DISCUSSION PAPER SERIES

DP16954
(v. 2)

The Use of Scanner Data for Economics Research

Pierre Dubois, Rachel Griffith and Martin O'Connell

INDUSTRIAL ORGANIZATION

MONETARY ECONOMICS AND FLUCTUATIONS

PUBLIC ECONOMICS

ISSN 0265-8003

The Use of Scanner Data for Economics Research

Pierre Dubois, Rachel Griffith and Martin O'Connell

Discussion Paper DP16954
First Published 25 January 2022
This Revision 22 March 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Industrial Organization
- Monetary Economics and Fluctuations
- Public Economics

The Use of Scanner Data for Economics Research

Abstract

The adoption of barcode scanning technology in the 1970's 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 typically either 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 pro and cons of this data source, and discuss some of the ways its availability to researchers has transformed the economics literature.

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

Keywords: scanner data, Demand estimation, market power, policy counterfactual, inflation

Pierre Dubois - pierre.dubois@tse-fr.eu
Toulouse School of Economics and CEPR

Rachel Griffith - rachel.griffith@ifs.org.uk
University of Manchester and CEPR

Martin O'Connell - moconnell9@wisc.edu
University of Wisconsin-Madison and CEPR

Acknowledgements
Prepared for Annual Review of Economics, https://doi.org/10.1146/annurev-economics-051520-024949. The authors would like to 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, and the ANR under grant ANR17-EURE-0010 (Investissements d'Avenir program).

The Use of Scanner Data for Economics Research

Pierre Dubois, Rachel Griffith and Martin O'Connell*

March, 2022

Abstract

The adoption of barcode scanning technology in the 1970's 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 typically either 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 pro and cons of this data source, and discuss some of the ways its availability to researchers has transformed the economics literature.

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

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

Acknowledgments: Prepared for Annual Review of Economics, https://doi.org/10.1146/annurev-economics-051520-024949. The authors would like to 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, and the ANR under grant ANR17-EURE-0010 (Investissements d'Avenir program).

*Correspondence: Dubois: Toulouse School of Economics, pierre.dubois@tse-fr.eu; Griffith: Institute for Fiscal Studies and University of Manchester, rgriffith@ifs.org.uk; O'Connell: University of Wisconsin-Madison and Institute for Fiscal Studies, martin_o@ifs.org.uk

1 Introduction

The advent of barcode scanner technology in the 1970's led to a new form of data known as scanner data. Researchers quickly recognized the value of using these data to learn about what factors influence consumers' choices and demands.[1] Scanner data has since given rise to a voluminous literature that seeks to use it 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 in macro and monetary economics (such as 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 check-outs. 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 include key product characteristics, such as brand, manufacturer, package type, flavors, etc., as well as the location (store) of purchase and date of purchase. They are often longitudinal, recording repeated transactions for the same household, individual or store over time. In addition, they often also provide further details about the transaction, such as whether the product was on promotion (for instance, subject to a temporary price reduction or 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]

[1] To the best of our knowledge the first paper to use scanner data to estimate a model of consumer choice is Guadagni and 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 and Steinsson (2008) and Eizenberg, Lach, and Oren-Yiftach (2021)). However, these data typically cover only a subset of products, and do not contain infor-

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 provide receipts to the market research firm to validate purchases. The data therefore contain a record of individual transactions (i.e., barcode X, was 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, 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 is also available from market research firms who 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 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 Base,[6] and the Dominik's supermarket Database.[7]

---

mation on disaggregate product sales or quantities. This motivates recent interest in harnessing scanner data in official inflation indices. We discuss this in Section 5.

[3]These can be accessed for research purposes from the University of Chicago; see https://www.chicagobooth.edu/research/kilts/datasets/nielsen.

[4]See https://www.kantarworldpanel.com/global/Coverage/worldpanel/United-Kingdom.

[5]These can be accessed for research purposes from the University of Chicago; see https://www.chicagobooth.edu/research/kilts/datasets/nielsen.

[6]See https://www.iriworldwide.com/en-us/solutions/academic-data-set.

[7]These can be accessed for research purposes from the University of Chicago; see https://www.chicagobooth.edu/research/kilts/datasets/dominicks.

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 empirically modeling product differentiation, heterogeneity in consumer preferences, consumer dynamics and the role of advertising in impacting 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, merger analysis, modeling strategic retailers and their vertical relations with manufacturers, and 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. Our aim is not to provide a comprehensive survey – there are both 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

Disaggregate price and quantity information

A key advantage of scanner data is that it provides 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, evaluating mergers, and assessing a range of government policies. As we also discuss below, detailed price and quantity information also facilitates 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 US or the Living Cost and Food Survey (LCFS) in the UK, 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, and industry dynamics, within specific (e.g., the breakfast cereal) markets. To address some research questions it may be desirable to work with data that is at a level more aggregate 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, O'Connell, and Smith (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 will have different tax liabilities.

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 consumers' choice over where to shop, how retailers interact with manufacturers, how the choice environment affects consumer decisions, and 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 in 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.

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. Since 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 US and Understanding Society in the UK, collect information from

households infrequently (quarterly, annually or biennially) and do not contain detailed spending information. Recently researchers have used high frequency bank and credit card transaction data (e.g., Gelman, Kariv, Shapiro, Silverman, and Tadelis (2014)), but this does not provide product level information. The panel structure of household scanner data (combined with the product level information it contains) is advantageous for a number of reasons.

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 it allows for identification of random coefficient distributions under weaker conditions than with market level data (e.g., see Berry and 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 consumer specific preferences. In addition, the time series dimension of household scanner data is sometimes sufficiently long that it is feasible to directly estimate household specific preference parameters in non-linear choice models. As we discuss below, this allows researchers to relax commonly imposed distributional assumptions and weakens the independence assumption placed on household preferences.

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 dependence of someone's choice today on what they chose in the past, or features of their past choice environment. This enables researchers to study behaviors such as habit formation, stockpiling, consumer choice inertia, and temptation.

