IARIW 2021

Monday 23 – Friday 27 August

Statistisk sentralbyrå
Statistics Norway

Quality Adjustment at Scale: Hedonic vs. Exact Demand-Based Price Indices

Gabriel Ehrlich

(University of Michigan)

John Haltiwanger

(University of Maryland and NBER)

Ron Jarmin

(U.S. Census Bureau)

David Johnson

(University of Michigan)

Ed Olivares

(University of Maryland)

Luke Pardue

(University of Maryland)

Matthew D. Shapiro

(University of Michigan and NBER)

Laura Yi Zhao

(University of Maryland and Bank of Canada)

Paper prepared for the 36th IARIW Virtual General Conference
August 23-27, 2021
IARIW Plenary Session: New Developments in Economic Measurement
Time: Monday, August 23, 2021 [14:00-18:30 CEST]

Quality Adjustment at Scale: Hedonic vs. Exact Demand-Based Price Indices*

Gabriel Ehrlich John Haltiwanger Ron Jarmin David Johnson

Ed Olivares Luke Pardue Matthew D. Shapiro Laura Yi Zhao

August 8, 2021

Preliminary: Comments Welcome

Abstract

We explore alternative methods for adjusting price indices for quality change at scale. By the latter we mean methods that can be used with large scale item-level transactions data that has been digitized with price, quantity and item-attribute data. We consider hedonic methods that take into account the changing valuation of both observable and unobservable characteristics in the presence of product turnover. We also consider demand based approaches that take into account changing product quality from product turnover and changing product appeal of continuing products. Using these methods, we find evidence of substantial quality-adjustment in prices for a wide range of goods including high-tech consumer products and food products.

*Ehrlich: University of Michigan; Haltiwanger: University of Maryland and NBER; Jarmin: U.S. Census Bureau; Johnson: University of Michigan; Olivares: University of Maryland; Pardue: University of Maryland; Shapiro: University of Michigan and NBER; Zhao: University of Maryland and Bank of Canada. Laura Zhao worked on this project when she was a doctoral student at the University of Maryland. We acknowledge financial support of the Alfred P. Sloan Foundation and the additional support of the Michigan Institute for Data Science, the Michigan Institute for Teaching and Research in Economics and the U.S. Census Bureau. The results here are in part based on researchers own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein. We also use the NPD data housed at the U.S. Census Bureau. All results using the NPD data have been reviewed to ensure that no confidential information has been disclosed (CBDRB-FY19-122, CBDRB-FY21-074.). Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the view of the U.S. Census Bureau.

1 Introduction

Retail businesses create item-level data on the prices and quantities of the goods that they sell. Such data form the basis for re-engineering key economic indicators by building consistent aggregates of value, volume, and price directly from item-level transactions. Aggregation of transactions data could supplant traditional surveys and enumerations conducted by statistical agencies (see Ehrlich et al., 2021). While ambitious, such re-engineering has many potential advantages. One is to address the issue of rapid product turnover that is associated with quality improvements. We find quarterly product-level entry and exit rates that range between 5 and 15 percent across product groups. Current statistical agency procedures for measuring prices inadequately address such turnover. This paper considers scalable procedures using item-level transactions data that can be used to measure quality change, and therefore account for entering and exiting goods as well as changing consumer valuations of product attributes.

Use of high-frequency, item-level sales data to produce accurate inflation measures also requires incorporation of advances in index number and economic theory. We consider two complementary approaches: hedonics and demand-based models. Both approaches suggest that quality improvement is widespread across a large range of consumer goods, including in categories in which technological progress is not immediately visible.

Our preferred hedonic approach builds on the insights of Erickson and Pakes (2011, hereafter "EP"), who develop a novel method of calculating hedonic price indices that accounts for product quality changes that are unobserved to the econometrician. Higher-frequency, item-level transactions data with prices, quantities, and attributes greatly facilitates the implementation of the EP methodology. These data permit implementing the hedonics approach with superlative price indices (such as the Fisher or Tornqvist) in real time using internally consistent expenditure weights. We compare and contrast the EP methodology with more commonly used alternative hedonic methods such as the time dummy method.

Our demand-based approach builds on the exact price indices developed from theoretical models of consumer demand: the Sato-Vartia price index (Sato, 1976; Vartia, 1976); the Feenstra (1994) adjustment to the Sato-Vartia index, which adjusts for quality change from product entry and exit (denoted the Feenstra index hereafter); and the CES Unified Price Index (CUPI) developed in Redding and Weinstein (2020). The demand-based approaches have the attractive feature that they yield exact price indices under certain sets of assumptions. Moreover, in principle these methods impose sufficient structure that they do not require attribute data beyond a basic product taxonomy to implement. Empirically implementing these methods at scale introduces a number of challenges, however, including classifying the goods in a manner so that the CES assumptions are valid. We have found that, in practice, the attribute data proves helpful in addressing these challenges.

A common feature of the frontier research methods using both hedonic and demand-based approaches is that they can account for changing consumer valuations of products or product characteristics. In principle, the CUPI of Redding and Weinstein (2020) captures both quality change due to product turnover and time-varying product appeal over the course of products' time in the marketplace, without directly using detailed product attributes. Likewise, the EP approach estimates changing consumer valuations of various product attributes over time in addition to adjusting price indices for product entry and exit.

We implement the hedonic and demand-based approaches at scale using item-level transactions data from two major sources. The first platform we use is from NPD, which covers a wide range of general merchandise goods from bricks and mortar and online retailers. In this paper, we construct data for a select number of product groups: memory cards, headphones, coffee makers, boys' jeans, and work and occupational footwear. The NPD data include rich product attributes, which facilitate the implementation of the EP methodology. The second platform we use is the Nielsen Marketing data provided by the Kilts Center for Marketing at the University of Chicago Booth School of Business, which covers a wide range of food products from grocery stores, discount stores, pharmacies and liquor stores.[1] Two

[1]We focus in this paper on product groups classified as "food" in the Nielsen data to sidestep potential concerns about the representativeness of the nonfood product groups.

challenges in the Nielsen data are that, first, it contains sparse data on product attributes, and second, it contains far more product groups (over 100) than the select group we used in the NPD data (five). We overcome those challenges by adapting the EP methodology to incorporate machine learning (ML) techniques that can relate product prices to seemingly limited characteristic data.

Consistent with the literature using scanner data, we find enormous product turnover at a quarterly frequency, along with rich post-entry product life-cycle dynamics. Products' market shares peak several quarters after entry, while prices decline monotonically after entry on average. We also find evidence of substantial quality adjustment in price indices using either hedonic or demand-based approaches across the full range of product groups we consider. The magnitude of the quality-adjustment is greater for high-tech goods such as memory cards and headphones, but we find that quality adjustment is pervasive for food product groups as well. While the latter might be surprising, our findings are consistent with the changing and increasing variety of food products available over time.

We find the EP method incorporating time varying unobservables systematically yields greater quality adjustment than the other hedonic methods we consider, including the time dummy method. Among the demand-based methods, we find that the Feenstra (1994) index, which adjusts the Sato-Vartia for product turnover, systematically yields lower price inflation than the Sato-Vartia. This result suggests that product turnover is associated with quality improvement.

The most general demand-based index we consider is the CUPI, which generalizes the Feenstra index to allow for changing product appeal over product life cycles. We find the CUPI implies substantial quality adjustment beyond what the Feenstra index implies. A challenge for implementing the CUPI is that two of its three components are unweighted geometric means. These terms are sensitive to the inclusion of goods with very small quantities or market shares, which is one reason that unweighted indices are generally discouraged in the index number literature. Redding and Weinstein (2020) employ a reallocation procedure, by which they move a subset of goods out of the CUPI's unweighted geometric mean terms and into the Feenstra (1994) adjustment term using what we term a *common goods rule* based on the durations of products' time in the marketplace.[2] Applying a common goods rule brings the CUPI's measurement of price changes closer in line with other indices. Our results suggest more research is needed to provide guidance about how to define a common goods rule.

We proceed as follows. Section 2 describes the underlying data. The conceptual framework we use for hedonic and demand-based indices is presented in section 3. Section 4 presents our main results. Section 4.3 discusses the advantages and drawbacks of the alternative methods of conducting quality adjustment at scale that we have considered, in light of our results. Section 5 provides concluding remarks.

2 Data

This section provides an overview of the two data sets that we use to compute price indices. The first comes from the NPD Group and the second comes from Nielsen.

2.1 NPD Data

We use proprietary data from the NPD Group provided to the U.S. Census Bureau consisting of monthly sales and quantity data at the product-store level from 2014 through 2018.[3] The NPD group tracks more than 65,000 retail stores, including online retailers. The retail stores

[2]Redding and Weinstein (2020) refer to these rules "as alternative definitions of common varieties." In his insightful discussion of Ehrlich et al. (2021), Robert Feenstra motivated a common goods rule as a necessary recognition that it takes time for goods to enter and exit markets, a concept he denoted *seasoning*. As we discuss below, one limitation of the Redding and Weinstein (2020) specification of the common goods rule is that it requires forward-looking information and therefore cannot be implemented in real time. We find we can mimic their approach using only backward-looking and real-time information.

[3]Month definitions follow the National Retail Federation (NRF) calendar. The NRF calendar is a guide for retailers that ensures sales comparability between years by dividing a year into months based on a 4 weeks-5 weeks-4 weeks format. The layout of the calendar lines up holidays and ensures the same number of Saturdays and Sundays in comparable months across years. The NRF calendar thus ensures the comparability of the aggregated sales over time.

cover a wide range of general merchandise products. The NPD data analyzed here consist of five broad product groups, within which we conduct our analysis separately: memory cards, coffee makers, headphones, boys' jeans, and work/occupational footwear (hereafter simply "occupational footwear"). The NPD data have unique item-level identifiers that are consistent cross-sectionally and over time. We aggregate the item-by-store level observations to the national product-quarter level and calculate total quantity sold and average price for each product-quarter. The item-level data cover tens of thousands of product-quarter level observations.

An attractive feature of the NPD data is that they contain detailed and organized information on the characteristics of each product. Beyond basic information such as product category and brand, these characteristics include details on different types of products within the broader categories (e.g., on-ear vs. in-ear headphones; coffee vs. espresso machines) and the features or attributes of different products (e.g. built-in grinders or auto-on/off settings for coffee makers). In some cases, the attributes include continuous variables, which facilitate estimation of hedonic models. We use the detailed product characteristics in the estimation of hedonic price indices and to group products into subcategories in our estimation of nested CES models.

Table 1 displays average item-level product turnover rates for each product group. Each of the five groups exhibits a high degree of product turnover–although these rates are lower on a sales-weighted basis, suggesting that turnover is more common among goods with smaller market shares.

Figure 1 presents life-cycle dynamics of product market shares and prices within these product groups. The illustrated statistics are mean log differences from the product-specific initial values upon entry. Prices decline steadily after entry, while market shares exhibit a hump-shaped pattern post-entry.

Taken together, these findings highlight two important features of the data. First, there is considerable item-level product turnover that is a potentially important source of changing product quality. Second, post-entry dynamics suggest that it may be important for methods for quality adjustment to account for time-varying product appeal. Both the hedonic and demand-based approaches we consider permit such variation.

2.2 Nielsen Data

We use the Nielsen Retail Scanner data (also referred to as RMS) from the Kilts Center for Marketing at the University of Chicago Booth School of Business for the period 2006 to 2015. The data consists of over 2.6 million products identified by the finest level of aggregation– 12-digit universal product codes (UPCs) that uniquely identify specific goods.[4] The RMS data are collected from over 40,000 individual stores from approximately 90 retail chains in over 370 metropolitan statistical areas (MSAs) in the United States. Total sales in Nielsen RMS are worth over $200 billion per year and represent 53% of all sales in grocery stores, 55% in drug stores, 32% in mass merchandisers, and 2% in convenience stores.

Nielsen organizes item-level goods into 10 departments, over 100 product groups, and over 1,000 product modules. A typical department is, for example, dry grocery, which consists of 41 product groups, such as baby food, coffee, and carbonated beverage. Within the carbonated beverage product group, there are product modules such as soft drinks and fountain beverage. The product groups are classified into food and nonfood sectors based on a concordance provided by the Bureau of Labor Statistics (BLS).

The RMS consists of more than 100 billion unique observations at the week-store-UPC level. We first aggregate the weekly data to the monthly frequency according to the NRF calendar and then aggregate the monthly data to quarterly. Before aggregating to the data to quarterly frequency, we drop outliers, defined as observations with prices above 3 times or below one-third the median for each UPC in a given month. We also drop product-month observations with quantities sold that are more than 24 times that product's median quantity sold per month. One feature of barcoded products is that goods of different sizes

[4]The Nielsen data contain both UPC codes and "UPC version codes." The unique product identifier used in the analysis is the combination of the UPC code and the UPC version code.

and packaging have different barcodes, even if the product contained in the packaging is the same. To ensure comparability between prices, we follow Hottman et al. (2016) and normalize UPC prices to the same units (e.g., ounces), utilizing the size and packaging information provided by Nielsen. Consistent with the literature, we winsorize monthly price changes at the top and bottom 1% of each product group.

We focus on the product groups in food in the main text. Using the Economic Census data from 2012, we have calculated that the type of retailers that the Nielsen scanner data tracks have very high coverage of food items (about 90%). Moreover, using a back-of-the-envelope calculation based on Nielsen's coverage of different types of retailers, we estimate that Nielsen scanner data accounts for about 41% of total food sales in the U.S. In contrast, the data's coverage is much weaker for several nonfood categories. The type of stores Nielsen tracks accounts for about 53% of small appliance sales. However, Nielsen's coverage of general merchandise stores is only 32%. These figures imply (using our back-of-the-envelope calculation) that Nielsen scanner data accounts for only about 19% of total small appliance sales in the U.S. Coverage in other categories is even lower. We calculate that the Nielsen scanner data accounts for only about 5% of total sales of hardware and tools. In unreported results, we have compared patterns of total expenditures for harmonized categories from Nielsen and Personal Consumption Expenditures data (PCE). We find evidence suggesting that the Nielsen Retail Scanner data's coverage of nonfood items deteriorated during our study period, potentially driven by the ongoing structural shifts in Retail Trade, especially towards e-commerce. In contrast, we find a much closer correspondence between total expenditure trend patterns from Nielsen and the PCE for harmonized categories of food items.

3 Conceptual Framework

The goal of any price index is to approximate the change in the cost of living between two time periods—that is, to calculate how much more or less expensive it is to achieve the same standard of living as in some base period given current prices. An important challenge in constructing price indices from item-level data is the substantial pace of product turnover documented in the previous section. Traditional "matched-model" price indices do not capture quality change from such product turnover. In contrast, the hedonic and demand-based indices we construct from the item-level data do incorporate quality change from product turnover and changing valuations of products or product characteristics.

3.1 Traditional Price Indices

We consider first the familiar formulas for traditional price indices, which aggregate the price changes of individual goods across periods to weighted or unweighted averages. These indices–which include the Jevons, Laspyeres, Paasche, Fisher, and Tornqvist–remain the basis of most statistical agencies' price index calculations. Every traditional price index has the following generic form:

$$\text{Arithmetic:} \quad \Psi_t^A = \sum_k w_{kt} \frac{p_{kt}}{p_{kt\text{-}1}}$$

$$\text{Geometric:} \quad \ln \Psi_t^G = \sum_k w_{kt} \ln \frac{p_{kt}}{p_{kt\text{-}1}},$$

where w_k is a weight assigned to product k–typically based on the product's market share– and the ratio of prices to be aggregated often is called a price relative. The use of different weights determines the index: the Jevons is the unweighted geometric mean of price relatives, the Laspeyres (arithmetic or geometric) uses lagged market shares as weights ($w_{kt} = s_{kt\text{-}1}$), and the Paasche uses current period market shares as weights ($w_{kt} = s_{kt}$). The Fisher and Tornqvist indices are defined as the geometric means of the arithmetic and geometric Laspeyres and Paasche indices, respectively. The Tornqvist index thus uses the average market shares in the current and base periods as weights.[5] An important advantage of the transactions-level data is that they allow the Tornqvist weights to be updated every period in an internally consistent manner.

Traditional price indices have a theory-free interpretation as weighted-average changes in product prices. While this statistical interpretation is valuable on its own, there is also an economic interpretation of these indices dating back to the seminal work of A.A. Konus (Konüs, 1939; Schultz, 1939). The Laspeyres and Paasche indices provide upper and lower bounds, respectively, on the exact change in the cost of living between two periods in the absence of product turnover.[6] Other so-called superlative indices, including Tornqvist and Fisher, have more desirable theoretical properties: they are the change in the unit expenditure function (i.e., the exact price index) that is the second-order approximation of that for a wide class of utility functions in the absence of product turnover and taste shocks (Diewert, 1978). We will generally use a superlative index, particularly the Tornqvist, when comparing traditional indices with hedonic or demand-based indices.

All of these price indices require both the *price* and *sales or expenditure share* of each good in each time period to calculate weighted price changes. While statistical agencies often only have data on sales and expenditure shares from disparate sources at higher levels of aggregation and lower frequency, high-frequency scanner data connect the prices and quantities sold of each product, allowing for the construction of weighted price indices at finer levels using internally consistent price and quantity data.[7]

These traditional price indices are all "matched-model" price indices: they calculate price changes across the goods that were sold both in the base and in the current period. The traditional indices therefore do not account directly for goods that enter or exit across

[5] A related but distinct index is the Sato-Vartia index, which uses the logarithmic mean of market shares in the current and base periods as weights. This index is based on assuming a CES demand structure. We discuss the the Sato-Vartia index in Section 3.3 below.

[6] In the case of strictly normal goods, the arithmetic Paasche is a lower bound of the equivalent variation, and the arithmetic Laspeyres is an upper bound to the compensated variation, so we have that Paasche $\leq$ EV $\leq$ CV $\leq$ Laspeyres. Paasche $<$ Laspeyres typically holds in the data, and will be the case when substitution is, on net, away from goods that have the highest change in price and towards those with the lowest.

[7] The limitations of the current system are discussed further in Ehrlich et al. (2021).

periods, which may be an important source of changing product quality.[8]

3.2 Hedonic Price Indices

In this section, we describe our use of hedonic methods to adjust price indices for quality changes, especially in the context of product turnover. The log-level hedonic price model common in the literature takes the form:

$$\ln p_{kt} = h_t(Z_k) + \eta_{kt}, \tag{1}$$

where Z_k is a vector of observable characteristics for good k. The function $h_t()$ is typically linear in parameters, and the hedonic equation is estimated with with ordinary or weighted least squares regression. An important feature of equation (1) is that the hedonic function varies over time, i.e., the function $h_t()$ is estimated separately period-by-period. The time-varying estimation allows the hedonic function to capture changing consumer valuations, markups, or other aspects of market structure (Pakes, 2003). Hedonic imputation price indices provide one way to correct for product turnover, by using observable characteristics to impute the "missing" prices for entering and exiting products.

A core limitation of the log-level estimation approach outlined in equation (1) is that there are likely to be product characteristics that the econometrician cannot observe. Erickson and Pakes (2011) introduce hedonic methods that can account for such unobserved characteristics. One simple, though effective, method is to estimate hedonic models of price changes rather than price levels. A log-difference hedonic model estimates the change in hedonic price coefficients directly, and allows for an item-level fixed unobservable characteristic. The most general form of the Erickson and Pakes (2011) approach also accounts for time-varying unobservable characteristics, so we will call this approach the *TV approach*

[8] Currently, the BLS uses hedonic quality adjustment for only about 7 percent of the products in the CPI, which is a Laspeyres-style index with weights that are typically two years old from disparate sources. The BLS uses a matched-model approach and other methods to adjust for quality change for those goods to which it does not apply hedonics, but these approaches face well-known limitations (e.g. Triplett, 2004).

for short. Implementing the TV approach requires two steps. First, we estimate the log-level hedonic specification in equation (1) for period $t-1$. Second, estimate a log-difference hedonic model, including the lagged residuals from the first stage. The second estimating equation is then:[9]

$$\Delta \ln p_{kt} = Z_k' \beta_t + \kappa \hat{\eta}_{kt\text{-}1} + v_{kt}. \tag{2}$$

Including the initial residuals from equation (1) in equation (2) allows the model to capture the influence of time-varying valuations of unobservable product characteristics to the extent that the initial residuals are correlated with price changes. The intermediate case of fixed unobservable characteristics simply uses first differences, which controls for unobserved fixed effects:

$$\Delta \ln p_{kt} = Z_k' \beta_t + v_{kt}. \tag{3}$$

In our analysis, we consider log-level, log first-difference, and TV approaches.

