Observational price variation in scanner data cannot reproduce experimental price elasticities*

Robert L. Bray† Robert Evan Sanders‡ Ioannis Stamatopoulos§

July 19, 2024

Abstract

We present experimental evidence that the observational price variation in typical supermarket scanner data is insufficient to recover true price elasticities. We analyze a large, brick-and-mortar field experiment that generated 389,890 random, in-store prices over 35 weeks, across 409 products, at 82 "test" stores of a Midwestern grocery retailer. We compare the demand elasticity estimates derived from these experimental price changes against those derived from observational price changes at 34 "control" stores. During the experiment, the average experimental elasticity is -0.34, whereas the average observational elasticity is about -2.0. The gap is even wider in the difference in differences, and replicates at the category and the product levels. We cannot reconcile this gap by controlling for promotions, conducting an event study around each price change, focusing on base-price changes, accounting for longer-term price effects, or instrumenting with the chain price, the lagged price, the raw input price, or the wholesale price. Our findings suggest that most meaningful variation in grocery prices is tainted by endogeneity.

Keywords: demand estimation, endogeneity, instrumental variables, pricing, promotions, field experiments

JEL Classification: C93, L11, M31, C52, L81

*We are grateful to Eversight for their partnership in this project. We are thankful for comments from Dan Ackerberg, Judith Chevalier, Günter Hitsch, Guido Imbens, Emi Nakamura, Peter Rossi, Jón Steinsson, Steve Tadelis, and Miguel Villas-Boas. We thank the participants of the Columbia Graduate School of Business, Stanford Graduate School of Business, the UC Berkeley, Haas School of Business, and the University of Virginia Darden School of Business. We also thank the participants in the Quantitative Marketing and Economics Conference (2023).

†Northwestern University, r-bray@kellogg.northwestern.edu

‡University of California, San Diego, r1sanders@ucsd.edu

§The University of Texas at Austin, yannis.stamos@mccombs.utexas.edu

1 Introduction

"Little can be said about the functioning of a market without a quantitative assessment of demand," write Berry and Haile (2021). Indeed, as the authors explain, addressing almost any counterfactual question about a market—e.g., assessing the impacts on social welfare of a tax, a subsidy, a tariff, a merger, or the introduction of a new good—requires an estimate of the price elasticity of demand. That is, a quantification of how demand responds to price changes, holding everything else constant.

Unfortunately, estimating price elasticities is also challenging. The primary challenge is that "clean" price variation is scarce in observational data because price changes reflect economic agents' beliefs about demand and supply. For example, suppose grocery-store managers systematically discount products during seasonal demand peaks, because they know demand becomes more elastic at these times (see, e.g., Butters et al. 2020). Then a naïve demand analysis that disregards the store managers' price-setting habits would likely overestimate a discount's effectiveness at stimulating demand—i.e., would over-estimate the price elasticity of demand.

To address this challenge, econometricians have built a rich arsenal of demand-estimation techniques with observational data. Most of these techniques use either (i) control variables that absorb simultaneous shocks to prices and demands, such as changes in advertising or promotional intensity (see, e.g., Rossi 2014), or (ii) instrumental variables that isolate shocks to prices with no direct effect on demands, such as changes in wholesale prices or changes in contemporaneous prices of the same product in other markets (see, e.g., Hausman 1996 and Nevo 2001). These techniques deliver elasticity estimates consistent with economic theory and are being ever refined to handle more variables and complex functional forms (see, e.g., Semenova and Chernozhukov 2021).

But do these observational techniques recover "true" price elasticities? Put differently, do these techniques successfully "reproduce the results of experiments that use random [price] assignment without incurring their costs" (LaLonde 1986)? We address this fundamental question in the context of grocery retailing. Grocery retailing is a fitting context for our question: it is a $717 billion industry that provides high-quality and readily available scanner data.[1]

We partnered with two companies—a Midwestern grocery retailer and a Silicon Valley pricing-solutions company—to conduct a large-scale, brick-and-mortar price experiment. The experiment involved 82 test stores and 34 control stores. At the control stores, we have "observational" prices from May 2, 2018, to March 10, 2020, and at the test stores, we have "observational" prices from May 2, 2018, to July 9, 2019, and experimental prices from July 10, 2019, to March 10, 2020. The pricing solutions company randomized the test-store prices at the store–product–week level, generating a store–product's price in the current week by multiplying the previous week's price with a random scalar (which could be 1 to allow random prices to persist). We observe 389,890 such random prices for 409 products.

We compare the test-store elasticities with the control-store elasticities before and after the experimental intervention. In the pre-experimental period, the test-store elasticities are systematically *more* negative than the control-store elasticities: the median ratio (across products) of a product's control-store estimate to the corresponding test-store estimate is 0.81 in the pre-period. However, in the post-experimental period, the test-store elasticities are systematically

less negative than the control-store elasticities: the analogous median ratio is 3.94 in the post-period. The means tell a similar story: from the pre- to the post-experimental period, the average pre-experimental elasticity estimate changes from -1.63 to -1.97 at the control stores, and from -2.05 to -0.34 at the test stores. This difference-in-differences highlights the effect of switching from observational to experimental prices, as we estimate test-store elasticities off of observational variation in the pre-period and off of experimental variation post-period. Further, note that the three observational means are less than -1, in accordance with the Lerner-index rule. In contrast, the experimental mean is larger than -1, in violation of this rule. Accordingly, it appears that retailers set base prices at the *inelastic* portion of the demand curve—a "paradox" that is obscured in the observational sample.

These patterns survive every potential correction we tried. For example, they hold at the category level, as well as the product level; they hold in the cross section, as well as in the difference in differences; they hold with one-week demand response times (short-run), as well as with three-month demand response times (long-run); they hold when using all observational price changes, as well as when using base-price changes only, and so on. Moreover, we cannot attribute our results to differences in statistical properties of observational and experimental price-setting processes, such as the sizes of the price changes, the price change frequencies, the price levels, or the number of coincident price changes. Likewise, our results are robust to our model choice, to controlling for promotion status, and to controlling for prices of substitute products.

We cannot find *any* instrumental variable that convincingly closes the gap between the experimental and non-experimental elasticity estimates.[2] We tried Hausman instruments, in the fashion of DellaVigna and Gentzkow (2019); we tried the orthogonal complement of Hausman instruments, in the fashion of Butters et al. (2022); we tried lagged prices, in the fashion of Villas-Boas and Winer (1999); we tried an event study design, in the fashion of Rossi (2014); we tried the producer price index, in the fashion of Chintagunta et al. (2005); and we tried wholesale prices, in the fashion of Fong et al. (2010). All of these potential corrections decisively fail.

One wonders whether economic theory led the community astray in the pursuit of a valid instrument. All the instrumental variables we tried, besides wholesale prices, yield estimates less than -1 and hence conform with basic pricing theory. However, the true elasticities lie far above this theoretical upper limit. Accordingly, if someone did find a valid instrument, they would likely file-drawer it. Indeed, Rossi calls for a pricing experiment such as ours against which observational IVs can be compared: "My hope is that this paper will spur a reexamination of the use of IV methods and generate a greater interest in [...] experimental methods that produce valid instruments by definition."

2 Related Literature

Our study contributes to a growing effort to evaluate observational methods with field experiments. LaLonde (1986) begin this line of research with a study comparing experimental and observational estimates of a training program's effect on participants' earnings. LaLonde demonstrated that standard observational techniques were unable to reproduce the experimental

results, which spawned a two-decade research program to employ and refine these techniques (Heckman et al. 1997, Heckman et al. 1998, Dehejia and Wahba 1999, Smith and Todd 2001, Dehejia and Wahba 2002, and Smith and Todd 2005). The articles of Eckles and Bakshy (2021), Lewis et al. (2011), Gordon et al. (2019), and Gordon et al. (2022) are the primary LaLonde works of the past few years. Eckles and Bakshy's article demonstrates that observational estimates of peer effects are sound, and the other three demonstrate that observational estimates of advertising effects are *not* sound. Interestingly, none of these latter three articles report a *systematic* bias as we do. Whereas we show that observational estimates are consistently below experimental estimates, they show that observational estimates are sometimes above and sometimes below experimental estimates.

To the best of our knowledge, we are the second to apply the LaLonde treatment to demand estimation. In an unpublished working paper, Fong et al. (2010) compare observational demand elasticities to those derived from an experiment that randomly shifted the prices of 192 products at 18 grocery stores over 17 weeks. Like us, Fong et al. suggest that (i) retailers set prices at the inelastic region of demand (or, more specifically, they fail to rule this out) and that (ii) observational elasticities are larger than experimental elasticities. Their latter claim, however, is unconvincing because their observational and experimental sample compositions differ drastically. Whereas their experimental sample comprises 192 products at 18 stores in 2009, their observational sample comprises 1,969 products at 81 stores between 2003 and 2009. Given that they compare different products at different stores in different years, it would be surprising if their experimental and observational elasticities did *not* differ. Simply put, their research design does permit an interpretable LaLonde comparison.

