18] If the consumer price and producer price are equal, this is also the firm's markup.

variety can respond endogenously to the policy change, subject to a free entry condition. In the case of a standard CES demand model, both of the short-run and long-run effects are linked together by a single elasticity parameter. What our framework highlights is that in order to separately identify the demand elasticity (holding variety constant) from the variety effect, one requires two separate sources of variation. Conceptually, one needs an instrument to trace out demand holding variety constant and an instrument for variety. Practically, finding plausibly exogenous shocks to variety is likely to be challenging. Therefore, instead, we trace out the "long run" demand curve that allows both prices and variety to respond to cost shifter and show that this can be combined with the "short run" demand curve to identify the variety effect.

With the variety effect in hand, we next consider the business-stealing effect. Totally differentiating the free-entry condition with respect to the tax, one can solve for $p-c$ and show that:

$$(p-c)J\frac{\partial q}{\partial J} = \frac{\frac{d(\tau pQ)}{d\tau} - Q\frac{dp}{d\tau} - \tau pq\frac{dJ}{d\tau}}{\frac{dQ}{d\tau} - q\frac{dJ}{d\tau}}\frac{\frac{dQ}{d\tau} - q\frac{dJ}{d\tau} - \frac{dQ}{d\tau}|_J}{\frac{dJ}{d\tau}} \tag{23}$$

This condition shows that we can recover the business-stealing effect using the short-run and long-run effects of the tax on firm output, prices, and firm expenditures. To see the intuition for this expression, note that if the zero profit condition holds before and after the policy change, then

$$-(p(1-\tau)-c)\frac{dq}{d\tau} = q\frac{d(p(1-\tau))}{d\tau}$$

Thus, if there is a business-stealing effect so that per-firm output goes down in response to the policy reform $\left(\frac{dq}{d\tau} < 0\right)$, in a long-run equilibrium, these losses have to be offset by higher revenue in the form of higher per-unit prices. By re-arranging this condition, we get $p-c = \frac{-q\frac{dp}{d\tau} + \frac{d(pq\tau)}{d\tau}}{\frac{dq}{d\tau}}$, which is the first term in equation (23). The second term follows naturally from the fact that $q = q(J(\tau),\tau)$ and $\frac{dQ}{d\tau} = q\frac{dJ}{d\tau} + J\frac{dq}{d\tau}$. Note that we do not require direct estimates of costs and markups which may be difficult to measure in practice. By comparing the estimated variety effect and business-stealing effect, we can determine whether there is too little or too much variety.

4.3 Marginal Welfare Gain or Loss of a Tax Change

We consider a government that imposes an ad-valorem tax (τ) on each product in the market (but not the outside good). This generates revenue $R = \tau pQ$. Welfare is defined as $W = CS + PS + R$ where CS and PS are aggregate consumer and producer surplus, respectively. The marginal welfare

gain is:

$$\frac{dW}{d\tau} = \frac{dCS}{d\tau} + \frac{dPS}{d\tau} + \frac{dR}{d\tau} \tag{24}$$

First, consider the case of competitive pricing $p(1-\tau) = c$ and socially optimal variety (given the tax τ and given competitive pricing) for all m. In this case, the marginal welfare gain from increasing the tax rate τ is $\frac{dW}{d\tau} = \tau p \left.\frac{dQ}{d\tau}\right|_J$.[19] Next, assume pricing decisions are left to firms and variety can be set optimally by the government, given the taxes (τ) and given firms' pricing decisions. The marginal welfare gain from increasing the tax is $\frac{dW}{d\tau} = (p-c) \left.\frac{dQ}{d\tau}\right|_J$.[20] Finally, assume that firms control both pricing and entry decisions. The marginal welfare gain becomes:

$$\frac{dW}{d\tau} = (\Lambda - f)\frac{dJ}{d\tau} + (p-c)\frac{dQ}{d\tau} \tag{25}$$

To see the intuition for this, consider the case where $\Lambda = 0$ which corresponds to homogeneous products.[21] We see that the new term added to the welfare formula is $-f\frac{dJ}{d\tau}$. In this case, a tax cut leads to inefficient entry since the new output produced as a result of the tax could have been produced more cheaply by incumbent terms. The only modification when there is product variety is that entry might not be inefficient if consumers place a sufficiently high value on the new products.[22]

Empirical Implementation The above expressions do not make use of the free-entry condition and may be difficult to implement empirically since they require estimates of firms' costs and markups. When the free-entry condition holds before and after the policy change the marginal firm $J(\tau)$ earns zero profits. In the symmetric model this implies that producer surplus $PS = J\pi = 0$ before and after the policy change. To derive a condition that is more easily implementable, we impose the assumption that after a tax change, profits are driven to zero to show that the long-run welfare gain from marginally increasing taxes is:[23]

$$\frac{dW}{d\tau} = -Q \left.\frac{dP}{d\tau}\right|_J \frac{\frac{dQ}{d\tau}}{\left.\frac{dQ}{d\tau}\right|_J} + PQ + \tau\frac{d(PQ)}{d\tau} \tag{26}$$

To implement the marginal welfare gain in (26), we only require behavioral responses to taxes.

[19] See Harberger (1964), Chetty (2009)

[20] See Auerbach and Hines (2001).

[21] This is considered in Besley (1989) and Auerbach and Hines (2001). Although these papers consider Cournot competition, our formulas are valid for a broader class of models.

[22] One can also show the following equivalent representation for the marginal welfare gain: $\frac{dW}{d\tau} = (\Lambda + \pi + \tau pq)\frac{dJ}{d\tau} + (p-c)J\frac{dq}{d\tau}$.

[23] In a model with asymmetric firms, we do not need to assume $PS = 0$ before and after the policy change, rather we would require $\frac{d\pi_j}{d\tau} = 0$ to get a similar formula. This is equivalent to $-(p_j(1-\tau) - c'(q_j))\frac{dq_j}{d\tau} = q_j\frac{d(p_j(1-\tau))}{d\tau}$, the condition we used to estimate markups in the previous section.

In particular, we do not need to estimate the distribution of random utility shocks, the fixed cost of entry f, the marginal cost of production c, or the market conduct parameter θ.

The empirical setting we consider below is about consumption of grocery products. Our data contains a classification which assigns products to "modules", which we may interpret as nests through the lens of our model. Here we briefly sketch out a more general model where each product belongs to a nest, $m \in M$, where we take the nesting structure as exogenous.[24] This modeling structure closely follows Sheu (2014). Products within a nest are taken to be more substitutable than products between nests. We assume that expenditures are exhausted across all grocery store products. Thus, while the random utility model considered above defined the "reference good" as the no purchase option, in this model, we assume that the outside option of not purchasing a product in some nest is choosing the most preferred option among products in all other nests. Thus, any effects of changing expenditures in a given nest would be captured by shifting expenditures to a different nest. In this setting, some nests are subject to taxes and other nests are not. We denote the set of taxable nests as M_T.

$$\frac{dW}{d\tau} = \sum_{m=1}^{M} \left(\Lambda_m \frac{dJ_m}{d\tau} - Q_m \frac{dp_m}{d\tau} \right) + \sum_{n \in M_T}^{M} \frac{d(\tau p_n Q_n)}{d\tau} \tag{27}$$

The first term shows that one needs to evaluate the effect of the tax on variety and prices in all nests. The second term shows that one needs to consider the fiscal externality which only requires measuring behavioral responses in taxable nests. Under two assumptions, we can retrieve the marginal welfare gain in (26). First, we assume symmetry within M_T and within $M \backslash M_T$. Second, we assume that prices, quantity and variety in untaxed nests do not respond to taxes.