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. It happens that sometimes information useful in economic research is either 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. Examples include Allcott, Lockwood, and Taubinsky (2019), who 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 sugary sweetened beverages, and Griffith, von Hinke, and Smith (2018), who estimate the probability households in their scanner data set are in receipt of benefits by matching on geographic location and household characteristics with the LCFS, which contains information on benefit claims. Other researchers have succeeded in securing the cooperation of scanner data collectors in conducting an experiment. One example is Chetty, Looney, and Kroft (2009), who randomly varied price posting between the sales tax exclusive price (as in normal on the US) and the tax inclusive price in a Northern Californian store to estimate how tax salience impacts the incidence and excess burden of taxation.

Spending on other goods and services. The vast majority of research using scanner data has been with data covering 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 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 non-alcoholic beverage purchases for consumption outside the home – including those made in dine-in restaurants (for instance, see O'Connell, Smith, and Stroud (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, DIY) in several countries (see Beck and Jaravel (2020) who uses this data set to measure differences in inflation and product entry across countries).

3 Demand

The answers to many empirical questions in economics rely upon on having credible estimates of consumer demand. These include, for instance, understanding how consumers will adjust their choices in response changes in firms' strategies or government policy (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

[8] See Investopedia.com

the implications of consumer choice behavior for profits and firms' pricing and marketing strategies, 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 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 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 (Gorman (1980), Lancaster (1966), Rosen (1974) and McFadden (1974)). 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 – which provides product level information on the key product characteristics, as well as prices and quantities – 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 the firms in the market. In addition to work that focuses on differentiation of products in characteristics space, scanner data, which crucially contain information on store of purchase, has given rise to research that studies differentiation across stores, which arises both through dif-

[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] See Bresnahan (1981) and Berry, Levinsohn, and Pakes (1995), who use data on product characteristics, sales and recommended list prices obtained from industry trade publications.

ferentiation in store characteristics (e.g., floorspace) and geographical differences in location, and how this influences where consumers choose to shop (e.g., Thomassen, Smith, Seiler, and Schiraldi (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 (a price fall for one product stimulates demand for another). A strand of literature seeks to maintain the disaggregate notion of a product (defined by its bundle of characteristics), but to model choice allowing for complementarities in demand. Recent examples include Lewbel and Nesheim (2019), who estimate a quadratic utility model of demand for fruit products (apples, oranges etc.) and Ruiz, Athey, and Blei (2019) who estimate demand over 5,500 UPCs based on a sequential probabilistic model that envisages the consumer making repeated but interacting discrete choices over all of the available options in a store.

Dubois, Griffith, and Nevo (2014) estimate a demand model that nests models with preferences defined in product space and models where preferences are defined in characteristics space. They use household scanner data from the US, UK 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, which may otherwise not be possible due to differences in the way that other data are collected. 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 heterogeneity in tastes (for instance, see the review by Browning and Carro (2007)). An important strength of scanner data is that they facilitate modeling 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 model by

interacting tastes for product characteristics with observable demographics, or by inclusion of consumer-specific preferences. The latter are typically included through random coefficients, where the researcher seeks to estimate the super-parameters 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 the latter case the choice model is known as either a random coefficient or mixed logit and, as shown by McFadden and Train (2000), if specified richly enough, 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 consumer income, p_{jt} 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, $U_{ijt} = \mathbf{x}'_{jt}\beta_i + \alpha_i(y_i - p_{jt}) + \epsilon_{ijt}$ and 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 distributed i.i.d. extreme value and η_i are drawn from some known parametric distribution (e.g., independent normal distributions), which means that the market share of product j in market t takes the form:

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

The model can be estimated my maximum likelihood or GMM (see Train (2003)). It can also accommodate an unobserved market varying product attribute, which, as shown by Berry, Levinsohn, and Pakes (1995), can be contracted out and often plays a key roll 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, or market share, functions that are tractable. This means 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., consumer-level preferences drawn from a parametrically specified distribution) preference heterogeneity. Berry and Haile (2014) provide a formal

identification argument for the use of market level data to uncover consumer preferences distributions in differentiated products markets.

Household scanner data are particularly useful for modeling heterogeneous consumer preferences. As discussed in Berry and Haile (2016), consumer level data allow the researcher to relax the formal conditions (relative to market level data) for non-parametric identification of differentiated products 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 the prices and product characteristics they face change, provides information about the strength of their individual preferences, leading to more robust identification of random coefficients distributions.