We also consider the related, but distinct, *time dummy* method that has been actively used in the research literature and by BLS. We follow the recent literature (e.g., Byrne et al., 2019) using adjacent-period, weighted least squares estimation with Tornqvist market-share weights. Specifically, we estimate hedonic regression equations pooling observations from the adjacent periods $t-1$ and t. Letting T denote the total number of periods in the data, we thus estimate $T-1$ separate pooled two-period regressions of the form:

$$\ln p_{k\tau} = \alpha_{t-1,t} + \delta_t + Z_k' \gamma_{t-1,t} + \epsilon_{k\tau}, \quad \tau = \{t-1, t\}, \tag{4}$$

where $\alpha_{t-1,t}$ is the constant, Z_k is the vector of characteristics for good k, $\gamma_{t-1,t}$ is a vector of estimated hedonic coefficients held fixed across periods $t-1$ and t, and δ_t is a fixed effect

[9]It can be shown that this characterization is equivalent to the time varying unobservables specification in Erickson and Pakes (2011). In the latter, they describe a closely related multi-step procedure. First, estimate the log levels hedonics and recover the residual. Second, estimate the first difference of the log price relative on characteristics. Third, estimate the change in the residuals from the the log levels on the characteristics. Using the sum of the predictions from the latter two steps (as described in Erickson and Pakes (2011)) is equivalent to using the predictions from equation 2.

for period t.[10]

Exponents of the resulting coefficients δ_t can be interpreted as the quality-adjusted change in the price level between period $t-1$ and t. Intuitively, the period-t fixed effect δ_t reflects the difference in average price of a "generic" good between $t-1$ and t because the contributions of all of the product characteristics have been partialled out. Importantly, the hedonic regressions include both entering and exiting goods, so goods contribute to the index in any period in which they have sales. A limitation of the time dummy method relative to the TV approach is that the former does not account for unobservable product characteristics. Further limitations of the time dummy approach are discussed in Pakes (2003) and Diewert et al. (2008). Our implementations of the TV approach and the time dummy method with the NPD data uses standard econometric methods to estimate the hedonic function $h_t()$. This approach is feasible with NPD data given the enormous value-added provided by the NPD group in terms of item-level attributes.

The product descriptions in the Nielsen data provided by the Kilts Center for Marketing at the University of Chicago are generally not coded to be human-intelligible. For instance, two product descriptions for soft drinks are `ZR DT LN/LM CF NBP CT` and `NATURAL R CL NB 12P`, while a product description for toilet paper is `DR W 1P 308S TT 6PK`. A human analyst could decipher portions of these descriptions : `DT` means "diet," `12P` means twelve pack, `1P` means one ply, `308S` means 308 sheets, etc. It would not be feasible for human analysts to encode such data at scale, however, and simple dictionaries would be fooled (e.g., the `P`-suffix means "pack" for soft drinks and "ply" for toilet paper).

An additional challenge in the Nielsen data is its sheer scale. The Retail Scanner data contains more than 100 product groups and over 1,000 product modules. It would be difficult for human analysts to specify sensible hedonic regression equations for so many product

[10]We specify the hedonic regression equation (4) using the same vector of characteristics Z_k in each pair of adjacent periods. Occasionally, new features are introduced to the data. In pairs of adjacent periods entirely prior to the introduction of a new characteristic, it will be omitted from the regression because of collinearity with the intercept term. In pairs of adjacent periods in which the new feature is absent during period $t-1$ and present during period t, the feature will be included in the estimated regression. Symmetric arguments apply for characteristics that exit.

groups. It would be even more difficult to update those regression equations over time as product mixes and characteristics change.

To address these challenges, we have implemented deep neural networks to predict product prices and price changes from the product descriptions in the Nielsen Kilts Center data. Our approach parallels the TV approach of Erickson and Pakes (2011), in that it first predicts price levels and then uses the prediction error in a second-stage neural net to capture time-varying unobservable effects. We provide more details of our approach in the appendix and in a companion paper (Cafarella et al. (2021, in progress)) that focuses on the machine learning methodology. Bajari et al. (2021) use an advanced machine learning approach that includes encoding image data as inputs into price predictions.[11]

In order to make our alternative hedonic approaches as comparable as possible, we use weighted estimation methods in all cases. We follow the recent time dummy literature by using Tornqvist expenditure weights, so that the time dummy method yields a quality-adjusted Tornqvist price index. We apply quantity-share weights for the imputation approaches using both econometric and ML methods for estimation of the hedonic pricing function. Bajari et al. (2021) also use quantity-share weights in their implementation of hedonic price indices with ML methods; our approach in both the Nielsen and NPD data thus facilitates comparisons to their results. Using quantity-share weights focuses the hedonic estimation procedure on accurately mapping the relationship between prices and characteristics for the market basket of goods purchased by the consumer.[12] The quantity-weighting is used only in the estimation of the hedonic relationship. The Tornqvist indices that draw upon these hedonics use expenditure-share weights consistent with the literature.

We focus on full-imputation versions of the hedonic imputation indices, which use pre-

[11]Bajari et al. (2021) provides novel methodology for encoding images via machine learning but do not incorporate the TV approach to constructing hedonic price indices.

[12]See for example the discussion in De Haan (2008). We provide further discussion of the motivation for using weighted results in the appendix. We also show that the first-difference methods we focus on are largely robust to using weighted or unweighted specifications. In unreported results, we have explored using expenditure weights and have also found similar results.

dicted prices for all observations, including for common goods.[13] Pakes (2003) shows that this form of hedonic imputation index provides a bound to the exact change in the cost of living under a weaker set of assumptions than those commonly used in the literature. The key assumption is that consumers have preferences over the characteristics embodied in goods, rather than over the goods themselves. Indeed, full-imputation indices can be interpreted as characteristic price indices (Hill and Melser, 2006; De Haan, 2008).[14] Using full-imputation indices also facilitates comparison with the time dummy method, as highlighted by De Haan (2008) and Diewert et al. (2008).[15] In addition, Erickson and Pakes (2011) observe that single- and double-imputation indices are subject to a form of selection bias, because they treat the hedonic estimation residuals for continuing, entering, and exiting goods in an asymmetric manner.[16] Full-imputation indices have also been used in Benkard and Bajari (2005), Diewert et al. (2008) and Bajari et al. (2021). Our implementation of hedonic indices builds on and integrates the insights of this literature.

3.3 Demand-Based Price Indices

In this section we describe our use of exact cost-of-living indices for Constant Elasticity of Substitution (CES) demand systems. The CES utility function yields a tractable demand system with several computable price indexes that correspond exactly to the theoretical unit cost function faced by a representative consumer in the presence of product turnover and time-varying product appeal. We explore the Sato-Vartia index (Sato, 1976; Vartia, 1976),

[13] That is we use the predicted price relatives for continuing goods and for the *missing* prices for entering and exiting goods. For the implementation of the TV approach, we assume the lagged residual for an entering good in the period prior to entry is zero.

[14] For example, Hill and Melser (2006) show that the full-imputation hedonic Tornqvist index estimated with a semi-log model has a dual representation as the Fisher index in characteristics space.

[15] De Haan (2008) argues that, in the absence of unobserved characteristics, these indices are "strikingly similar." Diewert et al. (2008) note the similarities and also derive the conditions under which they are identical. They note the full imputation approach is more flexible and in practice yields different results than the time dummy method. Neither of these papers highlights the importance of unobserved characteristics, as in Erickson and Pakes (2011). Incorporating the TV approach developed by Erickson and Pakes (2011) to address unobserved characteristics in the full-imputation indices produces additional advantages over the time dummy method. For these reasons, we favor the full-imputation TV approach of Erickson and Pakes (2011) in our hedonic indices.

[16] See footnote 3 of Erickson and Pakes (2011).

the Feenstra-Adjusted Sato-Vartia index (Feenstra, 1995), which we will call the "Feenstra index," and the CES Unified Price Index of Redding and Weinstein (2020), which we will call the "CUPI." We focus on CES demand systems as this structure has been developed to provide tractable, implementable price indices that can account for quality change and product turnover.

We start with the CUPI, as it nests the other two indices as special cases. Redding and Weinstein characterize the CES unit expenditure function as:

$$P_t = \left[\sum_{k \in \Omega_t} \left(\frac{p_{kt}}{\varphi_{kt}} \right)^{1-\sigma} \right]^{\frac{1}{1-\sigma}}, \tag{5}$$

where $\sigma > 1$ is the consumer's elasticity of substitution between products and Ω_t is the set of products sold in period t. φ_{kt} is a product-level appeal parameter with time-invariant and time-variant components:

$$\ln \varphi_{kt} = \ln \varphi_k + \ln \theta_{kt}. \tag{6}$$

φ_k captures differences in average levels of expenditure across goods, i.e. some goods have persistently higher quality or product appeal than others. θ_{kt} captures shocks to product appeal, which are assumed to be i.i.d. across goods and over time. Redding and Weinstein (2020) emphasize that including time-varying product appeal is essential to make the CES system consistent with the observed micro variation in prices and quantities. They specify a normalization on the changes in the appeal shocks so that there is no change in average tastes at the product group level. This assumption, combined with their assumption, which we also maintain, that consumers have Cobb-Douglas preferences across product groups, guarantees that product-level appeal shocks do not spill across product groups.

Consumers' optimally chosen expenditure shares in this system are given as:

$$s_{kt} \equiv \frac{p_{kt} c_{kt}}{\sum_l p_{lt} c_{lt}} = \frac{(p_{kt}/\varphi_{kt})^{1-\sigma}}{\sum_{l \in \Omega_t} (p_{lt}/\varphi_{lt})^{1-\sigma}} = \frac{(p_{kt}/\varphi_{kt})^{1-\sigma}}{P_t^{1-\sigma}}, \tag{7}$$

where c_{kt} is the quantity of good k purchased in period t.

Redding and Weinstein (2020) derive the exact-price index in this setting (the CUPI) as:

$$\Psi_{t-1,t}^{CUPI} = \left(\frac{\lambda_{t,t-1}}{\lambda_{t-1,t}}\right)^{\frac{1}{\sigma-1}} \frac{\tilde{P}_t^*}{\tilde{P}_{t-1}^*} \left(\frac{\tilde{S}_t^*}{\tilde{S}_{t-1}^*}\right)^{\frac{1}{\sigma-1}}. \tag{8}$$

The first term in the CUPI is the Feenstra (1994) adjustment factor for product turnover, with elements defined as:

$$\lambda_{t,t-1} = \frac{\sum_{k\in\mathbb{C}_t} p_{kt}c_{kt}}{\sum_{k\in\Omega_t} p_{kt}c_{kt}}, \qquad \lambda_{t-1,t} = \frac{\sum_{k\in\mathbb{C}_t} p_{kt-1}c_{kt-1}}{\sum_{k\in\Omega_{t-1}} p_{kt-1}c_{kt-1}}, \tag{9}$$

where $\mathbb{C}_t$ is the set of common goods (what Redding and Weinstein (2020) denote as "common varieties").[17] Denoting the sales-weighted product entry and exit rates as $ER_{t-1,t}$ and $XR_{t-1,t}$, the log Feenstra adjustment term can be approximated as: $\ln\left(\frac{\lambda_{t,t-1}}{\lambda_{t-1,t}}\right)^{\frac{1}{\sigma-1}} \approx \frac{1}{\sigma-1}\left(XR_{t-1,t} - ER_{t-1,t}\right)$. The Feenstra term thus indicates a downward adjustment to traditional matched price indices when the sales share of entering products is higher than the sales share of exiting products; it collapses to one in the absence of product turnover.[18]

The second term in the CUPI is the traditional Jevons index defined over the set of common goods. The third term is the what Redding and Weinstein (2020) call the "consumer valuation bias," with elements defined as the unweighted geometric average expenditure shares on common varieties in periods $t-1$ and t:

$$\tilde{S}_t^* = \left(\prod_{k\in\mathbb{C}_t} s_{kt}\right)^{\frac{1}{N_{\mathbb{C}_t}}}, \qquad \tilde{S}_{t-1}^* = \left(\prod_{k\in\mathbb{C}_t} s_{kt-1}\right)^{\frac{1}{N_{\mathbb{C}_t}}}, \tag{10}$$

where we have denoted the number of products in $\mathbb{C}_t$ as $N_{\mathbb{C}_t}$.

[17] The simplest definition of this set is products that were sold in both periods $t-1$ and t, but it is possible to re-define the set so that only goods that are sold in quantities above a threshold or which are sold in the market for a suitably long duration are included in this set, as in Redding and Weinstein (2020).

[18] We use the actual Feenstra term and not the approximation in our implementation.

It is instructive to consider the log version of the CUPI, which is given by:

$$\ln \Phi_{t-1,t}^{CUPI} = \frac{1}{\sigma - 1} \ln \left(\frac{\lambda_{t,t-1}}{\lambda_{t-1,t}} \right) + \frac{1}{N_{\mathbb{C}_t}} \sum_{k \in \mathbb{C}_t} \ln \left(\frac{p_{kt}^*}{p_{kt-1}^*} \right) + \frac{1}{\sigma - 1} \frac{1}{N_{\mathbb{C}_t}} \sum_{k \in \mathbb{C}_t} \ln \left(\frac{s_{kt}^*}{s_{kt-1}^*} \right). \quad (11)$$

Equation (11) clarifies that two of the CUPI's three terms (the Jevons index and the consumer valuation bias) are unweighted geometric means. As discussed below, this property is important for the CUPI's empirical implementation.

In the absence of time-varying product appeal, the CUPI collapses to the Feenstra (1994) index:

$$\Phi_{t-1,t}^{Feenstra} = \left(\frac{\lambda_{t,t-1}}{\lambda_{t-1,t}} \right)^{\frac{1}{\sigma-1}} \left(\Phi_t^{SV} \right), \quad (12)$$

where Φ_t^{SV} is the Sato-Vartia price index defined over common varieties. With no product turnover, the Feenstra index further collapses to the Sato-Vartia index (Sato, 1976; Vartia, 1976), defined as:

$$\ln \left(\Phi_t^{SV} \right) = \sum_{k \in \mathbb{C}_t} \omega_{kt} \ln \left(\frac{p_{kt}}{p_{kt\text{-}1}} \right), \quad \omega_{kt} = \frac{s_{kt} - s_{kt\text{-}1}}{\ln(s_{kt}) - \ln(s_{kt\text{-}1})} \Bigg/ \left(\sum_{k \in \mathbb{C}_t} \frac{s_{kt} - s_{kt\text{-}1}}{\ln(s_{kt}) - \ln(s_{kt\text{-}1})} \right). \quad (13)$$

Each of these price indexes is exact under different assumptions. The Sato-Vartia price index is exact if there is no product turnover and no time variation in product appeal. The Feenstra-adjusted Sato-Vartia index is exact in the presence of product turnover but the absence of time-varying product appeal. The CUPI is exact under the more general conditions of product turnover and time variation in product appeal. Since there is a good deal of product turnover in the item-level data, these generalizations of the Sato-Vartia index are empirically relevant.

Although the CUPI is the most general of CES exact price indices, its inclusion of two unweighted geometric mean terms contrasts with the Sato-Vartia and Feenstra indices, which include only expenditure-weighted terms. The CUPI's unweighted terms are sensitive to products with very small expenditure shares. The CUPI can therefore feature large measured price changes from what would appear to be economically minor products.

Redding and Weinstein (2020) adjust their empirical implementation of the CUPI by applying what we call a "common goods rule," which defines the set of goods over which the Jevons index and consumer valuation bias adjustment terms are calculated.[19] The goods excluded from the set of common goods are reallocated to the product turnover component (Feenstra adjustment factor), which is expenditure weighted. Redding and Weinstein (2020) restrict the set of common goods in their empirical CUPI to those that are sold for a sufficiently long duration both prior to period $t-1$ and subsequent to period t. They measure annual CUPI inflation from the fourth quarter of one year to the fourth quarter of the next year. Defining those quarters as periods $t-1$ and t, they define common goods as those sold in both of those quarters as well as in the 3 quarters prior to $t-1$ and the 3 quarters subsequent to t. In addition, they require the good be sold for at least 6 years total (although not necessarily consecutively). A common goods rule of this sort can be motivated by the argument that it takes time for goods to enter and exit the market. A limitation of this particular duration-based common goods rule is that it requires forward-looking information to implement and thus is not practical in real time. We find that we can mimick their results using a more practical backward-looking rule.

4 Results

In this section, we present and discuss the traditional, hedonic, and demand-based exact price indices we have calculated in the item-level data. We focus first on our results from the NPD data, because the richness of the data permits more exploration of alternative methods.

[19] That is, the goods designated as $\mathbb{C}_t$.

4.1 NPD Results

4.1.1 Hedonics

Figure 2 presents the results for alternative hedonic price indices for the five NPD product groups and compares these indices to the traditional Tornqvist index. We focus on the Hedonic Tornqvist time dummy, the Hedonic Tornqvist using fixed unobservables, and the Hedonic Tornqvist with time varying unobservable (the TV approach).[20] The values displayed in the figure are annual percent changes in the 4th quarter of each year from chained cumulative quarterly indices. All of the price indices track each other closely, but there are systematic differences in the patterns. For all product groups, the TV approach yields the lowest rate of price inflation compared to the traditional Tornqvist, the time dummy based index, or the first-difference based index. The gap between the traditional Tornqvist and the TV approach indices varies considerably across product groups, with the largest average differences for memory cards (-2.9 percentage points annually) and headphones (-2.5) and smaller differences for coffee makers (-0.70), boys' jeans (-1.30), and occupational footwear (-0.42).

The time dummy method does not yield systematic quality-adjustment differences relative to the traditional Tornqvist index. The time dummy method suggests a notable quality adjustment for coffee makers, but for other products the difference is modest or is positive rather than negative. Our finding of limited quality adjustment the time dummy method is broadly consistent with the discussion in Erickson and Pakes (2011). As they emphasize, traditional hedonic approaches cannot account for the changing valuations of unobservable product characteristics, and in particular, how those changing valuations interact with product turnover. For example, if entering goods have desirable unobserved characteristics and correspondingly high prices, then the time dummy method may erroneously suggest a higher index value relative to the traditional Tornqvist.[21] Additional limitations of the time dummy

[20] In the appendix, we provide more details about traditional price indices and additional hedonic estimation approaches such as log levels.

[21] For headphones, the traditional Tornqvist is notably lower in 2016 compared to the Hedonic Tornqvist

approach have been highlighted by Pakes (2003) and Diewert et al. (2008).[22]

The findings in Figure 2 also don't exhibit a systematic relationship between the Tornqvist and the Hedonic Tornqvist using fixed unobservables. Given that we find a systematic lower Hedonic Tornqvist with time varying unobservables, the somewhat erratic pattern of the first difference specification with fixed unobservables suggests it important to permit time varying valuation of the unobservables in order to systematically adjust for unobservables. Figure 3 provides further evidence on the efficacy of the TV approach. The results in this figure display how the TV approach works even if key observable characteristics are left out of the hedonic estimation. Specifically, for memory cards the memory size is omitted, and for the other product groups, the large brand dummy variables are omitted. In the appendix, we present additional analyses showing that leaving out unobserved characteristics has a much larger effect on hedonic indices using a log-level estimation approach.

Our results are broadly consistent with the findings in Erickson and Pakes (2011). They present examples (e.g., televisions) in which standard log-level hedonic estimation suggests higher rates of inflation than traditional matched models. However, with the same data, they find their time unobservables methodology (both using fixed valuation of unobservables and time varying unobservables) yields systematically lower inflation than the traditional Tornqvist. They also conduct a similar test of their methodology as we show in Figure 3. They find their time varying unobservables methodology is robust to leaving out important characteristics but standard (e.g., level hedonic) specifications are not robust to this leave out exercise.