5 Data

The previous section showed that the sufficient statistics for welfare are $\frac{dJ}{d\tau}$, $\frac{dp}{d\tau}$, $\frac{dp}{d\tau}\Big|_J$, $\frac{dQ}{d\tau}$, and $\frac{dQ}{d\tau}\Big|_J$. To estimate these objects, we rely on several data sets. For our measures of p, Q, and J, we rely on Nielsen's Retail Scanner Data. To measure τ, we rely on hand-collected local sales tax data in the U.S. This section describes how we constructed the final samples used in our empirical analysis. Further details on data construction are provided in the Appendix.

[24] For a more detailed analysis, we refer the reader to the Appendix.

5.1 Nielsen Retail Scanner Data

We obtained the Nielsen scanner data from the Kilts Marketing Data Center at the University of Chicago Booth School of Business. The micro data records weekly prices and quantities by product at the barcode level (Universal Product Code, UPC) for over 35,000 stores from approximately 90 retail chains across the United States (except for Hawaii and Alaska), covering the years 2006-2014.[25] Each store, geolocated at the county level, is assigned one of five possible store types ("channels"), and can be matched with its parent chain.[26] Products are organized in a hierarchical structure: There are over 2.5 million different UPCs, which are categorized into approximately 1,200 *product-module*s. Each module is then assigned to one of roughly 120 *product-groups*, which in turn is part of one of 10 broader *product-departments*. Appendix Table 1 shows a few examples of UPCs included in the retail data.

The Retail Scanner dataset's coverage of total U.S. sales volume varies across locations and store-types. For instance, it covers more than half of the total sales volume of U.S. grocery stores, but only 2 percent of sales in convenience stores. We impose several sample restrictions to address any potential bias caused by differential coverage of sales by type of products across regions. First, we restrict our sample of stores to grocery stores.[27] Second, we only keep modules sold in all 48 continental states. Finally, in our main specification, we restrict the sample to the top selling modules that rank above the 80th percentile of total U.S. sales in the distributions of food and non-food modules. These modules account for almost 80% of the total value of sales in grocery stores in the scanner data.

From the scanner data, we construct two samples that we use for our empirical analysis: 1) repeated cross-sections where the unit of observation is at the store-module-year level, 2) panel data where the unit of observation is at the store-module-quarter level. For each sample, we generate measures of price (p), expenditure (pQ) and product variety (J) at the module level. All of our regressions below will be based on expenditures, not output. We discuss below how we obtain demand elasticity estimates from our expenditure and price elasticity estimates.

To measure price for each module-store-period combination, we take several steps. First, for each UPC in each store, we aggregate weekly revenue and output over time (yearly for cross-section and quarterly for time-series) and compute the average price by taking the ratio of these. This delivers a price for each UPC-store combination per period. Second, for each module, we regress

[25] Products without a barcode such as random weight meat, fruits, and vegetables are not included in the data set.

[26] The five channels are grocery, drug, mass merchandise, convenience and liquor stores. Each store and each parent chain has a unique identifier. Retail chain names are confidential and unknown to researchers.

[27] The distribution of stores by store-type varies substantially across regions. Restricting our sample to grocery stores ensures that compositional changes across regions are not driving the results.

the log of average price on a set of UPC fixed effects and store-by-period fixed effects. Since these regressions are estimated separately for each module, the store-by-period effects by construction are permitted to vary across modules. The estimated module-store-year and module-store-quarter fixed effects serve as the long-run and short-run price indices, respectively. To measure expenditures, we aggregate revenue across all UPCs to the module-store-year level for the cross-sectional analysis and to the module-store-quarter level for the time-series analysis. Finally, product variety is measured as the count of UPCs with positive sales in a given store over the relevant period of time (year for long-run, quarter for short-run), separately for each module.[28] Finally, we normalize our time-series measures. For price, we divide by the average price in each store-quarter. For expenditures, we divide by total expenditures in each store-quarter. Thus, below we will refer to this measure as an expenditure share. For variety, we normalize by the store-quarter average number of varieties. As we describe more fully below, we normalize these variables to account for store-level trends in our regression analysis. An alternative way of addressing this is include store-by-period fixed effects, however this is computationally burdensome due to the large number of fixed effects that need to be estimated.

5.2 U.S. Sales Tax Exemptions and Rates

The second source of data we use is a hand-collected monthly panel of local (county and state) sales tax *rates* and state-level *exemptions*, which vary at the product-module level, covering the years 2006-2015.[29] In general, exemptions are set by states and are module-specific.[30][31] The general rule of thumb is that food products are tax-exempt and non-food products are taxable. However, there are important exceptions to this rule. First, several states tax food at the full rate or a reduced rate. Second, in a few states, food products are exempt from the state-level portion of the total sales tax rate, but remain subject to the county-level sales tax. Third, in some cases where food is tax-exempt, there is a tax that applies at the product-module level. For example, prepared foods are subject to sales taxes in many states. Finally, some states exempt some non-food products from sales taxes. Our final exemption sample is at the county-module-month level, however it should be noted that changes in exemptions over time are very rare during our sample period. For tax rates,

[28] The definitions of price and variety are similar to the corresponding definitions in Handbury and Weinstein (2015).

[29] We use sales tax rates from 2015 to test whether there is an anticipation response to changes in sales tax rates and do not find any evidence that there is. Results are available upon request.

[30] The Online Appendix describes the sources we used in determining exemptions and rates.

[31] There are a handful of exceptions to this. Colorado, for example, allows each county to decide whether to subject food to the county-level portion of the sales tax rate.

we collected monthly state-level and county-level rates.[32]

There are several possible sources of measurement error in our sales tax rates. First, we do not incorporate county-level exemptions or county-specific sales surtaxes that apply to specific products or modules, although our understanding is that these cases are uncommon. Second, there may be measurement error coming from our exemption definitions and how we assigned a taxability status to each module, which in some cases required a subjective judgment based on interpreting the text of the state sales tax law. While the bulk of the variation in taxes occurs at the module level or higher, there are some instances where taxability varies within module. For example, in New York, fruit drinks are tax exempt as long as they contain at least 70% real fruit juice, but are subject to the sales tax otherwise. Therefore, some products in Nielsen's module "Fruit Juice- Apple", may or may not be taxed in New York, but all are considered eligible for the sales tax exemption in our database since we cannot readily identify the real fruit juice content.[33]

As a final step, we merge the effective sales tax rates to the Nielsen scanner data. This requires aggregating the sales tax data to the level of the scanner data. For the cross-sectional analysis, we obtain yearly sales tax rates by relying on the effective rate on September 1 of a given year.[34] For the time-series analysis, we use the rate effective at the mid-point of each quarter (February for quarter 1, May for quarter 2, etc). We then merge the sales tax rates to the scanner data by product-module, county and time. Our final cross-sectional and panel samples cover over 10,000 grocery stores, and contain price, expenditures and variety for 198 modules in 1,625 counties.[35]

Table 1 presents the tax status of the top selling food and non-food modules in our sample. There are several noteworthy observations. First, modules such as soft drinks, ice cream, and candy are taxed in states that exempt food, like Connecticut, Florida, and Wisconsin. Second, several non-food modules are exempt from taxes. For example, toilet tissue and diapers are exempt in New Jersey and Pennsylvania and magazines are exempt in Maine, Massachusetts, New York and Oklahoma. This provides an additional source of variation in tax liabilities across states which is useful for identifying the long-run effect of taxation as we discuss more fully below.

[32]Some cities and other localities also impose an additional local sales tax rate. We do not incorporate rates that apply to areas smaller than counties.

[33]In cases where it is impossible to tell whether the majority of products in a given module are subject to the tax or not, we code the statutory tax rate as missing. This results in excluding less than 3% of the observations in our sample.

