ANNUAL REVIEWS

Annual Review of Economics

The Use of Scanner Data for Economics Research

Pierre Dubois,[1] Rachel Griffith,[2,3] and Martin O'Connell[2,4]

[1]Toulouse School of Economics, Toulouse, France; email: pierre.dubois@tse-fr.eu

[2]Institute for Fiscal Studies, London, United Kingdom; email: rgriffith@ifs.org.uk, martin_o@ifs.org.uk

[3]School of Social Sciences, University of Manchester, Manchester, United Kingdom

[4]University of Wisconsin-Madison, Madison, Wisconsin, USA

Annu. Rev. Econ. 2022. 14:723–45

First published as a Review in Advance on May 13, 2022

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-051520-024949

JEL codes: C80, D12, D22, E31, L10

Keywords

scanner data, demand estimation, market power, policy counterfactual, inflation

Abstract

The adoption of barcode scanning technology in the 1970s gave rise to a new form of data: scanner data. Soon afterwards, researchers began using this new resource, and since then a large number of papers have exploited scanner data. The data provide detailed price, quantity, and product characteristic information for completely disaggregate products at high frequency, and they typically track a panel of stores and/or consumers. Their availability has led to advances, inter alia, in the study of consumer demand, the measurement of market power, firms' strategic interactions and decision making, the evaluation of policy reforms, and the measurement of price dispersion and inflation. In this article we highlight some of the pros and cons of this data source, and we discuss some of the ways its availability to researchers has transformed the economics literature.

1. INTRODUCTION

The advent of barcode scanner technology in the 1970s gave rise to a new form of data, known as scanner data. Researchers quickly recognized the value of using these data to learn about which factors influence consumers' choices and demands.[1] Scanner data have since given rise to a voluminous literature that seeks to use them to study a wide range of economic behaviors. The data have been most widely used in industrial organization and marketing to study consumer choice, firms' strategic decisions, and market power. Increasingly, scanner data are also being used to study a range of questions in public economics (such as the impact and design of taxes and regulations), health economics (such as the drivers of rising obesity), and macro and monetary economics (such as the drivers of aggregate price fluctuations).

There are two main forms of scanner data. The first, store scanner data, are collected at the point of sale by the in-store scanners used at checkout. The second, household scanner data, are collected by individuals or households using scanner technology that is typically provided by a third-party company (for example, a market research firm). Both forms of scanner data share a number of key features. They provide information on quantities and prices at the level of disaggregate individual products [i.e., at the barcode or universal product code (UPC) level], and they include key product characteristics, such as brand, manufacturer, package type, flavors, etc., as well as the location (store) and date of purchase. They are often longitudinal, recording repeated transactions for the same household, individual, or store over time. In addition, they often provide further details about the transaction, such as whether the product was on promotion (for instance, subject to a temporary price reduction or a discount for purchasing multiple units). Scanner data are most commonly available for fast-moving consumer goods (which, approximately, correspond to those products typically available in supermarkets, including food, drinks, alcohol, toiletries, detergents, etc.). However, recent innovations are leading to the availability of scanner data covering a wider range of purchases, including those made in takeaways and dine-in restaurants.

A leading reason for the widespread adoption of scanner data in economics research is that they provide the only systematic source of information on prices and quantities for specific products that are disaggregated over retailers and time.[2] This information is especially useful when studying consumer and firm behavior in markets for differentiated products.

Household scanner data are usually collected by market research firms, who recruit people into their sample and provide them with scanner technology. Typically, participating households record all fast-moving consumer goods that they purchase and bring into the home, and they provide receipts to the market research firm to validate purchases. The data therefore contain a record of individual transactions (i.e., barcode X, purchased from retailer Y, on date Z), with information on quantities and prices. Households are tracked through time and often are present in the sample for many months or years. The data also typically include demographic information collected from households through survey questionnaires. Examples of household scanner data used for academic research are the Nielsen Homescan Consumer Panel,[3] which covers US households,

[1] To the best of our knowledge, the first paper to use scanner data to estimate a model of consumer choice is by Guadagni & Little (1983).

[2] There are alternative data sources that contain some of this information on disaggregate product prices—for example, the data collected by national statistical offices for official inflation measurement (e.g., see Nakamura & Steinsson 2008, Eizenberg et al. 2021). However, these data typically cover only a subset of products, and they do not contain information on disaggregate product sales or quantities. This motivates the recent interest in harnessing scanner data in official inflation indices. We discuss this in Section 5.

[3] These can be accessed for research purposes through the University of Chicago, at **https://www.chicagobooth.edu/research/kilts/datasets/nielsen**.

and the Kantar Fast-Moving Consumer Goods (FMCG) Purchase Panel,[4] which covers British households.

Store scanner data can sometimes be obtained directly from a retailer, but they are also available from market research firms that obtain and collate the data from several retailers. Usually, store scanner data contain information on quantities and prices of all products that are sold in each of the participating stores. Often this information will be available at weekly frequency. In some cases, this is supplemented with the menu of prices of all products available in the store at that time (including those that are not purchased). This form of scanner data has the advantage of providing comprehensive information on the sales and prices of products sold by a store or retail chain. In some cases, purchases made by loyalty card holders can be identified. This allows researchers to build a consumer-level database of purchases that can include limited demographic information. However, unlike household scanner data, these data do not typically link household purchases across retailers and therefore usually cover a single retailer's customer base. Store scanner data have been widely used in economic research, including the Nielsen Retail Scanner data set,[5] the IRI Infoscan data set,[6] and the Dominik's supermarket database.[7]

In this article we provide an overview of some of the ways in which scanner data have led to advances in economics research. We begin in the next section by discussing some of the main uses of scanner data in economics research. We then provide an overview of what we consider to be some of the more exciting strands of research that the availability of scanner data has stimulated. In Section 3, we focus on the estimation of models of consumer demand, highlighting how scanner data have played an instrumental role in developments in the empirical modeling of product differentiation, heterogeneity in consumer preferences, consumer dynamics, and the role of advertising in affecting choice. In Section 4 we discuss the related topic of modeling firm behavior and market equilibrium. This includes work on measuring the extent of market power exercised by firms, analyzing mergers, modeling strategic retailers and their vertical relations with manufacturers, and studying the extent of pass-through of exchange rate changes to equilibrium prices. In Section 5 we discuss how scanner data have led to better measurement of inflation, including variation in rates across households and accounting for the impact of product entry, and we discuss how national statistical offices are beginning to incorporate scanner data into official Consumer Price Indexes (CPIs). Our aim is not to provide a comprehensive survey—there are influential papers and entire research agendas that use scanner data and are not included in our discussion. Rather, we aim to illustrate some of the ways that scanner data have enabled the economics literature to progress.

2. THE MAIN FEATURES OF SCANNER DATA

2.1. Disaggregate Price and Quantity Information

A key advantage of scanner data is that they provide well-measured information on prices and quantities at the individual product level. As we discuss below, this is extremely useful for estimating consumer demand and firm supply in differentiated product markets, which in turn enables researchers to address a wide range of economics questions, including measuring market power,

[4]These can be obtained from Kantar FMCG, at **https://www.kantarworldpanel.com/global/Coverage/worldpanel/United-Kingdom**.

[5]This can be accessed for research purposes through the University of Chicago, at **https://www.chicagobooth.edu/research/kilts/datasets/nielsen**.

[6]The data set can be found at **https://www.iriworldwide.com/en-us/solutions/academic-data-set**.

[7]This can be accessed for research purposes through the University of Chicago, at **https://www.chicagobooth.edu/research/kilts/datasets/dominicks**.

evaluating mergers, and assessing a range of government policies. As we discuss below, detailed price and quantity information also facilitate the measurement of aggregate price fluctuations.

There is no other widely available single source of data containing disaggregate product-level prices and quantities. Data from consumer expenditure surveys, such as the Consumer Expenditure Survey (CEX) in the United States or the Living Cost and Food Survey (LCFS) in the United Kingdom, contain household spending information at a more aggregate level. For instance, they provide information on a household's spending on breakfast cereal. In contrast, scanner data provide information on purchases of individual cereal products (for example, a 720g package of Kellogg's Corn Flakes). This more detailed information facilitates the modeling of demand and supply, as well as industry dynamics, within specific (e.g., the breakfast cereal) markets. To address some research questions, it may be desirable to work with data that are at a more aggregate level than the individual product level. In this case, scanner data have the advantage of allowing the researcher to aggregate products in the most appropriate way to address the question at hand. For instance, Griffith et al. (2019) use Kantar FMCG Purchase Panel data to study consumer demand and tax design in the alcohol market. They are interested in how varying tax rates across different sources of alcohol can enable the tax system to better target the most socially costly consumption. To estimate a tractable model of consumer choice over alcohol types, they first aggregate the thousands of alcohol products available in the market up to dozens of alcohol varieties, taking care not to aggregate over products that have different tax liabilities.