We also contribute to a literature that seeks to make high-external-validity claims about the nature of observational retail price variation, retail price-setting practices, and their implications on demand estimation. For instance, researchers have long-debated the extent of endogeneity in scanner data (e.g., Villas-Boas and Winer 1999; Chintagunta et al. 2005) and what to do about it (e.g., Rossi 2014).[3] Our findings indicate that observational elasticity estimates are strikingly different from experimental elasticity estimates, and we suspect that residual endogeneity may be the cause. A concerning implication of our findings is that conclusions drawn from studies that rely on observational elasticity estimates may be unfounded, calling into question a large body of literature (e.g., Chevalier et al. 2003, DellaVigna and Gentzkow 2019, Butters et al. 2020; Hitsch et al. 2021; Döpper et al. 2022).

3 Data and Experiment

The pricing experiments took place at a large Midwestern grocery retailer, which has 82 stores in one geographic market and 34 stores in another. The experiment was executed in two waves. In both waves, a subset of the retailer's products received experimental prices in the larger market's stores. We label these experimental prices, products, and stores as "test" and the non-experimental prices, products, and stores as "control." The retailer randomly changed each test product–store's price every few weeks. In the second wave, the retailer held test prices longer, favored more extreme price changes (very high or low), and randomly chose about half of its test stores to halt the experiment and "optimize" prices. Importantly, throughout both waves, all

test products at the control stores received business-as-usual, retailer-decided, "observational" prices. Before providing additional details about the pricing experiments, we first describe our data and sample-construction process.

3.1 Data

We have five datasets:

1. **Store data**: Each observation is uniquely identified by a store ID. For each store, we observe: open date; close date; and physical street address. These data contain 136 unique stores, spanning 2 markets, 92 postal codes, and 5 states.

2. **Product data**: Each observation is uniquely identified by a Universal Product Code (UPC) and a date (product attributes can change). For each UPC–date, we observe: UPC name (e.g., "Johnsonvl fam pk links"); UPC description (e.g., "Family pack sausage links"); category name (e.g., meat, seafood, or grocery); sub-category name (e.g., smoked and cured fish, or yogurt); and "line-group" ID, which tracks UPCs that are required to share the same price (e.g., various flavors of Chobani Yogurt 5.3 oz). Henceforth, we refer to line-groups as "products" and to individual UPCs as "variants." (We further discuss this distinction in Appendix B.1.) This database spans June 25, 2019, to September 12, 2021, and contains 171,413 unique variants from 19 departments, 653 categories, and 3,623 sub-categories.

3. **Transaction-line data**: Each observation is uniquely identified by a basket ID, a variant, and a transaction-line type (regular purchase, return, employee discount, or coupon). For each transaction–variant–line-type we observe: date-time (the day, hour, and minute); store; quantity; and revenue. We derive prices from the revenues and quantities (for details, see Appendix B.1). This database contains the universe of the retailer's 260,174,529 transactions between May 2, 2018, and August 22, 2021.

4. **Promotion and wholesale-price data**: Each promotion observation is uniquely identified by a variant, a store, and a promotion ID. For each variant–store–promotion, we observe: the promotion start and end date; the promoted price; and the promotion type (e.g., "Print Advertising Sale," "Digital Advertising Sale," "Temporary price promotion—TPR"). This database spans June 25, 2019, to August 22, 2021. Before June 25, 2019, we impute promotion status using the duration and depth of observed price discounts (see Appendix B.2 for details).

   Each wholesale price is uniquely identified by a variant, a store, and a date. The wholesale-price data span July 1, 2019, until September 12, 2021. These data require a fair amount of cleaning before they can be used to construct cost-shifter instruments (see in Appendix B.3 for details).

5. **Experimental-price data**: Each observation is uniquely identified by a product, store, and start date. For each product–store–start date, we observe: the experimental price level and the end date. This database spans July 10, 2019, to March 10, 2020. The second wave

of the experiment, during which the retailer also began experimenting with "optimized" prices at some stores, started on September 25, 2019, and ended on March 10, 2020.

In summary, we observe quantities and prices for every variant sold at every store between May 2, 2018 and August 22, 2021, promotions and wholesale prices after July 1, 2019, and all assigned experimental prices for all test product–stores. Further, to evaluate an additional, commonly used cost-shifter instrument, we also obtained historical Producer Price Indices (PPIs) for each of our categories from the Bureau of Labor Statistics (described in Appendix B.3).

To derive our core sample from these raw data, we: (i) limit the sample to the nine categories that received price experiments (butter spreads, candy, shredded cheese, sliced cheese, other cheese, snacking cheese, facial tissue, pasta, pasta sauce), (ii) aggregate the transaction-level data to the variant–store–date level, (iii) impute missing prices and promotion statuses (forward and backward, dropping observations where imputed values disagree), and (iv) exclude observations after December 31, 2019 (unless otherwise specified). We drop the 2020 data to exclude the COVID-19 pandemic from our sample. More detailed steps are provided in Appendix B.1.

The unit of observation in our core sample is a variant–store–date. After applying the filters above, our sample comprises 42,352,031 observations; of which 28,388,810 involve test products; 20,148,823 involve test products and test stores; and 2,263,428 involve test products, test stores, and test prices. This core sample contains all 1,314 products in the 9 test categories, of which 409 received test prices. These experimental products represent the lion's share of sales, accounting for $117.5 million out of the $147.3 million in revenue associated with the test categories. The sample contains 116 stores, of which 82 receive test prices, and 100,474 product–stores, of which 26,213 receive test prices. These test products at test stores received 389,890 random, in-store prices over the 35 weeks of experiments. Henceforth, all statistics we report refer to this core sample unless otherwise specified.

The price-update processes were nearly identical at the test and control stores before the experiment (Table 1). For example, the median variant–store in both control and test stores held a single base price in the entire pre-experimental period and, taking TPRs into account, saw a median price duration of 14 days. Moreover, prices at both control and test stores were assigned uniformly across the retailer. After the experiment, the test stores' price-updating process deviated from the status quo: the median variant—store's base-price duration dropped to about a week, and prices became less spatially uniform. (For a breakdown by category, see Table A.)

— Insert Table 1 about here —

3.2 Price Experiments

The price experiments ran for 35 weeks, 11 weeks for the first wave, and 24 weeks for the second wave. During these two waves, 409 test products received test prices across 82 test stores. Test prices were reviewed weekly for each test product–store, whereupon the retailer either posted a new experimental price or randomly kept the current price.[4] With the exception of test products at test stores during the experiment, all other prices were business-as-usual, "observational" prices, which were independent of test-price updates and outcomes.

Notably, test products and test stores were not chosen at random: test products had higher revenue, and test stores came from the larger of the retailer's two markets. Instead, price randomization occurred at the product–store–week-year level. (In Section 4, we discuss how this experimental design informs our empirical strategy.)

First Wave. In the first wave, test prices were reviewed every Wednesday, for 11 consecutive weeks. Candy, shredded cheese, other cheese, pasta, and pasta sauce received their first test prices on July 10, 2019; butter spreads, sliced cheese, snacking cheese, and facial tissue received theirs one week later.

We do not have access to the pricing-solutions company's proprietary code for generating test prices. However, the company explained its pricing algorithm to us at a high level, and we verified its key statistical properties with the data. Approximately, a product–store's first test price was created by multiplying its last observational base price by a random number, drawn roughly uniformly from $[0.8, 0.98] \cup [1.02, 1.2]$. From the second test price onward, the next test price was set by drawing a multiplier from a similar nearly uniform distribution, but with a point mass at 1, which meant that there was only a 64.0% chance of a price change in a given week. This data-generating process held throughout the first wave, with two caveats: First, the test-price multipliers were slightly negatively autocorrelated within product–store (-0.2), so prices would not become too extreme over time.[5] Second, toward the end of the first wave (around weeks 7 and 8) and continuing into the second wave, the distribution of multipliers became more concentrated at the 0.8, 1, and 1.2 modes, so that price changes became more extreme and prices were held for longer.

To verify the exogeneity of the price multipliers, we conduct three checks. First, we confirm that the random price multipliers are roughly uniformly distributed, as described above (Figure 1). Second, we confirm that the multipliers used in the first week are uncorrelated with the historical revenues and price-elasticity estimates (Table B). Third, we confirm that the multipliers used after the first week are uncorrelated with the previous week's revenues (Table C).

— Insert Figure 1 about here —

Second Wave. In the second wave, the retailer continued with the first wave's data-generating process, but with one crucial difference: it introduced price optimization. Specifically, in the first week of the second wave, the retailer randomly relabeled 50 out of 82 experimental stores from "test" to "optimal" by category, and in each subsequent week, it randomly swapped the labels of three "optimal" and three "test" stores, also by category. The retailer then treated the "test" store–categories as before, but assigned the "optimal" store–categories prices that the pricing solutions company believed would maximize profits. This process started on September 25, 2019, and lasted until March 10, 2020, when the World Health Organization declared COVID-19 a global pandemic, and the retailer's prices were frozen in place because of emergency anti-price gouging laws.[6]

We drop the "optimal" prices, because they are not random, and we do not know how they are generated. However, before dropping these observations, we verify that both the original labeling and the label-swapping were correctly randomized (Tables D, E, and Table F).