[34]Most rate changes occur either on January 1 or July 1.

[35]The panel includes stores in 1,625 counties, but the number of stores and counties varies slightly across years.

5.3 Descriptive Statistics

Figure 5 shows the cross-sectional distribution of the total sales tax rate (state + county) in September 2008. There is substantial cross-sectional variation in sales tax rates ranging from zero in Montana, Oregon, New Hampshire and Delaware to a maximum rate of 9.75 percent in Tennessee. Table 2 compares the observable characteristics of low and high tax states. It presents annual descriptive statistics for the year 2008 for simplicity as the patterns are very similar in the other years of our sample. Column (1) reports means and standard deviations for all counties and columns (2) and (3) report results for high and low sales tax counties, respectively. The typical county in our sample has roughly $75 million (U.S. dollars) in yearly grocery store sales (for the top 20 percent selling modules) with about 6.5 stores per county. Food modules account for roughly 75 percent of total annual sales on average. There are roughly 100 varieties sold in a typical module in a typical grocery store over a year. Turning to taxes, the average combined county and state sales tax rate is 6.3 percent while the average tax rate on food products alone is 1.6 percent. Finally, the typical county has a population of about 165,000, a household median income of $44,000 and roughly 50 percent with a high school degree or less.

Turning to columns (2) and (3), we see that grocery stores are very similar between high-sales tax and low-sales tax counties on a number of dimensions although low-tax counties have larger sales per store ($10 million versus $9.3 million). Locations with sales tax rates above the median exhibit lower rates of excise taxes on alcohol and cigarettes, tend to be more populous, have more grocery stores, and cover smaller territories. In column (4), we regress the county characteristics on the sales tax rate. The reported coefficients indicate that sales tax rates are negatively associated with variety and price levels, but positively correlated with the food share of sales. These regressions also provide further evidence that counties with high sales tax rate have, on average, lower rates for other types of taxes.

Finally, in Appendix Figure 1, we present visual evidence on the distribution of food tax exemptions across states. In general we see that food taxability status is spatially correlated. For example, most states that tax food are located in the South or in the Midwest. In regressions below, we evaluate the robustness of our results to controlling for module fixed effects interacted with census region fixed effects.

6 Empirical strategy

In this section, we discuss the research design we use to study the effects of taxation. Later we discuss how we map our reduced-form empirical estimates to our sufficient statistics in order to implement our welfare formula in equations (22), (23) and (26).

6.1 Long-run Effects of Taxation

Grocery stores sell some products that are subject to the sales tax and other products that are exempt, generating within-store differences in after-tax prices between products. Our empirical strategy exploits cross-sectional variation in sales tax rates across counties interacted with tax status across modules. Since we are expoiting *cross-sectional* variation which corresponds to the steady-state, we interpret the resulting estimates as "long-run" elasticities. In particular, our estimates will incorporate the endogenous response of product variety to sales taxes.[36]

To see how this source of variation can be used to identify the long-run effects of taxation, consider the following example. Suppose that food modules are exempt in all locations and non-food products are taxed everywhere, and sales tax rates are set by legislators independently of local differences in sales/prices between food and non-food products. In this case, we can recover a consistent estimate of the long-run elasticity of taxation by estimating the following difference-in-differences (DD) regression model:

$$\log y_{mrcs} = \beta^{LR} \left(\log(1+\tau_{cs}) \times Nonfood_m\right) + \delta_r + \delta_m + \varepsilon_{mrcs} \tag{28}$$

where the outcome y_{mrcs} is either price, expenditures , or product variety . The terms δ_r and δ_m are store fixed effects and module fixed effects, respectively, $Nonfood_m$ is a dummy variable for non-food modules and τ_{cs} is the sales tax rate in county c.[37] Any county-level differences that do not vary across modules are absorbed by the store fixed effects. Any systematic differences in taxability across modules are soaked up by the module fixed effects. The coefficient of interest is β^{LR} . Under the assumptions stated above, we can use OLS to estimate the long-run causal effect of taxes on prices, expenditures, and variety.

Our preferred specification builds on the DD specification by additionally incorporating variation

[36]Similarly, Atkin, Faber and Gonzales (2016) use cross-sectional variation in store-level prices to estimate long-run elasticities of substitution across stores.

[37]Note that the main effects of the $Nonfood_m$ indicator and the sales tax rate τ_{cs} are absorbed by the inclusion of module and store fixed effects in the model.

in tax rates across modules within the broad categories of food and non-food products. This mainly arises due to product-specific exemptions, such as the taxation of candy products in some states or the exemption of diapers. In this case, the long-run estimating equation is given by:

$$\log y_{mrcs} = \beta^{LR} \log(1 + \tau_{mcs}) + \delta_r + \delta_m + \varepsilon_{mrcs} \tag{29}$$

The main difference between equations (28) and (29) is the definition of the sales tax rate. For the latter equation, taxes may vary across food (non-food) products within a store, hence the tax rate is also subscripted by m. The long-run parameter β^{LR} is identified under the assumption that the within-store *differences* in statutory rates across modules do not systematically vary across counties with within-store differences in unobservables. For example, our estimates of β^{LR} for expenditures would be biased upwards if jurisdictions where the consumption share of unhealthy food products (e.g., candy, soft drinks) is relatively high responded by specifically subjecting these goods to the sales tax.

Our final empirical strategy to estimate long-run effects is to implement a border-design following Holmes (1998), Dube, Lester and Reich (2010), and Hagedorn, Manovski and Mitman (2016). We restrict the sample of stores to those located in contiguous counties located on opposite sides of a state border. Two contiguous counties located in different states form a county-pair d, and counties are paired with as many cross-state counties they are contiguous with. The estimating equations are modified such that module fixed effects are now county-pair specific:

$$\log y_{mrcsd} = \beta^{LR} \log(1 + \tau_{mcs}) + \delta_r + \delta_{md} + \varepsilon_{mrcs} \tag{30}$$

To estimate equation (30), the original dataset is rearranged by stacking all pairs. For instance, a module-store cell located in county c appears as many times as the number of counties county c is paired with. Regressions are weighted by the inverse of the number of pairs a county is part of.

6.2 Short-run Effects of Taxation

The empirical strategy to estimate the short-run expenditure and price elasticities with respect to the sales tax exploits within-store time-series variation in tax rates over the 2006-2014 period. We interpret these as "short-run" because firm entry and exit is unlikely to adjust instantaneously to high-frequency variation in sales taxes, an assumption we empirically test below. The baseline short-run specification is the following:

$$\log y_{mrcst} = \beta^{SR} \log(1 + \tau_{mcst}) + \delta_t + \delta_{mr} + \delta_m \times t + \varepsilon_{mrcst} \tag{31}$$

where the unit of time (t) is a quarter, and δ_t and δ_{mr} are quarter and module-by-store fixed effects. In some specifications, we include a module-specific time trend $\delta_m \times t$ while in others we include module-by-quarter fixed effects, δ_{mt}.[38] The dependent variables used to estimate the elasticities of interest are normalized within store-time cells to account for module-invariant store-specific trends.[39] The identifying assumption is that states and counties do not *differentially* change effective sales tax rates across products endogenously with respect to changes in consumer demand. Additionally, we require that any quarter-to-quarter variation in product variety is unrelated to sales tax policy changes – an assumption we test below.

7 Results

In this section, we report our long-run and short-run elasticity estimates. We also use the empirical framework developed in the previous section to provide evidence in support of the symmetric price assumption.