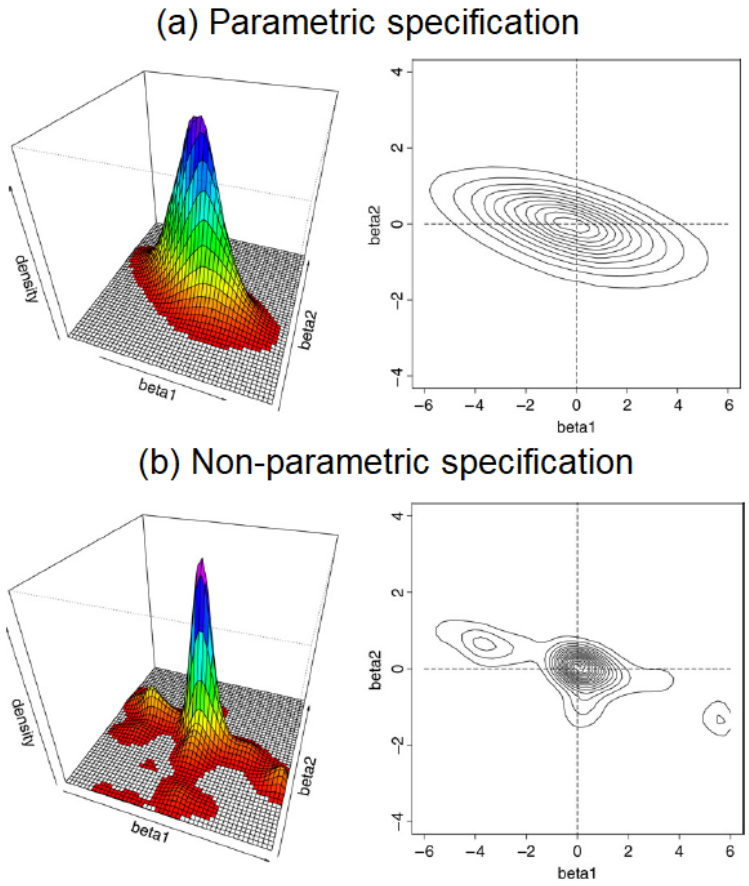
A number of papers have developed estimation methods that take advantage of consumer level data to relax parametric assumptions typically placed on random coefficients distributions (including Burda, Harding, and Hausman (2008) and Fox, Kim, Ryan, and Bajari (2011)).[11] Burda, Harding, and Hausman (2008) implement a Bayesian Markov Chain Monte Carlo estimation procedure applied to consumer store choice using household scanner data. They estimate consumer preferences over price (i.e., the price of a fixed basket of products) and travel distance and compare their non-parametric estimates of the distribution of consumer preferences to a parametric (normally distributed) one similar to those often used in practice. Figure 3.1 shows the comparison of the parametric and non-parametrically estimated consumer preference distribution. The latter indicates a significant departure from normality, with three modes, the largest one corresponding to consumers with moderate willingness to pay and travel, and two smaller ones corresponding to people who are either very price sensitive but highly willing to travel, or price insensitive but that place a very high value on convenience.

An alternative approach is taken in Dubois, Griffith, and O'Connell (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

[11] A prior literature develops identification arguments and estimators for semi-parametric 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 distributed type I extreme value. 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 Manski (1975), followed by Cosslett (1983), Manski (1987), Matzkin (1992), Matzkin (1993), Ichimura and Thompson (1998), and Lewbel (2000). Ichimura and Thompson (1998) were the first to develop a discrete choice model with a nonparametric distribution of random coefficients. Other notable exceptions are Briesch, Chintagunta, and Matzkin (2010) and Lewbel (2000); see also Greene (2009) for a survey.

(β_i, α_i) without assuming a parametric form and avoid the need to impose orthogonality assumptions between consumer preferences and other variables, including choices made in other markets. This paper focuses on estimating demand for soft drinks when purchased for immediate consumption outside of the home, and shows that consumers' preferences over the sugar in these products are stronger among the young, the low income and those with high overall dietary sugar, and that these correlations drive who responds to the price changes that would 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' choice relates to individual welfare.

Figure 3.1: *Distribution of consumer preferences, Burda, Harding, and Hausman (2008)*



Reproduction of Figures 13 and 14 in Burda, Harding, and Hausman (2008) under license.
Notes: β_1 denotes consumer preferences over price interacted with distance, and β_2 denotes consumer preferences over distance.

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 to the consumer internalizing the future effect of their behavior when making decisions. Here we focus on two forms of consumer dynamics where the availability of scanner data has contributed to significant advances.

The first area concerns the formation of consumers' brand preferences. This is important both for understanding the drivers of brand performance and market concentration, but also 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, Schanzenbach, and Almond (2016)).

Bronnenberg, Dhar, and Dubé (2009) document that for many consumer goods there is substantial geographical dispersion (across US cities) in market shares that is persistent over decades and 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, Dubé, and Gentzkow (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 collected for the same households on the location history of members of the household, and showing that current shopping behavior is associated with location of birth and the association becoming weaker the further in the past a consumer migrated from their birth location. This work is a good example of a scanner data provider been willing to collaborate with researchers in collecting valuable supplementary information. In their survey article Bronnenberg and Dubé (2017) discuss the likely mechanisms underlying these patterns and suggest informational frictions associated with learning about experiential brand characteristics is likely to be important in generating a kind of habit formation.

Scanner data has also led to significant progress in documenting and understanding a second source of consumer dynamics – stockpiling of storable products. Using store scanner data Pesendorfer (2002) documents that sales of ketchup increase with the time elapsed since the previous sale. Similar patterns are found by Hendel and 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 choosing to build up inventories

during sales periods and drawing them down when the good is not available at a discounted price.

Hendel and Nevo (2006a) build a model of consumer choice for storable products that captures the impact of consumer tastes for product characteristics, their cost of storage, the size of their inventories and their expectations of future price changes on their decision. They apply the model to laundry detergent, with each period the consumer choosing 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 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 over estimates of own price elasticities (as brand switching is conflating with intertemporal switching). Wang (2015) illustrates the importance of taking account of these effects 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 estimating them. Hendel and Nevo (2013) consider a simplified stockpiling model, in which consumers can store at no cost for a pre-specified number of periods. This means that consumers face a problem analogous to a static one, but where the effective price of a product is the minimum price seen in the set of periods immediately before and including the current one over which storage is costless. This set up avoids the need to solve a Bellman equation and, as the authors show, means 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, Weinberg, and Weiss (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 their most recent previous purchase to impact

their current decision. Subsequent work has highlighted the importance of allowing for the possible impact advertising has on the composition of consumers buying a product as well as on individual consumers' choice functions (Erdem, Keane, and Sun (2008)) and for the possibility of non-convexities in how advertising impacts demand (Dubé, Hitsch, and Manchanda (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 their history of advertising exposure. For instance, Dubois, Griffith, and O'Connell (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 adverts 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 motivated to tackle obesity – and find that a resulting direct reduction in overall potato chip consumption is partially unwound by firms lowering equilibrium prices.