Moreover, in the second wave (as at the end of the first wave), experimental prices continued to be randomly held for longer, and they were more skewed toward high and low values. Specifically, in the second wave, the probability of changing a test price in a given pricing decision (calculated over all product–stores and weeks) was 38.5%. Similarly, the probability a store–product's price was the maximum or minimum price (across all test stores) was 70.4%, whereas it was just 23.5% in the first wave. This difference between the first and second waves is important for our analyses. In particular, the first wave enables us to recover "short-run" experimental elasticities (from experimental prices that are held for weeks), and the second wave allows us to recover "medium-run" experimental elasticities (from experimental prices held for months).

Observational Verses Experimental Variation. Experimental price variation differs from observational price variation (Table 1 and Figure B). Before the experiment, prices were sticky. The median price duration was 0.5 months, and the median base-price duration was 7.1 months (these numbers are roughly in line with similar statistics in Nakamura and Steinsson 2008). Moreover, 93% of price changes came from temporary price promotions (TPRs). In contrast, the median base-price duration was 7 days during the first wave of the experiment and 14 days during the second wave. And only 36% of price changes were caused by TPRs. Finally, prices became more dispersed across stores after the experiment. We measure price dispersion with the fraction of stores that set a product's price to the modal price. In the pre-experimental period, the median fraction is 100% (i.e., more half of all product-dates have only one price across stores), whereas in the post-experimental period this median is only 25%. We explore the implications of these price process differences in Section 5.4.

4 Empirical Strategy

We aim to compare elasticities recovered from observational price variation to those recovered from experimental price variation. Because the test market is qualitatively different from the control market, we employ a "difference-in-differences" approach, comparing the changes in the test-store elasticities, before and after the experiment, to the corresponding changes in the control-store elasticities. To confirm the parallel trends that our difference-in-difference approach relies on, we partition the sample into a collection of time windows, nearly all of which are 77 days long, the duration of the first experimental phase. Partitioning time into 77-day windows will give us a sense of how the test and control elasticity estimates fluctuate over the primary experimental period.[7]

Following DellaVigna and Gentzkow (2019) and Butters et al. (2020), we estimate demand elasticities with the following specification, applied at the product–market–time-window-grouping level:

$$\log(q_{jst} + 0.1) = \alpha_{jgw} + \eta_{jgw} \log p_{jst} + \epsilon_{jst}, \tag{1}$$

where j is the product; s is the store; t is the date; $g = g(s)$ is the market, either "treated" or "control"; $w = w(s)$ is the time-window grouping, either the union of the pre-experimental time windows, the union of the post-experimental time windows, or an individual time window;

q_{jst} is the quantity sold (which we increase by 0.1 to ensure positivity); p_{jst} is the retail price; ϵ_{jst} is a mean-zero residual; α_{jgw} is a fixed effect coefficient; and η_{jgw} captures the average price elasticity of demand, weighted across stores and dates, for product j in market g during time window w, our primary coefficient of interest.

For our primary analyses, we estimate specification (1) at the jgw level in four ways. First, we estimate it with ordinary least squares (OLS). Second, we estimate it with OLS and temporary-price-reduction (TPR) dummies, which absorb changes in demand scale that coincide with changes in TPR status. Third, we estimate it with two-stage least squares (2SLS), with the following first-stage specification:

$$\log p_{jst} = \gamma_{jgw} + \beta_{jgw} z_{jst} + \nu_{jst}, \tag{2}$$

where z_{jst} is a price instrument, ν_{jst} is a mean-zero residual, and γ_{jgw} and β_{jgw} coefficients to estimate. Fourth, and finally, we estimate (1) with both the 2SLS specification above and with TPR dummies included in the first- and second-stage regressions.

We use several instruments across our various specifications, but the most important one is the experimental instrument, $\log \hat{p}_{jst} - \log p_{js0}$, where $\hat{p}_{jst}$ is the experimental price recommendation on day t and p_{js0} is the product's price the day before the experiment began. Recall that the experimental price recommendation, $\hat{p}_{jst}$, equals the last endogenous price, p_{js0}, multiplied by a series of random price shocks. Hence, $\hat{p}_{jst}$ is not fully exogenous because it depends on p_{js0}, but the ratio $\hat{p}_{jst}/p_{js0}$ is fully exogenous, as it responds only to the experimental shocks. Consequently, $\log \hat{p}_{jst} - \log p_{js0}$ represents the exogenous portion of $\log \hat{p}_{jst}$.

5 Findings

5.1 Primary Results

We now present our demand elasticity estimates—or, rather, the average of these estimates over products. Note, there are two ways to pool across products: (i) run aggregated regressions (i.e., estimate (1) at the gw-level) or (ii) run disaggregated regressions and then calculate the mean elasticity estimates over products (i.e., estimate (1) at the jgw-level and then report the average over j). Although we get similar results with both approaches, we report our estimates with the second approach only, as its outputs are more interpretable, being simple averages with equal weight given to all products. In contrast, the former approach gives more weight to products with more price variation, which complicates matters, as price elasticity could correlate with price variability.

First, we present our elasticity estimates, averaged over all products (Figure 2, "Aggregated"). We begin by pooling across time, setting time window group w either to the union of the pre-experimental time windows or to the union of the post-experimental, pre-COVID time windows. The observational elasticity estimates are significantly more negative than the experimental elasticity estimates: the average pre-experimental elasticity estimate (and standard error) is -1.63 (0.02) for the control stores and -2.05 (0.02) for the test stores, and the average post-experimental elasticity estimate is -1.97 (0.03) for the control stores and -0.34 (0.04) for the test stores. The first three of these estimates stem from OLS regressions implemented with observational data, whereas the last stems from 2SLS implemented with the experimen-

tal instrument. The difference-in-differences suggests an observational-estimate bias of (-2.05 − (-1.63)) − (-0.34 − (-1.97)) = -2.05, with corrresponding standard error 0.06. Adding TPR dummy variables decreases this difference-in-differences to (-1.17 − (-1.03)) − (-0.32 − (-1.08)) = -0.89, with corresponding standard error 0.08. But even in this case, the bias is large enough to make it appear that the retailer prices at the elastic part of the demand curve, when in fact, it prices at the *inelastic* part.[8]

— Insert Figure 2 about here —

To demonstrate that the jump in the test store elasticity estimates represents a structural break, we decompose them by time window (Figure 2, "Disaggregated"). The analysis reveals a deviation from the historical trend at the start of the experiment: before the experiment, the test-store estimates track, with parallel trends, slightly below the control-store estimates; but when the experiment starts, they jump well above the control-store estimates, and remain there for the experiment's entire duration. Without TPR dummies, the average (across products) test-control estimate differences are

$$-0.58, -0.87, -0.51, -0.45, -0.28, \text{ and } -0.41$$

in the six pre-experimental time windows, and are

$$2.09, 1.64, \text{ and } 1.59$$

in the three post-experimental time windows. By design, each estimate corresponds to nearly the same amount of data, as all time windows span about eleven weeks.

The aggregate result also replicates for all categories (Figure 3). Specifically, without TPR dummies, the pre-experimental average difference between test and control estimates across the butter-spreads, front-end-candy, shredded-cheese, sliced-cheese, snacking-cheese, other-cheese, facial-tissues, pasta, and pasta-sauce products are

$$-0.60, -0.29, -0.46, -0.40, -0.39, -0.23, -0.87, -0.18, \text{ and } -0.69,$$

and the post-experimental means are

$$1.23, 0.90, 1.77, 2.18, 2.18, 1.26, 0.25, 1.63, \text{ and } 2.30.$$

These patterns stand with TPR dummies, except all elasticity estimates are smaller in magnitude.

— Insert Figure 3 about here —

Further, the same pattern persists when we disaggregate to the product level (Figure 4). In particular, the percentage of products for which the test-store elasticity estimate is significantly more negative, at the $p < 0.05$ level, than the corresponding control-store elasticity estimate (across both TPR-dummy and TPR-dummy-free specifications) is

$$21\%, 27\%, 26\%, 19\%, 15\%, \text{ and } 22\%$$

in the six pre-experimental time windows, and is only

$$9\%, 9\%, \text{ and } 9\%$$

in the three post-experimental time windows. Conversely, the percentage of products for which the test-store elasticity estimate is significantly more positive than the corresponding control-store elasticity estimate is only

$$8\%, 7\%, 6\%, 6\%, 11\%$$

before the experiment, whereas it is

$$47\%, 37\%, \text{ and } 41\%$$

during the experiment. The fraction of control-store elasticities that are more negative than the corresponding test-store elasticities increases from 0.31 to 0.84 without TPR dummy controls, and increases from 0.48 to 0.63 with TPR dummy controls.