7.1 Long-run Estimates based on Cross-Sectional Variation in Sales Taxes

As a first step, we estimate this simplified DD model in equation (28) on yearly cross-sectional data by restricting the sample to counties where food products are fully exempt from the state sales tax, and to modules that are either taxed or exempt in all stores in this subset of counties.[40] We estimate the cross-sectional model separately for each year between 2006 and 2014, and then take a simple linear combination of these nine coefficients. Coefficients of interest are shown in Table 3.

In all specifications, standard errors are clustered at the state-module level, the broadest level at which sales taxes are determined. The dependent variable is expenditures in columns (1) and (2), consumer price in columns (3) and (4), and variety in columns (5) and (6). Note that because we

[38] All results reported in this paper hold if months are used as the unit of time instead of quarters. Results are available upon request.

[39] This normalization is similar but not equivalent to including store-time fixed effects, which is computationally burdensome. The exact normalization procedure is described in details in the Appendix. Note that because our measures are in logs, the cross-sectional framework with store fixed effects yields identical results whether outcomes are normalized within store or not.

[40] The selection criteria for the difference-in-differences estimation sample is based on state-level exemptions and therefore includes stores located in a handful of states where food products are exempt from the state's sales tax but may remain subject to some local taxes.

use consumer prices, a coefficient of one in columns (3) and (4) implies full pass-through of the sales tax to consumers. In columns (1), (3) and (5), the specifications include store and module fixed effects. Our estimates across these columns indicate that the tax elasticity of expenditures is -0.840 (s.e. 0.574), and that there is slight undershifting of taxes onto consumer prices with a coefficient of 0.894 (s.e. 0.074). The elasticity of product variety with respect to sales taxes is -0.987 (s.e. 0.299). To address the concern that sales taxes are spatially correlated across regions of the U.S. in ways that may endogenously reflect the geographic distribution of consumer preferences, we consider a specification that allows module fixed effects to vary across census regions (columns (2), (4), and (6)). All of our results gain in precision with the inclusion of module-by-region fixed effects, and the point estimates are robust.

Next, we report results from our preferred specification (equation (29)) in Table 4. Standard errors are clustered within state-module cells. For expenditures, the elasticity is -0.683 (s.e. 0.255) and for variety it is -0.848 (s.e. 0.148) which are qualitatively similar to the corresponding estimates shown in Table 3, both for the baseline specifications and when allowing for region-specific module fixed effects. The price coefficient suggests that sales taxes are slightly overshifted to consumers with an elasticity of 1.145 (s.e. 0.036). All of the estimates reported in Table 4 are statistically significant at the 1% level.

7.1.1 Robustness of Long-Run Estimates

We next explore a series of robustness checks in Table 5. Columns (1) and (2) report our benchmark results from Table 4 for comparison. First, to further address spatial correlation of taxes, we turn to a specification that relies exclusively on module-specific exemptions. More precisely, we exclude all observations included in the difference-in-differences model. This sample thus includes: (a) all observations in counties where food is subject to a sales tax, as well as (b) for counties where food is generally exempt, the subset of modules for which there is some between-state variation in taxability status. The results, shown in columns (3) and (4) of Table 5, are in line with the baseline estimates. Second, we examine the robustness of our results to dropping small counties (with a population below 150,000), for which few stores are observed and are therefore more likely to be subject to sampling issues. These are reported in in columns (5) and (6) and again we find that our results are qualitatively similar. Third, to verify that the estimated effects are not driven by *counties* setting their local sales tax rates endogenously with respect to local consumer preferences, we instrument the county-level effective sales tax rate with the state-level effective rate (columns (7) and (8)) and again find similar results. Finally, in column (9), we include all Nielsen's modules that are observed

in all of the 48 continental states. Our estimates are consistent across all these specifications.

Table 6 presents the results of the border-counties analysis.[41] For all specifications, to account for spatial auto-correlation, standard errors are clustered two-way – by state-module as well as by border-segment-module (Cameron et al. 2011). In columns (1), (3) and (5), the estimating equation is equivalent to equation (29), but the sample is restricted to border-counties. This sample restriction has little impact on the magnitude of the coefficient for consumer prices, but yields slightly smaller point estimates for both expenditures and variety. We then allow module fixed effects to vary by county pairs in columns (2), (4) and (6). The implied tax elasticity of expenditures -0.649 (s.e. 0.142) is in line with our baseline results, as is the coefficient for prices 1.036 (s.e. 0.016). The effect of taxes on variety -0.304 (s.e. 0.070) is smaller in this specification, but remains statistically different from zero at the 1% level.

Both pooled cross-sectional results and "border county" results suggest large effect of taxes on variety, with elasticity roughly between 0.3 and 0.8. This estimate can be compared to some the recent estimates of the elasticity of variety to population in Handbury and Weinstein (2014) and Schiff (2015) which range from 0.3 to 0.6. In both cases (change in taxes and change in population), there is a change in aggregate demand which in turn affects vareity.

7.2 Short-run Estimates based on Time-Series Variation in Sales Taxes

To estimate the short-run effects of taxation, we use our quarterly panel covering 2006-2014. The results based on our main specification in equation (31) are reported in Table 7. Standard errors are clustered at the state-module level to allow for correlation across stores located in a given state and to adjust for serial correlation. Columns (1) to (2) contain our estimates for expenditures and columns (3) to (4) contain our results for prices. In columns (1) and (3), we include module, store and time fixed effects, module effects interacted with store effects and module-specific linear time trends. By including module effects interacted with store effects, we effectively shut down the cross-sectional variation in taxes and rely exclusively on within-store, across module time variation. Our estimates indicate an expenditure elasticity of -0.315 (s.e. 0.111) and a price elasticity of 1.024 (s.e. 0.038), consistent with full pass-through. We cannot reject the null hypothesis that the coefficient for prices is equal to one at conventional levels of statistical significance. In columns (2) and (4) we control for module-specific trends more flexibly by including module effects interacted with quarter

[41] There is evidence that border counties set local sales tax rate strategically to compensate for cross-border difference in state-level sales tax rates (Agarwal 2015). In unreported regressions, we verified that results are not driven by this possible source of bias by instrumenting the statutory county-level tax rate (τ_{mcs}) with its state-level value (τ_{ms}).

fixed effects instead of linear time trends. The expenditure elasticity remains negative but is smaller in magnitude -0.165 (s.e. 0.103), and the price elasticity is again statistically indistinguishable from one. Finally, as a placebo test we report results using variety as the outcome variable in columns (5) and (6). Reassuringly, the coefficient is statistically insignificant and close to 0, supporting our interpretation of the evidence as representing a short-run effect of taxation.

Overall, we find larger effects of taxes on variety and expenditures in the long-run, while pass-through estimates are similar in the long-run and in the short-run. The coefficient on price remains relatively close to one and is considerably smaller than estimates reported in Besley and Rosen (1999). In Section 8, we employ these empirical estimates to implement the sufficient statistics approach to welfare analysis described in Section 2.

7.2.1 Robustness of Short-Run Estimates

In Table 8 we explore the sensitivity of our short-run estimates to alternative samples and specifications. Columns (1) and (2) show our baseline estimates from Table 7 for comparison. In column (3), we include store-by-time fixed effects to flexibly account for any location-specific time trends. The expenditure elasticity -0.426 (s.e. 0.128) is slightly larger under this specification, while the point estimates for prices and variety are barely affected by the inclusion of these additional fixed effects. In columns (4) to (6), we restrict our sample to large counties and obtain very similar results for all three dependent variables.

To further assess the robustness of our short-run estimates, we also report results from a state border sample. In column (7) of Table 8, our main specification is estimated on the restricted state border sample, and column (8) includes pair-specific module trends. Reassuringly, these results closely mirror our main estimates, with coefficients for expenditures and prices slightly smaller in this sample.