---

using the time dummy method. This is a year when the share-weighted average price per item increases substantially. This pattern is consistent with entering goods having higher prices than existing goods. The time dummy method still yields a negative price change in that year but not as negative as the standard Tornqvist. The Hedonic Tornqvist TV method yields a more negative price index than the standard Tornqvist.

[22]Pakes (2003) raises questions about the bound implied by time dummy method. Diewert et al. (2008) highlight that the time dummy method requires more restrictive assumptions.

4.1.2 CES Demand Based Price Index

We turn now to CES demand-based price indices. For the Feenstra (1994) price index and the CUPI, implementation requires estimates of the elasticities of substitution. Our baseline approach is to estimate a single elasticity for each of the NPD product groups. We employ the method used by Feenstra (1994) and Redding and Weinstein (2020) for this purpose.[23] Table 2 reports the estimated elasticities, which range from about 5.2 to 7.8, consistent with the literature. The table also reports estimates from nested specifications, which we discuss below.

Figure 4 plots the Sato-Vartia, Feenstra, and CES unified (CUPI) price indices, as well as the components of the latter two indices. The baseline CUPI is calculated without a common goods rule and without any nesting within product groups. The Feenstra index comprises the "Lamba Ratio" (Feenstra adjustment term) and the Sato-Vartia index, while the CUPI includes the identical Lambda Ratio, the "P* ratio," (or Jevons index), and the "S*" ratio (the consumer valuation bias term introduced to account for changes in consumer tastes over time). The Lambda Ratio and S* ratio components in the figure are scaled by $\frac{1}{\sigma-1}$ so that the CUPI is the sum of the three components; see equation (8). We find that the CUPI is low relative to the Feenstra index and quite low in absolute terms. In all goods but occupational footwear, the CUPI produces an estimate of 30%–40% declines in the price level annually, and it is often 10%–30% below the Feenstra Index.

The large differences between the Feenstra index and CUPI in these products arise from two sources. The first source is the difference between P* ratio (Jevons index) and the Sato-Vartia index. The Sato-Vartia is a weighted average price change among common goods, and the P* ratio is an unweighted average. In boys' jeans, for instance, the CUPI P* ratio is far below the Sato-Vartia. The difference between the weighted and unweighted price ratios for common goods suggests there are a large number of low-share goods experiencing price

[23] This method double-differences the demand and supply curves sweeping out time and product group effects. The double-differenced demand and supply shocks are assumed to be uncorrelated but heteroksedastic across products. This yields a GMM specification for estimation.

declines that are driving down the CUPI. The second source is the introduction of the S* ratio in the CUPI, intended to account for changing consumer tastes. Almost everywhere, the S* ratio contributes a large downward shift to the CUPI. It is also an unweighted geometric mean that is sensitive to low-share goods.

The CUPI's sensitivity to low-share goods led Redding and Weinstein (2020) to introduce a common goods rule (hereafter often denoted the CGR) to the index. The logic, as discussed above, is that it takes some time for goods to break into the market. A limitation of the forward-looking duration-based approach for the CGR in Redding and Weinstein (2020) is that it cannot be implemented in real time. Their approach requires information about goods' future presence in or absence from the marketplace. We implement a related but distinct methodology that can be implemented in real time using only current and backward looking information available in quarter t. For our NPD analysis, we specify a market share threshold for goods present in periods t and $t-1$ to be considered as common goods for the Jevons and the consumer valuation bias terms of the CUPI.[24] Goods below this threshold are used in the Feenstra adjustment term. For our NPD analysis, we consider alternative market share percentile thresholds. As we discuss below for our analysis of Nielsen data which is a longer panel, we consider further alternative approaches to define common goods.

Figure 5 illustrates the CUPI's sensitivity to the CGR for different market share thresholds. Specifically, we consider market share thresholds for continuing goods in t and $t-1$ of the 10th percentile, the 30th percentile and the 50th percentile. We depict the CUPI for these different CGR rules alongside the Feenstra index and the CUPI without a common goods rule. Implementing the restriction on the set of "common goods" by market share raises the CUPI by cutting off the low end of market share distribution from relative comparisons and shifting it to the entry/exit adjustment term (the Lambda ratio). In that sense, applying a stricter CGR moves the CUPI closer to the Feenstra-adjusted Sato-Vartia index, which combines a traditional matched model index with an adjustment for entry and exit.

[24]The details of the procedure are as follows. Compute in both $t-1$ and t the Xth percentile of the expenditure shares within product groups. A common good must exceed the Xth percentile in both periods.

The resulting price indices generally shift up as successively stricter definitions of common goods are imposed, ranging from the 10th percentile of market share to the 50th percentile. For some product groups like memory cards, the CUPI using the CGR at the 30th or 50th percentile yields patterns in the ballpark of the Feenstra index. However, even with a 50th-percentile threshold, the CUPI remains low compared to the Feenstra index for products groups such as headphones and boys' jeans.

These findings have a number of important implications. First, the CUPI is sensitive to the specific definition of the CGR, in a manner that varies across product groups. A 50th-percentile threshold for the market share of goods present in t and $t-1$ implies that an entering good does not count as a common good until it reaches thes top half of the market share distribution. Similarly, a good that is on its way to exit that falls below the 50th percentile of market share is put into the entry/exit group (and becomes part of the Feenstra adjustment term). Many factors may underlie these patterns, as we discuss below. For boys' jeans, the seasonal product turnover cycle in apparel is arguably at work. Late in any apparel product turnover cycle, bargain racks are often available with low prices but limited supply.

We are sympathetic to the view that some form of CGR is needed. The primary inference we draw at this point is that the CUPI is sensitive to the specification of the CGR and more research is needed. Part of this should be further research into the dynamics of the process of the entry and exit of goods. Our analysis in Figure 1 is a step in that direction. We think it is likely that process varies by product group, consistent with our results showing the CUPI's differential sensitivity to various CGRs across product groups. Further research, motivated by theory, is necessary to provide guidance about product-group specific common goods rules.

Martin (2020) notes that the S* ratio can reflect not only shifting preferences, but also any model misspecification, including a nested preference structure. The CUPI's assumed CES preference structure imposes an equal elasticity of substitution within product groups, and violations of this assumption could lead to biased measures of inflation. Furthermore, the CUPI is more vulnerable to this issue than other CES price indices.[25] We explore this issue by exploiting the detailed product attributes to define a nested product substitution structure using two methods. First, we define nests within product groups with a heuristic-based approach. With this method, we assign products to subgroups based on a set of key variables that we as analysts hypothesize define market strata. As this procedure is labor intensive and informal, we also construct alternative subgroups by allocating products to groups based on their decile of predicted price from a log-level hedonic model. Intuitively, in the first approach we implicitly assume that substitutability is constant within market strata (for example, drip coffee makers versus espresso machines), while in the second approach we assume that price tiers (for example, low-end versus high-end coffee makers) define the substitution structure.

The nested approach requires estimation of elasticities of substitution for products within the same nest and across nests. We follow Hottman et al. (2016) in estimating within- and between-nest elasticities for each product group. The within-group estimation uses a modified Feenstra (1994) estimator that double-differences market shares and prices with respect to time and a time-varying nest level mean.[26] The between-nest estimator of the elasticity of substitution uses an instrumental variable (IV) approach building on Hottman et al. (2016).[27]

[25]More precisely, Martin (2020) shows that the CUPI is not consistent in aggregation. Vartia (1976) defines consistency in aggregation as the equality of a single-stage or two-stage index number. In the single-stage of an index number, all goods are included in a single aggregation. In a two-stage construction, the index is computed for a number of subgroups, and the subgroups are aggregated using the same index number formula. Diewert (1978) shows that the Sato-Vartia index, which is also exact for CES preferences under stricter assumptions, is consistent in aggregation.

[26]The identifying assumption of the Feenstra (1994) estimator is that supply and demand shocks are independent when sales growth and price growth are differenced with respect to a time-varying mean. The (Hottman et al., 2016) assumption is arguably more natural, as differencing with respect to a within-nest mean more plausibly identifies orthogonal supply and demand shocks.

[27]We follow Hottman et al. (2016) by specifying the between-group relationship between the nest-level price index and expenditure share. The former is endogenous, and Hottman et al. (2016) overcome this by using variation in the nest-level price index caused by changes in within-nest expenditure share dispersion. We innovate on the procedure of Hottman et al. (2016) by using the S* ratio (i.e., changes in common goods expenditure share dispersion) from the within-nest CUPI as the instrument, which removes changes in expenditure-share dispersion induced by product turnover. The identifying assumption is that within-nest

The estimated elasticities for the nested specifications are reported in Table 2. The results are broadly similar across the two nested approaches, with the within-nest elasticities estimated to be larger than the between-nest elasticities, as expected. These within vs between elasticity differences are potentially quite important for implementing the Feenstra index and CUPI, which both require elasticities of substitution to implement.

Figure 6 plots the UPI with a 30th percentile CGR along with the nested CUPI, also using a 30th percentile CGR applied at the within-nest level.[28] The alternative nesting approaches yield similar results, with the nested CUPI tending to show slightly less deflation than the CUPI. In unreported results, we find that the relationship between the nested and CUPIs is robust to using alternative CGR cutoffs.

4.1.3 Comparing Traditional, Hedonic, and Exact Price Indices

Figure 7 presents the main traditional, hedonic, and demand-based price indices for all five product groups. Since the CUPI indices are an outlier for some groups (even using the 30th percentile CGR), Figure 8 displays price indices without the CUPIs but with the addition of the Laspeyres index. Price indices in these product groups all follow a roughly similar pattern of relative orders: the Laspeyres index is the highest, the Tornqvist, Sato-Vartia and the Feenstra are in the next group, the Hedonic Tornqvist using the TV method is systematically lower, and the CUPI (both baseline and nesting with characteristics or p-hats) are at the bottom, especially for headphones and boys' jeans. The Feenstra index is systematically lower than the Sato-Vartia index, consistent with the quality adjustment for product turnover lowering the estimated rate of inflation.

The substantial gap between the Laspeyres and Tornqvist indices for all product groups highlights the advantages of using item-level scanner data, which permits construction of a superlative price index with internally consistent prices and expenditure shares in adjacent

demand shocks are uncorrelated with between-nest demand shocks. This innovation integrates the insights of Hottman et al. (2016) with Redding and Weinstein (2020).

[28]Nests are weighted by the number of products to adjust for differential product group sizes.

periods. Thus, using scanner data can produce substantial improvements in price measurement even without quality adjustment. Along these lines, it is notable that the gap between the Laspeyres and the Tornqvist indices varies over time, consistent with the Laspeyres index exhibiting a time-varying substitution bias.

Figures 9 and 10 present analogous plots of price levels to the plots of price changes in Figures 7 and 8. Figures 9 and 10 thus provide perspective on the cumulative effects of the differences between the various indices. Table 3 provides the indices in 2018:4 reflecting the cumulative effects since 2014:4 where all indices are set equal to one. The hedonic Tornqvist using the TV approach yields systematically larger cumulative declines in prices than the traditional Tornqvist and the demand-based indices other than the CUPI. The range of differences varies across product groups, with larger differences for memory cards, headphones, and boys' jeans. The Feenstra index yields systematically larger cumulative declines than the Sato-Vartia index, but the differences are smaller than the differences between the traditional and hedonic Tornqvist indices. The CUPI (with a 30th percentile CGR) yields substantially larger cumulative price declines for headphones and boys' jeans, while for memory cards and coffee makers, the CUPI yields similar declines to the hedonic Tornqvist. For occupational footwear, the cumulative declines from the CUPI are larger than from the other indices, but the gap is relatively modest.

Taking stock of the results from the NPD data, the most robust methodology yielding systematic quality adjustment is the hedonic Tornqvist using the TV method. The Feenstra index is also a useful point of comparison given its systematic relationship with the Sato-Vartia index. The CUPI is the most general demand-based index, but, even with a seemingly strict CGR, it yields very large price declines for some product groups.

The hedonic Tornqvist using the TV method yields systematically lower rates of inflation relative to the Feenstra. Both methods incorporate quality adjustments from product turnover, albeit in different ways. The hedonic Tornqvist with the TV method also incorporates time-varying changes in the relationships between prices and characteristics, including unobservable characteristics. The largest gaps between the hedonic Tornqvist and the Feenstra are for memory cards, headphones, and boys' jeans.[29]

4.2 Nielsen Results

For the Nielsen scanner data, we focus in the main text results for food product groups. The Feenstra index and CUPI require estimates of the elasticity of substitution. As with the NPD data, we use the Feenstra (1994) procedure to estimate those elasticities. The estimated elasticities for the 50+ product groups in food display considerable variation. While the median is about 6, the 10th percentile is about 4 and the 90th percentile is 12. These patterns are similar to those reported in Redding and Weinstein (2020).

For the hedonic TV approach, we combine the insights of the Erickson and Pakes (2011) with machine learning as described above and in the appendix.[30] The machine learning approach is necessary given the Neilsen scanner data's sparse yet complex information on item-level attributes.

We again explore CGRs to implement the CUPI. For the Nielsen data, we consider a related but distinct approach for defining the CGR rule compared to the approach explored with the NPD data. For Nielsen, we compute percentiles of the pooled sales distribution within a narrow product group for pooled sales in $t-1$ and t. Common goods are defined as goods present in both $t-1$ and t and in period t have sales above the Xth percentile of this pooled sales distribution. We implement this alternative CGR approach for Nielsen since as discussed below this approach provides flexibility to consider longer duration based alternatives. The Nielsen data is a longer panel and facilitates more exploration of alternatives that depend on the duration of how long goods have been in the market.[31]

[29] A limitation of our analysis with the NPD data is that the five product groups are narrow and do not provide sufficient coverage to compare with the BLS CPI for price indices or expenditure patterns from the PCE. The Nielsen data we consider in the following section has broad coverage of food products, which permits us to make such comparisons.

[30] Our companion paper, Cafarella et al. (2021, in progress) provides further detail.

[31] In unreported results, we have found that the Nielsen results using the identical CGR used in the NPD data yields very similar results to those reported here using a two-quarter horizon.

Figure 11 shows the results for the aggregated food categories of the CUPI and its components using different CGRs under the sales based percentiles. For Nielsen, it is the S* ratio (i.e., the consumer valuation bias) that is especially sensitive to the CGR. Recall this is an unweighted geometric mean sensitive to small market shares. The sensitivity of the S* ratio yields sensitivity of the CUPI to the CGR percentile thresholds. The CUPI without a CGR percentile threshold (the baseline) has average quarterly price inflation about 10 percentage points below the Feenstra. Using a 50th percentile for the CGR yields a price index that is much closer to the Feenstra index.

In the appendix, we consider alternative specifications of the CGR using market thresholds using percentiles of sales pooled over over current and prior 4 quarters. Using a longer horizon puts more weight on goods present for the longer horizon and thus this is a step in the direction of the duration based CGR rule used by Redding and Weinstein (2020) but does not require forward looking information. Using this longer horizon approach for computing sales percentiles, we find that a CGR between the 25th and 50th percentile yields results similar to those in Redding and Weinstein (2020) for the CUPI.[32] Again, our primary inference at this point is that the CUPI is sensitive to the CGR and more research is needed to provide guidance about such rules on a product group by product group basis.

Figure 12 presents a full set of price indices in change and level forms. The panels of the figure include the BLS CPI computed (thanks to BLS) for the Nielsen product groups. We find that the CPI and the traditional Laspeyres index track each other closely in Nielsen's Food product groups, with a growing discrepancy towards the end of the sample. The Tornqvist and Sato-Vartia indices are lower than the Laspeyres, and the quality-adjusted indices (Feenstra, hedonic Tornqvist using the TV approach, and CUPI) are even lower.

The cumulative level implications highlight that the hedonic Tornqvist is about 5 percentage points lower in 2015 than the traditional Tornqvist, and the Feenstra index is about

[32]We find comparable results even though Redding and Weinstein (2020) use the Nielsen Consumer Panel. In the appendix, we also explore the use of the Nielsen Consumer Panel and CGR sales based percentile rules. We find we can also closely approximate their results for the CUPI using the CGR rules we consider. Details are discussed in the appendix.

5 percentage points lower than the Sato-Vartia. These substantial cumulative differences for the Food product groups suggest notably that quality improvement via product turnover has not been limited to products where technological progress is most visible. Using a 25th percentile CGR, the CUPI is more than 40 percentage points lower than the Feenstra index in 2015; using a 50th percentile CGR reduces the difference to 20 percentage points.[33]

We consider the patterns in the Nielsen data to be broadly similar to the patterns in the NPD data. Quality adjustment, either via hedonic approaches or the Feenstra product turnover adjustment, imparts a substantial downward adjustment on price indices. The more general CUPI suggests an even larger quality adjustment, but we note again its sensitivity to the CGR. This sensitivity is robust to alternative approaches to defining the CGR thresholds for common goods.

4.3 Discussion

Item-level transactions data with prices, quantities, and attributes have considerable advantages for computing quality-adjusted price indices compared to traditional BLS methods, which use disparate sources for price quotes for common goods and expenditure shares. Even traditional matched model price indices such as the Laspeyres, and especially the Tornqvist, constructed from item-level data dominate the current approach to constructing indices from many disparate sources. The expenditure shares from the item-level data are internally consistent with the price data, and they are also available in real time. This permits constructing superlative price indices such as the Tornqvist in real time. We find that the Tornqvist index measures systematically lower inflation than the Laspeyres, and the gap is time-varying. This pattern again highlights that an important gain from using scanner data can be realized by using traditional superlative price indices with internally consistent

[33] Results for nonfood in the appendix show substantially greater departures between the BLS CPI and the Nielsen Laspeyres consistent with our concerns about the representativeness of Nielsen for the nonfood groups. The CUPI for nonfood is very low. With a 30th (50th) percentile CGR, the CUPI price level (indexed to 2006) is almost 70 (40) percentage points lower in 2015 compared to the Feenstra. This may be due to the limited coverage of nonfood items in the Nielsen scanner data.

prices and expenditure shares.

Assuming the item-level data contain information on product attributes, hedonic methods can be used at scale in real time, as the hedonic estimation methods can be based on the current (and in the case of using log price relatives for first-difference approaches) and lagged prices. We have found that the most robust approach for implementing hedonics at scale is to use the time-varying unobservables approach from Erickson and Pakes (2011). Our results provide ample support for their argument that it is important to correct for the reevaluation of the unmeasured characteristics of continuing, entering, and exiting goods.

Demand-based indices offer a useful alternative for comparison to hedonic indices. These indices are exact under certain sets of assumptions, and in the most general case (the CUPI), they can account both for quality change via product turnover and for time-varying product appeal for continuing goods. The limitation of the CES-based approaches we have considered is their sensitivity to the strong assumptions imposed by their underlying models, which may not be consistent with the data. When these indices are constructed from pooled national data, these indices rely on an important assumption that all goods within each product group or *nest* are equally substitutable. That assumption is arguably at odds with the data, especially for goods that have recently entered the market, which may take time to "roll out" geographically across the market (the same argument works for goods on the way out).

Figure 13 displays one way to characterize the challenge of imposing these assumptions. We pool the Nielsen item-level data at the weekly frequency from 2006–15 and compute the market penetration of items in the pooled data both on an unweighted (i.e., all items get the same weight) and sales-weighted basis.[34] Market penetration is defined as the share of Nielsen metro areas in which the item-level week is observed to have positive sales. On an unweighted basis, the distribution is very skewed to the left, with most item-level week observations having very low market penetration. Almost all of the unweighted distribution has less than a 20 percent market penetration. This pattern alone raises questions about applying a

[34]This figure uses both food and nonfood items. Future drafts will break this out for food vs. nonfood.

national market based CES price index for most items. These patterns also provide insights as to why common goods rules are needed for the CUPI. In unreported results, we find that the mass of the unweighted distribution with the lowest market penetration reflects entering and exiting goods.

The weighted distribution also raises questions about a national CES preference-based model, even for the price indices that rely only on expenditure weighted terms. Only 15 percent of sales are for items with a truly national market, although much of the mass of the sales-weighted distribution has market penetration of over 80 percent of metro areas. These findings provide some reassurance for the pooled national Sato-Vartia and Feenstra indices, both of which implicitly assume a national CES market but in contrast to the CUPI contain only expenditure-weighted terms. Traditional and hedonic superlative indices such as the Tornqvist are also expenditure-weighted indices. Such indices have been shown to be approximations or bounds for exact price indices under the assumption of a national market for the goods in question. The expenditure weighting of these indices makes this assumption more palatable.