2.2. Information on the Choice Environment

It is common for scanner data to contain information on the retailer and location in which a transaction occurs and whether the transaction entailed a promotion. As we discuss below, this has led to advances in our understanding of how consumers choose where to shop, how retailers interact with manufacturers, and how the choice environment affects consumer decisions, and it has contributed to shedding light on intertemporal aspects of consumer choice, such as the timing of purchases when products are on sale. In addition, as scanner data are high frequency (usually either at the daily or weekly level) it is often possible to match them with other relevant information about the environment—for example, advertising, news stories, the weather, and other events. This has enabled researchers to explore how various aspects of the choice environment impact the decisions that consumers make.

2.3. Panel Structure

Household scanner data track households through time. It is common for households to be present in the data for several months or years. Because participating households record all (usually, fast-moving consumer) goods they purchase and bring into the home, the data contain a large amount of information about each household's choice behavior. In contrast, consumer expenditure surveys tend to be repeated cross-sections, and official longitudinal studies, like the Panel Survey of Income Dynamics in the United States and Understanding Society in the United Kingdom, tend to collect information from households infrequently (quarterly, annually, or biennially) and to not include detailed spending information. Recently, researchers have used high-frequency bank and credit card transaction data (e.g., Gelman et al. 2014), but these do not provide product-level information. The panel structure of household scanner data (combined with the product-level information they contain) is advantageous for a number of reasons.

2.3.1. Identification of preferences. In the choice models that researchers most commonly estimate with scanner data, preferences are typically modeled as household specific and drawn

from a random coefficient distribution. An advantage of micro-level panel data is that they allow for identification of random coefficient distributions under weaker conditions compared to market-level data (e.g., see Berry & Haile 2020). In particular, observing the same set of consumers making repeated choices while exposed to different prices and choice sets is informative about the degree of dispersion in consumers' specific preferences. In addition, the time-series dimension of household scanner data is sometimes sufficiently long that it is feasible to directly estimate households' specific preference parameters in nonlinear choice models. As we discuss below, this allows researchers to relax commonly imposed distributional assumptions and weaken the independence assumption placed on household preferences.

2.3.2. Dynamic behaviors. By tracking the behavior of the same decision maker over time, household scanner data are well suited to the study of dynamic consumer behaviors. For instance, they allow researchers to model the dependence of someone's choice today on what they chose in the past or on features of their past choice environment. This enables researchers to study behaviors such as habit formation, stockpiling, consumer choice inertia, and temptation.

2.3.3. Demographic information. Household scanner data generally provide rich information on demographics. Market research firms collect these data and sell them commercially, and therefore the demographic information reflects the business needs of their clients. Sometimes information useful in economic research either is not collected or, when it is, is not very detailed—for instance, on labor supply, work status, wealth, and benefit receipt. This places limitations on the use of the data for some applications. However, researchers have had success in complementing the information commonly available in scanner data. For example, Allcott et al. (2019b) survey participants in the Neilsen Homescan data to obtain measures of nutritional knowledge and self-control as a basis for estimating the extent of consumer misoptimization in their choice of consumption of sugar-sweetened beverages. Griffith et al. (2018b) estimate the probability that households in their scanner data set are in receipt of benefits by matching data on geographic location and household characteristics with LCFS data, which contain information on benefit claims. Other researchers have succeeded in securing the cooperation of scanner data collectors in conducting an experiment. For example, Chetty et al. (2009) randomly varied price posting between the sales tax-exclusive price (as is normal in the United States) and the tax-inclusive price in a Northern Californian store to estimate how tax salience impacts the incidence and excess burden of taxation.

2.3.4. Spending on other goods and services. The vast majority of research using scanner data has been on fast-moving consumer goods (e.g., food, drinks, alcohol, toiletries, cleaning supplies, etc.). As discussed below, access to granular data on this segment of the economy has led to many advances in economic research. Fast-moving consumer goods account for around half of consumer good expenditure.[8]

Increasingly, scanner data providers are beginning to collect data outside of the fast-moving consumer goods brought into the home. One example is the Kantar Out of Home Purchase Panel collected for UK individuals, which includes all food and nonalcoholic beverage purchases for consumption outside the home, including those made in dine-in restaurants (see O'Connell et al. 2021, who use these data to track expenditures and calories over the COVID-19 pandemic). Another is the GfK Point Of Sale panel, which covers purchases of slow-moving consumer goods (e.g., electronics, do-it-yourself products) in several countries (see Beck & Jaravel 2020, who use these data to measure differences in inflation and product entry across countries).

[8]For information on fast-moving consumer goods, readers may consult Investopedia (at **https://www.investopedia.com/terms/f/fastmoving-consumer-goods-fmcg.asp**).

3. DEMAND

The answers to many empirical questions in economics rely upon having credible estimates of consumer demand. These include, for instance, understanding how consumers will adjust their choices in response to changes in firms' strategies or government policies (e.g., pricing, product redesign, taxes, regulations that restrict availability, provision of information) and the consequent impact on their welfare. In addition, answering many questions about the supply side of a market, including the implications of consumer choice behavior for profits and firms' pricing and marketing strategies, often relies first on obtaining estimates of demand. The availability of scanner data has led to significant advances in our ability to estimate rich models of consumer demand.

3.1. Product Differentiation

Nevo (2011) provides a comprehensive discussion of the development of the literature on the estimation of demand for differentiated products. An earlier literature on demand estimation focused on the estimation of aggregate demand for a set of J goods of the form $\mathbf{q} = D(\mathbf{p}, \mathbf{z}, \boldsymbol{\epsilon})$, where $\mathbf{q}$, $\mathbf{p}$, $\mathbf{z}$ and $\boldsymbol{\epsilon}$ are $J \times 1$ vectors of quantities, prices, exogenous demand shifters, and random shocks (see Deaton 1986). Nevo (2011) cites a number of limitations of this framework for estimating differentiated product demand, including a dimensionality problem—in many markets J is large (sometimes >100), so for any reasonable parametric specification there are too many parameters to estimate[9]—and the fact that these models cannot be straightforwardly used to predict demand for a new good.

A solution to these drawbacks is offered by treating products as bundles of characteristics over which consumers have preferences (Lancaster 1966, McFadden 1974, Rosen 1974, Gorman 1980). This reduces the dimensionality of the problem to the number of characteristics that define a product and enables the researcher to simulate the effects of the introduction of a new good (i.e., a new bundle of characteristics). A set of papers pioneered the application of characteristics models to the estimation of demand and supply in differentiated products markets using (non-scanner) data on the automobile industry.[10]

The increasing availability of scanner data has led to a flourishing of research estimating demand in differentiated product markets. In a seminal paper, Nevo (2001) uses store-level scanner data to estimate demand over breakfast cereal brands, where consumers derive utility from brand characteristics. As we discuss below, he uses this demand model as an input into the measurement of the degree of market power exercised by breakfast cereal manufacturers. In addition to work that focuses on differentiation of products in characteristics space, scanner data, which contain information on the stores of purchase, have given rise to research that studies differentiation across stores—in terms of both store characteristics (e.g., floorspace) and geographical locations—and how this influences where consumers choose to shop (e.g., Thomassen et al. 2017).

Papers that estimate differentiated products demand in a single market—e.g., breakfast cereal—commonly make the (in this context, often mild) assumption that all products are substitutes. However, when considering choice among a broader set of products, it is plausible that some of them are complements (i.e., a price fall for one product stimulates demand for another). A strand of literature seeks to maintain the disaggregate notion of products, defined by

[9] For instance, even if demand is constrained to be linear, after imposing Slutsky symmetry, there are $\frac{1}{2}J(J+1)$ price parameters.

[10] Bresnahan (1981) and Berry et al. (1995) use data on product characteristics, sales, and recommended list prices obtained from industry trade publications.

their bundles of characteristics, while also incorporating demand complementarities into consumer choice models. Recent examples include work by Lewbel & Nesheim (2019), who estimate a quadratic utility model of demand for fruit products, and by Ruiz et al. (2019), who estimate demand over 5,500 UPCs based on a sequential probabilistic model that envisages the consumer as making repeated but interacting discrete choices over all the available options in a store.

Dubois et al. (2014) estimate a demand model that nests models with preferences defined in product space and models with preferences defined in characteristics space. They use household scanner data from the United States, the United Kingdom, and France to study differences in demand patterns and preferences across countries. As scanner data are collected in similar ways across countries, often by the same market research firm, they facilitate cross-country comparisons that may not be possible otherwise. The authors also exploit the fact that data on nutrients from the back-of-package labels have been matched at the product level in each of these countries, allowing them to obtain accurate measures of the nutritional characteristics of households' shopping baskets.

3.2. Flexible Preference Heterogeneity

A consistent finding from consumer-level choice data is that there is great variation in the choices consumers make, and that this is driven both by differences in income and by heterogeneity in tastes [see, e.g., the review by Browning & Carro (2007)]. An important strength of scanner data is that they facilitate the modeling of consumer preference heterogeneity. This is true of scanner data in general, and particularly so for household scanner data.