7.2.2 Testing for Symmetric Pass-through

Section 3.1 showed that the key assumption for the validity of the sufficient statistics formula for the variety effect in the case of asymmetric prices is symmetric pass-through. We now empirically test this assumption by comparing pass-through rates across classes of products within each module. Our approach consists in partitioning each module into several categories of UPCs and estimating short-run pass-through parameters separately for each of these categories. This is equivalent to testing for heterogeneous effects of sales tax changes on consumer prices across categories of products.

We first test for symmetric pass-through across products with different price levels. To classify UPCs as either high- or low-price products, we first use our panel dataset from 2006-2014 to estimate UPC fixed effects by regressing pre-tax prices on UPC and store fixed effects, separately for each module. For each module, we define the average value of the UPC fixed effects that is time-invariant. We denote UPCs with a fixed effect above their module-specific mean as "high-price" products, and below-mean UPCs as "low-price" products. Next, we regress log prices on UPC fixed effects and store-by-quarter fixed effects interacted with the high/low price dummy variable. This delivers a quarterly panel data set of store-module price indices with two price indices for each module-store-quarter cell.

Our second test of symmetric pass-through is based on a comparison of products across brands with different market shares. Formally, for each module, we aggregate UPCs into brands and compute the market share for each brand. We then classify brands on the basis of their market shares in each module and define the top selling brand, the second selling brand and all remaining brands. Each UPC in the data is assigned to one of the categories based on that UPC's brand. Finally, for each module, we regress log price on UPC fixed effects and store-by-quarter fixed effects interacted with a categorical variable representing the brand popularity. This delivers a quarterly panel data set of store-module price indices with three price indices for each module-store-quarter cell.

Results for theses specification checks are reported in Table 9. The estimates are obtained by estimating our main short-run specification that includes module-store fixed effects and module-specific linear time trends (equation (31)). In columns (1) and (2) we calculate pass-through rates separately for high- and low-price UPCs. We fail to reject the null of equal pass-through rates (p-value>0.1). Similarly, in columns (3) to (5), we compare pass-through rates between top-selling brands and brands with relatively smaller market shares. We test for equality of the three estimates and again we cannot reject symmetric pass-through (p-value>0.17).[42]

[42]In unreported regressions, we re-estimate these pass-through rates but substitute module-time fixed effects for module-specific linear time trends. In these cases, the p-values for the null of symmetric pass-through are 0.235 and 0.252 for the price-based and the brand-based tests, respectively.

8 Welfare calibration

As a final step in the paper, we numerically implement our welfare formulas using our empirical estimates. First, note that we can re-write equation (26) as:

$$\frac{1}{pQ}\frac{dW}{d\tau} = -\frac{dlogP}{d\tau}\Big|_J \frac{\frac{dlogQ}{d\tau}}{\frac{dlogQ}{d\tau}\Big|_J} + 1 + \tau\frac{dlogPQ}{d\tau} \tag{32}$$

In order to implement our welfare formula, we need to measure $\frac{dlogQ}{d\tau}$ and $\frac{dlogQ}{d\tau}\Big|_J$. We get this by applying the chain rule and using our expenditure elasticity and price elasticity estimates, $\frac{dlogQ}{d\tau} = \frac{dlogpQ}{d\tau} - \frac{dlogp}{dt}$ and $\frac{dlogQ}{d\tau}\Big|_J = \frac{dlogpQ}{d\tau}\Big|_J - \frac{dlogp}{d\tau}\Big|_J$. Our calibrations are contained in Tables 10 and 11.[43] First, consider Table 10. Columns (1) to (5) provide a set of illustrative calibrations for our welfare formulas. The values of the inputs in these columns do not correspond to our empirical estimates and are intended only as illustrations to provide intuition. Across all columns, the marginal welfare gains/losses $\frac{J}{pQ}\frac{dW}{dJ}$ and $\frac{1}{pQ}\frac{dW}{d\tau}$ are the main objects of interest. For comparison, we also report the marginal welfare gain of taxes assuming that variety is fixed or efficient and prices are set competitively.

The first two columns illustrate the scenarios where either the variety effect or the business-stealing effect is 0. In column (1), we see that the short-run and long-run effects of taxes on prices and output are the same and so the variety effect is 0. In this case, the reduced-form estimates reveal that consumers do not value variety. Since the shift in the aggregate demand curve depends jointly on the variety response to taxes and the variety effect, there is no downward shift in the aggregate demand curve, despite the fact that taxes nevertheless affect entry/exit of firms since they affect profits given the fixed cost of entry. In this case, additional entry only reallocates output among firms, but does not cause an overall expansion of output. Clearly in this case, there is too much variety at current level of taxes since the business-stealing effect dominates (i.e., $dW/dJ < 0$). Turning to column (2), here we see that the business-stealing effect is 0 so that increase in variety only increases total quantity and does reduce quantity per firm; in this case only the variety effect remains. Clearly, in this case, variety is inefficiently low (e.g., $dW/dJ > 0$). The next set of columns illustrate the

[43] For the variety effect and business-stealing terms reported in Tables 10 and 11, we scale each term by J/pQ. Thus, the estimates reported in the table are rescaled versions of equations (22) and (23), respectively. Namely

$$\frac{J}{pQ}\Lambda = -\left[\frac{dlog(p)}{d\tau}\Big|_J \frac{\frac{dlogQ}{d\tau}}{\frac{dlogQ}{d\tau}\big|_J} - \frac{dlog(p)}{d\tau}\right]\frac{1}{\frac{dlog(J)}{d\tau}}$$

$$\frac{J}{pQ}(p-c)J\frac{\partial q}{\partial J} = \frac{p-c}{p}\frac{\partial log(q)}{\partial log(J)} = \frac{\tau\frac{dlog(pQ)}{d\tau} + 1 - \frac{dlogp}{d\tau} - \tau\frac{dlogJ}{d\tau}}{\frac{dlogQ}{d\tau} - \frac{dlogJ}{d\tau}}\frac{\frac{dlogQ}{d\tau} - \frac{dlogJ}{d\tau} - \frac{dlogQ}{d\tau}|_J}{\frac{dlogJ}{d\tau}}$$

relative importance of the variety effect and business-stealing. We see that for a plausible range of short-run and long-run estimates, we may have either insufficient or excessive entry; ultimately, it is an empirical question depending on the short-run and long-run price elasticities and the magnitude of the variety response to taxes. Turning to the welfare effect of taxes ($dW/d\tau$), we see that the role played by the gap between the variety and business-stealing effects. When this gap is large, the welfare losses from a change in taxes are larger (in magnitude) compared to the case where the gap is small. We also see that a larger behavioral response of taxation leads to a larger welfare loss, consistent with standard intuition that taxes have a fiscal externality on the government's budget.

Using Table 10 as a template, we next implement the same formulas with our preferred set of reduced-form estimates from empirical analysis. We report these results in Table 11. Our estimates imply that the variety effect is around 1.2-1.6, depending on whether we use preferred pooled cross-sectional results or instead the "border pair" results. The scaling of the variety effect gives a unit-free measure that can be interpreted as the elasticity of the inverse aggregate demand curve with respect to variety. In other words, the magnitude of estimated variety effect suggests that an exogenous decrease in variety of 1 percent will decrease average willingness-to-pay by 1.2-1.6 percent. This is a large variety effect, although it is smaller that what is implied by logit model that constrains variety effect to be related to the price elasticity of demand. In addition to the variety effect, we can also recover estimate of business-stealing effect, and we find it is smaller in magnitude than the variety effect , suggesting that there is too little variety relative to the social optimum (i.e., $dW/dJ > 0$).