We believe that the demand-based indices that incorporate quality adjustment (specifically the Feenstra and the CUPI) provide useful benchmarks that should be used for purposes of comparison with indices such as the hedonic Tornqvist using the TV approach. However, the national CES market assumption is too strong, especially for the CUPI as it is currently implemented. Future research could presumably make progress by developing a framework to distinguish between national and local goods and aggregate indices from local *nests* as appropriate. Likewise, progress could be made in better understanding the dynamics of product entry and exit, so that common goods rules could be disciplined by the nature of this process.

We think these topics should be high priorities for future research. In a companion paper (Ehrlich et al., 2021, in progress), we have been exploring the properties of traditional, hedonic, and demand-based indices in a simulated data environment. It is straightforward to show the attractive features of the CUPI in a rich environment of demand and cost shocks such as that developed in Hottman et al. (2016). The challenge is to enrich the model environment in sufficiently rich ways to mimic the limitations that the CUPI (and other indices) face in confronting the framework with the actual data. We have found that if we introduce some form of quantity rationing (as the bargain rack characterization of Boys' Jeans above suggests), price stickiness, or segmented markets, the demand-based approaches, and the CUPI in particular, can yield substantial departures from the exact price index.

5 Concluding Remarks

Using item-level transactions data with price, quantity, and attribute information enables quality adjustment of prices at scale. This paper employs such an approach to present evidence that traditional matched-model methods overstate the rate of inflation and understate the rates of real expenditure and real output growth substantially. We find that these patterns are pervasive, that is, not limited to goods such as electronics where technological progress is most visible.

We have explored and evaluated two alternative approaches for quality adjustment at scale with item-level transactions data, hedonic methods and demand-based methods. For hedonics, we have found that it is critically important to use the methodology developed by Erickson and Pakes (2011) that takes into account time-varying changes in the valuation of unobservable characteristics of continuing, entering, and exiting goods. Using this methodology, we have found that traditional matched model indices overstate the rate of inflation a wide range of goods.

We have focused on CES frameworks for demand-based indices, building in particular on the path-breaking work of Redding and Weinstein (2020). The CES unified price index they develop is quite general, incorporating quality change from product turnover (consistent with Feenstra, 1994) but also time-varying product appeal for continuing goods. A challenge in implementing the CUPI is that, contrary to the sales patterns in the item-level data, it is based on a national market for each CES-based nest of goods. Current implementations of the CUPI address these limitations by imposing strict common goods rules. While this approach is promising, our results indicate that the construction and suitability of common goods rules is an open area requiring further research.

This paper is a step in demonstrating that using item-level transactions data at scale can lead to a re-engineering of key national indicators. The current paper shows the promise for price measurement. A next step is to explore the promise of using the improved price index measurement with internally consistent measures of sales to improve measurement of real output growth. This is a high priority for future research.

References

P. Bajari, Z. Cen, V. Chernozhukov, M. Manukonda, J. Wang, R. Huerta, J. Li, L. Leng, G. George Monokroussos, S. Vijaykunar, and S. Wan. Hedonic prices and quality adjusted price indices powered by ai. Technical report, The Institute for Fiscal Studies Department of Economics, UCL, CENMAP working paper CWP 04/21, 2021.

C. L. Benkard and P. Bajari. Hedonic price indexes with unobserved product characteristics, and application to personal computers. *Journal of Business and Economic Statistics*, 23 (1):61–75, jan 2005. ISSN 07350015. doi: 10.1198/073500104000000262.

C. Broda and D. Weinstein. Globalization and the gains from variery. *Quarterly journal of economics*, (May), 2006. URL https://academic.oup.com/qje/article-abstract/121/2/541/1884019.

D. M. Byrne, D. E. Sichel, and A. Aizcorbe. Getting Smart About Phones: New Price Indexes and the Allocation of Spending Between Devices and Services Plans in Personal Consumption Expenditures. *Finance and Economics Discussion Series*, 2019(012), 2019. ISSN 19362854. doi: 10.17016/feds.2019.012.

M. J. Cafarella, G. Ehrlich, M. D. Shapiro, and L. Y. Zhao. Using machine learning to account for quality change and product turnover in price indexes. Technical report, Census, Maryland and Michigan, 2021, in progress.

J. De Haan. Hedonic price indexes: a comparison of imputation, time dummy and other approaches. 2008.

E. Diewert. Hedonic Regressions: A Review of Some Unresolved Issues. 2002. URL https://www.researchgate.net/publication/228404555.

E. Diewert, S. Heravi, and M. Silver. Hedonic Imputation versus Hedonic Time Dummy Indexes. Technical report, NBER WP 14018, 2008.

W. E. Diewert. Superlative Index Numbers and Consistency in Aggregation. *Econometrica*, 46(4), 1978. URL https://about.jstor.org/terms.

G. Ehrlich, J. Haltiwanger, R. Jarmin, D. Johnson, and M. D. Shapiro. Re-engineering key national economic indicators. *Big Data for Twenty-First Century Economic Statistics*, 2021.

G. Ehrlich, J. Haltiwanger, E. Olivares, M. D. Shapiro, and L. Y. Zhao. Hedonics vs demand based quality-adjusted price indices: Theory vs evidence. Technical report, Census, Maryland and Michigan, 2021, in progress.

T. Erickson and A. Pakes. An experimental component index for the CPI: From annual computer data to monthly data on other goods. *American Economic Review*, 101(5): 1707–1738, aug 2011. ISSN 00028282. doi: 10.1257/aer.101.5.1707.

R. C. Feenstra. New Product Varieties and the Measurement of International Prices. *The American Economic Review*, 84(1):157–177, 1994.

R. C. Feenstra. Exact Hedonic Price Indexes. 77(4):634–653, 1995.

R. J. Hill and D. Melser. The Hedonic Imputation Method and the Price Index Problem. Technical report, 2006.

C. J. Hottman, S. J. Redding, and D. E. Weinstein. Quantifying the sources of firm heterogeneity. *The Quarterly Journal of Economics*, 131(3):1291–1364, 2016.

A. A. Konüs. The problem of the true index of the cost of living. *Econometrica*, pages 10–29, 1939.

R. Martin. Changing Tastes Versus Specification Error in Cost-of-Living Measurement. BLS Working Paper 531, 2020.

A. Pakes. A reconsideration of hedonic price indexes with an application to PC's. *American Economic Review*, 93(5):1578–1596, 2003. ISSN 00028282. doi: 10.1257/000282803322655455.

S. J. Redding and D. E. Weinstein. Measuring Aggregate Price Indices with Taste Shocks: Theory and Evidence for CES Preferences. *The Quarterly Journal of Economics*, 135(1): 503–560, 2020. ISSN 0033-5533. doi: 10.1093/qje/qjz031.

K. Sato. The Ideal Log-Change Index Number. *The Review of Economics and Statistics*, 1976. ISSN 00346535. doi: 10.2307/1924029.

H. Schultz. A Misunderstanding in Index-Number Theory: The True Konus Condition on Cost-of-Living Index Numbers and Its Limitations. *Econometrica*, 1939. doi: 10.2307/1906996.

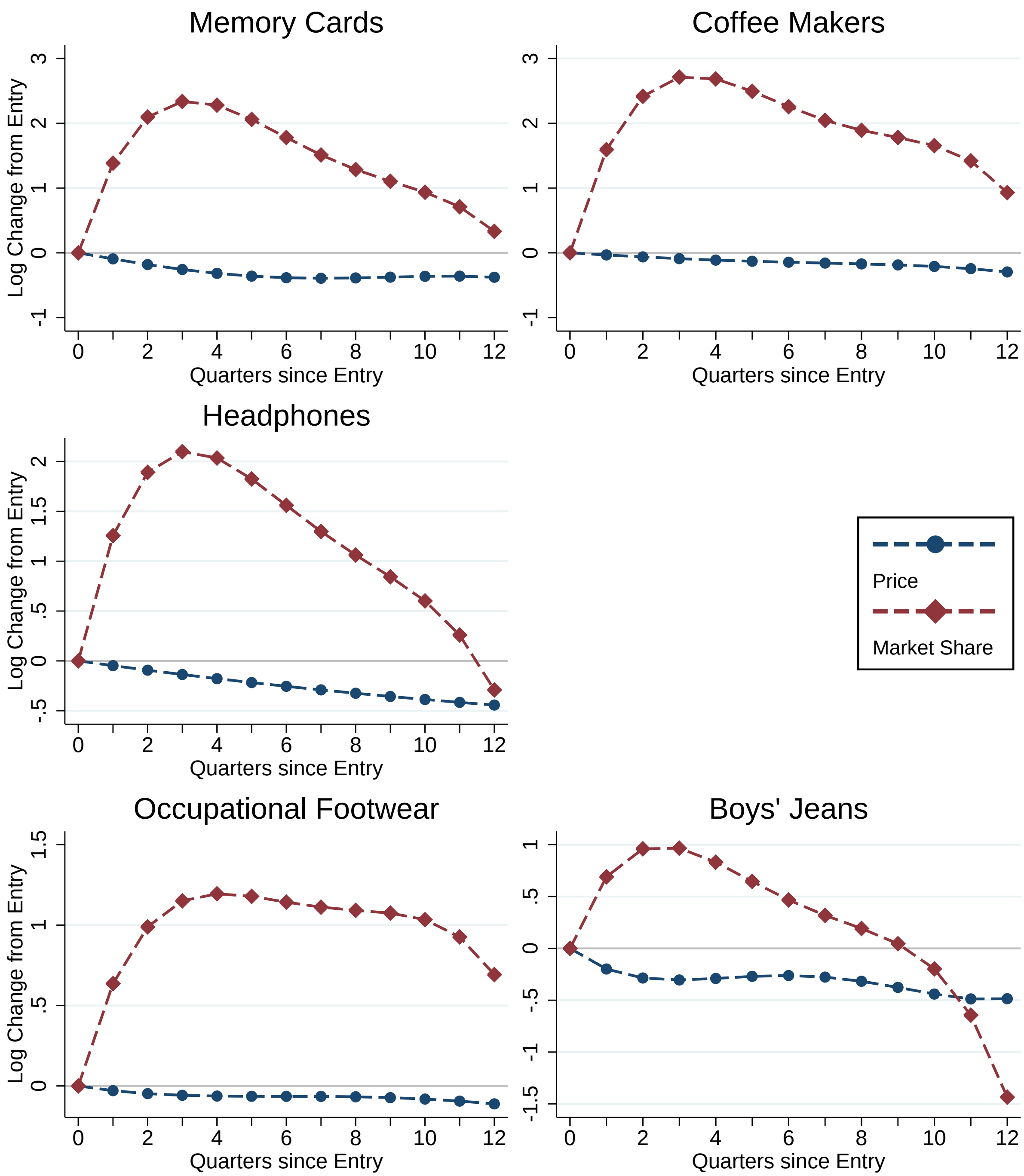
M. Silver. The use of weights in hedonic regressions: the measurement of quality adjusted price changes. 2003.

J. Triplett. Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes. 2004. doi: 10.1787/643587187107. URL http://dx.doi.org/10.1787/643587187107.

Y. Vartia. Ideal Log-Change Index Numbers. *Scandinavian Journal of Statistics*, 3(3): 121–126, 1976.

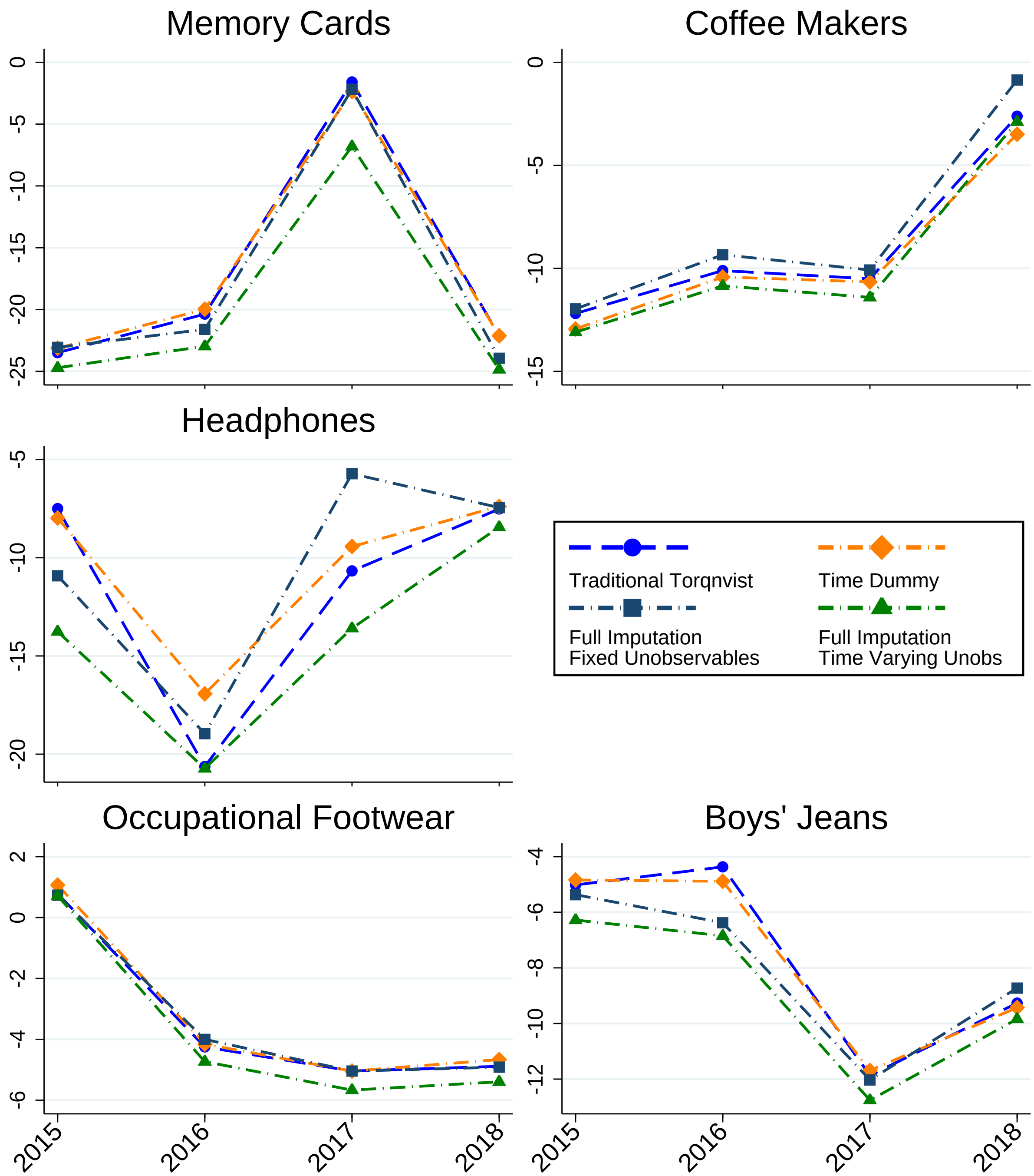
6 Figures

Figure 1: Product Lifecycle Dynamics



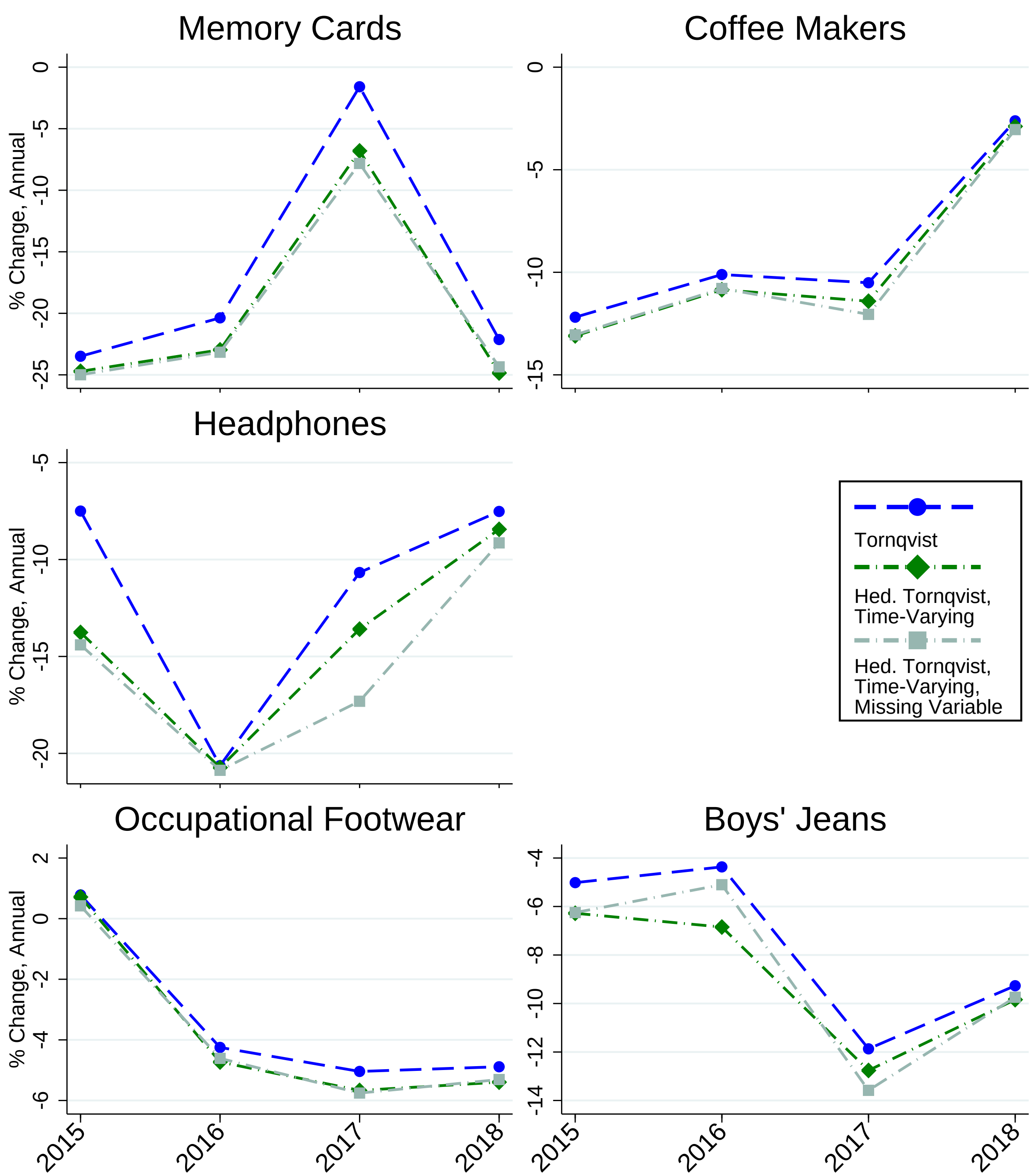
Notes: Unweighted average market share and prices relative to their value in the period of their initial entry. Entry occurs in period 0. All series are smoothed with a quartic spline. Data comes from the NPD Group.