Another advantage scanner data offers is that it contains information on dozens of different industries in which advertising expenditures are high. Shapiro, Hitsch, and Tuchman (2021) exploit this by estimating the relationship between the quantity of a particular brand 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 suggest that large advertising expenditures represent a misallocation of resources. Griffith, Krol, and Smith (2018) 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 (it benefits a specific product rather than the entire category), advertising is typically over-provided 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 so 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 and 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 summarize 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 identification of marginal costs. This idea was originally implemented using (non-scanner) data on the automobile market (see Bresnahan (1987), Berry, Levinsohn, and Pakes (1995)), but 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 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, $\mathbf{p} = \{p_j\}_{j\in\mathcal{J}}$ prices and $\mathbf{q}(\mathbf{p}) = \{q_j(\mathbf{p})\}_{j\in\mathcal{J}}$ 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 necessary 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-

[12] Scanner data has also been used to extend the trade and macroeconomic literature on firm heterogeneity. For instance, Hottman, Redding, and Weinstein (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 all firm growth, can be attributed to firm appeal (i.e., quality or taste).

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 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 typically are available either at the brand or UPC level, and contain well measured prices, they are ideally suited to this kind of supply analysis.

In the preceding section we refer to the seminal paper by Nevo (2001), who estimates consumer demand in the breakfast cereal market using store scanner, which has been 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 are sold by single product firms, and that firms in the market engage in joint profit-maximization. Each of these correspond 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 firms derive from internalizing pricing externalities between 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 differentiated product demand and supply modeling have become an increasingly important element in the toolkit of competition authorities.[14] Rather than simulating the effects of a merger, Miller and Weinberg (2017) use store-week-UPC level scanner data in the beer market that covers a time when a merger (or joint venture) between two firms took place to test the Bertrand

[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.

[14] See, for example, Commission and US Department of Justice (2006), European Commission Directorate General for Competition (2015), Wang and Vistnes (2013).

model of competition.[15] They reject the hypothesis that the post-merger joint venture fails to internalize pricing externalities at all, but their estimates also suggest the joint venture does not fully internalize these. Recently Backus, Conlon, and Sinkinson (2021) propose a test, based on those in Vuong (1989) and Rivers and Vuong (2002), for alternative models of firm conduct and 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 includes 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) extends the canonical differentiated product supply model to incorporate a strategic retailer, considering several alternative models of linear pricing (using scanner data on yogurt and peanut butter purchases from two local regional chains in the US). 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 that could be obtained under non-linear pricing, with each implying different vectors of wholesale and retailer price-cost margins. She then estimates the relationship between these and prices and uses it as the basis to conduct non-nested Cox type tests of the best fitting model. Relatedly, Bonnet and Dubois (2010) formalize the different possible vertical contracts between manufacturers and retailers including two-part tariffs contracts with or without Resale Price Maintenance as a possible rationale for different price equilibrium. They use household level scanner data in the French bottle water market and apply the non nested test proposed by Rivers and 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.

[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 pre-merged firms equals κ and they estimate this parameter.

In contrast, Thomassen, Smith, Seiler, and Schiraldi (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 10,000s 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 if the commodities were all priced by independent category managers.

Another strand of research on retailer behavior that the availability of scanner data has contributed to is the dynamics of store entry. Holmes (2011) studies the geographical rollout of Wal-Mart stores in the US.[16] Understanding how it grew so rapidly, from its entry in Arkansas in 1962, is of interest to competition authorities and firm strategists, as well as policymakers concerned with food habits and dietary diseases. Figure 4.1 illustrates that Wal-Mart store openings radiated from its 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 advantage of lower distribution costs and the ability to quickly respond to changes in demand, with the costs of sales cannibalization from nearby stores. The comprehensive sales information available for individual stores contained in the AC Nielsen scanner data is a key input into modeling consumer store choice and hence extent of sales cannibalization.

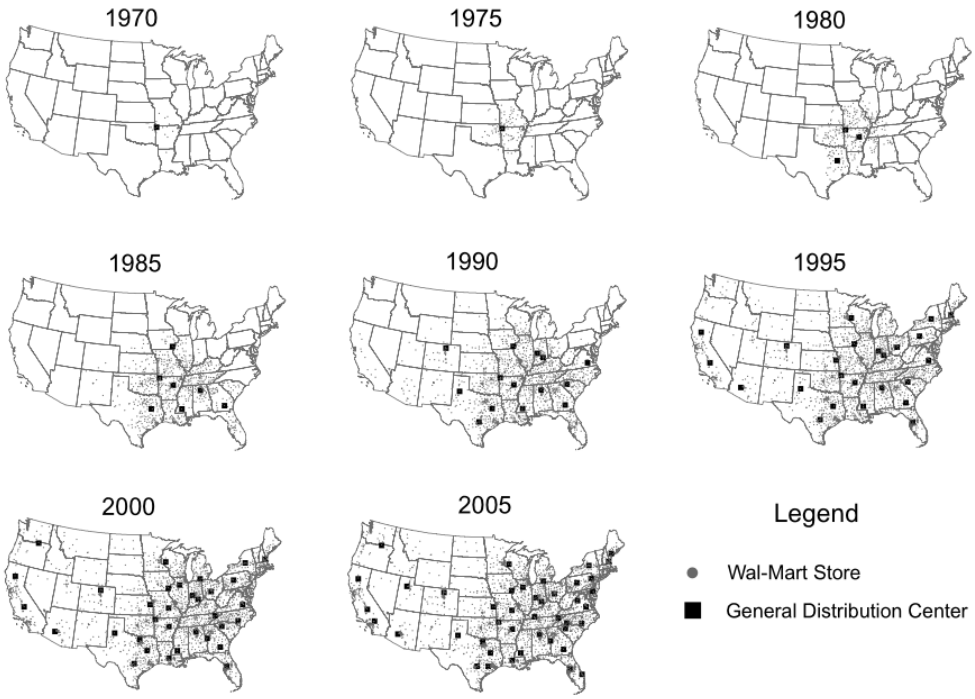
Many of the papers that we have discussed so far in this section have focused on the measurement of market power and identifying firm conduct, thereby addressing questions at the heart of the study of industrial organization. However, supply-side decisions are relevant in many other areas of economics, 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 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 – often there is a less access to healthy foods (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, Diamond,

[16] Wal-Mart is by far the largest retailer in the US, accounting for around like a quarter of all groceries sold in the US, see Foodindusty.com.