To give some intuition behind magnitude of the variety effect estimate, consider the following example: assume that an exogenous 1 percent decrease in variety results in all consumers who previously purchased the (removed) varieties deciding to purchase the outside good. Assuming identical market shares initally, then in this example a 1 percent decrease in variety will reduce quantity demanded by 1 percent (holding price constant). This is identical to an inward shift in the inverse aggregate demand curve by the reciprocal of the price elasticity of demand, or approximately 2 given the short-run estimates in Table 11. Alternatively, we can assume that the individuals who purchased the (removed) varieties instead purchase second choice, which they value 10 percent less on average. In this case, we would calculate an inward shift in inverse aggregate demand curve by 0.1 percent. The combination of the large variety effect and small business stealing effect is consistent with many consumers not choosing an alternative variety when their preferred variety is elimianted due to taxes.

Turning to the welfare effect of taxes, we find that since taxes further distort variety below the socially optimal level, the welfare effect of taxes with endogenous variety is larger than the welfare

effect of taxes ignoring variety, with most of the social costs due to the additional cost of reduced variety. Thus, we see that it's important to account for the variety distortion. The final columns of Table 11 reproduces these calibrations instead imposing the logit assumption. A common theme across the columns is the implied estimate of the variety effect is significantly larger under the logit specification compared to using our approach. This can be seen by comparing columns (1) to (2) with columns (3) to (4). The variety effect is significantly larger, which likely relates to the well-known result that the logit model tends to overstate the welfare gains from new goods (Hausman 1996; Petrin 2002).

These estimates are intended to be illustrative of our new revealed-preference approach, and given the uncertainty in our reduced-form estimates, the findings should be interpreted tentatively. However, our formulas also provide a way to recover implied markups (as a share of consumer price), which we find to be between 0.15 and 0.39. These values are broadly similar to other estimates of average markups (above wholesale costs) estimated using more detailed price and cost data (see, e.g., Gopinath et al. 2010). This suggests that the combination of our formulas and reduced-form estimates provide sensible markups, even thought we do not observe markups or costs directly. We thus view our results as a useful first step at valuing gains to variety using reduced-form estimates.

9 Conclusion

Understanding how changes in product variety affect consumer welfare is critical for a host of questions. In this paper, we develop a sufficient statistics method for valuing changes in product variety and we apply it to two classic questions. First, we consider whether product variety is insufficient or excessive relative to the social optimum. Second, we consider the welfare effects of taxation with endogenous product variety. Central to both applications are reduced-form estimates of taxes on prices and quantities where variety is held constant and where variety responds to a change in taxes, subject to a free-entry condition, along with estimates of taxes on varieties.

We implement our approach by combining rich retail scanner data from grocery stores in the U.S. with detailed state and county sales tax data. Our empirical results indicate that at current sales tax rates and entry costs, product variety is below the social optimum. Additionally, we estimate a large effect of sales taxes on product variety, and we find that the marginal excess burden of sales taxes is significantly larger than estimates ignoring the indirect effect of taxes on product variety. These results are of course specific to this setting, but we see this "proof of concept" as demonstrating applicability to other settings in IO and International Trade.

Our analysis can be extended and generalized in several dimensions. Most directly, one could use the sufficient statistics formula for the variety effect developed here to study the impact of tariffs on welfare or other government policies that result in changes in equilibrium varieties, such as price controls. One could also use a similar approach to study the welfare effects of mergers, taking into account the variety effect in addition to the price effect. Finally, it would be useful to extend the methods developed here to allow for firm heterogeneity and multi-product firms when modeling socially optimal product variety and to investigate endogenous changes in product quality in response to government policies.

10 References

1. Agrawal, David R. (2015), "The Tax Gradient: Spatial Aspects of Fiscal Competition," *American Economic Journal: Economic Policy*, 7(2): 1-29.

2. Anderson, Simon P., Andre de Palma, and Yurii Nesterov (1995), "Oligopolistic Competition and the Optimal Provision of Products," *Econometrica*, 63(6): 1281-1301.

3. Anderson, Simon P., Andre De Palma, and Jacques-Francois Thisse (1987) "The CES is a discrete choice model?," *Economics Letters*, 24(2): 139-140.

4. Anderson, Simon P., Andre De Palma and Brent Kreider (2001a), "The Efficiency of Indirect Taxes under Imperfect Competition," *Journal of Public Economics*, 81(2): 231-251.

5. Anderson, Simon P., Andre De Palma and Brent Kreider (2001b), "Tax Incidence in Differentiated Product Oligopoly," *Journal of Public Economics*, 81(2): 173-192.

6. Arkolakis, Costas, Arnaud Costinot and Andres Rodriguez-Clare (2012), "New Trade Models, Same Old Gains?" *American Economic Review*, 102(1): 94-130.

7. Atkin, D., Benjamin Faber and Marco Gonzalez-Navarro (2016), "Retail Globalization and Household Welfare: Evidence from Mexico," *Journal of Political Economy*.

8. Auerbach, Alan J. and James R. Hines Jr. (2001). "Perfect Taxation with Imperfect Competition," NBER Working Paper #8138.

9. Baily, Martin N. (1978), "Some Aspects of Optimal Unemployment Insurance," *Journal of Public Economics*, 10 (December): 379-402.

10. Berry, Steven T. and Joel Waldfogel (1999), "Free entry and social inefficiency in radio broadcasting," *RAND Journal of Economics*, 30(3): 397-420.

11. Berry, Steven T., Alon Eizenberg, and Joel Waldfogel (2015), "Optimal Product Variety in Radio Markets," Working Paper.

12. Besley, Timothy J. (1989). "Commodity taxation and imperfect competition: A note on the effects of entry," *Journal of Public Economics*, 40(3): 359-367.

13. Besley, Timothy J. and Harvey S. Rosen (1999). "Sales taxes and prices: an empirical analysis,"*National Tax Journal*, 52(2): 157-158.

14. Bhattacharya, Debopam (2015), "Nonparametric Welfare Analysis for Discrete Choice," *Econometrica* 83(2): 617–649.

15. Broda, Christian and David E. Weinstein (2006) "Globalization and the Gains From Variety," *The Quarterly Journal of Economics*, 121(2): 541-585.

16. Bulow, Jeremy I., and Paul Pfleiderer (1983). "A Note on the Effect of Cost Changes on Prices." *Journal of Political Economy*, 91 (1): 182–85.

17. Cameron, Colin, Jonah Gelbach, and Douglas Miller (2011), "Robust Inference With Multiway Clustering." *Journal of Business and Economic Statistics*, 29(2): 238-249.

18. Cardell, N. Scott (1997), "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," *Econometric Theory*, 13(2): 185-213.

19. Chetty, Raj (2006), "A General Formula for the Optimal Level of Social Insurance," *Journal of Public Economics* 90 (November): 1879-1901.

20. Chetty, Raj (2008), "Moral Hazard versus Liquidity and Optimal Unemployment Insurance," *Journal of Political Economy*, 116(2): 173-234.

21. Chetty, Raj (2009), "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Reviews of Economics*, Volume 1: 451-488.

22. Delipalla, Sophia and Michael Keen (1992), "The Comparison between Ad Valorem and Specific Taxation under Imperfect Competition," *Journal of Public Economics*, 49(3): 351-361.

23. Dixit, Avinash K. and Joseph E. Stiglitz (1977), "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, 67(3): 297-308.

24. Dube, Arindrajit, T. William Lester, and Michael Reich (2010), "Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties," *The Review of Economics and Statistics*, 92(4): 945–964.

25. Einav, Liran, Amy Finkelstein, and Mark R. Cullen (2010), "Estimating Welfare In Insurance Markets Using Variation in Prices," *The Quarterly Journal of Economics*, 125(3): 877-921.