— Insert Figure 4 about here —

In the following sections, we present our attempts to reconcile the discrepancy between our experimental and non-experimental elasticity estimates. In summary, no matter how we slice our data or which instruments we use, we cannot reliably replicate our experimental elasticities with non-experimental data—we cannot find any purely observational method to estimate the causal effect of changing prices on demands.

5.2 Zero Sales

In our main specification, line (1), we add 0.1 to sales before we log them. This arbitrary correction is not ideal (e.g., see Cohn et al. 2022 and Chen and Roth 2024), and we now examine whether it meaningfully impacts our findings (Figure 5). First, we decrease the sales-offset term from 0.1 to 0.01 ("Smaller Sales Offset"). Second, we decrease the relative influence of this sales-offset term by aggregating our daily data to weekly data, which decreases the fraction of zero observations ("Weekly Aggregation"). Third, we replace our primary specification with the standard multiplicative-error-term count model (see Cameron and Trivedi 2013, p. 399), which corresponds to moment conditions $\mathrm{E}\left((q_{jst}/\exp(\alpha_{jgw}+\eta_{jgw}\log p_{jst})-1)z_{jst}\right)=0$ ("Count Model"). Since these moment conditions do not depend on the logarithm of q_{jst}, this specification can seamlessly accommodate the zeros in the sales series. These three alternative specifications replicate our primary finding.

We next address a second issue raised by zero-sales days: we observe a product's price only when it sells, which means we do not become aware of a price change until the first transaction under the new price. To remove observations with ambiguous prices, we exclude from our primary sample strings of zero-sales that start and end with different prices (see Appendix B.1 for more detail). This correction eliminates measurement error but introduces a potential selection bias, as our sample includes a day with a price change only if it has at least one sale. More specifically, the correction skews our sample by compelling each price series's first and last observation to have non-zero sales. To eliminate this skew, we drop each price series' first and last observations—i.e., those that must have non-zero sales (Figure 5, "Selection Bias Correction"). This specification mitigates the potential selection bias, ensuring that an observation's sales quantity does not influence whether it is included in our sample. The corresponding estimates replicate our primary result.

— Insert Figure 5 about here —

5.3 TPR Effects

As Hendel and Nevo (2006) explain, static-demand elasticity estimates will be exaggerated if some consumers stockpile during TPRs and (consequently) purchase less after the TPRs. To

check our results' sensitivity to these potential post-TPR dips, we run a specification with eight additional dummy variables, flagging TPRs in each of the preceding eight weeks (Figure 6, "Inventory Stockpiling"). This alternative specification also replicates our main finding. (Other researchers have used similar approaches to account for stockpiling, e.g., Hendel and Nevo 2003, Butters et al. 2020, and Strulov-Shlain 2019.)

— Insert Figure 6 about here —

Further, as Varian (1980) points out, TPRs serve as a mechanism for retailers to share "informed" consumers. And if this is the case, then observational elasticities could describe a weighted average of *two* demand curves—one characterizing "uninformed" consumers who buy at full price, and one characterizing "informed" consumers who buy on promotion. To disentangle this potential blend, we recreate our estimates with separate TPR-free and TPR-only non-experimental subsamples (Figure 6, "Base Price" and "Base Price Inverse"). We then estimate elasticities with the variation in each subsample, gleaning the "uninformed" customers' price sensitivity from base-price fluctuations and the "informed" customers' price sensitivity from discount-magnitude fluctuations. The results suggest that the observational-elasticity bias holds across both "informed" and "uninformed" customers: it is -0.85 (0.14) in the "Base Price" subsample and -1.13 (0.09) in the "Base Price Inverse" subsample. In other words, even if there are two relevant demand curves, the observational elasticities of both curves are more negative than the corresponding experimental elasticities.

We further refine the "Base Price" specification to estimate distinct elasticities for price increases and decreases. We do so because price increases are more likely to be cost-driven, owing to a systematic rise in wholesale prices. To isolate the price-increase and price-decrease variation, we cluster each product's data in the "Base Price" sample by price so that the ith cluster comprises the observations that correspond to the $(i-1)$th and ith base prices.[9] Next, we assign a product's ith cluster to the "Base Price Increase" subsample if the ith base price exceeds the $(i-1)$th base price and otherwise assign it to the "Base Price Decrease" subsample. Finally, we rerun our observational-elasticity regressions on these two subsamples with additional price-cluster dummies to wash out the between-cluster price variation. We find an even stronger bias with the base-price decreases but a weaker bias with base-price increases (Figure 6, "Base Price Decrease" and "Base Price Increase"). Unfortunately, estimating elasticities off of base-price increases does not appear to be a robust correction, as the bias reemerges when we restrict the pre- and post-period samples to the products that are present in both (Figure 6, "Base Price Increase, Robust").

5.4 Price-Process Differences

The experimental and observational price processes yield different: price levels (because the experimental prices sometimes lie outside of the typical range); price-change sizes (because the new-price-to-old-price ratio distribution does not resemble a Uniform[0.8, 1.2] in the non-experimental sample); number of coincident price changes (because the experiment updates many test products at once); and price change frequencies (because test prices change somewhat frequently). In theory, such pricing-process differences could generate different elasticity

estimates. For example, lower prices could attract more price-sensitive consumers, resulting in more negative elasticities. Or larger price changes could be more noticeable. Or price drops of competing products could make a given price drop less salient, since changes are more conspicuous against a constant background. Or more frequent price changes could make a current price change less salient.

We test our results' sensitivity to these price-process differences by focusing on different sub-samples of the experimental data (Figure 7). First, we separately estimate elasticities that correspond to lower-than-baseline experimental price levels ("Price Level is Low") and to higher-than-baseline experimental price levels ("Price Level is High"), where we use the last pre-experimental base price as the baseline. Second, we separately estimate elasticities that stem from smaller-than-control experimental price changes ("Price Change is Small") and from larger-than-control experimental price changes ("Price Change is Large"), where we represent control-store price changes at the product level by the median price change size. Third, we separately estimate elasticities that correspond to experimental price changes that are less synchronized than control ("New Prices are Few") and more synchronized than control ("New Prices are Many"), where we measure price synchronization for each price change with the number of other products' price changes in its category–store within a week of its date, and we classify experimental price changes using the median synchronization in its category at control stores. All subsamples reproduce our main findings.

— Insert Figure 7 about here —

Next, we control for the influence of other products' prices in the same product category at the same store. We run two different specifications, one incorporating the quantiles of these prices as regression control variables, which captures the full competing price distribution (similar to Strulov-Shlain 2019), and another incorporating the prices of the five top-selling other products (Figure 7, "Other Price Quantiles" and "Other Price Top Products"). Both alternative specifications replicate our main finding.

Lastly, we examine whether our results stem from the higher frequency of the experiment's price updates. We do so with two checks. In the first check, we limit the experimental sample to the first experimental price of each product at each store (Figure 7, "First Price Only"). The rationale behind this check is that there is nothing unusual about the timing of a product's first experimental price; it is only the subsequent price changes, which follow in rapid secession, that appear amiss. This specification's estimates resemble our primary estimates, which suggests that our results are not a relic of the short time between experimental price changes. In the second check, we add previous prices as controls in our regressions (Figure 8). Specifically, we add the log prices from 1, 2, 3, 4, 5, and 6 weeks prior as additional independent variables, and we do the same for the promotion dummy variable when we control for TPRs. For the experimental estimates, we instrument for the six lagged prices with six analogously lagged experimental instruments. These experimental estimates should capture the causal relationship between the sales on day t and the prices on days t, $t-7$, $t-14$, $t-21$ $t-28$, $t-35$, and $t-42$. The estimates suggest that prior prices have negligible lingering effects.

— Insert Figure 8 about here —

5.5 Demand Response Times

In the previous section, we grappled with one potential problem posed by our frequent test price changes: the effect of one week's experimental prices bleeding into the following week. But the short duration of our test prices also raises another issue: it may not give customers enough time to adjust to the new prices fully —à la the LeChatelier principle (Milgrom and Roberts 1996). Hence, we may be benchmarking short-run experimental elasticities to long-run observational elasticities. We address this issue in several ways.

First, we provide evidence that short-run elasticity estimates are actually *more negative* than long-run elasticity estimates, in which case our frequent experimental price changes could *underestimate* the observational bias. To assess how demand's price sensitivity changes with the response time, we estimate observational base-price elasticities off of the consumer response τ weeks after each base-price change for $\tau \in \{0, \cdots 15\}$ (Figure 9). To do so, we: (i) cluster each store–product's data by price change so that the ith cluster comprises the pre-change observations that correspond to the $(i-1)$th base price and the post-change observations that correspond to the ith base price; (ii) limit the post-change observations to those that follow between τ and $\tau + 1$ weeks after the given price change; (iii) run separate treated-store and control-store elasticity regressions in which we include price-cluster dummy variables (to restrict attention to within-cluster price variation), for each product and τ value; and, finally, (iv) average across these estimates. The estimates increase with τ, which suggests that, if anything, elasticities become less negative with the price incubation period (as Rossi 2014 predicted).