[17] For instance, see review in Bitler and Haider (2011).

Dube, Handbury, Rahkovsky, and Schnell (2019) use information on supermarket entry and household moves that are contained in scanner data to test whether the store offering in poorer neighborhoods is a driver of nutritional inequality. They find that food deserts only contribute 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 influencing retailer supply decisions.

Figure 4.1: *Entry of Wal-Mart stores in the US, Holmes (2011)*



Reproduction of Figure 1 in Holmes (2011) under license.
Notes: Illustrates diffusion of Wal-Mart stores and distribution centers.

4.3 Exchange rate pass-through

Another example of where the supply-side modeling facilitated by scanner data has led to important advances 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 that operated in the Chicago metropolitan area. This widely used data set contains information at the UPC-week level, and 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 reoptimizing their prices) both contribute equally to incomplete pass-through of exchange rates to prices.

By applying a static supply-side model week-by-week to the beer market Hellerstein (2008) assumes that firms optimally set prices each week. However, a key empirical fact to emerge from scanner data is that prices are sticky, with non-sale prices often remaining unchanged for at least a year (see Eichenbaum, Jaimovich, and Rebelo (2011)). Goldberg and Hellerstein (2013) extend the beer supply-side model by incorporating costs of adjusting prices (menu-costs). They assume that when price changes the adjustment is optimal, but when price remains unchanged this is a consequence of firms choosing to not pay the menu-cost. This enables them to bound the size of adjustment costs. They find that after these are accounted for mark-up 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 and 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 and Hellerstein (2013), they find mark-up 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 dataset, 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 what drives variation in inflation across different households. Traditionally national statistical offices have produced official measures of consumer inflation – Consumer Price Indexes (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 necessarily has focused on a subset of the economy.

Heterogeneity in spending and prices paid

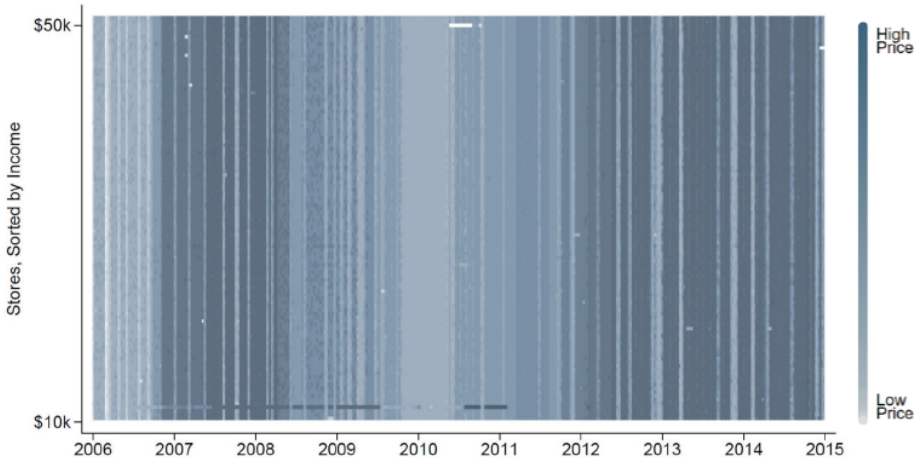
A central contribution of this literature has been to document the heterogeneity in inflation across households. This may arise due to differences in expenditure patterns across households, or differences in prices paid for the same good. The data underlying CPIs provide limited scope for capturing this 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 only sample one price for each product in the basket at each time period and therefore do not record differences in prices paid for the same product. Scanner data has been used to demonstrate that, in the US at least, such differences can be large. For instance, Kaplan, Menzio, Rudanko, and Trachter (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 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 and Gentzkow (2019) show that within a retail chain, while the price charged for a specific good can vary a lot from week-to-week, it is close to uniform across stores. Figure 5.1 illustrates 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 \$10,000s 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 – prices vary a lot over

time. The authors show that this patterns holds across many products and many chains, which suggests that the differences in prices offered 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, Sacks, and Seo (forthcoming) argue that the tendency for retailers to set uniform prices is mainly driven by their insensitivity to local demand conditions, and show, by exploiting variation in local taxes, that firms do adjust prices in response to local cost shocks.

Figure 5.1: *Retailer uniform pricing, DellaVigna and Gentzkow (2019)*



Reproduction of Figure 1(A) in DellaVigna and Gentzkow (2019) under license.
Notes: Figure shows log price in store s (plotted on the vertical axis) and week t (plotted on the horizontal axis) for a particular orange juice product. Stores belong to the same retail chain. Darker colors indicate higher price, and the figure is blank if price is missing. Each column is a week. Each row is a store, and stores are sorted by store-level income per capita.

Aguiar and Hurst (2007) highlight the importance of differences in prices paid for identical goods between those working and those retired for interpretation of the retirement-saving puzzle. They use household scanner data to show that households in their late 60s pay 4 percent less for a basket of identical products than households 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) why drops in total expenditure at retirement are unlikely to translate into large consumption falls. Griffith, O'Connell, and Smith (2016) and Nevo and Wong (2019) show that over the Great Recession households increased their shopping intensity, consistent with a fall in the opportunity cost of time, and that they also switched to buying in bulk and more on promotions, both associated with large savings (see Griffith, Leibtag, Leicester, and Nevo (2009)). Coibion, Gorodnichenko, and Hong (2015) quantify the importance of cyclical adjustment in shopping behavior for inflation using IRI store level scanner data for the US. 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.

Heterogeneity in inflation rates

The use of scanner data has helped illuminate just how big dispersion in inflation rates are across households. Kaplan and 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 US (using AC Nielsen household scanner data). They show that two-thirds of this variation is due to differences in prices paid for the same good, with the rest mainly being 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 AC Nielsen data to measure inflation rates by income groups and shows that between 2004-2015 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 the product level nature of the data, 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 percent higher utility per dollar expenditure in wealthy cities relative to poor cities.