26. Fabinger, M. and G. Weyl, (2016). "The Average-Marginal Relationship and Tractable Equilibrium Forms". Working Paper.

27. Feenstra, Robert C. (1994), "New Product Varieties and the Measurement of International Prices," *American Economic Review*, 84(1): 157-177.

28. Gabaix, Xavier, David Laibson, Deyuan Li, Hongyi Li, Sidney Resnick, and Casper G. de Vries. (2016). "The Impact of Competition on Prices with Numerous Firms." *Journal of Economic Theory,* Vol. 165, p. 1-24.

29. Gentzkow, Matthew, Jesse M. Shapiro and Michael Sinkinson (2014), "Competition and Ideological Diversity: Historical Evidence from US Newspapers," *American Economic Review*, 104(10): 3073-3114.

30. Gillitzer, Christian, Henrik J. Kleven and Joel Slemrod (2015), "A Characteristics Approach to Optimal Taxation: Line Drawing and Tax-driven Product Innovation," *Scandinavian Journal of Economics*.

31. Hagedorn, Marcus, Iourii Manovskii and Kurt Mitman (2016), "The Impact of Unemployment Benefit Extensions on Employment: The 2014 Employment Miracle?", Working paper.

32. Handbury, Jessie, and David E. Weinstein (2015), "Goods Prices and Availability in Cities," *Review of Economic Studies*, 82(1): 258-296.

33. Harberger, Arthur (1964). "The Measurement of Waste," *American Economic Review* 54(3): 58-76.

34. Hart, O. (1982). "Reasonable Conjectures". Discussion Paper. London School of Economics.

35. Hausman, Jerry (1996), "Valuation of New Goods under Perfect and Imperfect Competition," in *The Economics of New Goods* (eds. Timothy F. Bresnahan and Robert J. Gordon).

36. Hausman, Jerry and Gregory K. Leonard (2002), "The Competitive Effects of a New Product Introduction: A Case Study," *Journal of Industrial Economics*, 50(3): 237-263.

37. Holmes, Thomas J. (1998), "The Effects of State Policies on the Location of Industry: Evidence from State Borders," *Journal of Political Economy*, 106(4): 667-705.

38. Hsieh, Chang-Tai, Nicholas Li, Ralph Ossa, and Mu-Jeung Yang (2015) "Accounting for the Gains from Trade Liberalization," Working Paper.

39. Kroft, Kory and Matthew J. Notowidigdo (2016), "Should Unemployment Insurance Vary with the Unemployment Rate? Theory and Evidence," *Review of Economic Studies*.

40. Mankiw, N. Gregory, and Michael Whinston (1986) "Free Entry and Social Inefficiency," *RAND Journal of Economics*, 17 (Spring): 48-58.

41. McFadden, D. (1981). "Econometric models of probabilistic choice". In Manski, C. and McFadden, D., editors, Structural Analysis of Discrete Data with Econometric Applications. MIT Press.

42. Melitz, Marc and Stephen Redding (2015), "New Trade Models, New Welfare Implications," *American Economic Review*, 105(3): 1105-1146.

43. Milgrom, Paul and John Roberts (1996). "The LeChatelier Principle". *American Economic Review* 86(1): 173-179.

44. Myles, Gareth D. (1989), "Product Variety and Tax Policy," Working Paper.

45. Nocke, Volker, and Nicolas Schutz (2017). "Multiproduct-Firm Oligopoly: An Aggregative Games Approach". Working Paper.

46. Petrin, Amil (2002), "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110(4): 705-729.

47. Resnick, S. I. (1987): "Extreme Values, Regular Variation and Point Processes". Springer, New York.

48. Riordan, M. (1985). "Imperfect Information and Dynamic Conjectural Variations". *Rand Journal of Economics* 16: 41-50.

49. Seade, Jesus (1987), "Profitable Cost Increases and the Shifting of Taxation: Equilibrium Responses of Markets in Oligopoly," Working Paper.

50. Sheu, Gloria (2014), "Price, Quality, and Variety: Measuring the Gains from Trade in Differentiated Products," *American Economic Journal: Applied Economics*, 6(4): 66-89.

51. Spence, Michael (1976a), "Product Differentiation and Welfare," *American Economic Review*, 66(2): 407-14.

52. Spence, Michael (1976b), "Product Selection, Fixed Costs, and Monopolistic Competition," *Review* of Economic Studies, Vol. 43, No. 2, pp. 217-235.

53. Stern, Nicholas (1987), "The Effects of Taxation, Price Control, and Government Contracts in Oligopoly and Monopolistic Competition," *Journal of Public Economics*, 32: 133-158.

54. Trajtenberg, Manuel (1989), "The Welfare Analysis of Product Innovations, with an Application to Computed Tomography Scanners," *Journal of Political Economy*, 97: 444-479.

55. Vives, X. (2001), "Oligopoly Pricing: Old Ideas and New Tools". MIT Press.

56. Weyl, Glen and Michal Fabinger (2013), "Pass-Through as an Economic Tool: Principle of Incidence under Imperfect Competition," *Journal of Political Economy*, 121(3): 528-583.

57. Wollmann, Thomas (2016), "Trucks without Bailouts: Equilibrium Product Characteristics for Commercial Vehicles," Working Paper.

Proofs of Propositions, Derivations of Formulas, and Model Extensions

Observe Q_m can be written in terms of cdf of max shock

$$Q_m = 1 - \mathbb{P}(\max_{j \in m} \varepsilon_j \leq p + \varepsilon_0)$$

For any family $(Q_m(p, J))_J$ which is decreasing in p and increasing in J, we can use induction to construct cdfs of shocks from cdf of max. For example:

$$\mathbb{P}(\varepsilon_2 \leq x) = \mathbb{P}(\max\{\varepsilon_1, \varepsilon_2\} \leq x) + \mathbb{P}(\varepsilon_2 \leq x | \varepsilon_1 > x)(1 - \mathbb{P}(\varepsilon_1 \leq x))$$

Where the first and third terms are fixed, but the second is free. another way to see the construction is that basically we can let ε_2 have distribution $1 - Q(\cdot, 2)$, and so on.

Gumbel is the unique iid distribution satisfying parallel demands

Assuming the unique attribute is price and these are symmetric, the inverse demands when there are J and $J+1$ varieties are parallel iff there exists t such that for all p then $Q(p, J) = Q(p + t, J + 1)$, that is

$$\mathbb{P}(\varepsilon_{0m} < -p + \nu_m(1 - \sigma_m) + (\sigma_m) \max_{1 \leq j \leq J} \varepsilon_j) = \mathbb{P}(\varepsilon_{0m} < -p + t + \nu_m(1 - \sigma_m) + (\sigma_m) \max_{1 \leq j \leq J+1} \varepsilon_j)$$

Since ε_{0m} and $\nu_m(1-\sigma_m)$ are independent of $\max_{1 \leq j \leq J} \varepsilon_j$ this can only be true if the distribution of the maxima is the same, that is

$$\max_{1 \leq j \leq J} \varepsilon_j \quad \stackrel{d}{=} \quad t + \max_{1 \leq j \leq J+1} \varepsilon_j$$

Let F be the cdf of ε, then the equation above implies there exist $t(n)$ such that for all x:

$$F(x) = F^n(x + t(n))$$

Iterating on both sides implies

$$F^{nm}(x + t(nm)) = F^{nm}(x + t(n) + t(m))$$

we recognize an instance of Hamel's functional equation $t(nm) = t(n) + t(m)$ which has solution $t(n) = c\log(n)$.[44] therefore

$$F(x) = F^y(x + c\log y),$$

letting $s = c\log y$,

$$F(0) = F^{e^{s/c}}(s),$$

and so

$$F(s) = e^{\log F(0)e^{-s/c}},$$

which is a Gumbel distribution with location parameter $c\log\log F(0)$ and dispersion parameter c.