— Insert Figure 9 about here —

Second, we juxtapose short-run, one-week experimental elasticities against analogous observational elasticities. Specifically, we limit our experimental and observational samples to a set of two-week time clusters, which comprise the week leading up to and the week following each price change (Figure 10, "Event Study"). We then run our demand regressions with time-cluster dummy variables to estimate elasticities using the within-period variation. We further replicate this analysis with only time clusters with price increases (Figure 10, "Event Study Increase") and only the time clusters with price decreases (Figure 10, "Event Study Decrease"). All these event study estimates mirror our primary estimates.

— Insert Figure 10 about here —

Third, we attempt to estimate longer-run experimental elasticities by extracting the longer-run variation in our test prices. Specifically, we wash out the short-run variation in our experimental prices by either replacing the experimental instruments with those assigned on the first day of the experiment, or by lagging them by four weeks (Figure 10, "Delayed Response to First Test Price" and "Delayed Response to Lagged Test Price"). These instruments are sound because test prices are highly autocorrelated (recall that the test price in one week equals that in the preceding week, multiplied by a random shock). These alternative specifications yield similar results.

Fourth, we estimate longer-run experimental elasticities by focusing on experimental price paths that are sufficiently stable for sufficiently long stretches (Figure 11). More specifically, we

partition each product–store's experimental price series into distinct *price regimes*—collections of consecutive prices that differ by no more than 5%. Whereas a specific experimental price can last a few weeks, at most, a price regime can last months. And since the price is nearly stable within each regime, we can use the longer price regimes to estimate longer-run price elasticities. To do so, we limit our sample to the price regimes that last at least i weeks, for $i \in \{1, \cdots, 12\}$, and then further limit the sample to the products whose within-regime price variation accounts for no more than 2% of its total price variation. Overall, 80%, 55%, 38%, 28%, 20%, 15%, 11%, 9%, 7%, 4%, 4%, and 3% of our experimental prices belong to the ith sample, for $i \in \{1, \cdots, 12\}$. The estimates of all twelve samples reiterate the customers' insensitivity to experimental price changes, even those that they have *months* to respond to.[10]

— Insert Figure 11 about here —

Finally, recall that we partitioned the sample into a collection of 77-day windows to estimate test and control elasticities with the same sample horizon (Figure 2, "Disaggregated"). Standardizing the sample horizon to 77 days roughly standardizes the demand response time to 77/3 days: if you draw a price innovation time and a measured demand time uniformly from $[0, 77]$, the expected demand response time is $\int_{x=0}^{77} \int_{y=0}^{77} |x - y|/77^2 dydx = 77/3$ days. (This analysis hinges on the experimental price shocks being *cumulative*, so that each price innovation persists to the end of the sample horizon.) Estimating elasticities across 77-day windows does not weaken our results.

5.6 Observational Instruments

Next, we examine whether commonly used instrumental variables reduce the difference between experimental and observational elasticity estimates.

First, we use the Hausman-type price instrument of DellaVigna and Gentzkow (2019) and Butters et al. (2020) (Figure 12, "Hausman Instrument"). This specification isolates chain-level price variation, which might not be driven by local demand shocks. In particular, for each product–store–date, we define the value of the instrument, z_{jst}, as the modal price for product j on date t across all stores that are more than 20 miles away from store s in its market g. Price uniformity makes this instrument quite strong: this specification's median first-stage R^2 value is 0.82 without TPR dummies and is 0.61 with TPR dummies. We also examine the elasticity estimates produced by recasting z_{jst} as a control variable instead of an instrumental variable (Figure 12, "Hausman Inverse"). This specification isolates the orthogonal price variation—i.e., local deviations from the chain price—which Butters et al. (2022) argues often arise from local cost differences, such as those from regulations and taxes. Neither specification meaningfully reduces the experimental-observational discrepancy.

— Insert Figure 12 about here —

Second, we use the one-week-lagged price instrument of Villas-Boas and Winer (1999) (Figure 12, "'Lagged Price Instrument"). This specification washes out price variations that are less than a week old, which may coincide with predictable, transient demand shocks. Price rigidity makes this instrument sufficiently strong: this specification's mean first-stage R^2 value is 0.30

without TPR dummies and is 0.56 with TPR dummies. (These R^2 values are not higher because most price variation comes from TPRs, and most TPRs are short-lived.) We also examine the elasticity estimates produced by recasting last week's price as a control variable instead of an instrumental variable (Figure 12, "Lagged Price Inverse"). This specification isolates the orthogonal price variation—i.e., washes out price variation that is *more* than one week old—simulating an event study around each price change. Neither specification meaningfully reduces the experimental-observational discrepancy.

Third, we examine cost-shifter instruments, specifically producer price indices (PPI) and wholesale prices.[11] These instruments isolate cost-driven price variation, which could represent exogenous supply shocks. We construct the PPI instruments with the corresponding BLS data for our nine categories (e.g., for butter spreads, we use the BLS's "PPI Commodity data for Processed foods and feeds–Butter, not seasonally adjusted"), and our wholesale-price instruments with an additional database that spans from July 1, 2019 to December 31, 2019 (see Appendix B.3 and Table G's caption for additional detail). Unfortunately, both instruments are weak in our sample. For PPI instruments, we regress log base price on PPI, one-month lagged PPI, two-month lagged PPI, and three-month lagged PPI (as suggested by Villas-Boas and Winer 1999). Out of 409 test products, only 99 yield full regression estimates, only 37 yield positive regression estimates, and only 11 yield positive regression estimates and an R^2 greater than 0.3 (Table G). Similarly, for wholesale prices, we regress the log base price on the log wholesale price.[12] Out of 409 test products, only 71 yield full regression estimates, only 57 yield positive regression estimates, and only 35 yield positive regression estimates and an R^2 greater than 0.3 (Table G).

6 Conclusion

There is a sizable, systematic difference between our observational elasticity estimates, which suggest that the retailer prices at the elastic part of its demand curve, and our experimental elasticities, which indicate the retailer prices at the inelastic part. We cannot attribute this difference to properties of our estimators, nor to differences between the observational and experimental price processes, nor to a contrast between short- and long-run elasticities, and we cannot correct this difference with promotion dummies, base-price variation, or commonly used instruments. All told, our experiment suggests that nature has played a devilish trick on us: she has arranged the market so that prices lie in the theoretically impossible demand region, but then she has covered her tracks by introducing a countervailing endogeneity bias that situates naïve demand-elasticity estimates right were theory tells us to expect them. Hence, our results raise two questions: (i) why does the retailer set prices at the inelastic portion of its demand curve, and (ii) why do observational elasticities not reflect this fact?

Regarding pricing at the inelastic part of the demand curve, theory offers two explanations: *cross-selling* and *consumer churn*. For cross-selling, Cournot (1838) and Thomassen et al. (2017) explain that positive price externalities can increase the elasticity at the optimal price above minus one. Simply put, lowering a product's price not only makes the product more attractive but also makes the entire store more attractive, increasing foot traffic down every aisle. Thomassen et al. (2017) empirically support this pricing behavior with a structural model of

category demand. For customer churn, Rotemberg (2002) explains that optimizing for customer lifetime value can also push the local elasticity at the optimal price above -1. Pricing too aggressively risks not only losing the sale but also losing the customer for good—foregoing decades of future sales. Moreover, cross-selling and customer churn exacerbate one another. (Accommodating these two factors will require significant effort, as they currently make demand models intractable.[13])

Regarding the negative bias in the observational demand estimates, theory offers two explanations: *counter-cyclical pricing* and *unobserved store conditions*. For counter-cyclical pricing, Kuksov and Villas-Boas (2008) explain that a negative correlation between demand scale and demand slope could make observational elasticity estimates lie below the truth. Simply put, demand could become more elastic as it grows, which would compel stores to lower prices before demand surges, which in turn would accentuate the apparent effect of the price reduction. Butters et al. (2020) empirically support this explanation with data from multiple retail chains. Although plausible, counter-cyclical pricing cannot fully explain our findings, which persist when focusing locally around base-price changes (unless one believes such price changes are perfectly timed). For unobserved store conditions, Chintagunta et al. (2005) explain that if even some price reductions co-occur with unobservable, demand-boosting actions, then observational elasticities could also be biased downward. For example, if stores sometimes stimulate demand by lowering the price and moving it to the eye-level shelf (see, e.g., Dreze et al. 1994) then observational elasticises will misattribute the sales arising from the conspicuous positioning to the markdown. And because traditional scanner data do not report planograms and other shelf allocation information, such endogeneity would generally be uncorrectable.