Figure 2: Hedonic Specifications: Fixed vs. Time-Varying Unobserved Characteristics



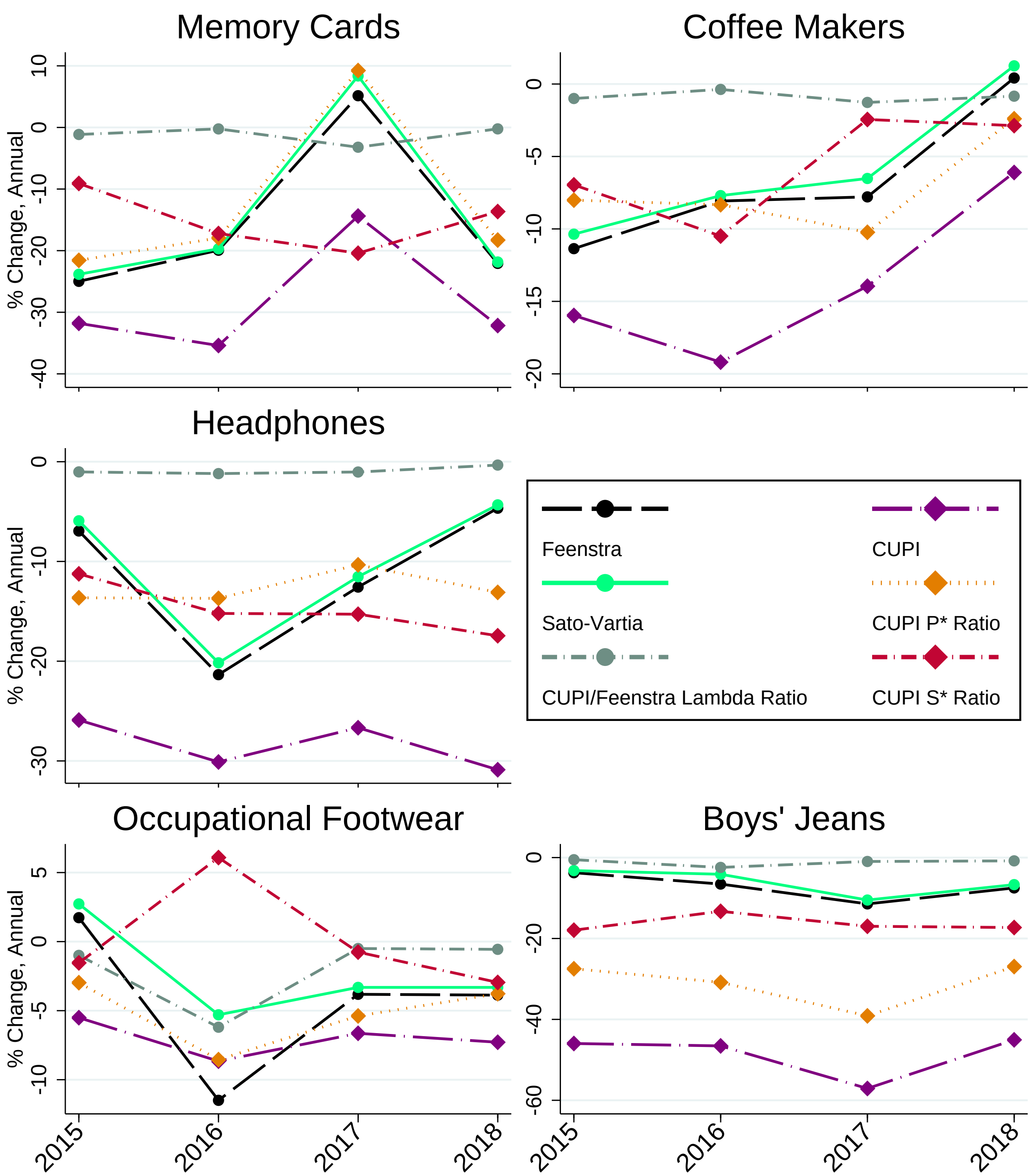
Notes: Traditional and hedonic Tornvqvist indices estimated with different methods. Values are percent change on an annual basis, aggregated from chained quarterly indices. The time-dummy Tornqvist index uses adjacent period estimation with Tornqvist market share weights. The fixed unobservables model estimates hedonic models of log change in price using WLS and average quantity-share weights. The time-varying unobservables model adds lagged hedonic level residuals to the log-difference specification. Data comes from the NPD Group.

Figure 3: Hedonic Specifications: Test of Time-Varying Unobservable Specification



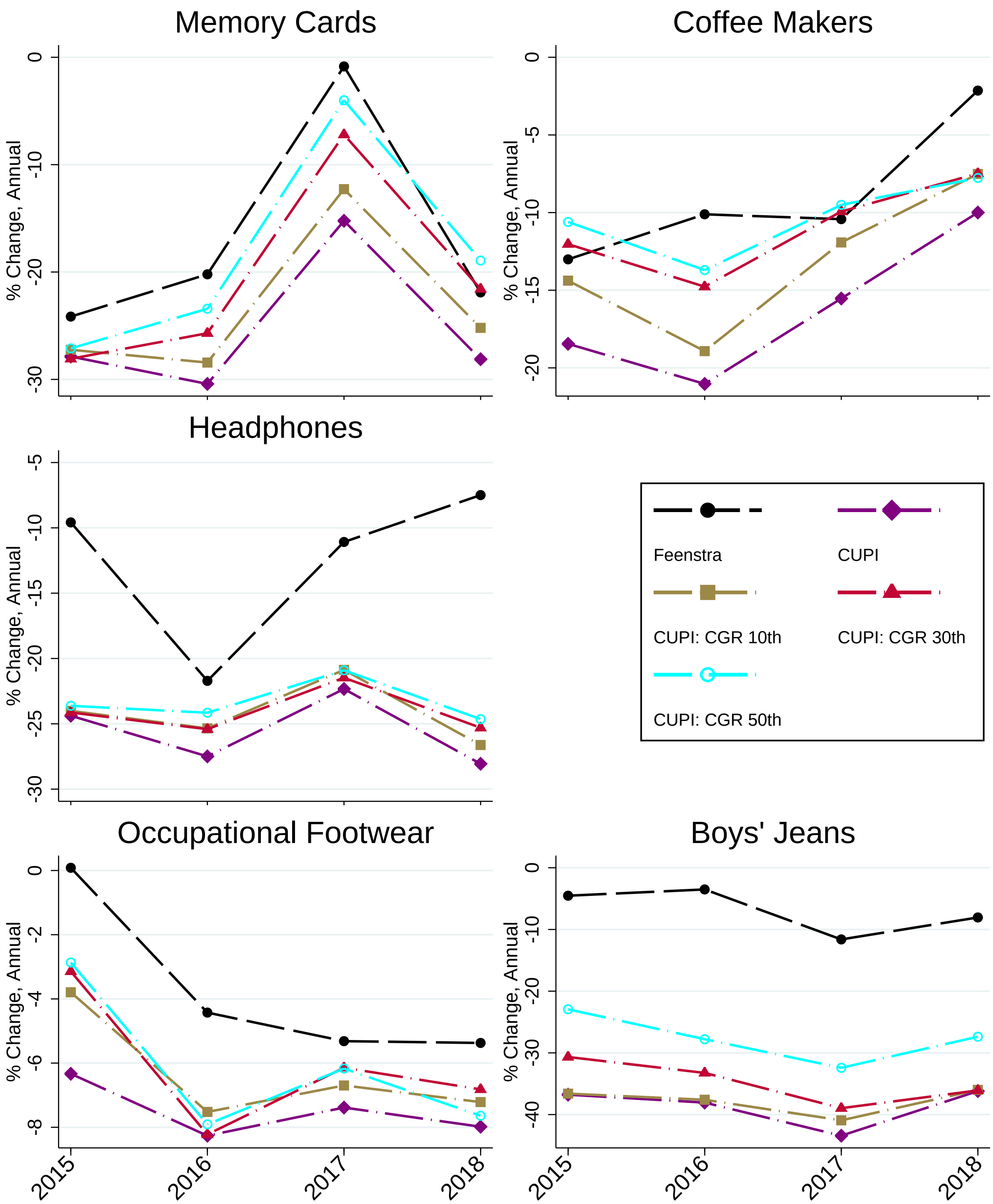
Notes: Values are percent changes on an annual basis, aggregated from chained quarterly indices. The time-varying unobservable model estimates hedonic models of log change in price using WLS and average quantity-share weights, including lagged hedonic level residuals. The "Missing Variable" series displays full imputation hedonic Tornqvist indices estimated using the time-varying unobservables approach, omitting key variables from the estimation. Data comes from the NPD Group.

Figure 4: Components of Feenstra and UPI



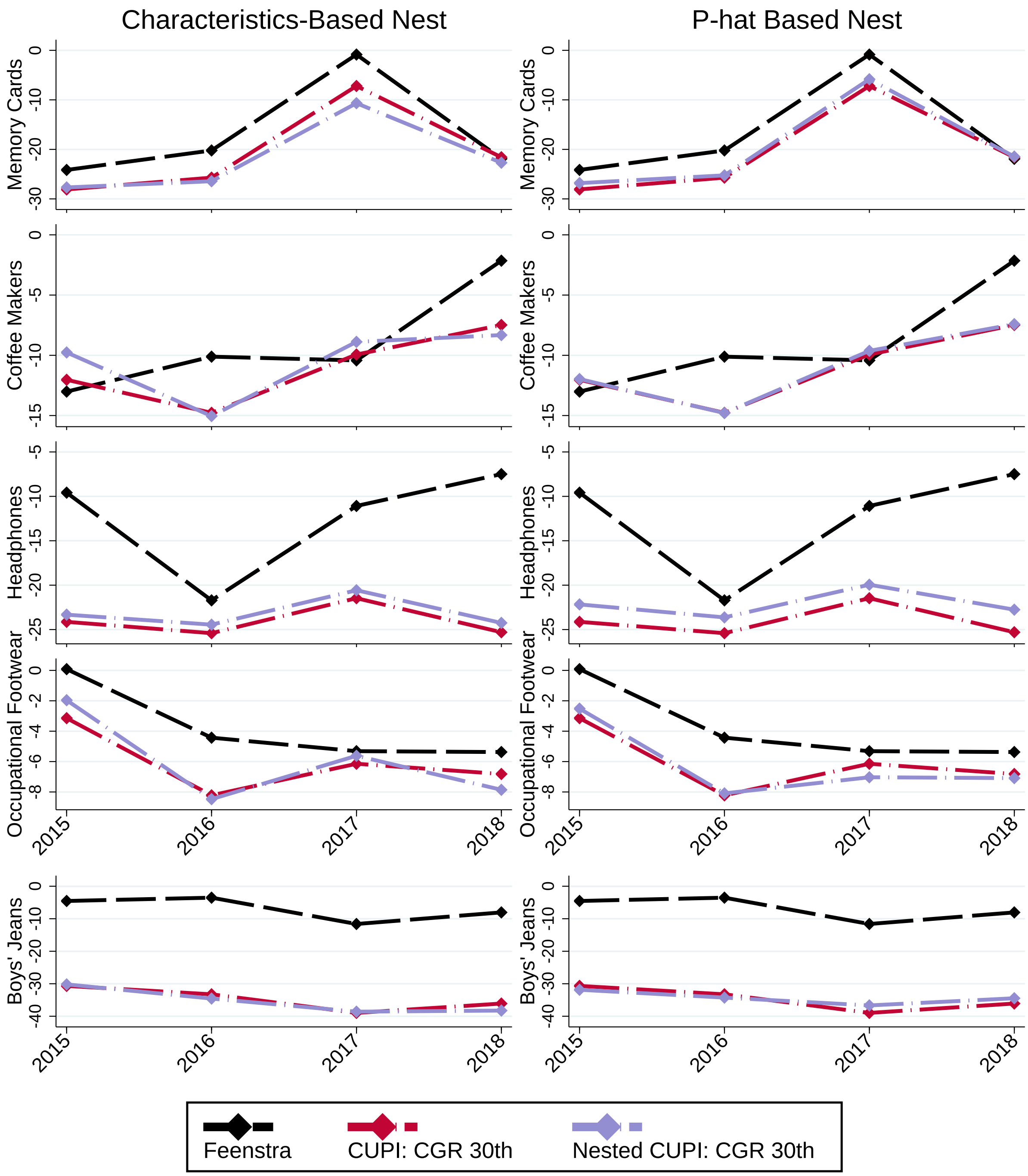
Notes: Values are percent change on an annual basis, aggregated from chained quarterly indices. The Feenstra index is the sum of the Sato-Vartia and CUPI/Feenstra Lambda Ratio. The CUPI is the sum of the Lambda ratio, P^*-ratio, and S^*-ratio. Data comes from the NPD Group.

Figure 5: CUPI: Common Goods Market Share Rules



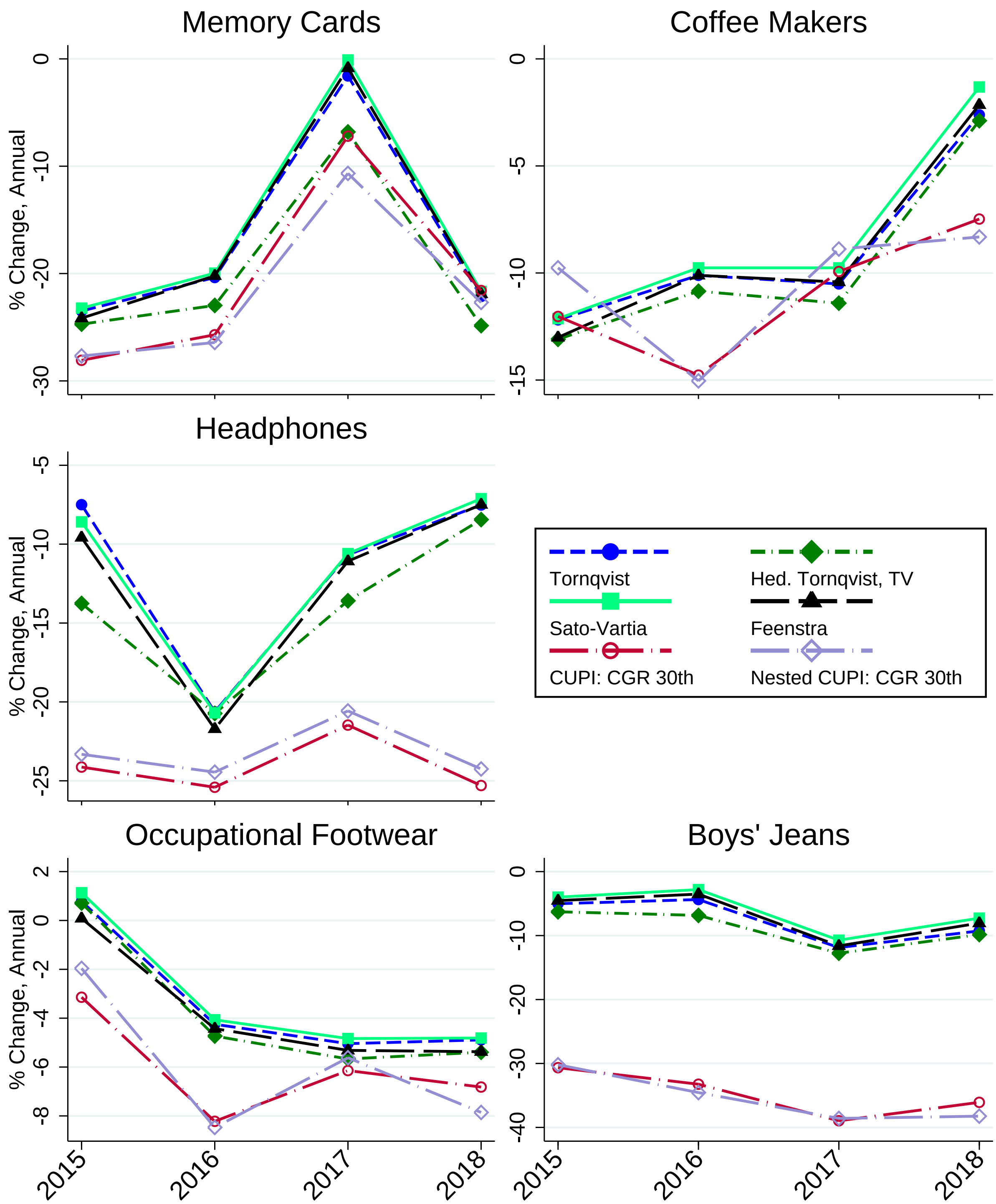
Notes: Values are percent change on an annual basis, aggregated from chained quarterly indices. Common goods market share rules for the CUPI exclude from the group of common goods those products with market shares below the noted percentile in both periods. The Feenstra-adjusted Sato-Vartia index is included for reference. Data comes from the NPD Group.

Figure 6: Nested UPI: Characteristics- and P-Hat- Based Nests



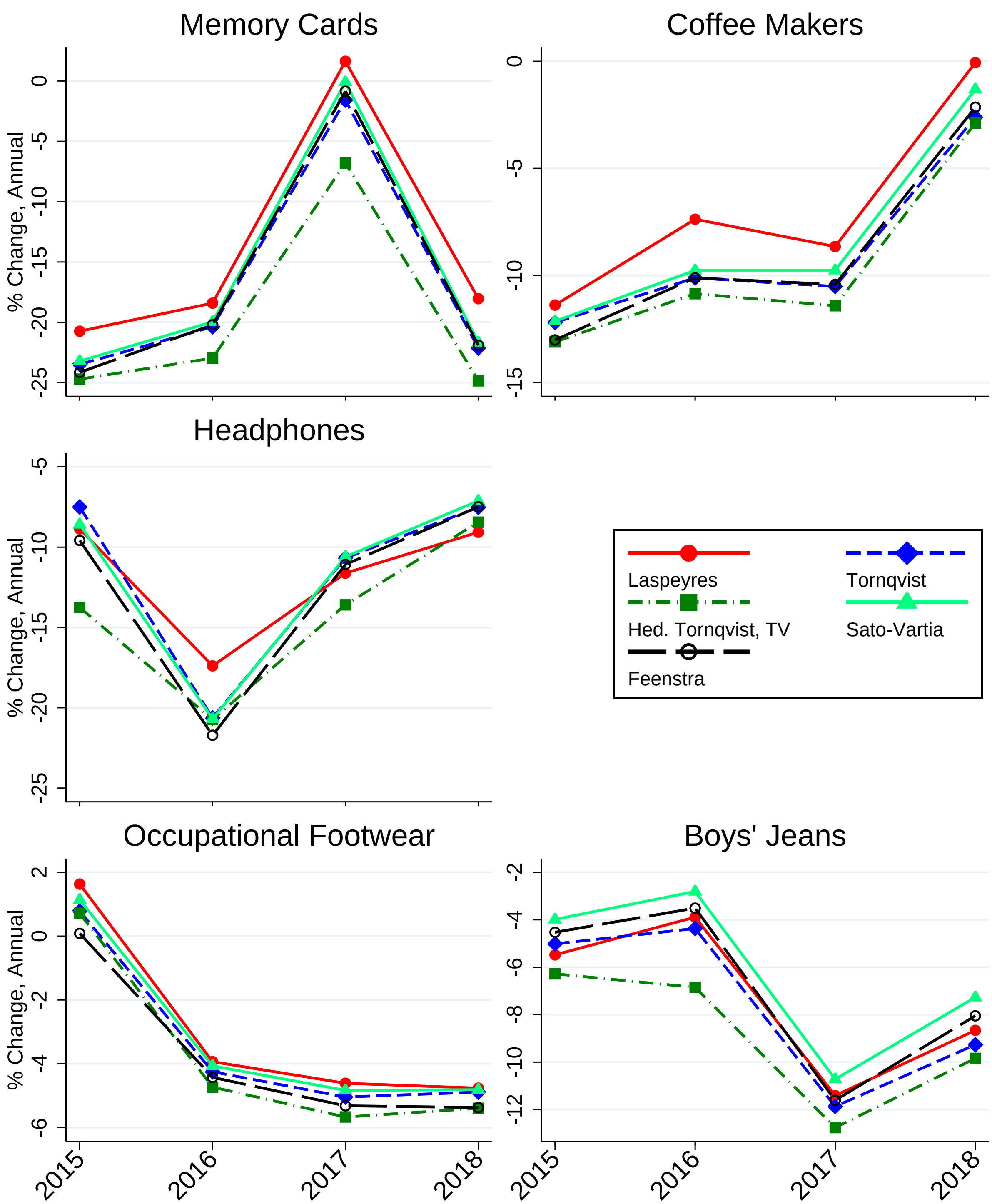
Notes: Values are percent change on an annual basis, aggregated from chained quarterly indices. For the characteristics-based nests, we assign items to groups based on shared observable characteristics. The p-hat based nests are based on the decile of predicted prices from unweighted hedonic log-level models. We estimate period-by-period hedonic models and assign items their most common decile over all periods. Data comes from the NPD Group.