By far the leading choice model used by researchers working with scanner data is the discrete choice random utility model, pioneered by McFadden (1974, 1978, 1980, 1984). Consumer preference heterogeneity can be included in this class of models by allowing tastes for product characteristics to vary with observable demographics or by including consumer-specific preferences. The latter are typically included through random coefficients, whereby the researcher seeks to estimate the superparameters governing the distribution of consumer preferences (for instance, the mean and standard deviation of a normal distribution of consumer preferences for a product's sugar content). In this case the choice model is known as either random coefficient or mixed logit, and, as shown by McFadden & Train (2000), if specified richly enough, it can approximate any random utility model to an arbitrary degree of accuracy.

In particular, such models typically assume that consumer i in market (period and/or region) t solves the choice problem $\max_{\{0,1,\ldots,J\}} U(\mathbf{x}_{jt}, y_i - p_{jt}, \epsilon_{ijt}; \theta_i)$, where $j = \{1, \ldots, J\}$ denotes different options available to the consumer ($j = 0$ indicates choosing not to purchase; $\mathbf{x}_{0t} = 0$, $p_{0t} = 0$); $\mathbf{x}_{jt}$ are product characteristics, y_i is consumer income, p_{jt} is product price, ϵ_{ijt} is an idiosyncratic shock to utility, and θ_i denotes the consumer's preferences. Researchers commonly assume that $U(.)$ takes the additively separable form $U_{ijt} = \mathbf{x}'_{jt}\beta_i + \alpha_i(y_i - p_{jt}) + \epsilon_{ijt}$; the consumer preferences, $\theta_i \equiv (\beta_i, \alpha_i)$, take the form $\theta_i = \bar{\theta} + \theta_z z_i + \sigma\eta_i$, where z_i denotes consumer demographics and η_i are unobserved consumer attributes. Hence, preferences may vary across consumers with observable and unobservable traits. In applied work it is common to assume that ϵ_{ijt} is an independent and identically distributed extreme value, and η_i are drawn from some known parametric distribution (e.g., independent normal distributions); this means that the market share of product j in market t takes the form

$$s_{ijt} = \int \frac{\exp(\mathbf{x}'_{jt}\beta_i - \alpha_i(y_i - p_{jt}))}{1 + \sum_{k=1,\ldots,J}(\exp(\mathbf{x}'_{kt}\beta_i - \alpha_i(y_i - p_{kt}))} dF(z_i, \eta_i).$$

The model can be estimated by maximum likelihood or generalized method of moments (GMM) (see Train 2003). It can also accommodate an unobserved attribute of each product that varies

across markets, which, as shown by Berry et al. (1995), can be contracted out and often plays a key role in identifying the parameter determining the impact of price on demand ($\bar{\alpha}$).

A key feature of choice models of this sort is that they readily aggregate from the utility maximizing decision rules of heterogeneous consumers to aggregate demand and market share functions that are tractable. This means that these models can be estimated with market-level data (i.e., data that aggregate over the individual choices of consumers). Many papers (including Nevo 2001, cited above) use store-level scanner data to estimate mixed logit choice models, which incorporate observed and unobserved (i.e., drawn from a parametrically specified distribution) preference heterogeneity. Berry & Haile (2014) provide a formal identification argument for the use of market-level data to uncover consumer preference distributions in differentiated products markets.

Household scanner data are particularly useful for modeling heterogeneous consumer preferences. As discussed by Berry & Haile (2016), consumer-level data (relative to market level data) allow the researcher to relax the formal conditions for nonparametric identification of differentiated product demand models. Household scanner data contain repeated observations for the same individual over time. The degree of correlation in a decision maker's choices as they face changing prices and product characteristics provides information about the strength of their individual preferences, leading to more robust identification of random coefficient distributions.

A number of papers have developed estimation methods that take advantage of consumer-level data to relax the parametric assumptions typically placed on random coefficient distributions (including Burda et al. 2008, Fox et al. 2011).[11] Burda et al. (2008) use household scanner data to implement a Bayesian Markov chain Monte Carlo estimation procedure applied to consumer store choice. They estimate consumer preferences over price (i.e., the price of a fixed basket of products) and travel distance and compare their nonparametric estimates of the distribution of consumer preferences with a parametric (normally distributed) one similar to those often used in practice. **Figure 1** shows the comparison of the parametrically and nonparametrically estimated consumer preference distributions. The latter indicates a significant departure from normality, with three modes: The largest one corresponds to consumers with moderate willingness to pay and to travel, and the two smaller ones correspond to people who are either very price sensitive but highly willing to travel, or price insensitive but less willing to travel long distances.

An alternative approach is taken by Dubois et al. (2020), who exploit the long time dimension of panel scanner data to estimate consumer-specific preference parameters (in contrast to treating them as random draws from a flexible distribution). This means that they can estimate the distribution of preferences (β_i, α_i) without assuming a parametric form and without the need to impose orthogonality assumptions between consumer preferences and other variables, including choices made in other markets. Their paper focuses on estimating demand for soft drinks when purchased for immediate consumption outside of the home, and it shows that consumers' preferences over the sugar in these products are stronger among the young, low-income individuals, and those with high overall dietary sugar; the paper also shows that these correlations drive who will respond to

[11] A prior literature develops identification arguments and estimators for semiparametric discrete choice models in which the distribution of the idiosyncratic shock, ϵ_{ijt}, is left unspecified. The majority of applied work using scanner data has maintained the assumption that ϵ_{ijt} is drawn from either a type I or a generalized extreme value distribution. We thank a referee for pointing out that the first estimator for discrete choice models that did not require a parametric distribution for the random terms was proposed by Manski (1975), followed by Cosslett (1983), Manski (1987), Matzkin (1992, 1993), Ichimura & Thompson (1998), and Lewbel (2000). Ichimura & Thompson (1998) were the first to develop a discrete choice model with a nonparametric distribution of random coefficients. Other notable exceptions are Briesch et al. (2010) and Lewbel (2000) (see also Greene 2009 for a survey).

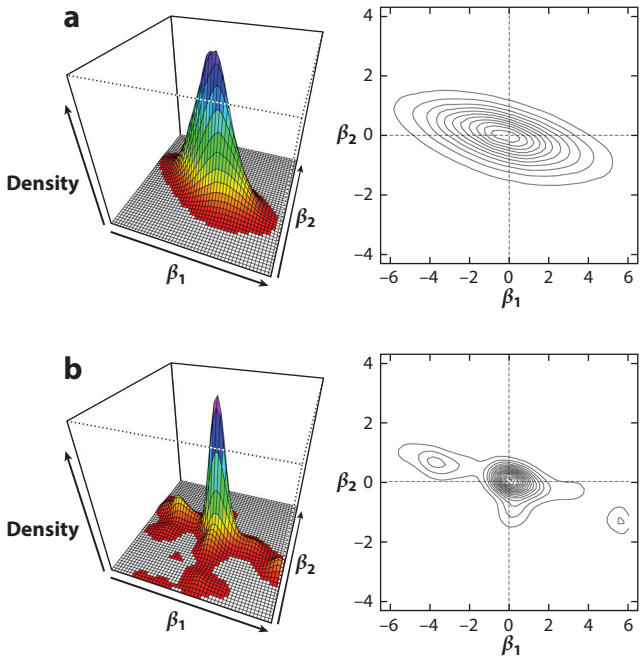

Figure 1

Distribution of consumer preferences. (*a*) Parametric specification. (*b*) Nonparametric specification. β_1 denotes consumer preferences over price interacted with distance, and β_2 denotes consumer preferences over distance. Figure adapted with permission from Burda et al. (2008).

the price changes that follow the introduction of a tax. Another novel feature of this paper is that it exploits data collected on individuals and therefore avoids making implicit assumptions about how households' choices relate to individual welfare.

3.3. Dynamics: Habits and Stockpiling

Dynamics in demand can arise if a consumer's current choice is influenced by decisions they have made or experiences they have been exposed to in the past. This, in turn, can lead the consumer to internalize the future effect of their behavior when making decisions. Here we focus on two forms of consumer dynamics that scanner data have been instrumental in allowing researchers to study.

The first area concerns the formation of consumers' brand preferences. This is important both for understanding the drivers of brand performance and market concentration and because of the significant evidence that consumption patterns (particularly for food) established early in life influence consumption and health outcomes in later life (e.g., Hoynes et al. 2016).

Bronnenberg et al. (2009) document that for many consumer goods there is substantial geographical dispersion across US cities in market shares, which is persistent over decades and is related to the original order of entry among surviving brands. This points to an early mover advantage, whereby early entrants into a geographical market are able to build up larger market shares relative to markets they enter later. Bronnenberg et al. (2012) provide evidence that, at the consumer level, brand tastes are highly persistent and evolve slowly over people's lifetimes. They do this by combining the rich brand-level purchase information contained in household

scanner data with survey information on the location history of members of the same households for which the scanner data are available; they show that current shopping behavior is associated with location of birth, and that the association becomes weaker the further in the past a consumer migrated from their birth location. This work is a good example of a scanner data provider being willing to collaborate with researchers in collecting valuable supplementary information. In their survey article, Bronnenberg & Dubé (2017) discuss the likely mechanisms underlying these patterns, and they suggest that informational frictions associated with learning about experiential brand characteristics are likely to be important in generating a kind of habit formation.