Overall, we are satisfied with our answers to the first question, but not with those to the second. Indeed, the standard Lerner-index rule may be a straw man, as it lacks salient aspects of retailing—cross-selling and customer churn primary among them. So it is not too surprising that the data violate this rule—i.e., that the equilibrium demand elasticity is above -1. In fact, this inelastic demand equilibrium is not new, as Hoch et al. (1994), Rotemberg (2002), Bijmolt et al. (2005), Fong et al. (2010), Levitt (2016), and Hitsch et al. (2021) have identified it. Rather, what is unusual is not the presence of inelastic demand but that essentially *all* observational methods fail to detect it. We explain that this universal bias could stem from counter-cyclical pricing or unobserved store conditions, but these explanations are speculative. Thus, we find ourselves in the awkward position of presenting a finding for which we cannot fully account. It is natural to discount results as aberrant as ours, but such a powerful and unambiguous experiment warrants scrutiny. Our results could be anomalous, but if they are not, they suggest that observational demand elasticities grossly differ from the truth, which could have fundamental implications for demand estimation.

Endnotes

1. This industry is highly studied: 14 out of the 85 papers that mention "price elasticity" in the *Journal of Political Economy, Econometrica, American Economic Review, Review of Economic Studies*, or *The Quarterly Journal of Economics* since 2019 use grocery-retail data. See USDA, "Retail Trends" (2023), https://www.ers.usda.gov/topics/food-markets-prices/retailing-wholesaling/retail-trends/ for an estimate of the market size.
2. See Berry and Haile (2021) for a recent discussion such instruments.
3. These debates are motivated, in part, from evidence of seemingly rigid pricing documented in the grocery retail industry (e.g., spatial uniformity by DellaVigna and Gentzkow 2019 and Hitsch et al. 2021, temporal rigidity by Nakamura and Steinsson 2008, and promotion rigidity by Anderson et al. 2017). That is, researchers have argued that if grocery-retail pricing is so inflexible, how much endogeneity can there be when using within-product–season–promotion-status variation? For example, Rossi 2014 argued: "In my view, [endogeneity] concerns are of utmost importance with cross-sectional data where the unobservable could be an unobserved product or market characteristic. With relatively high-frequency time series data (such as weekly data), the notion that there exists some demand shock that varies from week to week (and possibly also from brand to brand) and that this unobservable also drives a non-negligible portion of the variation in marketing mix variables [such as prices] is strained. The evidence for the existence of endogeneity biases in time series or panel data consists entirely of model-based evidence via comparison of the results with and without IVs. There is no direct evidence from the firm side (for example, from pricing experiments) that endogeneity biases are large in panel or times series data. [...] I believe that a strong argument must be made that endogeneity problems are of the first order. I do not see convincing arguments along these lines in our empirical literature."
4. Prices were reviewed approximately weekly. For example, in the second wave, prices of the candy category, which has many products, were reviewed every 3 to 4 weeks.
5. In addition, some products have hard cutoffs for minimum and maximum prices. These constraints are commonplace and present in observational prices, as well. For example, the price of a 1-liter bottle of Coca-Cola cannot be higher per unit than a 16-ounce bottle of Coca-Cola's price.
6. After March 10, 2020, pre-planned promotions still occurred unless the promoted price exceeded the currently posted base price. This quirk could occur, for example, in product–stores that had a very low experimental multiplier (e.g., 0.8) just before COVID-19 hit. Second, although this freezing of experimental prices is an interesting "quasi-experiment within the price experiment," we opt not to use data after March 10, 2020, because of potential changes in consumer behavior that likely coincided with the COVID-19 pandemic.
7. Individual time windows are mostly 77 days long, which is the duration of the first wave of the experiment. Specifically, time window 0 comprises the 77 days of the first experimental phase, time window -1 comprises the 77 days before that, -2 comprises the 77 days before that, and so forth until time window -6, which comprises the 49 days between the start of the sample and the beginning of time window -5. And since the second experimental phase began immediately after the first experimental phase, we define time window 1 as this phase's 2019 dates (98 days) and define time window 2 as its 2020 dates (69 days). Finally, we combine our time windows into a pre-experimental group, which comprises time windows -6 to -1, and a post-experimental group, which comprises time windows 0 and 1 (we exclude time window 2 from the post-experimental group, as it could be influenced by COVID). We will estimate demand elasticities by individual time window and time-window group.
8. We are not the first researchers to find inelastic product-level demand. This result has been bubbling up in papers for a while, but difficulty reconciling it with simple economic models likely leads to file-drawering of such results. Some examples of documented inelastic demand (or, even with observational data, a significant share of products with inelastic demand): Hoch et al. (1994), Rotemberg (2002), Bijmolt et al. (2005), Fong et al. (2010), Levitt (2016), and Hitsch et al. (2021). Interestingly, Hoch et al. (1994) is the most recent brick-and-mortar pricing experiment that is as large as ours, and they find (albeit at the category level) inelastic demand of about -0.33, and elastic observational demand of about -1.06. More significant may be the studies we *do not* see: researchers have divulged to us that they have discarded analyses that yield inelastic demand, as such estimates would be harder to publish.
9. This grouping scheme effectively doubles our sample, but our clustered standard errors account for the duplication.
10. We are not the first to find experimental elasticities are largely invariant to the duration of the posted prices:

Hoch et al. (1994)—who held category-level experimental prices for more than forty weeks—find the same.

11. Note that Hausman instruments, which we tested above, can also be thought of as cost-shifter instruments (Berry and Haile 2021), but we examine them separately because they are less direct and do not require additional data, as the instruments we examine here do.
12. We focus on base-price changes in light of the findings in Anderson et al. (2017), who show promotions do not respond to changes in wholesale prices.
13. Researchers are working on estimating static demand for a large number of products within a store. For example, Ruiz et al. (2017) and Smith and Griffin (2023).

References

Anderson E, Malin BA, Nakamura E, Simester D, Steinsson J (2017) Informational rigidities and the stickiness of temporary sales. *Journal of Monetary Economics* 90:64–83. 18, 19, 43

Berry ST, Haile PA (2021) Foundations of demand estimation. *Handbook of Industrial Organization*, volume 4, 1–62 (Elsevier). 2, 18, 19

Bijmolt TH, Van Heerde HJ, Pieters RG (2005) New empirical generalizations on the determinants of price elasticity. *Journal of marketing research* 42(2):141–156. 17, 18

Butters RA, Sacks DW, Seo B (2020) Why do retail prices fall during seasonal demand peaks? *Forthcoming (RAND Journal of Economics)* (19-21). 2, 4, 8, 12, 15, 17, 44, 45

Butters RA, Sacks DW, Seo B (2022) How do national firms respond to local cost shocks? *American Economic Review* 112(5):1737–1772. 3, 15

Cameron AC, Trivedi PK (2013) *Regression analysis of count data*, volume 53 (Cambridge university press). 11, 27

Chen J, Roth J (2024) Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics* 139(2):891–936. 11

Chevalier JA, Kashyap AK, Rossi PE (2003) Why don't prices rise during periods of peak demand? evidence from scanner data. *American Economic Review* 93(1):15–37. 4

Chintagunta P, Dubé JP, Goh KY (2005) Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models. *Management Science* 51(5):832–849. 3, 4, 17

Cohn JB, Liu Z, Wardlaw MI (2022) Count (and count-like) data in finance. *Journal of Financial Economics* 146(2):529–551. 11

Cournot A (1838) Researches into the mathematical principles of the theory of wealth. *Forerunners of Realizable Values Accounting in Financial Reporting*, 3–13 (Routledge). 16

Dehejia RH, Wahba S (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association* 94(448):1053–1062. 4

Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84(1):151–161. 4

DellaVigna S, Gentzkow M (2019) Uniform pricing in us retail chains. *The Quarterly Journal of Economics* 134(4):2011–2084. 3, 4, 8, 15, 18

Döpper H, MacKay A, Miller N, Stiebale J (2022) Rising markups and the role of consumer preferences. *Harvard Business School Strategy Unit Working Paper* (22-025). 4

Dreze X, Hoch SJ, Purk ME (1994) Shelf management and space elasticity. *Journal of retailing* 70(4):301–326. 17

Eckles D, Bakshy E (2021) Bias and high-dimensional adjustment in observational studies of peer effects. *Journal of the American Statistical Association* 116(534):507–517. 4

Fong NM, Simester DI, Anderson ET (2010) Private label vs. national brand price sensitivity: Evaluating non-experimental identification strategies. *MIT Working Paper*. 3, 4, 17, 18

Gordon BR, Moakler R, Zettelmeyer F (2022) Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *arXiv preprint arXiv:2201.07055* . 4

Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D (2019) A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science* 38(2):193–225. 4

Hausman JA (1996) Valuation of new goods under perfect and imperfect competition. *The economics of new goods*, 207–248 (University of Chicago Press). 2

Heckman JJ, Ichimura H, Smith JA, Todd PE (1998) Characterizing selection bias using experimental data. *Econometrica* . 4