Gumbel Approximation Theorem

Again, fixing p_n and J_n for all $n \neq m$ and ε_j for all $j \notin m$ then:

If $(\varepsilon_j)_{j\in m}$ are i.i.d. and the distribution is in the domain of attraction of the Gumbel distribution, then for big enough K and all $J_m \geq K$

$$\max_{j\in m} \varepsilon_j \overset{approx}{\sim} Gumbel(\mu(J_m), \sigma(J_m)),$$

for some $(\mu(J_m), \sigma(J_m))$, so the inverse demands are approximately parallel for $J_m \geq K$ and we can get the variety effect with our sufficient statistics formula.

For example, if $\varepsilon_j \sim N(0, \sigma^2)$ iid for all j the above theorem applies.

Measurement

To link the data with the theoretical framework, we construct module-level measures of prices, expenditures and variety from the scanner data. For each variable, we produce both a cross-sectional dataset in which the unit of time (t) is a year and a quarterly panel.[45]

[44] It is easy to extend the formula for real numbers through rationals, note

$$F(x) = F^n(x + t(n)) = F^m(x + t(m))$$

implies

$$F(x) = F^{n/m}(x + t(n) - t(m)),$$

so we can consistently define $t(n/m) = t(n) - t(m)$.

[45] We use statutory tax rates effective on September 1 of a given year as yearly rates. The rates effective on February 1, May 1, August 1, and November 1 are used as quarterly rates.

Expenditures Let q_{jmrcsw} denote the number of units of product (UPC) j in module m sold in store r, located in county c in state s, in week w. Similarly, let p_{jmrcsw} be the associated per-unit average weekly price. We denote weekly revenue from sales of product j in store r by $R_{jmrcsw} = q_{jmrcsw} \times p_{jmrcsw}$ and module-level measures are obtained by aggregating across UPCs and weeks: $R_{mrcst} = \sum_{j \in m} \sum_{w \in t} R_{jmrcsw}$. The unit of time, t, is either a year (for equation (29)) or a quarter (for equation (31)). Finally, we calculate expenditure *shares* within each store-time cells: $E_{mrcst} = R_{mrcst}/R_{rcst}$ for each period, where $R_{rcst} = \sum_m R_{mrcst}$.

Prices First, we average (pre-tax) prices across weeks to obtain either quarterly or yearly measures:

$$p_{jmrcst} = \frac{\sum_{w \in t} R_{jmrcsw}}{\sum_{w \in t} q_{jmrcsw}}.$$

We then average prices across UPCs to obtain module-level price indices. Handbury and Weinstein (2015) show that comparing standard indices across locations can be problematic if consumer preferences are heterogeneous across locations, and if some varieties are unavailable in some places. For example, if consumers in a given location tend to buy larger packages of a given beverage than in other locations, the average *per-unit* price will be higher in that location even though the *per liter* average price is likely lower. To correct for these sources of bias, we follow Handbury and Weinstein (2015) and adjust prices by estimating the following regression separately for each module:

$$\log p_{jmrcst} = \alpha_j + \alpha_{mrcst} + \varepsilon_{jmrcst}$$

where α_j and α_{mrcst} are UPC and module-store-time fixed effects, respectively.[46] We keep the estimated module-store-time fixed effects, $\hat{\alpha}_{mrcst}$ as pre-tax price indices (which are in logs). Adjusted consumer prices are then given by $\tilde{p}_{mrcst} = \exp\left(\hat{\alpha}_{mrcst} + \log(1 + \tau_{mcs})\right)$.

We then normalize the consumer price indices within store-time cells by dividing by the store-time-specific average *pre-tax* price.[47] Note that we normalize using pre-tax prices rather than consumer prices so that the *mechanical* relationship between the price measure and sales taxes is effectively one-to-one. To see this, note that $\log\left(\tilde{p}_{mrcst}/\bar{p}_{mrcst}\right) = \hat{\alpha}_{mrcst} + \log(1 + \tau_{mcs}) - \log \bar{p}_{mrcst}$.

[46] Observations are weighted by expenditures, R_{jmrcst}. Both a cross-sectional and a time series version of this equation are estimated. In the former,t is a year and only one year of data is used. For time-series, all periods (quarters) are simultaneously included the regression.

[47] The within store-time average is $\bar{p}_{mrcst} = \frac{1}{N_{rcst}} \sum_m \exp(\hat{\alpha}_{mrcst})$, where N_{rcst} is the number of modules with positive sales in store r at time t.

Variety Finally, variety is measured by counting the number of unique UPCs per module sold each period t in store r: $J_{mrcst} = \{j \in J_m | q_{jmcsrt} > 0\}$. As for prices, we normalize variety by dividing by the store-time average.

Robustness of Long-Run Estimates - Border-design

Year:	2006	2007	2008	2009	2010	2011	2012	2013	2014	Average
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent Variable:										
Log of Expenditure Share	-0.620**	-0.769***	-0.627**	-0.783***	-0.584**	-0.493*	-0.459*	-0.280	-0.238	-0.539
	(0.267)	(0.249)	(0.261)	(0.244)	(0.252)	(0.256)	(0.255)	(0.249)	(0.249)	(0.230)
Log of Average Consumer Price	1.099***	1.082***	1.107***	1.161***	1.124***	1.120***	1.116***	1.026***	1.002***	1.093
	(0.0419)	(0.0372)	(0.0367)	(0.0343)	(0.0348)	(0.0348)	(0.0355)	(0.0346)	(0.0350)	(0.030)
Log of Variety (# of UPCs)	-0.581***	-0.684***	-0.679***	-0.708***	-0.547***	-0.439***	-0.502***	-0.558***	-0.376**	-0.564
	(0.168)	(0.151)	(0.162)	(0.152)	(0.156)	(0.157)	(0.161)	(0.157)	(0.166)	(0.139)
Specification:										
Store Fixed Effects	y	y	y	y	y	y	y	y	y	y
Module fixed effects	y	y	y	y	y	y	y	y	y	y

Notes: All standard errors are clustered at the state-module level.

Robustness of Long-Run Estimates - Border-design

Year:	2006	2007	2008	2009	2010	2011	2012	2013	2014	Average
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent Variable:										
Log of Expenditure Share	-0.334**	-0.401***	-0.777***	-0.975***	-0.815***	-0.614***	-0.655***	-0.757***	-0.514***	-0.649
	(0.137)	(0.112)	(0.177)	(0.168)	(0.175)	(0.161)	(0.154)	(0.161)	(0.165)	(0.142)
Log of Average Consumer Price	1.022***	1.057***	1.047***	1.070***	1.051***	1.028***	1.013***	1.005***	1.035***	1.036
	(0.0199)	(0.0166)	(0.0210)	(0.0197)	(0.0214)	(0.0206)	(0.0190)	(0.0190)	(0.0196)	(0.016)
Log of Variety (# of UPCs)	-0.163**	-0.160**	-0.294***	-0.491***	-0.416***	-0.282***	-0.212***	-0.422***	-0.300***	-0.304
	(0.0808)	(0.0660)	(0.0915)	(0.0854)	(0.0877)	(0.0870)	(0.0797)	(0.0898)	(0.0957)	(0.070)
Specification:										
Store Fixed Effects	y	y	y	y	y	y	y	y	y	y
Module × Pair fixed effects	y	y	y	y	y	y	y	y	y	y

Notes: All standard errors are clustered at the state-module level.