Scanner data have also led to significant progress in documenting and understanding a second source of consumer dynamics: the stockpiling of storable products. Using store scanner data, Pesendorfer (2002) documents that sales of ketchup during a period of temporarily low price are higher the longer the time since the previous discounted price. Similar patterns are found by Hendel & Nevo (2006b), who also use household scanner data to show that a household's propensity to buy on sale is negatively correlated with measures of storage costs, and that the duration to the next purchase is longer after buying on sale. This evidence is consistent with households' choice to build up inventories during sales periods and to draw them down when the good is not available at a discounted price.

Hendel & Nevo (2006a) build a model of consumer choice for storable products that captures the impact of consumers' tastes for product characteristics, their costs of storage, the size of their inventories, and their expectations of future price changes. They apply the model to laundry detergent, with the consumer choosing in each period how much (if any) to buy, which brand to purchase, and how much to consume in order to maximize the present expected value of the flow of future utility. A key challenge they face is that the consumer's inventory is unobserved. Using the panel structure of the scanner data, they are able to generate an initial distribution of inventories, updating it over time based on observed purchases and on the estimated optimal consumption decision rule. An important empirical finding from this work is that a static demand model estimated in the presence of stockpiling dynamics and temporary price reductions leads to overestimates of own price elasticities (as brand switching is conflating with intertemporal switching). Wang (2015) illustrates the importance of taking these effects into account when simulating the impact of the introduction of a tax.

A barrier to the use of these models is the substantial computational cost associated with their estimation. Hendel & Nevo (2013) consider a simplified stockpiling model in which consumers can store at no cost for a prespecified number of periods. This means that consumers face a problem analogous to a static one, but in which the effective price of a product is the minimum price seen in the set of periods immediately before the purchase decision, and including the current one over which storage is costless. This setup avoids the need to solve a Bellman equation and, as the authors show, means that the parameters of the model are identified based on the information available in store scanner data.

3.4. Advertising

Economists have long been interested in how advertising affects consumer choice and hence market structure and welfare (see Bagwell 2007). The availability of scanner data has led to important advances in our understanding of how exposure to advertising affects consumer choice.

To the best of our knowledge, Kanetkar et al. (1992) were the first to use scanner data to estimate the impact of advertising on demand. They match scanner data to household television advertising exposure, measured by set-top boxes, and estimate a choice model that allows the amount of advertising that a household has been exposed to since its most recent purchase to impact the current decision. Subsequent work has highlighted the importance of allowing for

the possible impact advertising has on the composition of consumers buying a product and on individual consumers' choice functions (Erdem et al. 2008), as well as allowing for the possibility of nonconvexities in how advertising impacts demand (Dubé et al. 2005).

An important strength of scanner data for uncovering the impact of advertising on demand is that price and quantity information is disaggregated by brand and region, key dimensions over which advertising varies. Household scanner data offer the additional advantage that they can be used to link a household's choices to its history of advertising exposure. For instance, Dubois et al. (2018) combine the detailed household-level television viewing information contained in the Kantar FMCG Purchase Panel with data on the universe of food and drink advertisements shown on television (including brand, time, show, and station). This enables them to compute household-level measures of exposure to brand advertising. They include this in a model of choice of potato chips. Controlling for the demographics that advertisers commonly target allows them to isolate exposure differences driven by variation in viewing habits within targeted groups. They find evidence that advertising of a given brand raises demand for it, flattens the demand curve, and typically lowers demand for alternatives. They use the model to simulate the impact of restricting advertising—a policy option intended to tackle obesity—and find that a resulting direct reduction in overall potato chip consumption is partially unwound by a reduction in equilibrium prices.

Another advantage of scanner data is that they contain information on dozens of different industries in which advertising expenditures are high. Shapiro et al. (2021) exploit this by estimating the relationship between how much a particular brand is sold and how much it is advertised. They find that across the 288 major brands they consider, the impact of an increase in advertising on quantity sold is modest, and they suggest that large advertising expenditures represent a misallocation of resources. Griffith et al. (2018a) use scanner data across 60 product categories to study the welfare implications of retailers' pricing and advertising strategies for their own (store) brand products, showing that the presence of store brands can increase aggregate consumer surplus. When advertising is rivalrous (i.e., it benefits a specific product rather than the entire category), advertising is typically overprovided by the market, because firms do not account for the negative externalities of their advertising on other firms. When making decisions for a store brand, the retailer internalizes some of the negative externalities from rivalrous advertising and therefore spends less on advertising.

4. SUPPLY AND MARKET EQUILIBRIUM

Understanding firms' supply decisions is key for addressing a number of important economic questions. These include measuring the extent of market power exercised by firms, testing different models of firm conduct, and undertaking counterfactual analysis of the effects of changes to the market environment—such as changes in input costs, mergers, and the introduction of new taxes—on equilibrium prices, quantities, and ultimately welfare.

4.1. Measuring Market Power and Merger Analysis

The majority of markets, and certainly those covered by scanner data, are characterized by differentiated products. In the preceding section we summarized some of the key papers that have used scanner data to estimate differentiated products demand. Obtaining credible demand estimates is almost always a necessary step for estimating a model of supply of differentiated products. A key challenge in supply-side estimation is that the marginal costs of products are almost always unobservable. However, combining demand estimates with assumptions about the form of firm conduct allows for the identification of marginal costs. This idea was originally implemented

using (nonscanner) data on the automobile market (see Bresnahan 1987, Berry et al. 1995), but it has been widely used and extended in studies exploiting scanner data.[12]

The canonical supply-side Bertrand model entails a set of firms $f = 1, \ldots, F$ and of products $j \in \mathcal{J}$, where each firm owns some subset of the products, $\mathcal{J}_f \subset \mathcal{J}$. Letting $\mathbf{c} = \{c_j\}_{j \in \mathcal{J}}$ denote marginal costs, $\mathbf{p} = \{p_j\}_{j \in \mathcal{J}}$ denote prices, and $\mathbf{q}(\mathbf{p}) = \{q_j(\mathbf{p})\}_{j \in \mathcal{J}}$ denote quantities produced, each firm is assumed to choose prices to maximize their profits:

$$\Pi_f = \sum_{j \in \mathcal{J}_f} (p_j - c_j) q_j(\mathbf{p}).$$

Stacking the J first-order conditions, marginal costs can be written as

$$\mathbf{c} = \mathbf{p} + \left[\Omega \circ \left(\frac{\partial \mathbf{q}}{\partial \mathbf{p}} \right) \right]^{-1} \times \mathbf{q}(\mathbf{p}),$$

where Ω is the $J \times J$ ownership matrix [the (j, k) element equals 1 if product j and k are owned by the same firm and zero otherwise], and $\circ$ denotes element-by-element matrix multiplication. The important point is that, under the maintained assumption of Bertrand competition, it is possible to use demand estimates and observed prices to back out the implied marginal costs. This enables researchers to measure the extent to which prices are marked up above marginal costs ($(p_j - c_j)/p_j$), a measure of market power called the Lerner index, and to use the estimates of demand and supply primitives to undertake counterfactual analysis. As scanner data are typically available at either the brand or the UPC level, and contain well-measured prices, they are ideally suited to this kind of supply analysis.

In the preceding section we referred to the seminal paper by Nevo (2001), who estimates consumer demand in the breakfast cereal market using store scanner data aggregated to the city-quarter-brand level. He uses these estimates to identify brand-level marginal costs under three alternative assumptions: that firms engage in Bertrand competition, that products are priced as if they were sold by single-product firms, and that firms in the market engage in joint-profit maximization. Each of these corresponds to a different configuration of the ownership matrix.[13] By comparing the three alternative sets of Lerner indices with approximate measures from accounting data, he concludes that Bertrand competition best fits the data. The study highlights that high equilibrium Lerner indices (averaging around 0.4) can be sustained without collusion, in large part due to the market power that firms derive from internalizing pricing externalities among the several brands that they own.

In a related paper, using data on the same market, Nevo (2000) uses the supply framework to simulate the impact of a series of mergers on prices and welfare. Having estimated demand and marginal costs, he solves for the counterfactual price vector, $\tilde{\mathbf{p}}$, using the system of modified first-order conditions

$$\tilde{\mathbf{p}} = \mathbf{c} - \left[\widetilde{\Omega} \circ \left(\frac{\partial \mathbf{q}}{\partial \mathbf{p}} \right) \right]^{-1} \times \mathbf{q}(\tilde{\mathbf{p}}),$$

where $\widetilde{\Omega}$ denotes the counterfactual post-merger ownership matrix. Following this work, the use of scanner data and the modeling of differentiated product demand and supply have become an

[12] Scanner data have also been used to extend the trade and macroeconomic literature on firm heterogeneity. For instance, Hottman et al. (2016) develop a model, in the tradition of Melitz (2003), with heterogeneous multiproduct firms, estimate it using scanner data, and provide evidence that at least half of the heterogeneity in firm size, and almost the entirety of firm growth, can be attributed to firm appeal (i.e., quality or taste).