Redding and Weinstein (2020) propose a new price index that allows for taste shocks (thereby relaxing the assumption of time-invariant tastes underlying much preceding work on price indices). They use AC Nielsen 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, which allows us to measure differences in spending and price paid for disaggregate products across

households of different incomes and in different locations, has been key to the recent progress of this literature.

Real-time inflation measurement

Another important advantage of scanner data for inflation measurement is that it is available in close to real-time. This can be particularly helpful for tracking what is happening to prices in times of crisis. Jaravel and O'Connell (2020) use scanner data for the UK to document inflation at the beginning of the COVID-19 pandemic. This was a period when there was the threat of major disruption to supply chains, and sector shut downs and home working led to large changes in spending patterns. It was also a time when policymakers were required to take immediate decisions, including over how to support households subject to financial pressures. The paper documents that the onset of the UK's national lockdown 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.

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 price quotes for. They include expenditure weights that are at the disaggregate product level, allowing for weighting of the importance of products *within* broad commodity groups. They are available in close to real-time, meaning the expenditure weights are up to date (in contrast to those typically used in CPIs, which are from consumer expenditure surveys and available with at least a year's lag). The International Labour Organization's Consumer Price Manual (2004) recommends using chained price indices when high frequency data are available. This entails updating expenditure weights used in the index each period (usually month) and ensures the inflation measure reflects up to date expenditure patterns (rather than patterns a year or two 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, Fox, and Diewert (2011) illustrate empirically that the use of multi-lateral price indices, where the price level in one period is computed with comparison to the level in many other periods rather than only the preceding one,

can help solve this problem. As evidence in the academic literature on the merits of using scanner data for inflation measurement has grown, 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, 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 decisions making, the equilibrium implications of policy interventions and the measurement of prices and inflation. Despite researchers having access to these data for a few decades, the numbers 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 of differences in food availability and in price and product variety changes on various forms of across household inequalities. Scanner data are now part of the toolkit 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, 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 might differ to in the more commonly studied supermarket and grocery store context.

Advances in computational and statistical methods are enabling researchers to exploit the richness of scanner data in a number of new ways. For example, combined with the longitudinal nature of the data, these methods open up the possibility to estimate richer dynamics models. Most studies focus on a single or narrow range of products, but scanner data contain information on 100,000s of products with po-

tentially 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 US and only a handful of papers making cross-country comparisons. There are rich opportunities to exploit data in different countries to better understand how institutional and cultural differences drives the differences in market outcomes.

References

Aguiar, M. and E. Hurst (2007). Life-cycle prices and production. *American Economic Review 97*(5), 1533–1559.

Aguirregabiria, V. (1999). The Dynamics of Markups and Inventories in Retailing Firms. *Review of Economic Studes 66*(2), 275–308.

Allcott, H., R. Diamond, J.-P. Dube, J. Handbury, I. Rahkovsky, and M. Schnell (2019). Food Deserts and the Causes of Nutritional Inequality. *Quarterly Journal of Economics 134*(4), 1793–1844.

Allcott, H., B. B. Lockwood, and D. Taubinsky (2019). Regressive Sin Taxes, with an Application to the Optimal Soda Tax. *Quarterly Journal of Economics 134*(3), 1557–1626.

Backus, M., C. Conlon, and M. Sinkinson (2021). Common Ownership and Competition in the Ready-to-Eat Cereal Industry. Technical Report w28350, National Bureau of Economic Research, Cambridge, MA.

Bagwell, K. (2007). The Economic Analysis of Advertising. In *Handbook of Industrial Organization*, Volume 3, pp. 1701–1844. North-Holland.

Beck, G. W. and X. Jaravel (2020). Prices and Global Inequality: New Evidence from Worldwide Scanner Data. *SSRN Electronic Journal*.

Berry, S. and P. Haile (2014). Identification in Differentiated Products Markets Using Market Level Data. *Econometrica 82*(5), 1749–1797.

Berry, S. and P. Haile (2016). Identification in Differentiated Products Markets. *Annual Review of Economics 8*(1), 27–52.

Berry, S. and P. Haile (2020). Nonparametric Identification of Differentiated Products Demand Using Micro Data. Technical Report w27704, National Bureau of Economic Research, Cambridge, MA.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile Prices in Market Equilibrium. *Econometrica 63*(4), 841–890.

Bitler, M. and S. J. Haider (2011). An economic view of food deserts in the united states. *Journal of Policy Analysis and Management 30*(1), 153–176.

Bonnet, C. and P. Dubois (2010). Inference on vertical contracts between manufacturers and retailers allowing for nonlinear pricing and resale price maintenance. *RAND Journal of Economics 41*(1), 139–164.

Bresnahan, T. F. (1981). Departures from marginal-cost pricing in the American automobile industry. *Journal of Econometrics 17*(2), 201–227.

Bresnahan, T. F. (1987). Competition and Collusion in the American Automobile Industry: The 1955 Price War. *The Journal of Industrial Economics 35*(4), 457.

Briesch, R., P. Chintagunta, and R. Matzkin (2010). Nonparametric Discrete Choice Models With Unobserved Heterogeneity. *Journal of Business & Economic Statistics 28*, 291–307.

Bronnenberg, B. J., S. K. Dhar, and J. P. H. Dubé (2009). Brand History, Geography, and the Persistence of Brand Shares. *Journal of Political Economy 117*(1), 87–115.

Bronnenberg, B. J. and J.-P. Dubé (2017). The Formation of Consumer Brand Preferences. *Annual Review of Economics*, 47.

Bronnenberg, B. J., J.-P. H. Dubé, and M. Gentzkow (2012, October). The Evolution of Brand Preferences: Evidence from Consumer Migration. *American Economic Review 102*(6), 2472–2508.