Figure 7: Comparison of Main Price Index Specifications



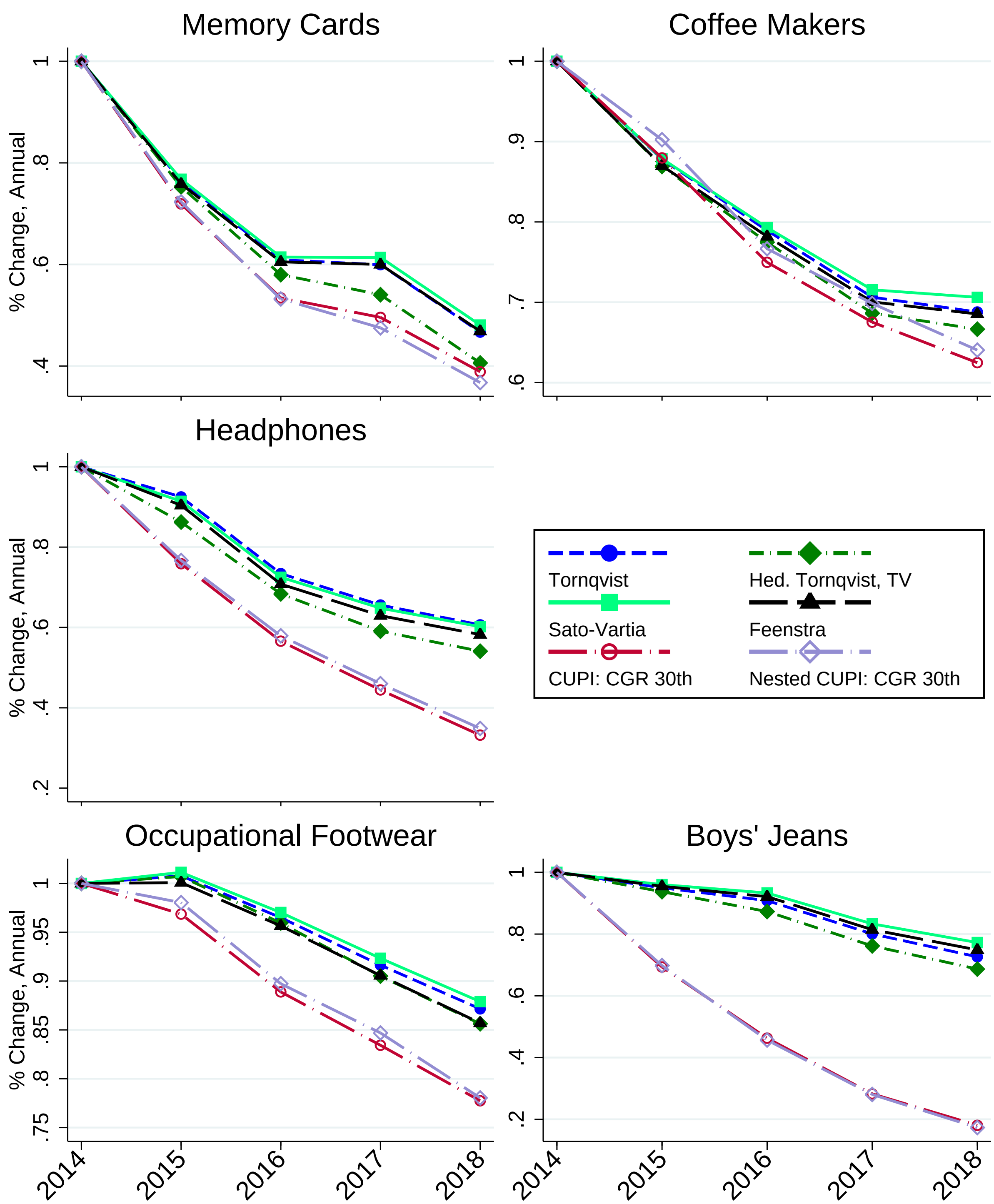
Notes: Values are percent change on an annual basis, aggregated from chained quarterly indices. The hedonic time-varying unobservables model is estimated over log price differences using WLS and with weights that are average quantity-shares in adjacent periods. Data comes from the NPD Group.

Figure 8: Main Price Index Specifications, Without CUPI



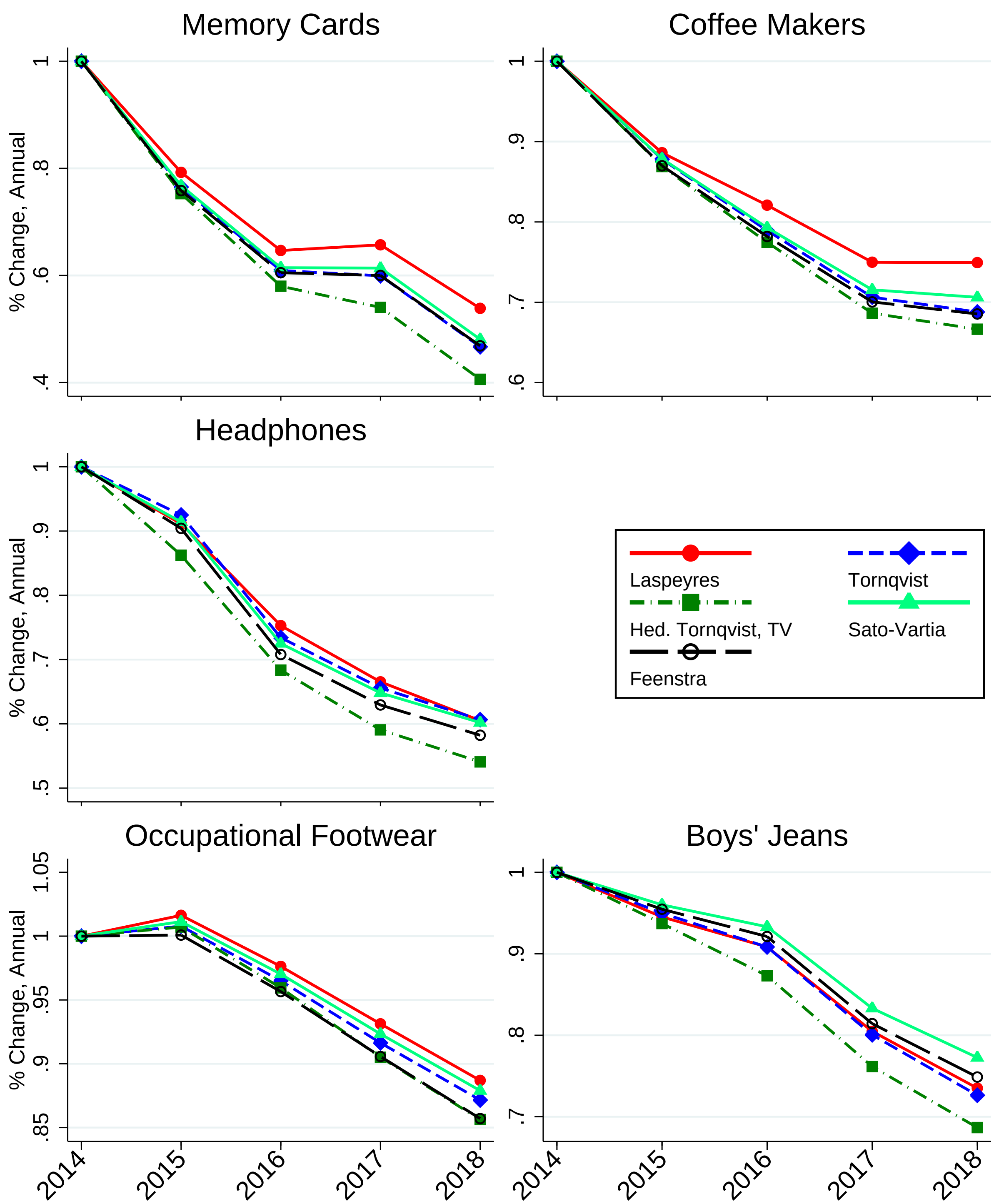
Notes: Values are percent change on an annual basis, aggregated from chained quarterly indices. The hedonic time-varying unobservables model is estimated over log price differences using WLS and with weights that are average quantity-shares in adjacent periods. Data comes from the NPD Group.

Figure 9: Main Price Index Specifications: Cumulative Price Level Changes



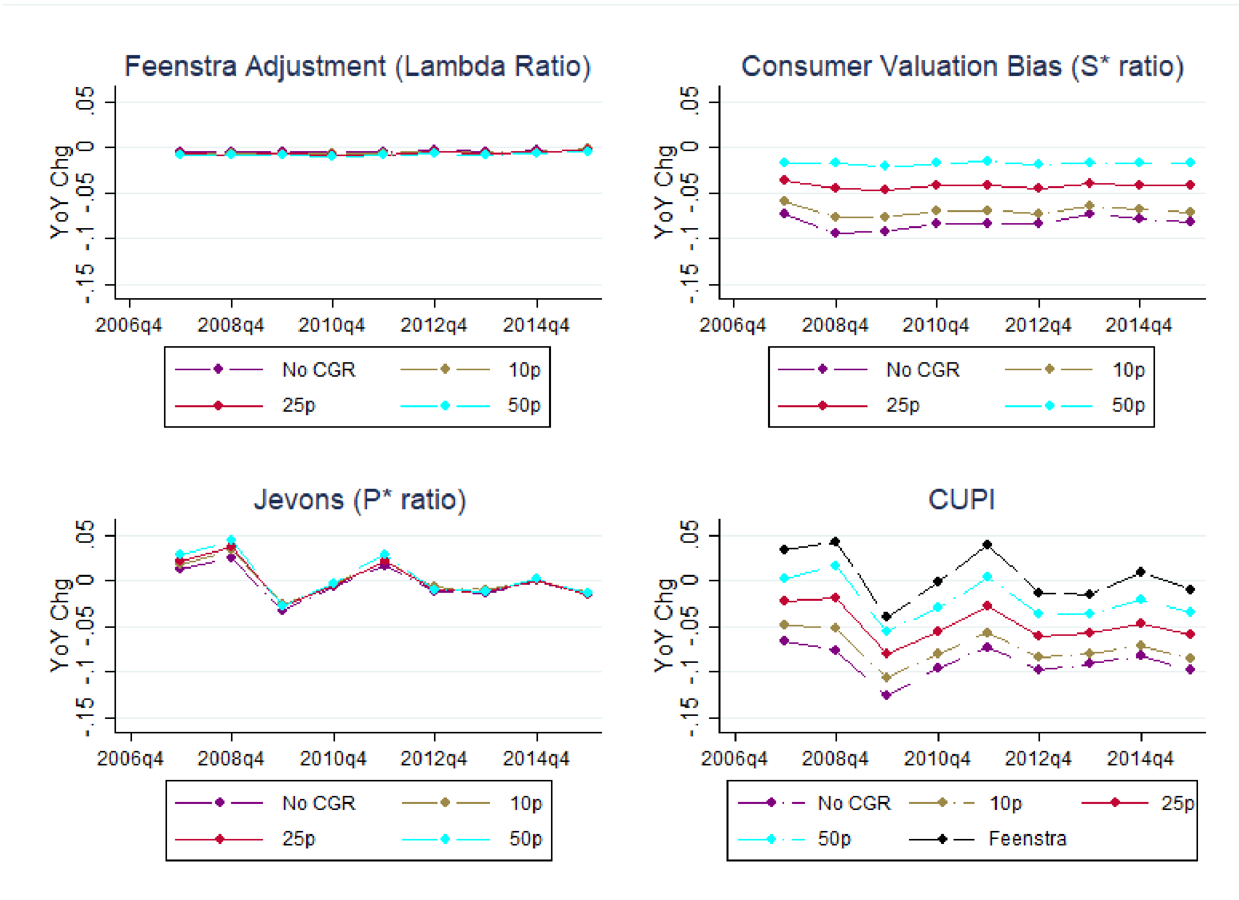
Notes: Values are cumulative changes relative to the 2014 price level, with 2014 price level set to 1. The hedonic time-varying unobservables model is estimated over log price differences using WLS and with weights that are average quantity-shares in adjacent periods. Data comes from the NPD Group.

Figure 10: Main Price Index Specifications: Cumulative Price Level Changes, No CUPI



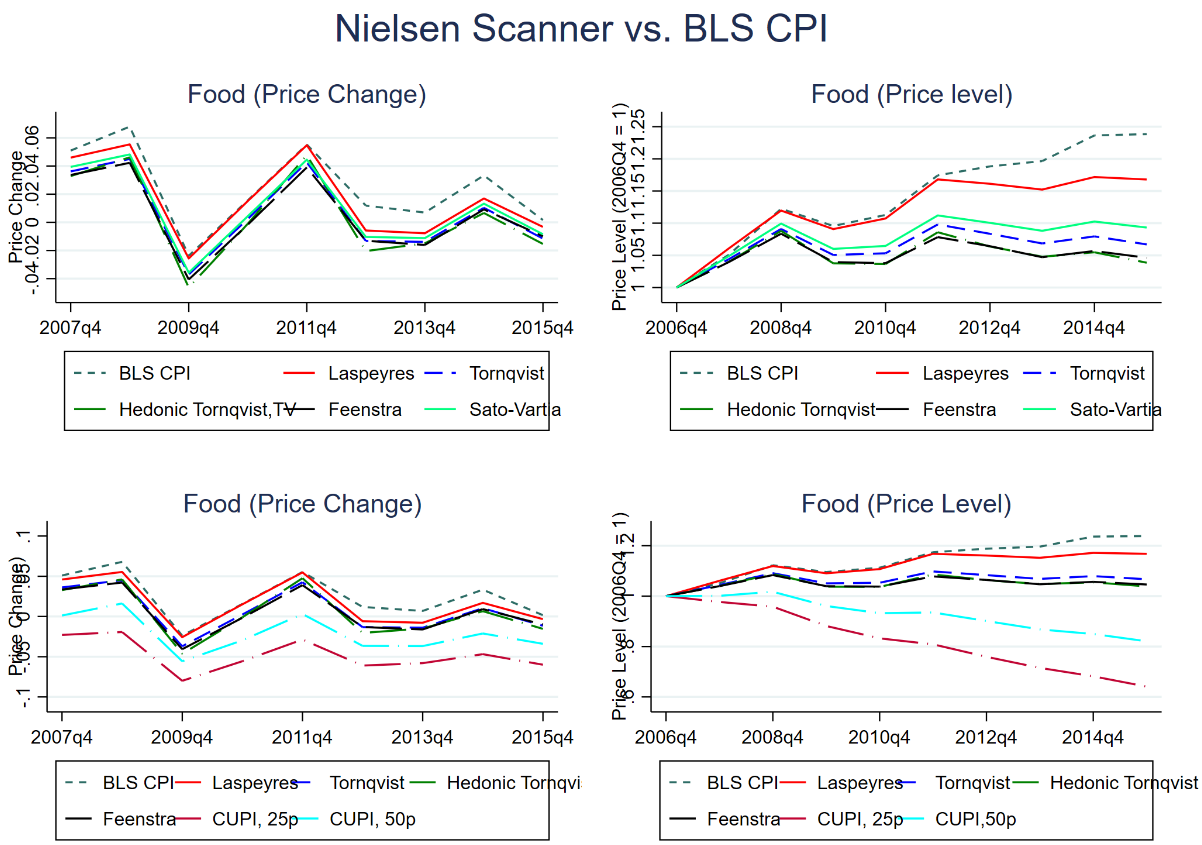
Notes: Values are cumulative changes relative to the 2014 price level, with 2014 price level set to 1. The hedonic time-varying unobservables model is estimated over log price differences using WLS and with weights that are average quantity-shares in adjacent periods. Data comes from the NPD Group.

Figure 11: Sensitivity of CUPI to CGR, Nielsen Food



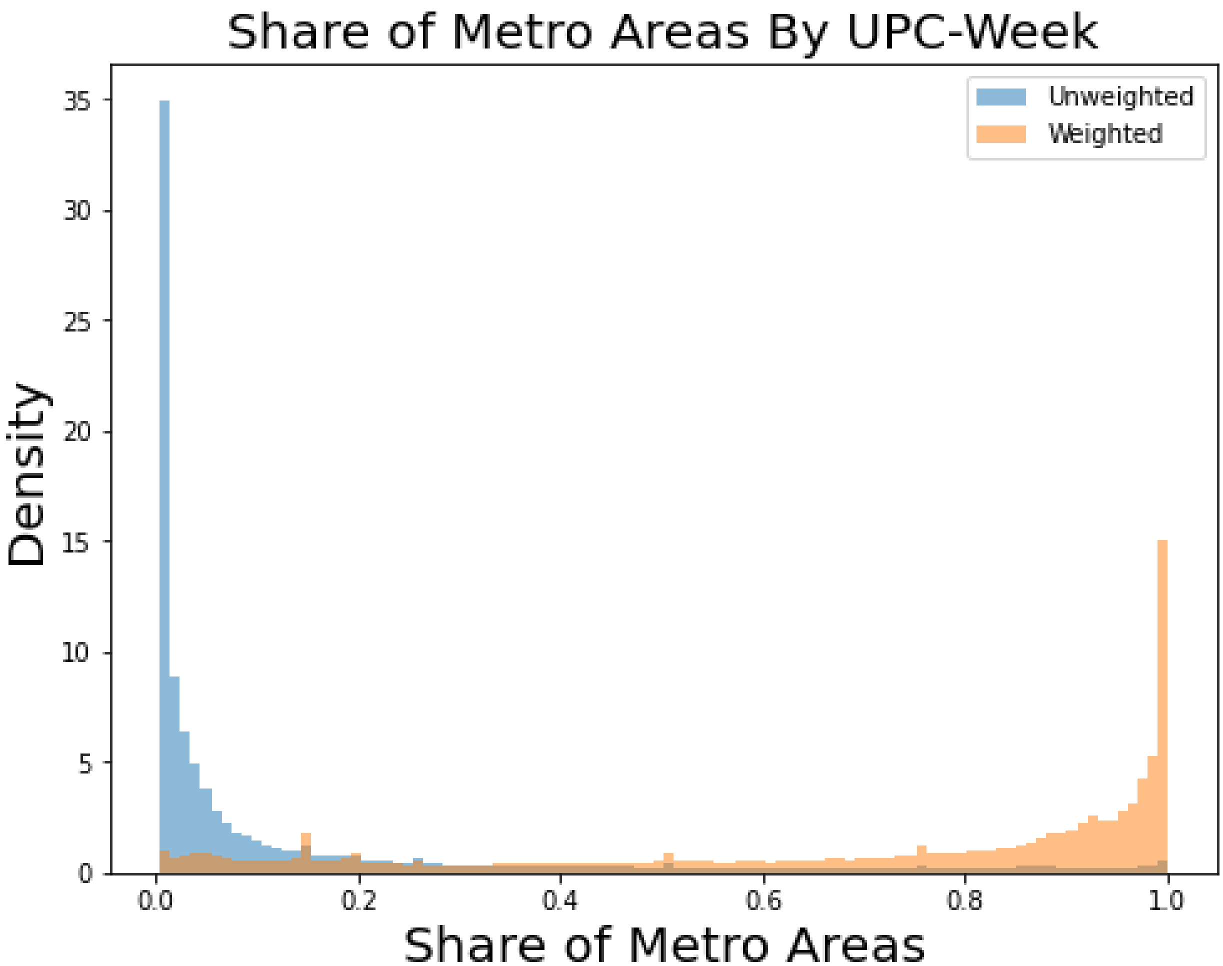
Notes: Values are annual changes from cumulative chained quarterly indices. Figures use Nielsen Retail Scanner data for food product groups.

Figure 12: Main Price Index Specifications: Price Changes and Levels, Nielsen food



Notes: Price changes are are annual changes from cumulative chained quarterly indices. Price levels reflect cumulative changes relative to the 2006 price level, with 2006 price level set to 1. Figures use Nielsen Retail Scanner data for food product groups.

Figure 13: Sales-weighted and Unweighted Distributions of Market Penetration of Items in Nielsen Data



Notes: All UPC items at a weekly frequency are used from 2006-2015. Unweighted shows the market penetration at the metro area of the unweighted pooled distribution. Sales-weighted shows the equivalent using sales weights. Figure uses Nielsen Retail Scanner data for food and nonfood product groups.