[13] In particular, this entails replacing the Bertrand ownership matrix, Ω, with the identity matrix (single-product firms) or a matrix of 1s (joint-profit maximization) in the marginal cost expression.

increasingly important element in the tool kit of competition authorities.[14] Rather than simulating the effects of a merger, to test the Bertrand model of competition Miller & Weinberg (2017) use scanner data at the store-week-UPC level in the beer market that cover a time when a merger (or joint venture) between two firms took place.[15] They reject the hypothesis that the post-merger joint venture entirely fails to internalize pricing externalities, but their estimates also suggest the joint venture does not fully internalize them. Recently, Backus et al. (2021) have proposed a test, based on those by Vuong (1989) and Rivers & Vuong (2002), for alternative models of firm conduct. They apply it to the breakfast cereal market, finding that Bertrand competition is more consistent with the data than a model in which firms internalize common-ownership effects due to overlapping shareholders.

4.2. Retailer Behavior and Vertical Relations

A criticism that has sometimes been levied at differentiated product demand and supply models is that they often do not explicitly incorporate retailer behavior. Scanner data, which typically include information on the retailer in which a product was purchased, have enabled researchers to tackle this issue by incorporating strategic retailers into supply models. It is rare for researchers to observe the details of manufacturer-retailer contracts; therefore, the supply model with strategic retailers and manufacturers is a function of additional unknowns. Sudhir (2001), using scanner data on yogurt and peanut butter purchases from two regional chains in the United States, extends the canonical differentiated product supply model to incorporate a strategic retailer considering several alternative models of linear pricing. Using scanner data in a Midwestern city, available at the week-UPC-retailer level, Villas-Boas (2007) considers several models of vertical relations, including linear pricing, vertical integration, and a pricing equilibrium obtained under nonlinear pricing, with each implying different vectors of wholesale and retailer price-cost margins. She then estimates the relationship between these vectors of price-cost margins and prices and uses it as the basis to conduct non-nested Cox-type tests of the best-fitting model. Relatedly, Bonnet & Dubois (2010) formalize the different possible vertical contracts between manufacturers and retailers, including two-part tariffs contracts with or without resale price maintenance, as possible rationales for different price equilibrium. They use household-level scanner data in the French bottled water market and apply the non-nested test proposed by Rivers & Vuong (2002). By studying a narrow market in detail, these papers shed light on the role that strategic interactions between retailers and manufacturers play in determining equilibrium prices.

In contrast, Thomassen et al. (2017) consider consumer choice over which supermarket to shop at, and how much expenditure to allocate across the goods available in the store, in order to measure the extent of market power exercised by UK supermarkets. They use household-scanner data and aggregate the tens of thousands of products offered by supermarkets into eight broad categories. They show that fixed shopping costs (e.g., due to time and travel costs) give rise to complementarities in demand for categories sold by the same retailer, and that this acts to lower equilibrium prices set by strategic retailers relative to a scenario in which the commodities are all priced by independent category managers.

Another strand of research on retailer behavior that the availability of scanner data has contributed to is on the dynamics of store entry. Holmes (2011) studies the geographical rollout of

[14]Readers are referred to, for example, US Dep. Justice & Fed. Trade Comm. (2006), Wang (2013), Eur. Comm. Dir. Gen. Compet. (2015).

[15]They do this by specifying that the (j, k) elements of the ownership matrix that correspond to pairs of products owned by either of the premerged firms equal κ and they estimate this parameter.

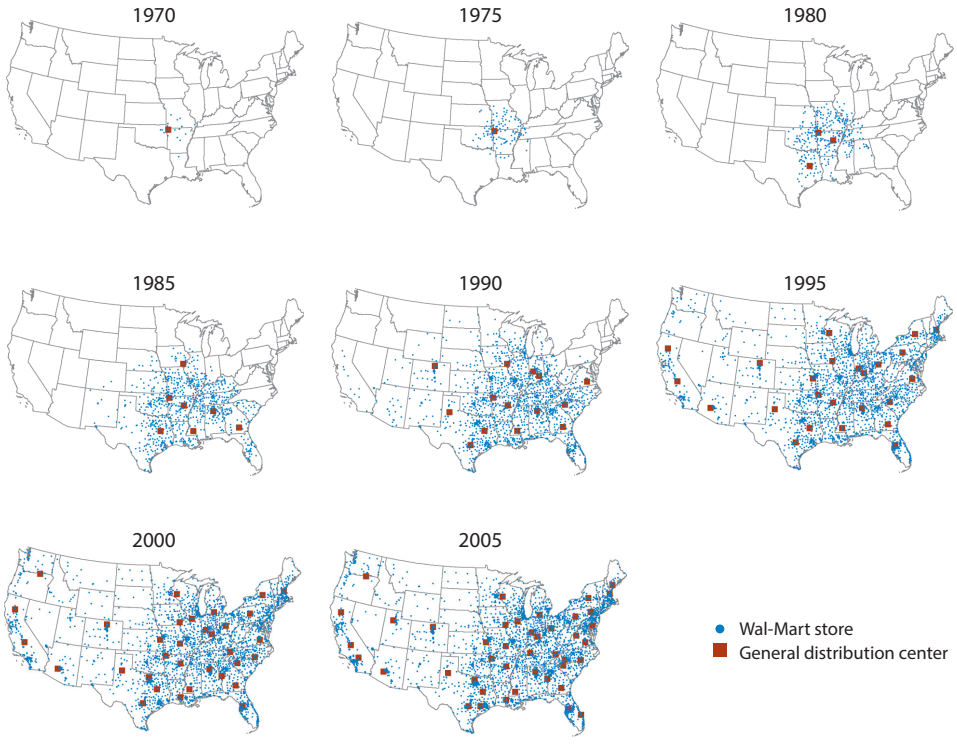

Figure 2

Entry of Wal-Mart stores in the United States. Red squares indicate the location of Wal-Mart distribution centers; blue dots indicate the location of Wal-Mart stores. Figure adapted with permission from Holmes (2011).

Wal-Mart stores in the United States.[16] Understanding how the retailer grew so rapidly since its entry in Arkansas in 1962 is of interest to competition authorities and firm strategists, as well as to policy makers concerned with food habits and dietary diseases. **Figure 2** illustrates that Wal-Mart store openings radiated from the retailer's initial presence in Arkansas, with new stores always located near areas where it already had a presence. Holmes (2011) estimates the net benefits of following this strategy of maintaining a dense network of stores, accounting for the trade-off between the logistical advantages of lower distribution costs and the ability to quickly respond to changes in demand on the one hand, and the costs of sales cannibalization from nearby stores on the other. The comprehensive sales information available for individual stores contained in the Nielsen Retail Scanner data is a key input into modeling consumer store choice and hence the extent of sales cannibalization.

Many of the papers discussed so far in this section focus on the measurement of market power and the identification of firm conduct, thereby addressing questions at the heart of the study of industrial organization. However, supply-side decisions are relevant in many other areas of economics, and by enabling better measurement and modeling of the supply side of markets, scanner data have contributed to important advances beyond industrial organization. One example of this

[16]Wal-Mart is by far the largest retailer in the United States, accounting for around a quarter of all groceries sold in the country (see **https://www.foodindustry.com/articles/top-10-grocers-in-the-united-states-2019/**).

is the study of food deserts. A number of public health researchers contend that food retailers' entry decisions are key drivers of nutritional inequalities, because in poorer neighborhoods—where diet quality and health outcomes tend to be well below average—access to healthy foods is often difficult (a situation known as food deserts).[17] An alternative explanation for this pattern is that these differences in supply are the equilibrium response to differences in consumer preferences. Allcott et al. (2019a) use information on supermarket entry and household moves contained in scanner data to test whether store offerings in poorer neighborhoods are drivers of nutritional inequality. They find that food deserts contribute only marginally to inequalities in diet quality, and that differences in preferences for food are much more influential in driving the inequalities. They conclude that policy would be better served by focusing on the targeted subsidization of healthy foods than by influencing retailer supply decisions.

4.3. Exchange Rate Pass-Through

Another example of how the supply-side modeling facilitated by scanner data has led to important advances in the economics field is the study of pass-through of exchange rate movements to consumer prices, a question of central importance in international economics. Hellerstein (2008) investigates this question using data from Dominick's Finer Foods, a retailer operating in the Chicago metropolitan area. This widely used data set contains information at the UPC-week level, and it unusually includes wholesale as well as retail prices. She uses a manufacturer-retailer supply model for the beer market and estimates the relationship between inferred marginal costs and exchanges rates. She finds that local-cost components and mark-up adjustments (i.e., firms re-optimizing their prices) contribute equally to incomplete pass-through of exchange rates to prices.