Browning, M. and J. Carro (2007). Heterogeneity and microeconometrics modelling. In *Advances in Economics and Econometrics*, Volume 3 of *Edited by Richard Blundell, Whitney Newey and Torsten Persson*, pp. 39. Cambridge University Press.

Burda, M., M. Harding, and J. Hausman (2008). A Bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics 147*(2), 232–246.

Butters, R. A., D. W. Sacks, and B. Seo (forthcoming). How Do National Firms Respond to Local Cost Shocks? *American Economic Review*.

Chetty, R., A. Looney, and K. Kroft (2009). Salience and Taxation: Theory and Evidence. *American Economic Review 99*(4), 1145–1177.

Coibion, O., Y. Gorodnichenko, and G. H. Hong (2015). The Cyclicality of Sales, Regular and Effective Prices: Business Cycle and Policy Implications. *American Economic Review 105*(3), 993–1029.

Commission, F. T. and US Department of Justice (2006). Commentary on horizontal merger guidelines. Technical report.

Cosslett, S. (1983). Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model. *Econometrica 51*(3), 765–782.

Deaton, A. (1986). Demand analysis. In *Handbook of Econometrics*, Volume 3, pp. 1767–1839. Amsterdam: Elsevier.

DellaVigna, S. and M. Gentzkow (2019). Uniform pricing in US retail chains. *Quarterly Journal of Economics 134*(4).

Dubé, J.-P., G. J. Hitsch, and P. Manchanda (2005). An empirical model of advertising dynamics. *Quantitative Marketing and Economics 3*, 107–144.

Dubois, P., R. Griffith, and A. Nevo (2014, March). Do Prices and Attributes Explain International Differences in Food Purchases? *American Economic Review 104*(3), 832–867.

Dubois, P., R. Griffith, and M. O'Connell (2018). The Effects of Banning Advertising in Junk Food Markets. *Review of Economic Studies 1*(1), 396–436.

Dubois, P., R. Griffith, and M. O'Connell (2020). How well targeted are soda taxes? *American Economic Review 110*(11), 3661–3704.

Eichenbaum, M., N. Jaimovich, and S. Rebelo (2011). Reference Prices, Costs, and Nominal Rigidities. *American Economic Review*, 37.

Eizenberg, A., S. Lach, and M. Oren-Yiftach (2021). Retail Prices in a City. *American Economic Journal. Economic Policy 13*(2), 175–206.

Erdem, T., M. Keane, and B. Sun (2008). The impact of advertising on consumer price sensitivity in experience goods markets. *Quantitative Marketing and Economics 6*(2), 139–176.

European Commission Directorate General for Competition (2015). *A Review of Merger Decisions in the EU: What Can We Learn from Ex Post Evaluations?* LU: Publications Office.

Fox, J. T., K. Kim, S. P. Ryan, and P. Bajari (2011). A Simple Estimator for the Distribution of Random Coefficients. *Quantitative Economics 2*(3), 381–418.

Gelman, M., S. Kariv, M. D. Shapiro, D. Silverman, and S. Tadelis (2014). Harnessing naturally occurring data to measure the response of spending to income. *Science 345*(6193), 212–215.

Goldberg, P. and R. Hellerstein (2013, January). A Structural Approach to Identifying the Sources of Local Currency Price Stability. *Review of Economic Studies 80*(1), 175–210.

Gorman, W. M. (1980). A Possible Procedure for Analysing Quality Differentials in the Egg Market. *Review of Economic Studies 47*(5), 843–856.

Greene, W. (2009). Discrete choice modeling. In *Palgrave Handbook of Econometrics*, pp. 473–556. London: Palgrave Macmillen.

Griffith, R., M. Krol, and K. Smith (2018). Why do retailers advertise store brands differently across product categories? *Journal of Industrial Economics LXVI*(3), 519–569.

Griffith, R., E. Leibtag, A. Leicester, and A. Nevo (2009). Consumer Shopping Behavior: How Much Do Consumers Save? *Journal of Economic Perspectives 23*(2), 99–120.

Griffith, R., M. O'Connell, and K. Smith (2016). Shopping Around: How Households Adjusted Food Spending Over the Great Recession. *Economica 83*(330), 247–280.

Griffith, R., M. O'Connell, and K. Smith (2019). Tax design in the alcohol market. *Journal of Public Economics 172*, 20–35.

Griffith, R., S. von Hinke, and S. Smith (2018). Getting a healthy start: The effectiveness of targeted benefits for improving dietary choices. *Journal of Health Economics 58*, 176–187.

Guadagni, P. M. and J. Little (1983). A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science 2*(3), 203–238.

Handbury, J. (2021). Are Poor Cities Cheap for Everyone? Non-Homotheticity and the Cost of Living Across U.S. Cities. *Econometrica forthcoming*, 94.

Hellerstein, R. (2008). Who bears the cost of a change in the exchange rate? Pass-through accounting for the case of beer. *Journal of International Economics 76*(1), 14–32.

Hendel, I. and A. Nevo (2006a). Measuring the implications of sales and consumer inventory behavior. *Econometrica 74*(6), 1637–1673.

Hendel, I. and A. Nevo (2006b). Sales and Consumer Inventory. *RAND Journal of Economics 37*(3), 543–561.

Hendel, I. and A. Nevo (2013). Intertemporal Price Discrimination in Storable Goods Markets. *American Economic Review 103*(7), 2722–2751.

Holmes, T. J. (2011). The Diffusion of Wal-Mart and Economies of Density. *Econometrica 79*(1), 253–302.

Hottman, C., S. Redding, and D. Weinstein (2016). Quantifying the Sources of Firm Heterogeneity. *Quarterly Journal of Economics 131*(3), 1291–1364.

Hoynes, H., D. W. Schanzenbach, and D. Almond (2016). Long-Run Impacts of Childhood Access to the Safety Net. *American Economic Review 106*(4), 903–934.

Ichimura, H. and S. Thompson (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics 86*(2), 269–295.