Heckman JJ, Ichimura H, Todd PE (1997) Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies* 64(4):605–654. 4

Hendel I, Nevo A (2003) The post-promotion dip puzzle: What do the data have to say? *Quantitative Marketing and Economics* 1(4):409–424. 12

Hendel I, Nevo A (2006) Measuring the implications of sales and consumer inventory behavior. *Econometrica* 74(6):1637–1673. 11

Hitsch GJ, Hortaçsu A, Lin X (2021) Prices and promotions in us retail markets. *Quantitative Marketing and Economics* 19(3):289–368. 4, 17, 18, 44, 45, 46

Hoch SJ, Dreze X, Purk ME (1994) Edlp, hi-lo, and margin arithmetic. *The Journal of Marketing* 16–27. 17, 18, 19

Kuksov D, Villas-Boas JM (2008) Endogeneity and individual consumer choice. *Journal of Marketing Research* 45(6):702–714. 17

LaLonde RJ (1986) Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* 604–620. 2, 3, 4

Levitt SD (2016) Bagels and donuts for sale: A case study in profit maximization. *Research in Economics* 70(4):518–535. 17, 18

Lewis RA, Rao JM, Reiley DH (2011) Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. *Proceedings of the 20th international conference on World wide web*, 157–166. 4

Milgrom P, Roberts J (1996) The lechatelier principle. *The American Economic Review* 173–179. 14

Nakamura E, Steinsson J (2008) Five facts about prices: A reevaluation of menu cost models. *The Quarterly Journal of Economics* 123(4):1415–1464. 8, 18

Nevo A (2001) Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69(2):307–342. 2

Rossi PE (2014) Even the rich can make themselves poor: A critical examination of iv methods in marketing applications. *Marketing Science* 33(5):655–672. 2, 3, 4, 14, 18

Rotemberg JJ (2002) Customer anger at price increases, time variation in the frequency of price changes and monetary policy. 17, 18

Ruiz FJ, Athey S, Blei DM (2017) Shopper: A probabilistic model of consumer choice with substitutes and complements. *arXiv preprint arXiv:1711.03560* . 19

Semenova V, Chernozhukov V (2021) Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* 24(2):264–289. 2

Smith AN, Griffin JE (2023) Shrinkage priors for high-dimensional demand estimation. *Quantitative Marketing and Economics* 21(1):95–146. 19

Smith JA, Todd PE (2001) Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review* 91(2):112–118. 4

Smith JA, Todd PE (2005) Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics* 125(1-2):305–353. 4

Strulov-Shlain A (2019) More than a penny's worth: Left-digit bias and firm pricing. *Chicago Booth Research Paper* (19-22). 12, 13

Thomassen Ø, Smith H, Seiler S, Schiraldi P (2017) Multi-category competition and market power: a model of supermarket pricing. *American Economic Review* 107(8):2308–2351. 16

Varian HR (1980) A model of sales. *The American economic review* 70(4):651–659. 12

Villas-Boas JM, Winer RS (1999) Endogeneity in brand choice models. *Management science* 45(10):1324–1338. 3, 4, 15, 16, 43

Table 1: Summary statistics: pre-experiment, first wave, and second wave (433, 77, and 97 days).

	Control stores			Test stores		
	Pre-experiment	First wave	Second wave	Pre-experiment	First wave	Second wave
Observations						
products	394	399	401	393	401	397
product–stores	12,010	10,494	10,552	29,980	28,991	25,933
product–store–dates	3,399,261	558,106	690,793	8,097,009	1,354,902	1,017,478
variant–store–dates	6,023,223	983,163	1,233,601	14,404,139	2,514,876	1,833,083
Median (across variant–stores)						
median price duration in days	14	26	15	14	7	8
% dates at market price	100	100	100	100	58	87
median *base*-price duration	214	63	81	217	7	14
% dates at market *base* price	100	100	100	100	25	50

Referenced on page(s) 6, 6, 8. We restrict our core sample to test products to construct this table. The unit of observation is a variant–store–date, where the variant is an individual UPC (e.g., Raspberry Chobani Yogurt 5.3 oz), subsumed in a product (e.g., Chobani Yogurt 5.3 oz). We first calculate all variables at the variant–store level and then aggregate by computing their medians. "Median price duration in days" values are censored in the post-experimental duration of base prices at the control stores. "% dates at market price" is the percentage of dates where the focal variant–store's price equals the modal price (across stores) for that variant in that market (i.e., test versus control stores) on that date.

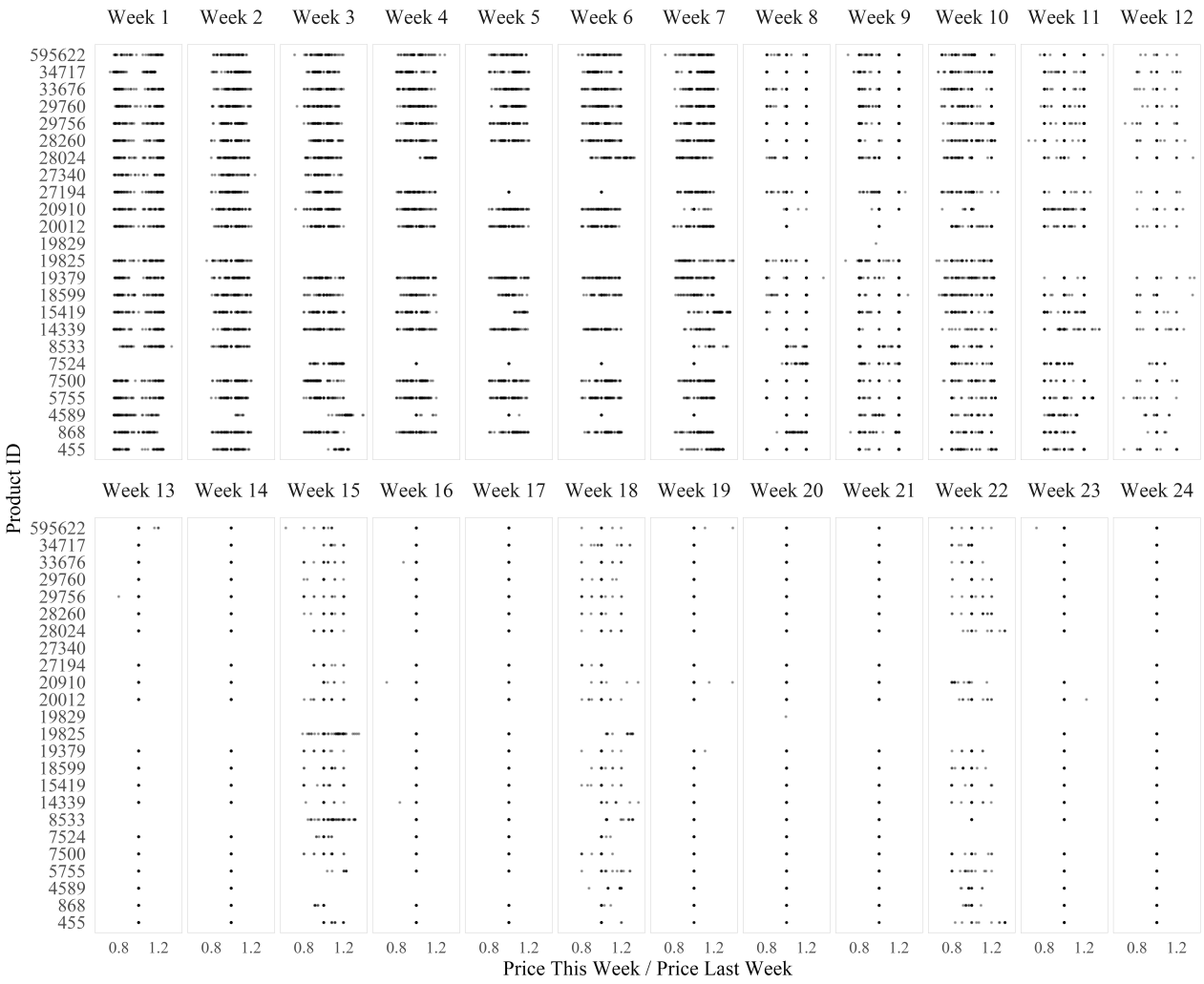


Figure 1

Referenced on page(s) 7, 7. This figure illustrates the experimental price variation for a representative product category, front-end candy (analogous plots for the other product categories are in the online appendix). The 24 panels correspond with the 24 consecutive Wednesdays that experimental prices were updated, starting on Wednesday, July 10, 2019, and terminating on December 25, 2019. The dots depict product–store price ratios: the first panel depicts the ratio of the experimental price assigned on the first Wednesday and the last non-experimental price, and the ith panel, for $i > 1$, depicts the ratio of the experimental prices assigned on the ith and $(i-1)$th Wednesdays. The hard edges at 0.8 and 1.2 are not relics of the plot, as we did not truncate or winsorize the price ratios in any way. Note that, although most experimental prices lasted from one Wednesday to the next, some spanned multiple weeks, especially those assigned after the 13th Wednesday.