7 Tables

Table 1: Rates of Product Turnover: NPD Data

	Entry Rate		Exit Rate	
	All	Initial	All	Final
Memory Cards	5.8%	3.0%	6.0%	3.3%
Coffee Makers	5.7%	3.4%	4.5%	2.1%
Headphones	6.4%	3.8%	5.5%	2.9%
Boys' Jeans	11.5%	8.3%	7.8%	4.3%
Occupational Footwear	13.5%	9.1%	10.6%	5.5%

Average quarterly rates of product turnover. "Initial" entries are those for which the product was never observed in the data prior to the quarter. "All" entries include entries in which the product was previously observed prior to a spell of absence and the re-entered the data (i.e., "re-entries"). "Final" exits are those for which the product was never again observed in the data after the quarter. "All" exits include exits for which the product is subsequently observed after a temporary spell of absence (i.e., "temporary exits"). Transition quarter between data vintages excluded. Data come from NPD Group.

Table 2: Estimated Elasticities of Substitution: NPD Data

Product	Estimator	Groups	Elasticity of Substitution			
			Within		Across	
Headphones	Feenstra	-	7.634	(0.748)		
	HRW plus	Manual	8.609	(0.544)	7.704	(0.491)
	HRW plus	Hedonic	9.537	(0.969)	8.958	(0.423)
Memory Cards	Feenstra	-	5.623	(0.484)		
	HRW plus	Manual	6.31	(0.675)	4.534	(0.298)
	HRW plus	Hedonic	6.621	(0.657)	5.25	(0.586)
Coffeemakers	Feenstra	-	5.183	(1.289)		
	HRW plus	Manual	5.495	(0.791)	3.42	(0.63)
	HRW plus	Hedonic	5.345	(0.99)	5.306	(0.374)
Work/Occ Footwear	Feenstra	-	7.31	(0.533)		
	HRW plus	Manual	5.545	(0.509)	3.057	(0.493)
	HRW plus	Hedonic	6.199	(0.548)	4.135	(0.769)
Boy's Jeans	Feenstra	-	7.861	(0.565)		
	HRW plus	Manual	7.439	(1.5)	3.234	(0.734)
	HRW plus	Hedonic	8.156	(1.82)	3.418	(0.657)

Estimated elasticities of substitution for CES and nested CES models. Standard errors in parentheses. Data come from NPD Group.

Table 3: Alternative Price Indices, Levels in 2018q4 Relative to 2014q4: NPD Data

	Memory Cards	Coffeemakers	Headphones	Boys Jeans	Work/Occ Footwear
Tornqvist	0.467	0.688	0.607	0.726	0.872
Hed. Tornqvist,TV	0.406	0.667	0.541	0.687	0.856
Sato-Vartia	0.481	0.706	0.602	0.773	0.879
Feenstra	0.469	0.685	0.582	0.749	0.857
CUPI, CGR 30p	0.389	0.625	0.332	0.181	0.777
CUPI-N, CGR 30p	0.367	0.640	0.349	0.173	0.780

Notes: Values are cumulative changes in 2018:4 relative to the 2014 price level, with 2014 price level set to 1. CUPI-N is nested CUPI using characteristics approach. Data come from the NPD Group.

Appendix

A Hedonic Imputation Indices

A.1 Hedonic Estimation: Levels vs. Difference and Weighted vs. Unweighted Results

We estimate hedonic models in both log-levels and log-differences. We also consider weighted and unweighted approaches. Figure D.1 presents results from these alternative estimation approaches for the five product groups we have explored in the NPD data. The log-level specifications, whether weighted or unweighted, yield more erratic patterns than the log-difference specifications. The log first-difference results are similar whether weighted or unweighted. Importantly, the log first-difference specification controls only for time-invariant unobservable characteristics. Our preferred TV approach, which we illustrate in Figure 7 in the main text, can also controls for time-varying valuations of unobservable product characteristics.

The log-level specifications are sensitive to omitted unobservable characteristics. To illustrate this point clearly, Figure D.2 presents an enhanced version of Figure 3 that shows the sensitivity of the levels specification to intentionally omitted key observable characteristics. Unlike the TV approach, the log-levels specification is very sensitive to omitting these observable characteristics.

We report goodness of fit statistics for the alternative specifications in Table E.1. As expected, the log-level estimation models account for a large share of variation in product price levels, as measured by R^2. This high explanatory power reflects the fraction of the cross-sectional variation in prices accounted for by the observable characteristics. Those same models account for a small fraction of the variation in price relatives. The EP methods (EP1 is first differences and EP2 is the TV approach) yield much higher R^2s for the price relatives, especially for the weighted specifications.

In the main text, we focus on weighted hedonic specifications. As we have noted, the time dummy method inherently calls for a weighted specification, as the estimation weights determine the type of price index produced. The hedonic imputation specifications we consider also use weighted specifications in the hedonic estimation procedure. Using weights promotes consistency with the the time dummy results. Moreover, in the context of scanner data, there are additional reasons to prefer weighted hedonic models. First, scanner data is generally inclusive of products with a wide range of availability. The objective is to obtain the quarter-by-quarter mapping between prices and characteristics. With scanner data, the sample will typically include goods that are not widely geographically available or that few consumers actually purchase, partly because they have recently entered or are about to exit the marketplace. These low-quantity goods will have an outsized influence on unweighted hedonic estimates, and may therefore lead to hedonic price indices that do not reflect the environment faced by the representative consumer. Intuitively, weighted regression coefficients should be interpreted as the implicit prices of characteristics that consumers actually purchased (see Silver (2003) and Diewert (2002) for motivation of using weighted hedonic specifications along these lines).

We use quantity-share weights in our hedonic specifications, but we have found broadly similar results using market-share weights. De Haan (2008) advocates for quantity weights to be used in estimation. He notes that in the context of scanner data in particular, we do not observe prices but average unit values. Given that consumers purchase items from different stores, at different times during the aggregated periods over which average unit values are calculated, and perhaps with different bargaining power, it is likely that these unit values are likely to be measured with heteroskedastic errors across different items. Importantly, the variance of unit values is inversely proportional to the (square root of) the number of units sold, rather than the total value sold. Quantity weights are frequently used in the trade literature, which also frequently depends on unit values of imports or exports (e.g., Broda and Weinstein, 2006).

We use quantity weights in the results presented in Figure D.1. For single-period log-level estimation, we use contemporaneous quantity shares. Intuitively, this specification only uses information from the current period to produce hedonic estimates. For estimation of the specifications proposed by Erickson and Pakes (2011), in which the dependent variable is the change in log prices, we use weights that are the average of the quantity shares in the previous and current periods. The results using the EP method presented in the main text take the same approach.

B Using the Nielsen Data

B.1 Common Goods Rules – Consumer Panel and Retail Scanner

This section presents sensitivity results to alternative common good rule approaches for both the Nielsen Scanner and Nielsen Consumer Panel data. Using the scanner data, Figure D.3 compares the results of imposing common goods rules using the two quarter horizon as in the main text (i.e., using percentiles from sales pooled over the current and prior period vs a 5 quarter horizon (i.e., computing percentiles for sales pooled over the current and prior 4 quarters). It is evident that with a longer horizon that the CUPI at the 25th and 50th percentile based CGR is higher using the longer horizon. The longer horizon puts more weight on the goods that have been present in all five quarters which moves us in the direction of the duration based CGR approach of Redding and Weinstein (2020).

To facilitate comparison of our results to Redding and Weinstein (2020), Figure D.4 shows various price indices calculated using all product groups in the Nielsen consumer panel data. We have found we can closely mimic their results using a market share common goods rule, in which goods that have a market share above the 5th percentile calculated using the five quarter horizon discussed above. In this figure, We label the CUPI with the 5th-percentile CGR as "CUPI RW CGR" in Figure D.4 and label the version without a CGR as "CUPI No CGR."[35] Comparing the results for the CUPI RW CGR to those in Figure III in Redding and Weinstein (2020) shows a close correspondence for the overlapping years. We note that we do not impose a CGR in computing the other price indices shown in Figure D.4. This contrasts with the approach taken in Redding and Weinstein (2020), who apply the same common goods rule for all of the price indices they display. It is not clear that application of a CGR makes sense for other price indices and we did not do this so the demand based indices could be readily compared to the traditional and hedonic indices.

[35]In other words, the CUPI No CGR treats goods present in periods $t-1$ and t as common goods.

Figure D.4 shows that the CUPI applying the RW-mimicking (5th percentile) CGR is substantially higher than the CUPI with no CGR. In this respect, the results for the Nielsen Consumer Panel data are similar to those from the Retail Scanner data. For purposes of comparison, Figure D.5 shows related indices using the Nielsen Scanner data pooling all product groups and focusing on a CGR based sales percentiles computing over the 5 quarter horizon. The CUPI with no CGR is very low, suggesting deflation of 10 percent or more per year. Even the CUPI using a 25th-percentile cutoff rule shows persistent deflation in the Retail Scanner data; imposing a 50th-percentile CGR brings the CUPI closer in line with the Laspeyres index. The series labeled "CUPI, RW CP" shows results from applying the market share threshold from the Consumer Panel to the Scanner Panel data, rather than calculating a percentile-based threshold directly from the Scanner Panel data. Using the Consumer Panel share threshold for the CGR produces results similar to using the 50th-percentile CGR calculated directly in the Scanner Panel data.

The much lower inflation rates the CUPI measures in the Nielsen Retail Scanner data relative to the Consumer Panel data highlight the scanner data's large number of very low-market share products, which disproportionately impact the CUPI. In contrast, the Laypeyres and Feenstra indices are much more similar between the Nielsen Consumer Panel and Nielsen Retail Scanner data. Figure D.4 again displays the Bureau of Labor Statistics' Consumer Price Index for all of the product groups included in the Nielsen data as a point of reference. The Laspeyres index calculated from the Nielsen Scanner data tracks the BLS CPI more closely than does the Laspeyres from the Nielsen Consumer Panel data, with a slightly higher correlation and a substantially smaller mean difference.

The closer correspondence between the Laspeyres index measured in the Nielsen scanner data with the CPI (relative to the Consumer Panel Laspeyres index's correspondence with the CPI) is driven entirely by the food product groups. Figure D.6 displays for the nonfood product groups the analogous plots to Figure 12, which displays results for food product groups. For comparability purposes, the CGR rules in this figure are based sales percentiles over the two quarter horizon. The relationship between the Laspeyeres index and the CPI measured in the Nielsen Retail Scanner data is much weaker in the nonfood product groups. For food product groups, the correlation in the price changes for Laspeyres and the BLS CPI is 0.97; the analogous correlation is only 0.49 for nonfood. There is also a much larger gap in the mean price change differences for nonfood than food. Figure D.6 also shows that the CUPI yields implausibly low rates of inflation for nonfood product groups, even with the 50th-percentile CGR.[36]

In unreported results, we have also found that expenditures for harmonized food groups tracks the BEA PCE closely (also see (see Ehrlich et al., 2021) where we show this holds at an aggregated total food expenditures level of aggregation). These patterns suggest to us that the Nielsen Scanner data are representative of prices and expenditures for food products. We find that expenditures for harmonized nonfood groups differ importantly from the BEA PCE for many groups. These findings are consistent with the concerns discussed in the main text about Nielsen's coverage of nonfood items. Taken together, these findings reinforce our

[36]Figure D.6 also displays other indices, including the hedonic Tornqvist index we calculate using machine learning methods. Interpreting these indices is challenging given the apparent non-representativeness of the nonfood product groups in the Nielsen data. Further research is needed to understand the behavior of the nonfood product groups in the Nielsen data more clearly.

concern that the Nielsen data may be non-representative for many of the nonfood product groups.

Finally, in unreported results, we have also compared the Laspeyres indices for food and nonfood product groups from the Nielsen Consumer Panel with the BLS CPI. We find broadly similar patterns to those we found with the Retail Scanner data—the Laspeyres index for the food product groups tracks the CPI much more closely than does the Laspeyres index for the nonfood product groups. Laspeyres index inflation runs an average of about 2 percentage points per year higher in the Consumer Panel relative to the corresponding CPI for both food and nonfood product groups. The Laspeyres index for the food product groups in the Consumer Panel has a correlation of 0.98 with the CPI, however, while the correlation for the nonfood product groups is only 0.54. Thus, the Nielsen Consumer Panel also tracks food products in a similar way to the CPI, but the correspondence is weaker for nonfood products.

The main message from this sensitivity analysis is that the CUPI is sensitive to the specification of the CGR both in the Nielsen Consumer Panel and in the Nielsen Scanner data. This sensitivity applies both to the market share threshold used and to the horizon over which the threshold is computed. Using the longer horizon market share threshold moves the CGR towards the Redding and Weinstein (2020) duration-based approach. It is worth noting that any duration based approach has greater data requirements for practical implementation.

This section also provides further evidence of the non-representative nature of the Nielsen nonfood data.

B.2 Machine Learning and Hedonics

This appendix summarizes our procedure for incorporating machine learning into hedonic estimation. Our companion paper, Cafarella et al. (2021, in progress), provides further details.

Using machine learning (ML) methods to estimate hedonic price indices requires making several practical choices regarding the architecture of the ML system used for prediction and the conversion of those predictions into price indices. As discussed in the main text of this paper, our preferred approach to constructing hedonic price indices is the "time-varying unobservables" hedonic imputation approach of Erickson and Pakes (2011). The core of this method is to estimate price *levels* for each product in each period in a first step. In a second step, this approach estimates price *changes*, using the hedonic residual (or prediction error) from the first step as a predictor. This methodology allows the hedonic predictions partially to capture unobserved product characteristics' influence on price changes.

In many ways, the "TV" approach of Erickson and Pakes (2011) can incorporate ML methods quite naturally. The key innovation is to use ML methods rather than standard regression techniques to estimate the hedonic functions for log price levels and changes in equations (1) and (2). Another important difference from the more standard econometric procedures we employ in the NPD data is that the Nielsen data available from the Kilts Center does not include pre-coded item-level product attributes. Attribute information is limited to short, non-standard text descriptions. We use deep neural networks to predict product prices and price changes from these product descriptions.

Several features of our methodology merit particular discussion. First, to convert text-based product descriptions into numerical characteristic representations, we use a hybrid feature encoding architecture that allows the system to incorporate "pre-trained" word embeddings (numerical representations) trained from an external corpus of text as well as specifically trained or "text-tailored" embeddings trained specifically on the product descriptions in the Nielsen Kilts Retail Scanner Data set. Second, our architecture does not predict prices or price changes directly, but rather predicts a set of probabilities that the price or price change lies in each of a set of 10 bins that partition the observed range. Third, the ML system minimizes the weighted cross-entropy loss function for the products' true price and price change distributions in the hedonic estimation.[37] Both steps are weighted using products' unit sales (quantities) shares in a product-group quarter. Fourth, because of the noise in the estimated probabilities, it may not be optimal to form price predictions as the simple probability-weighted expected price. We use a receiver operating characteristic (ROC) curve procedure to determine the optimal number of bins to include in the price prediction.

Figure D.7 shows the ML procedure's performance as measured by the prediction "accuracy" and "near accuracy" across every product group-quarter.[38] We define the model's accuracy as the proportion of products for which the model assigns the highest probability to the correct price or price change bin in a pseudo-out of sample exercise, and the model's near accuracy as the proportion of products for which it assigns the highest probability to the correct or an adjacent bin. Panel (A) shows the results for the log-level model (first stage), and panel (B) shows the results for the log-differences model using the first-stage residuals (second stage). The median log-level model achieves an accuracy of 56.6% and a near accuracy of 87.9%. In other words, the median-performing model predicts the correct price bin more than half the time from among 10 possibilities and predicts the correct bin or an adjacent bin nearly 90% of the time. Predicting price changes is an inherently harder task than predicting price levels. Nonetheless, the median-performing log-difference model has an accuracy of 24.2% and a near accuracy of 54.1%. We view these model performances as remarkable: in the median product group-quarter, the system is able to closely predict a product's price change over half the time based on the short, nonstandard product descriptions discussed in Section 3.2.

[37]In this application, the cross-entropy loss objective function is equivalent to a maximizing the likelihood of assigning the highest probability to the correct price or price change bin.

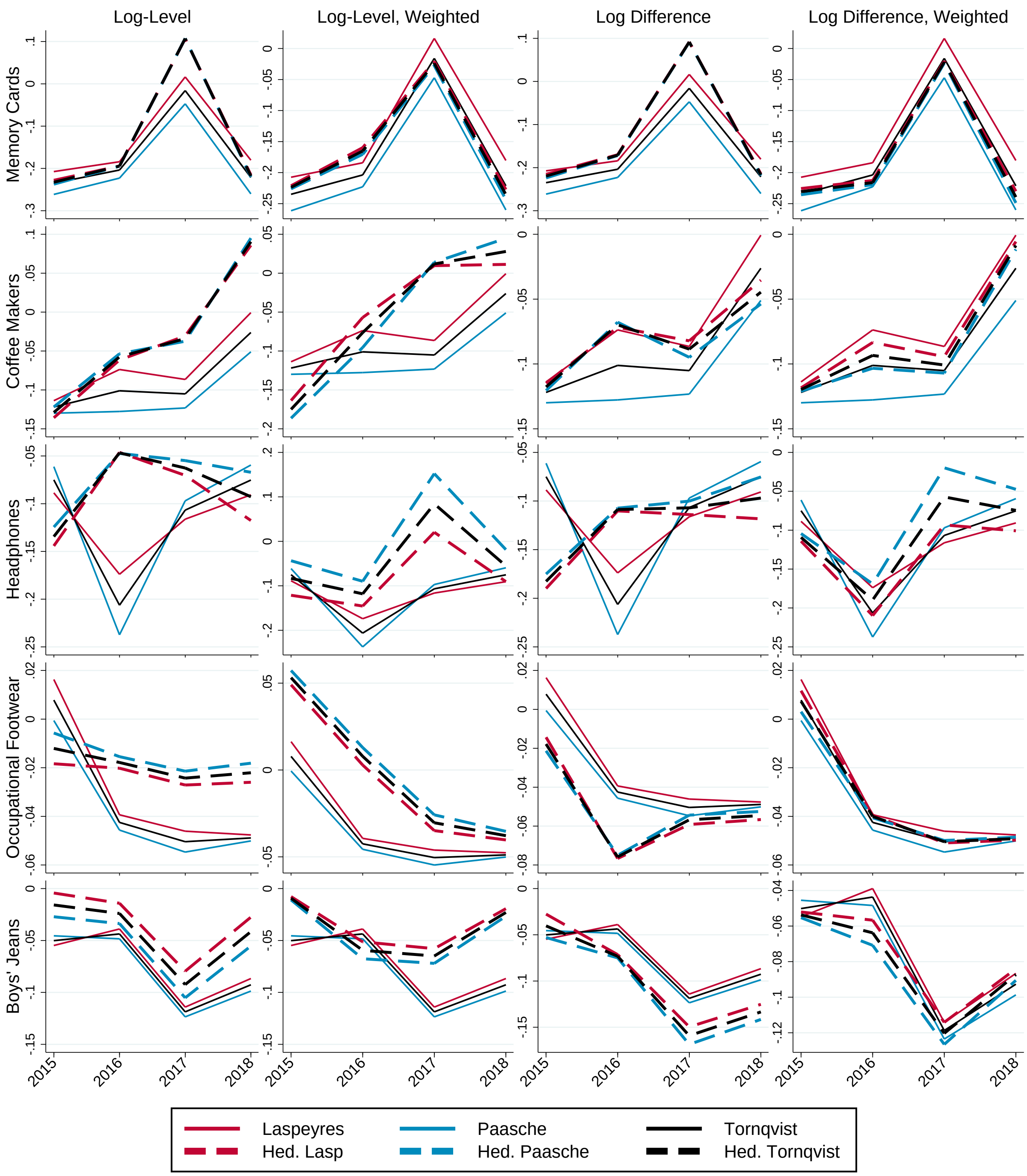
[38]This figure is reproduced from Cafarella et al. (2021, in progress).