International Labour Organization (2004). *Consumer Price Index Manual: Theory and Practice*. Geneva: International Labour Organization.

Ivancic, L., K. J. Fox, and E. W. Diewert (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics 161*(1), 24–35.

Jaravel, X. (2019). The Unequal Gains from Product Innovations: Evidence from the U.S. Retail Sector*. *The Quarterly Journal of Economics 134*(2), 715–783.

Jaravel, X. (2021). Inflation Inequality: Measurement, Causes, and Policy Implications. *Annual Review of Economics 13*.

Jaravel, X. and M. O'Connell (2020). Real-time price indices: Inflation spike and falling product variety during the Great Lockdown. *Journal of Public Economics 191*, 104270.

Kanetkar, V., C. B. Weinberg, and D. L. Weiss (1992). Price Sensitivity and Television Advertising Exposures: Some Empirical Findings. *Marketing Science 11*(4), 359–371.

Kaplan, G., G. Menzio, L. Rudanko, and N. Trachter (2019). Relative Price Dispersion: Evidence and Theory. *American Economic Journal: Microeconomics 11*(3), 68–124.

Kaplan, G. and S. Schulhofer-Wohl (2017). Inflation at the Household Level. *Journal of Monetary Economics*.

Lancaster, K. (1966). A New Approach to Consumer Theory. *Journal of Political Economy 74*(2).

Lewbel, A. (2000). Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables. *Journal of Econometrics 97*, 145–177.

Lewbel, A. and L. Nesheim (2019). Sparse demand systems: Corners and complements. *CEMMAP Working Paper CWP45/19*.

Manski, C. (1987). Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data. *Econometrica 55*(2), 357–362.

Manski, C. F. (1975, August). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*(3), 205–228.

Matzkin (1993). Nonparametric identification and estimation of polychotomous choice models. *Econometrica 58*(1-2), 137–168.

Matzkin, R. (1992). "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models. *Econometrica 60*(2), 239–270.

McFadden, D. (1974). *Conditional Logit Analysis of Qualitative Choice Behavior*. New York: in P. Zarembka, eds. Frontiers of Econometrics, Academic Press.

McFadden, D. (1978). Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments. In *Behavioural Travel Modelling*, D. Hensher and P. Stopher (Eds.). London: Croom Helm London.

McFadden, D. (1980). Econometric Models for Probabilistic Choice Among Products. *The Journal of Business 53*(3), S13–S29.

McFadden, D. and K. Train (2000). Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics 15*, 447–470.

McFadden, D. L. (1984). Econometric analysis of qualitative response models. In *Handbook of Econometrics*, Volume 2, pp. 1395–1457. Elsevier.

Melitz, M. J. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. *Econometrica 71*(6), 1695–1725.

Miller, N. H. and M. C. Weinberg (2017, November). Understanding the Price Effects of the MillerCoors Joint Venture. *Econometrica 85*(6), 1763–1791.

Nakamura, E. and J. Steinsson (2008). Five Facts About Prices: A Reevaluation of Menu Cost Models. *Quarterly Journal of Economics 123*(4), 1415–1464.

Nakamura, E. and D. Zerom (2010). Accounting for Incomplete Pass-Through. *Review of Economic Studies 77*(3), 1192–1230.

Nevo, A. (2000). Mergers with differentiated products: The case of the ready-to-eat cereal industry. *The RAND Journal of Economics 31*(3), 395–421.

Nevo, A. (2001). Measuring Market Power in the Ready-to-Eat Cereal Industry. *Econometrica 69*(2), 307–342.

Nevo, A. (2011). Empirical Models of Consumer Behavior. *Annual Review of Economics 3*, 51–75.

Nevo, A. and A. Wong (2019). The Elasticity of Substitution Between Time and Market Goods: Evidence from the Great Recession. *International Economic Review 60*(1), 25–51.

O'Connell, M., K. Smith, and R. Stroud (2021). The dietary impact of the COVID-19 pandemic. *IFS Working Paper 21/18*.

Pesendorfer, M. (2002). Retail Sales: A Study of Pricing Behavior in Supermarkets. *The Journal of Business 75*(1), 33–66.

Redding, S. J. and D. E. Weinstein (2020). Measuring Aggregate Price Indices with Taste Shocks: Theory and Evidence for CES Preferences. *The Quarterly Journal of Economics 135*(1), 503–560.

Rivers, D. and Q. Vuong (2002). Model selection tests for nonlinear dynamic models. *The Econometrics Journal 5*(1), 1–39.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy 82*(1), 34–55.

Ruiz, F. J. R., S. Athey, and D. M. Blei (2019). SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements. *arXiv:1711.03560 [cs, econ, stat]*. Comment: Published at Annals of Applied Statistics. 27 pages, 4 figures.

Shapiro, B. T., G. J. Hitsch, and A. E. Tuchman (2021). TV Advertising Effectiveness and Profitability: Generalizable Results from 288 Brands. *SSRN Electronic Journal*, 46.

Sudhir, K. (2001). Structural Analysis of Manufacturer Pricing in the Presence of a Strategic Retailer. *Marketing Science 20*(3), 244–264.

Thomassen, Ø., H. Smith, S. Seiler, and P. Schiraldi (2017). Multi-Category Competition and Market Power: A Model of Supermarket Pricing. *American Economic Review 107*(8), 2308–2351.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Villas-Boas, S. B. (2007). Vertical Relationships between Manufacturers and Retailers: Inference with Limited Data. *The Review of Economic Studies 74*(2), 625–652.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica 57*(2), 307.

Wang, E. X.-R. and G. Vistnes (2013). Economic-Tools for-Evaluating-Competitive-Harm-in-Horizontal-Mergers.pdf. Note, Charles River Associates.

Wang, E. Y. (2015). The impact of soda taxes on consumer welfare: Implications of storability and taste heterogeneity. *The RAND Journal of Economics 46*(2), 409–441.