By applying a static supply-side week-by-week model to the beer market, Hellerstein (2008) assumes that firms optimally set prices each week. However, a key empirical fact emerging from scanner data is that prices are sticky, with nonsale prices often remaining unchanged for at least 1 year (see Eichenbaum et al. 2011). Goldberg & Hellerstein (2013) extend the beer supply-side model by incorporating the costs of adjusting prices (i.e., menu costs). They assume that when price changes the adjustment is optimal, but when price remains unchanged it is a consequence of firms choosing not to pay the menu cost. This enables them to bound the size of adjustment costs. They find that after these are accounted for, markup adjustment is much less important for explaining incomplete exchange rate pass-through; quantitatively, this channel is largely replaced by the influence of price rigidities. In related work, Nakamura & Zerom (2010) study the sources of incomplete exchange rate pass-through in the market for coffee. They incorporate menu costs by explicitly modeling firms' decisions over when to adjust prices. They do this by building on earlier work with scanner data that extends the static supply model to a dynamic setting (see Aguirregabiria 1999). In contrast to Goldberg & Hellerstein (2013), they find that markup adjustment remains an important source of incomplete pass-through and that menu costs play only a minor role. The availability of product-level price and quantity information in scanner data, in addition to the wholesale prices in the Dominick's data set, have played an important role in the development of this literature.

5. MEASUREMENT OF INFLATION

Scanner data have played an important role in facilitating advances in the measurement of inflation and in our understanding of what drives variation in inflation across different households.

[17] Readers are referred to the review by Bitler & Haider (2011).

Traditionally, national statistical offices have produced official measures of consumer inflation—CPIs—that are constructed using price quotes, collected in person, and expenditure weights for broad commodity groups that are based on consumer expenditure surveys. The resulting index provides a picture of inflation experienced by the representative consumer. Studies that have harnessed scanner data for inflation measurement have been able to address a number of the limitations inherent in official CPIs. However, because scanner data typically cover only fast-moving consumer goods, work on inflation measurement with scanner data has necessarily focused on a subset of the economy.

5.1. Heterogeneity in Spending and Prices Paid

A central contribution of this literature has been to document the heterogeneity in inflation across households. This may arise due to differences in expenditure patterns across households or differences in prices paid for the same good. The data underlying CPIs provide limited scope for capturing this variation. One reason for this is that information on expenditure shares is collected only for relatively broad commodities; typically CPIs give equal weight to all products within each of these groups. Yet, as documented by Jaravel (2019), variation in expenditure shares across income groups principally arises within broad commodity groups, across very disaggregate products.

A second reason is that official statistics generally sample only one price for each product in the basket at each time period, and therefore they do not record differences in prices paid for the same product. Scanner data have been used to demonstrate that, in the United States at least, such differences can be large. For instance, Kaplan et al. (2019) show that the standard deviation of prices offered for the same product in the same geographical area and week is 15.3 percent, and it is mainly caused by dispersion in the price of a particular good relative to the price of other goods across different stores, and not by dispersion in the average price of goods across different stores.

DellaVigna & Gentzkow (2019) show that within a retail chain, the price charged for a specific good can vary a lot from week to week but it is close to uniform across stores. **Figure 3** illustrates

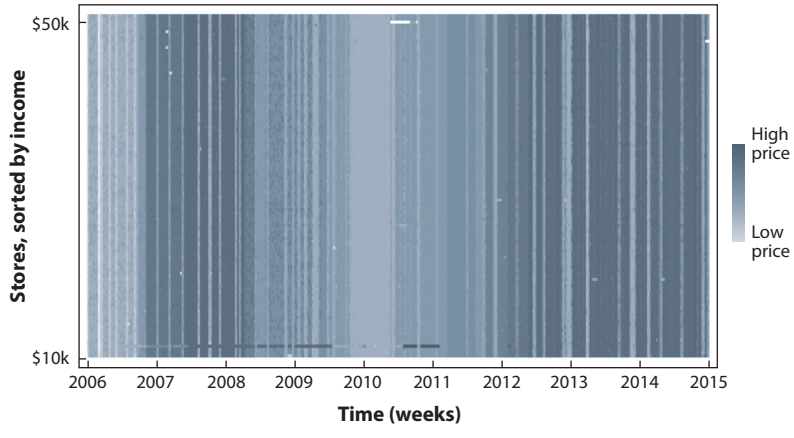

Figure 3

Retailer uniform pricing. The color shading in the figure indicates the log price in a store-week for a particular orange juice product sold in the same retail chain. Darker colors indicate higher price, and the figure is blank if price is missing. The horizontal axis indicates the week. The vertical axis indicates the store, where stores are sorted by income per capita of shoppers in that store. Figure adapted with permission from DellaVigna & Gentzkow (2019).

this. The figure plots the log price of a single orange juice product (a 59 oz. bottle of pulp-free Simply Orange juice) in one retail chain. Along the horizontal axis, each column corresponds to a week. Along the vertical axis, each row corresponds to one of 108 stores in the chain, and stores are sorted by store-level income per capita (divided into $10,000 differences in the per-capita income measure). Darker colors indicate higher week-store prices; blanks correspond to missing prices. The figure shows that in any time period there is little price variation across stores (the colors tend to be uniform in the columns) but there is a lot of variation in the horizontal dimension—that is, prices vary a lot over time. The authors show that this pattern holds across many products and many chains, which suggests that the differences in prices for the same good that have been highlighted in the literature are primarily due to dispersion in the price charged by different retail chains. Butters et al. (2022) argue that the tendency for retailers to set uniform prices is mainly driven by their insensitivity to local demand conditions, and by exploiting variation in local taxes, they show that firms do adjust prices in response to local cost shocks.

Aguiar & Hurst (2007) highlight the importance of differences in prices paid for identical goods between those working and those retired for understanding the retirement-saving puzzle. They use household scanner data to show that households with members in their late 60s pay 4% less for a basket of identical products compared to households with members in their 40s. They show that this can largely be explained by the fact that older households shop more frequently and are more likely to use coupon discounts, which they attribute to a lower opportunity cost of time. They suggest that this is an additional reason (along with an increase in home production at retirement) that drops in total expenditure at retirement are unlikely to translate into large consumption falls. Griffith et al. (2016) and Nevo & Wong (2019) show that over the Great Recession, households increased their shopping intensity, which is consistent with a fall in the opportunity cost of time, and that they also switched to buying in bulk and buying more on promotions, both of which are associated with large savings (see Griffith et al. 2009). Coibion et al. (2015) quantify the importance of cyclical adjustment in shopping behavior for inflation using IRI Infoscan store-level scanner data for the United States. They exploit regional variation in unemployment rates to show that inflation in the prices households pay is substantially more cyclical than inflation in posted prices, and that this is driven by consumers reallocating expenditures across retailers.

5.2. Heterogeneity in Inflation Rates

The use of scanner data has helped illuminate just how big dispersion in inflation rates is across households. Kaplan & Schulhofer-Wohl (2017) focus on measuring household-level inflation rates and show that the annual interquartile range in household inflation rates is 6.2–9.0 percentage points in the United States (based on Nielsen Homescan Consumer Panel data). They show that two-thirds of this variation is due to differences in prices paid for the same good, with the rest being mainly due to differences in spending patterns within broad categories; differences in expenditure shares across broad categories play only a minor role.

Jaravel (2019) also uses Nielsen Homescan Consumer Panel data to measure inflation rates by income groups and shows that between 2004 and 2015, the bottom income quintile experienced annual inflation rates 0.66 percentage points higher than the top quintile. Increases in product variety, well measured in scanner data due to their product-level nature, are more pronounced for higher-income groups. He uses plausibly exogenous changes in market size driven by differences in population growth by sociodemographic groups to show that this inequality in inflation and product variety is driven by faster demand growth among high- relative to low-income consumers, which leads to more innovation among products preferred by those with high incomes. These differences in inflation experiences across households and income groups have important

implications for the measurement of poverty and the design of welfare benefits and tax brackets. As highlighted by Handbury (2021), differences in cost-of-living measures by income interact with regional variation in product prices and availability. She shows that, relative to low-income households, high-income households enjoy 40% higher utility per dollar expenditure in wealthy cities relative to poor cities.

Redding & Weinstein (2020) propose a new price index that allows for taste shocks (thereby relaxing the assumption of time-invariant tastes that underlies much preceding work on price indices). They use Nielsen Homescan Consumer Panel data to quantify a substantial taste-shock bias that tends to lead to an upward bias in inflation measurement under standard price indices.

Jaravel (2021) provides a recent review of this literature. A key theme is that the granular quantity and price information provided by scanner data have been key to the recent progress of this literature, as they allow researchers to measure differences in spending and price paid for disaggregate products across households of different incomes and in different locations.

5.3. Real-Time Inflation Measurement

Another important advantage of scanner data for inflation measurement is that the data are available almost in real time. This can be particularly helpful for tracking what is happening to prices in times of crisis. Jaravel & O'Connell (2020) use scanner data for the United Kingdom to document inflation at the beginning of the COVID-19 pandemic. This was a period characterized by the threat of major disruption to supply chains, and sector shutdowns and home-working led to large changes in spending patterns. It was also a time during which policy makers were required to take immediate decisions, including over how to support households subject to financial pressures. The paper documents that the onset of the national lockdown in the United Kingdom coincided with a large spike in the month-to-month inflation rate for fast-moving consumer goods, which rose to 2.4%. This was primarily driven by a significant withdrawal of promotions by the major retailers.