Table E.1: Hedonic Models: Goodness of Fit

R^2 for:	Log Price Level		Log Price Relative				
Model:	Log-Level		Log-Level		EP1		EP2
Coffee Makers	0.69	0.62	0.09	0.05	0.14	0.20	0.22
Headphones	0.20	0.89	0.04	0.24	0.11	0.43	0.47
Memory Cards	0.65	0.71	0.02	0.05	0.03	0.09	0.13
Work/Occ Footwear	0.58	0.73	0.08	0.10	0.21	0.37	0.39
Boy's Jeans	0.34	0.72	0.08	0.22	0.13	0.43	0.47
Weighted:		y		y		y	y

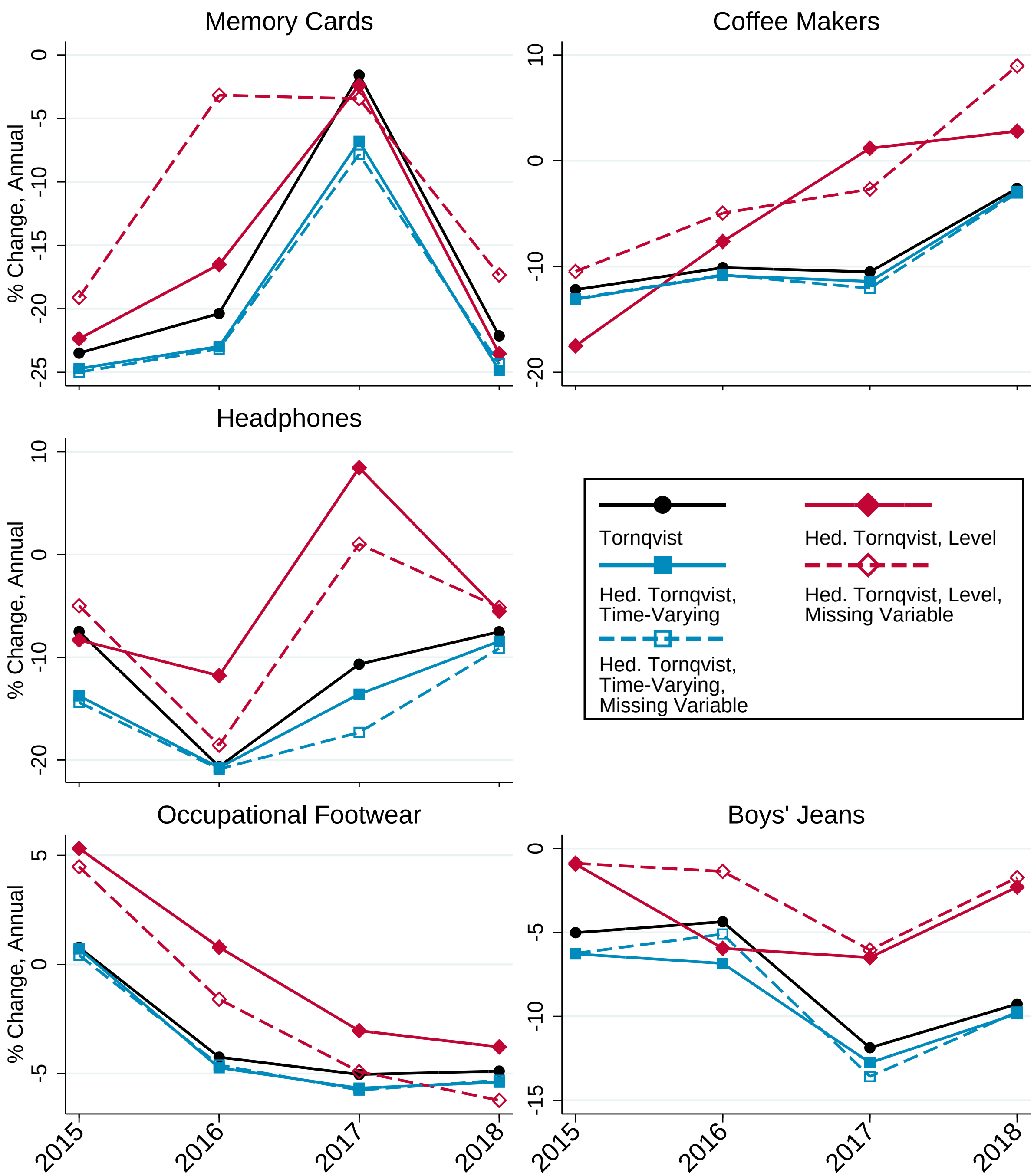
Average quarterly R^2 for hedonic regression models. For cases where the outcome variable (price level or price relative) does not match the LHS variable from the hedonic model, we report the R^2 from a regression of transformed predicted values from the hedonic model on actual values. For example, the price-relative R^2 for the log-level model is the R^2 from a regression of price relatives constructed from a log-level hedonic model on actual price relatives. Weights used in regressions are consistent in hedonic estimation and construction of R^2 measures. For the log-level model, weights are the quantity shares in the current period. For the log-difference and time-varying unobservables model, weights are the average quantity shares in the current and lagged periods. The time-varying unobservable model includes lagged residuals from a log-level hedonic regression.

Figure D.1: Levels, Differences, Weighted and Unweighted for Hedonics, NPD



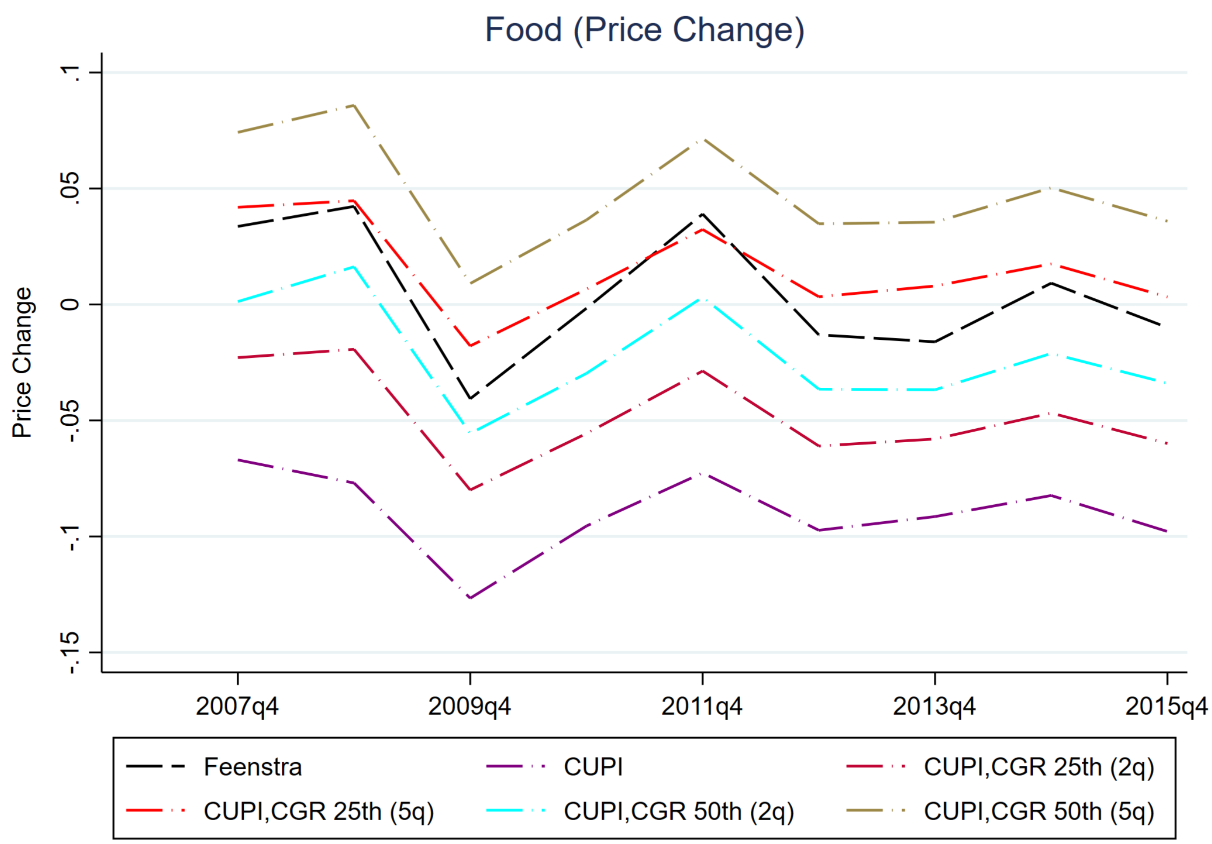
Notes: Values are annual changes from cumulative chained quarterly indices. Data comes from the NPD Group.

Figure D.2: Hedonic Specifications: Test of Time-Varying Unobservable Specification, with Levels



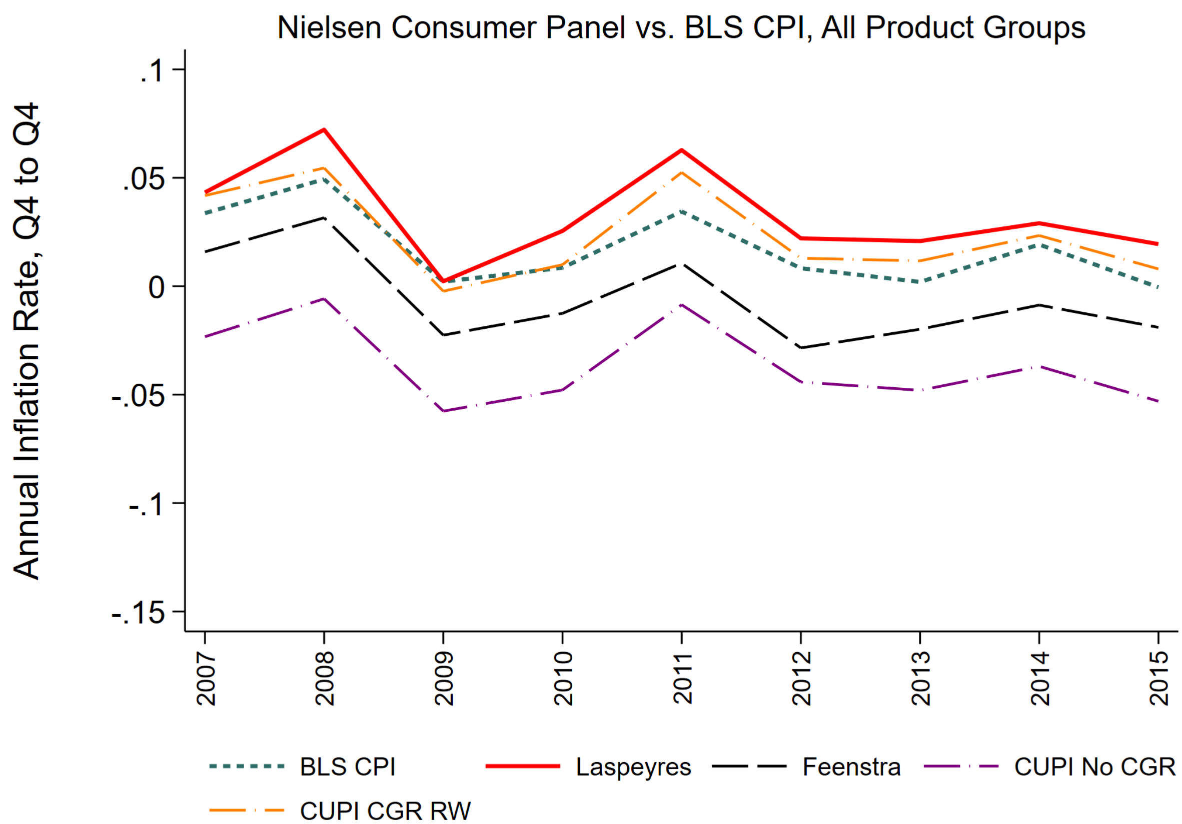
Notes: Values are log differences on an annual basis, aggregated from chained quarterly indices. Data comes from the NPD Group.

Figure D.3: Common Goods Rules – 2 quarter vs 5 quarter horizons



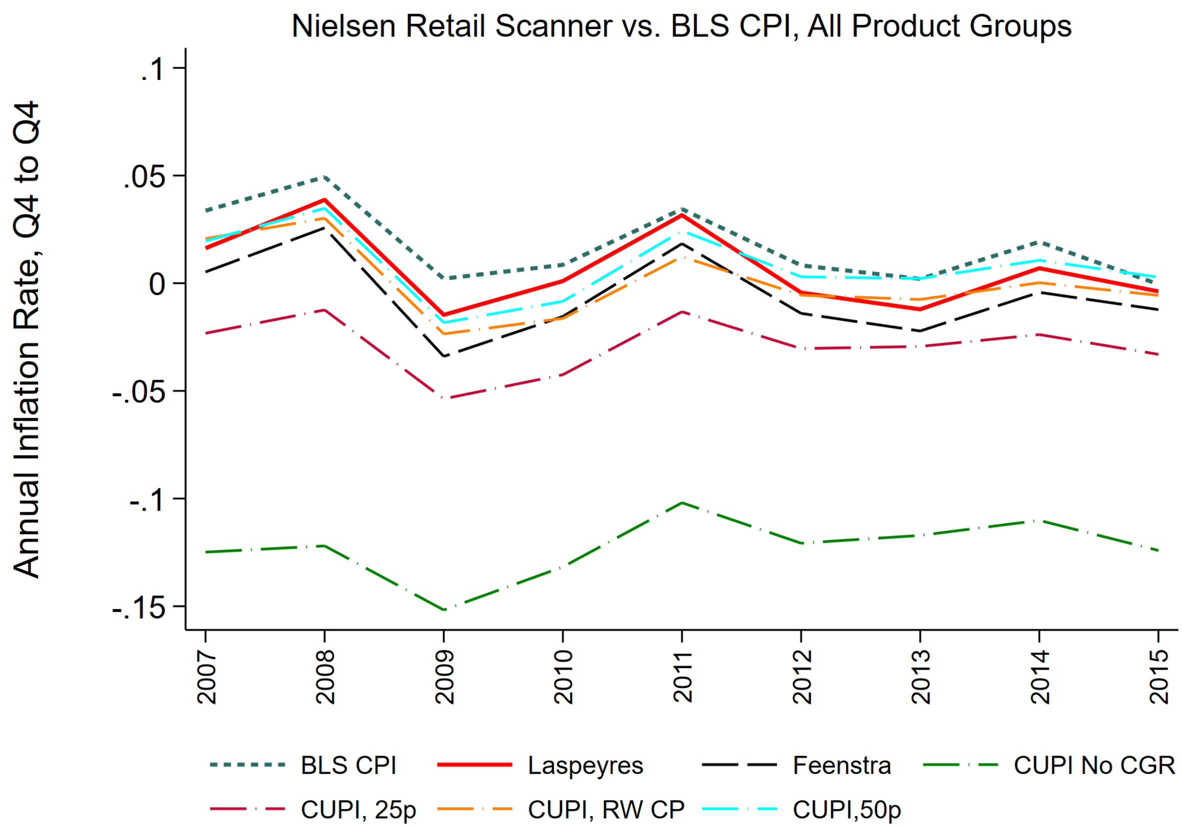
Notes: Figure uses Nielsen Scanner data for food. The $2q$ CUPI computes CGR percentile thresholds using sales pooled over a two quarter horizon (t and $t-1$). The $5q$ CUPI computes CGR percentile thresholds using sales pooled over a 5 quarter horizon (current and prior 4 quarters).

Figure D.4: Common Goods Rules – Nielsen Consumer Panel



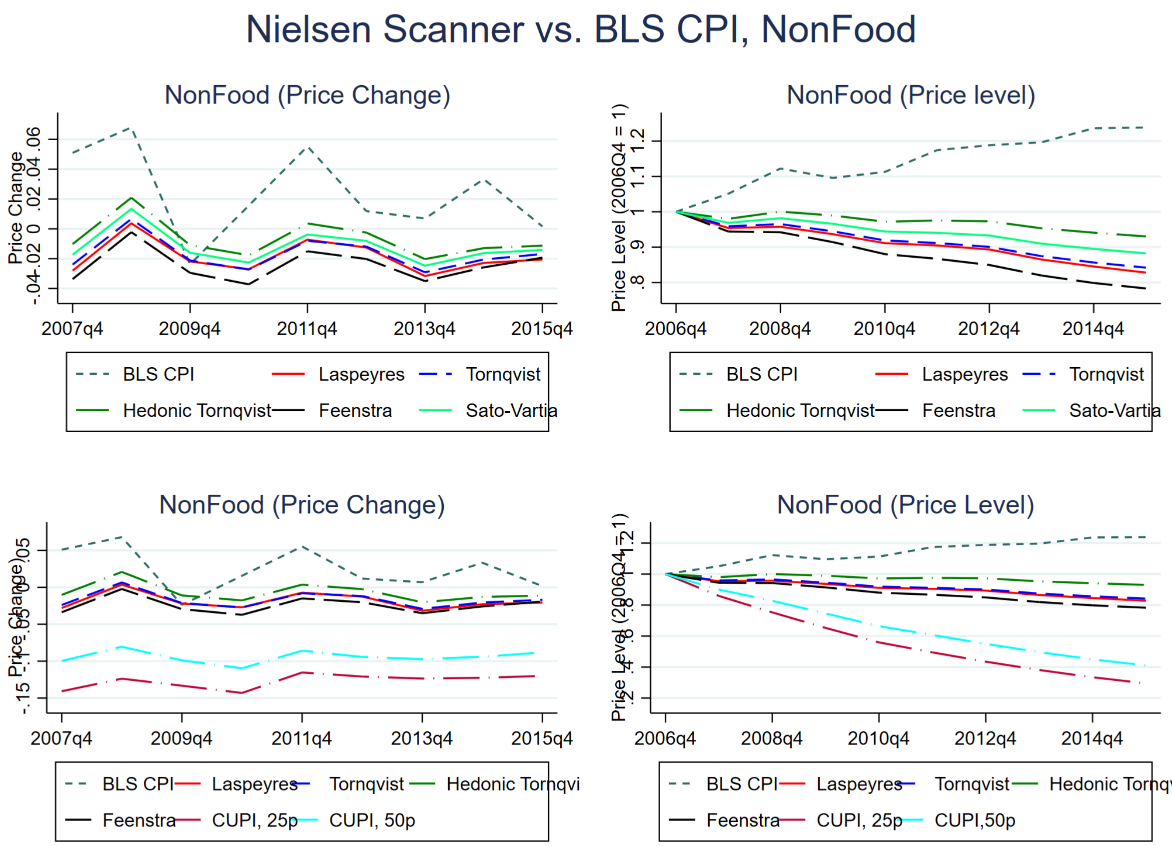
Notes: Figure uses Nielsen Consumer Panel data for food and nonfood product groups. The series "CUPI CGR RW" uses a 5th-percentile sales cutoff for the common goods rule. Percentile computed from sales pooled over 5 quarter horizon (current and prior 4 quarters).

Figure D.5: Common Goods Rules – Nielsen Scanner Panel



Notes: Figure uses Nielsen Retail Scanner data for food and nonfood product groups. The "CUPI, 25p" and "CUPI, 50p" series use 25th- and 50th-percentile cutoffs for the common goods rule, respectively. The series "CUPI, RW CP" uses the CGR 5th percentile threshold from the consumer Panel data for the common goods rule. Percentiles based on sales pooled over 5 quarter horizon (current and prior 4 quarters).

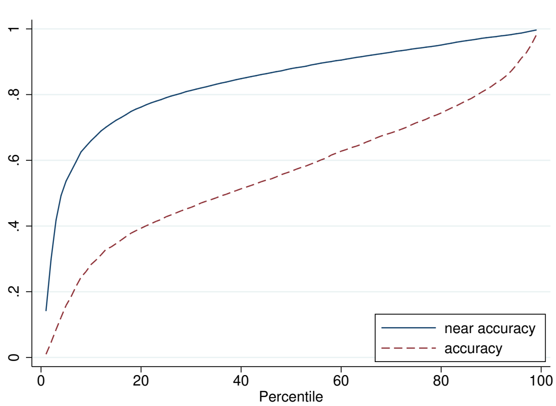
Figure D.6: Main Price Index Specifications: Price Changes and Levels, Nielsen Nonfood



Notes: Price changes are are annual changes from cumulative chained quarterly indices. Price levels reflect cumulative changes relative to the 2006 price level, with 2006 price level set to 1. Figure uses Nielsen Retail Scanner data for nonfood product groups.

Figure D.7: Out-of-sample performance of the deep neural network model

(A) Log Level



(B) Log First Difference