5.4. Improving Official Inflation Measurement

Scanner data offer the possibility of improving official inflation measurement for the sectors of the economy that they cover. They provide information on the prices of many more products than it would be feasible to collect in person. They include expenditure weights that are at the disaggregate product level, allowing for weighting of the importance of products within broad commodity groups. The data are available almost in real time, meaning that the expenditure weights are up to date—in contrast with those typically used in CPIs, which are from consumer expenditure surveys and are available with at least 1 year's lag. The International Labour Organization's (2004) *Consumer Price Index Manual* recommends using chained price indices when high-frequency data are available. This entails updating the expenditure weights used in the index for each period (usually 1 month) and ensures that the inflation measure reflects up-to-date expenditure patterns (rather than patterns 1 or 2 years in arrears, as in traditional CPI measurement). A challenge that has hindered incorporating high-frequency chained indices into CPIs is that they can suffer from the chain-drift problem, whereby a high-frequency relationship between prices and quantities leads the inflation index to become increasingly biased over time. However, Ivancic et al. (2011) illustrate empirically that the use of multilateral price indices, where the price level in one period is computed in comparison to the price level in many other periods, rather than only the preceding one, can help solve this problem. As evidence on the merits of using scanner data for inflation measurement has grown in the literature, national statistical offices have begun to

integrate scanner data into official inflation measures, with the Australian Bureau of Statistics, which introduced scanner data into its CPI in 2014, being at the vanguard of this move.

6. FINAL COMMENTS

The availability of scanner data to researchers has enabled many important advances in economics, including the study of consumer choice, firms' strategic decision making, the equilibrium implications of policy interventions, and the measurement of prices and inflation. Researchers have had access to these data for only a few decades, and the number of papers that use scanner data continues to grow over time. In the coming years, research using scanner data promises to deliver further frontier contributions. These data have enabled researchers to address a host of policy questions, including assessing the impact of mergers, advertising restrictions, and taxes on prices, profits, consumer surplus, and nutrition, as well as the impacts on household inequalities of differences in food availability, price, and product variety changes. Scanner data are now part of the tool kit of competition authorities and, increasingly, of national statistic authorities tasked with measuring consumer price inflation.

The technology used to collect product-level information on prices and quantities is now being extended from fast-moving consumer goods to other sectors. One example is the collection of data covering dine-in restaurants, fast foods, takeaways, and other food and drinks consumed outside of the home. There is relatively little work on consumer choice and firm behavior in these markets, and yet they are of interest because they account for a substantial share of consumer spending, because choices in these markets have important implications for health and well-being, and because there are good reasons to think that the choice environment, consumers' decision-making processes, and firm behaviors in this context might differ from those in the more commonly studied supermarkets and grocery stores.

Advances in computational and statistical methods are enabling researchers to exploit the richness of scanner data in a number of new ways. For example, exploiting the longitudinal nature of the data, these methods open up the possibility to estimate richer dynamic models. Most studies focus on a single or a narrow range of products, but scanner data contain information on hundreds of thousands of products with potentially interrelated demand and supply curves. New methods open the possibility of better understanding these relationships.

Scanner data are collected in many countries in a similar way. This aspect of the data has not been very well exploited, with the majority of papers focusing on the United States and only a handful of studies making cross-country comparisons. There are rich opportunities to exploit data in different countries to better understand how institutional and cultural differences drive differences in market outcomes.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from the Economic and Social Research Council (ESRC) under the Centre for the Microeconomic Analysis of Public Policy (CPP), grant number ES/T014334/1, and under the Open Research Area (ORA), grant number ES/VO13513/1, as well as from the Agence Nationale de la Recherche under grant ANR17-EURE-0010 (Investissements d'Avenir program).

LITERATURE CITED

Aguiar M, Hurst E. 2007. Life-cycle prices and production. *Am. Econ. Rev.* 97(5):1533–59

Aguirregabiria V. 1999. The dynamics of markups and inventories in retailing firms. *Rev. Econ. Stud.* 66(2):275–308

Allcott H, Diamond R, Dube JP, Handbury J, Rahkovsky I, Schnell M. 2019a. Food deserts and the causes of nutritional inequality. *Q. J. Econ.* 134(4):1793–844

Allcott H, Lockwood BB, Taubinsky D. 2019. Regressive sin taxes, with an application to the optimal soda tax. *Q. J. Econ.* 134(3):1557–626

Backus M, Conlon C, Sinkinson M. 2021. *Common ownership and competition in the ready-to-eat cereal industry*. NBER Work. Pap. 28350

Bagwell K. 2007. The economic analysis of advertising. In *Handbook of Industrial Organization*, Vol. 3, ed. M Armstrong, R Porter, pp. 1701–844. Amsterdam: North-Holland

Beck GW, Jaravel X. 2020. *Prices and global inequality: new evidence from worldwide scanner data*. Work. Pap., Siegen Univ., Siegen, Ger.

Berry S, Haile P. 2014. Identification in differentiated products markets using market level data. *Econometrica* 82(5):1749–97

Berry S, Haile P. 2016. Identification in differentiated products markets. *Annu. Rev. Econ.* 8:27–52

Berry S, Haile P. 2020. *Nonparametric identification of differentiated products demand using micro data*. NBER Work. Pap. 27704

Berry S, Levinsohn J, Pakes A. 1995. Automobile prices in market equilibrium. *Econometrica* 63(4):841–90

Bitler M, Haider SJ. 2011. An economic view of food deserts in the United States. *J. Policy Anal. Manag.* 30(1):153–76

Bonnet C, Dubois P. 2010. Inference on vertical contracts between manufacturers and retailers allowing for nonlinear pricing and resale price maintenance. *RAND J. Econ.* 41(1):139–64

Bresnahan TF. 1981. Departures from marginal-cost pricing in the American automobile industry. *J. Econom.* 17(2):201–27

Bresnahan TF. 1987. Competition and collusion in the American automobile industry: the 1955 price war. *J. Ind. Econ.* 35(4):457–82

Briesch R, Chintagunta P, Matzkin R. 2010. Nonparametric discrete choice models with unobserved heterogeneity. *J. Bus. Econ. Stat.* 28:291–307

Bronnenberg BJ, Dhar SK, Dubé JPH. 2009. Brand history, geography, and the persistence of brand shares. *J. Political Econ.* 117(1):87–115

Bronnenberg BJ, Dubé JP. 2017. The formation of consumer brand preferences. *Annu. Rev. Econ.* 9:353–82

Bronnenberg BJ, Dubé JPH, Gentzkow M. 2012. The evolution of brand preferences: evidence from consumer migration. *Am. Econ. Rev.* 102(6):2472–508

Browning M, Carro J. 2007. Heterogeneity and microeconometrics modelling. In *Advances in Economics and Econometrics*, Vol. 3, ed. R Blundell, W Newey, T Persson, pp. 47–74. Cambridge, UK: Cambridge Univ. Press

Burda M, Harding M, Hausman J. 2008. A Bayesian mixed logit-probit model for multinomial choice. *J. Econom.* 147(2):232–46

Butters RA, Sacks DW, Seo B. 2022. How do national firms respond to local cost shocks? *Am. Econ. Rev.* 112:1737–72

Chetty R, Looney A, Kroft K. 2009. Salience and taxation: theory and evidence. *Am. Econ. Rev.* 99(4):1145–77

Coibion O, Gorodnichenko Y, Hong GH. 2015. The cyclicality of sales, regular and effective prices: business cycle and policy implications. *Am. Econ. Rev.* 105(3):993–1029

Cosslett S. 1983. Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* 51(3):765–82

Deaton A. 1986. Demand analysis. In *Handbook of Econometrics*, Vol. 3, ed. Z Griliches, M Intriligator, pp. 1767–839. Amsterdam: North-Holland

DellaVigna S, Gentzkow M. 2019. Uniform pricing in US retail chains. *Q. J. Econ.* 134(4):2011–84

Dubé JP, Hitsch GJ, Manchanda P. 2005. An empirical model of advertising dynamics. *Quant. Mark. Econ.* 3:107–44

Dubois P, Griffith R, Nevo A. 2014. Do prices and attributes explain international differences in food purchases? *Am. Econ. Rev.* 104(3):832–67
Dubois P, Griffith R, O'Connell M. 2018. The effects of banning advertising in junk food markets. *Rev. Econ. Stud.* 1(1):396–436
Dubois P, Griffith R, O'Connell M. 2020. How well targeted are soda taxes? *Am. Econ. Rev.* 110(11):3661–704
Eichenbaum M, Jaimovich N, Rebelo S. 2011. Reference prices, costs, and nominal rigidities. *Am. Econ. Rev.* 101(1):234–62
Eizenberg A, Lach S, Oren-Yiftach M. 2021. Retail prices in a city. *Am. Econ. J. Econ. Policy* 13(2):175–206
Erdem T, Keane M, Sun B. 2008. The impact of advertising on consumer price sensitivity in experience goods markets. *Quant. Mark. Econ.* 6(2):139–76
Eur. Comm. Dir. Gen. Compet. 2015. *A review of merger decisions in the EU: What can we learn from ex post evaluations?* Rep., Eur. Comm., Luxemb.
Fox JT, Kim K, Ryan SP, Bajari P. 2011. A simple estimator for the distribution of random coefficients. *Quant. Econ.* 2(3):381–418
Gelman M, Kariv S, Shapiro MD, Silverman D, Tadelis S. 2014. Harnessing naturally occurring data to measure the response of spending to income. *Science* 345(6193):212–15
Goldberg P, Hellerstein R. 2013. A structural approach to identifying the sources of local currency price stability. *Rev. Econ. Stud.* 80(1):175–210
Gorman WM. 1980. A possible procedure for analysing quality differentials in the egg market. *Rev. Econ. Stud.* 47(5):843–56
Greene W. 2009. Discrete choice modeling. In *Palgrave Handbook of Econometrics: Applied Econometrics*, Vol. 2, ed. TC Mills, K Patterson, pp. 473–556. London: Palgrave Macmillan
Griffith R, Krol M, Smith K. 2018a. Why do retailers advertise store brands differently across product categories? *J. Ind. Econ.* 66(3):519–69
Griffith R, Leibtag E, Leicester A, Nevo A. 2009. Consumer shopping behavior: How much do consumers save? *J. Econ. Perspect.* 23(2):99–120
Griffith R, O'Connell M, Smith K. 2016. Shopping around: how households adjusted food spending over the Great Recession. *Economica* 83(330):247–80
Griffith R, O'Connell M, Smith K. 2019. Tax design in the alcohol market. *J. Public Econ.* 172:20–35
Griffith R, von Hinke S, Smith S. 2018b. Getting a healthy start: the effectiveness of targeted benefits for improving dietary choices. *J. Health Econ.* 58:176–87
Guadagni PM, Little J. 1983. A logit model of brand choice calibrated on scanner data. *Mark. Sci.* 2(3):203–38
Handbury J. 2021. Are poor cities cheap for everyone? Non-homotheticity and the cost of living across U.S. cities. *Econometrica* 89(6):2679–715
Hellerstein R. 2008. Who bears the cost of a change in the exchange rate? Pass-through accounting for the case of beer. *J. Int. Econ.* 76(1):14–32
Hendel I, Nevo A. 2006a. Measuring the implications of sales and consumer inventory behavior. *Econometrica* 74(6):1637–73
Hendel I, Nevo A. 2006b. Sales and consumer inventory. *RAND J. Econ.* 37(3):543–61
Hendel I, Nevo A. 2013. Intertemporal price discrimination in storable goods markets. *Am. Econ. Rev.* 103(7):2722–51
Holmes TJ. 2011. The diffusion of Wal-Mart and economies of density. *Econometrica* 79(1):253–302
Hottman C, Redding S, Weinstein D. 2016. Quantifying the sources of firm heterogeneity. *Q. J. Econ.* 131(3):1291–364
Hoynes H, Schanzenbach DW, Almond D. 2016. Long-run impacts of childhood access to the safety net. *Am. Econ. Rev.* 106(4):903–34
Ichimura H, Thompson S. 1998. Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *J. Econom.* 86(2):269–95
Int. Labour Organ. 2004. *Consumer Price Index Manual: Theory and Practice.* Geneva, Switz.: Int. Labour Organ.
Ivancic L, Fox KJ, Diewert EW. 2011. Scanner data, time aggregation and the construction of price indexes. *J. Econom.* 161(1):24–35
Jaravel X. 2019. The unequal gains from product innovations: evidence from the U.S. retail sector. *Q. J. Econ.* 134(2):715–83

Jaravel X. 2021. Inflation inequality: measurement, causes, and policy implications. *Annu. Rev. Econ.* 13:599–629
Jaravel X, O'Connell M. 2020. Real-time price indices: inflation spike and falling product variety during the Great Lockdown. *J. Public Econ.* 191:104270
Kanetkar V, Weinberg CB, Weiss DL. 1992. Price sensitivity and television advertising exposures: some empirical findings. *Mark. Sci.* 11(4):359–71
Kaplan G, Menzio G, Rudanko L, Trachter N. 2019. Relative price dispersion: evidence and theory. *Am. Econ. J. Microecon.* 11(3):68–124
Kaplan G, Schulhofer-Wohl S. 2017. Inflation at the household level. *J. Monet. Econ.* 91:19–38
Lancaster K. 1966. A new approach to consumer theory. *J. Political Econ.* 74(2):132–57
Lewbel A. 2000. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *J. Econom.* 97:145–77
Lewbel A, Nesheim L. 2019. *Sparse demand systems: corners and complements*. CEMMAP Work. Pap. CWP45/19, Cent. Microdata Methods Pract., London
Manski CF. 1975. Maximum score estimation of the stochastic utility model of choice. *J. Econom.* 3(3):205–28
Manski CF. 1987. Semiparametric analysis of random effects linear models from binary panel data. *Econometrica* 55(2):357–62
Matzkin R. 1992. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica* 60(2):239–70
Matzkin R. 1993. Nonparametric identification and estimation of polychotomous choice models. *Econometrica* 58(1–2):137–68
McFadden D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, ed. P Zarembka, pp. 105–42. New York: Academic
McFadden D. 1978. Quantitative methods for analyzing travel behaviour of individuals: some recent developments. In *Behavioural Travel Modelling*, ed. D Hensher, P Stopher, pp. 279–318. London: Croom Helm
McFadden D. 1980. Econometric models for probabilistic choice among products. *J. Bus.* 53(3):S13–29
McFadden D. 1984. Econometric analysis of qualitative response models. In *Handbook of Econometrics*, Vol. 2, ed. Z Griliches, M Intriligator, pp. 1395–457. Amsterdam: North-Holland
McFadden D, Train K. 2000. Mixed MNL models for discrete response. *J. Appl. Econom.* 15:447–70
Melitz MJ. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71(6):1695–725
Miller NH, Weinberg MC. 2017. Understanding the price effects of the MillerCoors joint venture. *Econometrica* 85(6):1763–91
Nakamura E, Steinsson J. 2008. Five facts about prices: a reevaluation of menu cost models. *Q. J. Econ.* 123(4):1415–64
Nakamura E, Zerom D. 2010. Accounting for incomplete pass-through. *Rev. Econ. Stud.* 77(3):1192–230
Nevo A. 2000. Mergers with differentiated products: the case of the ready-to-eat cereal industry. *RAND J. Econ.* 31(3):395–421
Nevo A. 2001. Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69(2):307–42
Nevo A. 2011. Empirical models of consumer behavior. *Annu. Rev. Econ.* 3:51–75
Nevo A, Wong A. 2019. The elasticity of substitution between time and market goods: evidence from the Great Recession. *Int. Econ. Rev.* 60(1):25–51
O'Connell M, Smith K, Stroud R. 2021. *The dietary impact of the COVID-19 pandemic.* IFS Work. Pap. 21/18, Inst. Fiscal Stud., London
Pesendorfer M. 2002. Retail sales: a study of pricing behavior in supermarkets. *J. Bus.* 75(1):33–66
Redding SJ, Weinstein DE. 2020. Measuring aggregate price indices with taste shocks: theory and evidence for CES preferences. *Q. J. Econ.* 135(1):503–60
Rivers D, Vuong Q. 2002. Model selection tests for nonlinear dynamic models. *Econom. J.* 5(1):1–39
Rosen S. 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *J. Political Econ.* 82(1):34–55
Ruiz FJR, Athey S, Blei DM. 2019. SHOPPER: a probabilistic model of consumer choice with substitutes and complements. arXiv:1711.03560 [stat.ML]

Shapiro BT, Hitsch GJ, Tuchman AE. 2021. TV advertising effectiveness and profitability: generalizable results from 288 brands. *Econometrica* 89(4):1855–79
Sudhir K. 2001. Structural analysis of manufacturer pricing in the presence of a strategic retailer. *Mark. Sci.* 20(3):244–64
Thomassen Ø, Smith H, Seiler S, Schiraldi P. 2017. Multi-category competition and market power: a model of supermarket pricing. *Am. Econ. Rev.* 107(8):2308–51
Train KE. 2003. *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge Univ. Press
US Dep. Justice, Fed. Trade Comm. 2006. *Commentary on the horizontal merger guidelines*. Tech. Rep., US Dep. Justice, Fed. Trade Comm., Washington, DC
Villas-Boas SB. 2007. Vertical relationships between manufacturers and retailers: inference with limited data. *Rev. Econ. Stud.* 74(2):625–52
Vuong QH. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(2):307–33
Wang EXR. 2013. *Economic tools for evaluating competitive harm in horizontal mergers*. Note, Charles River Assoc., Boston, MA
Wang EY. 2015. The impact of soda taxes on consumer welfare: implications of storability and taste heterogeneity. *RAND J. Econ.* 46(2):409–41