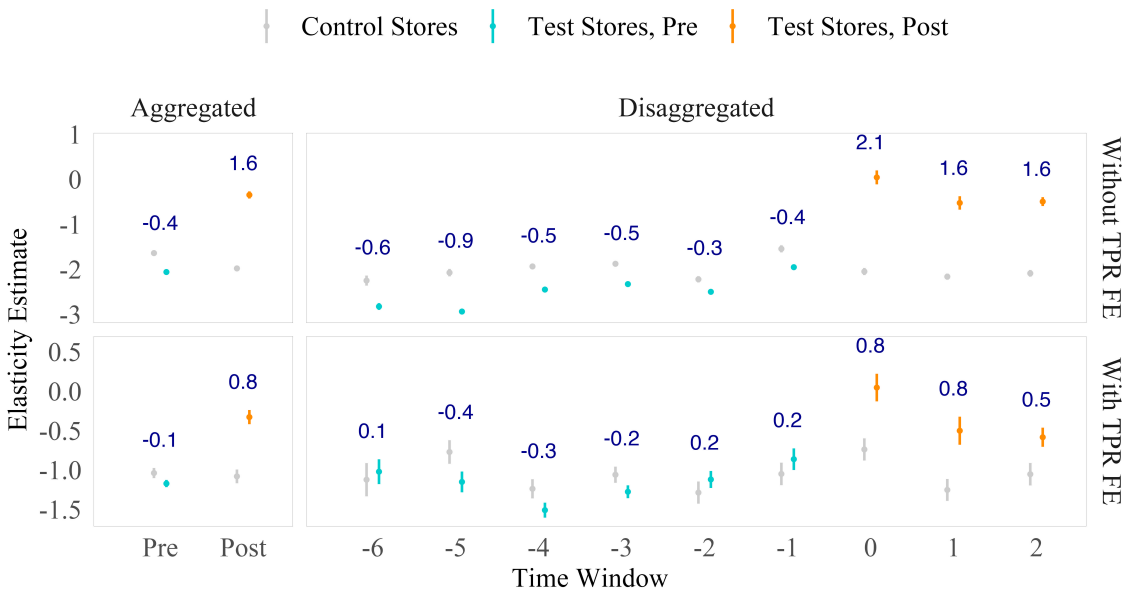


Figure 2

Referenced on page(s) 9, 10, 10, 15, 25, 25, 26, 26, 27, 27, 28, 28, 29, 29, 30, 31, 31, 32, 32, 33, 33, 36, 36. This figure depicts the cross-sectional means of our elasticity estimates. We estimate a separate elasticity for each combination of product, market, and time-window grouping, where the market is either "Test Stores" or "Control Stores," and the time-window grouping is either (i) an individual time window, a number between -6 and 2, (ii) "Pre," a flag that denotes time windows -6 through -1, or (iii) "Post," a flag that denotes time windows 0 and 1. (To avoid COVID, we exclude time window 2, which runs from January 1, 2020 to March 9, 2020, from the "Post" time-window grouping and all further analyses.) The dots denote the across-product average elasticity estimates, and the vertical bars denote their 95% confidence intervals (which we derive from store–week-clustered standard errors). The blue numbers denote the differences between the test- and control-store estimates. We run our regressions both with and without a temporary-price-reduction (TPR) dummy variable. We use OLS for all but the test stores in the experimental time windows, in which case we instrument for a given price with the exogenous portion of the experimental price recommendation, $\log \hat{p}_{jst} - \log p_{js0}$, where $\hat{p}_{jst}$ is the experimental price recommendation on day t and p_{js0} is the product's price the day before the experiment began. We drop elasticity estimates that (i) exceed 10 in absolute value, (ii) have a first-stage F-statistic less than 10, or (iii) stem from a regression with nearly collinear regressors (as indicated by a regression Gram matrix eigenvalue less than 0.2). Before averaging across products, we ensure a balanced product composition for each time window by excluding control-store products that lack corresponding test-store estimates, and vice versa. However, the product composition changes slightly over time, as not all products have price changes in all time windows.



Figure 3

Referenced on page(s) 10, 10. This figure is the same as Figure 2's "Aggregated" panel, except it averages over all products in a category, rather than all products in the sample. The filters discussed in Figure 2's captions still apply.

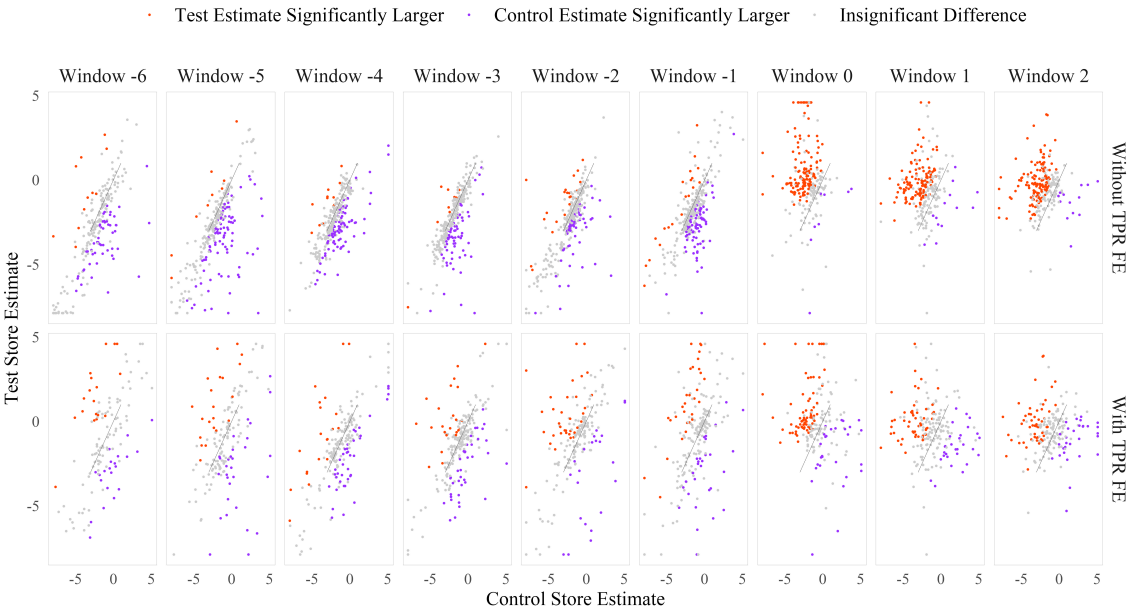


Figure 4

Referenced on page(s) 10, 11. This figure depicts our individual elasticity estimates. The horizontal axis measures the control-store elasticity estimates, and the vertical axis measures the test-store elasticity estimates. A dot is red if the test-store estimate significantly exceeds the control-store estimate (at the $p < 0.05$ level), it is purple if the control-store estimate significantly exceeds the test-store estimate, and it is gray otherwise. We winsorize all estimates at the 1% and 99% levels. The filters discussed in Figure 2's captions still apply.



Figure 5

Referenced on page(s) 11, 11, 11. This figure recreates the "Aggregated" elasticity estimates of Figure 2 with four alternative specifications. The first specification changes the dependent variable from $\log(q_{jst} + 0.1)$ to $\log(q_{jst} + 0.01)$. The second specification estimates elasticities with weekly data. We coarsen our sample from daily to weekly data by aggregating from each Wednesday to the following Tuesday. To perform this temporal aggregation, we calculate the total sales, sales-quantity-weighted mean log price, sales-quantity-weighted mean experimental instrument, and maximum TPR dummy variable, by store, product, and week. The third specification estimates elasticities with Cameron and Trivedi's (2013, p. 399) textbook count model with multiplicative heterogeneity. The last specification accounts for our sample reflecting a new price not when it is first posted, but when it is first associated with a purchase. This specification addresses this potential measurement error by excising the observations surrounding a price change, between and including the last day the product sold under the old price and the first day the product sold under the new price.



Figure 6

Referenced on page(s) 11, 12, 12, 12, 12, 30. This figure recreates the "Aggregated" elasticity estimates of Figure 2 with six alternative specifications. The observational estimates—those pertaining to the control stores or the pre-experimental period—are new, but the experimental estimates are the same as before. The orange dots vary across specifications, however, since we drop the experimental estimates that do not have a corresponding observational estimate. The first specification adds to the non-experimental regressions eight dummy variables that indicate which of the preceding eight weeks had TPR. The second specification limits the non-experimental sample to the observations that are not within a week of a TPR. The third specification limits the non-experimental sample to the observations with a TPR. The fourth specification drops the non-experimental observations that are within a week of a TPR and then clusters the remaining data by base price so that a store–product's ith cluster comprises the observations that lie between its $(i-1)$th and ith base price changes. By design, each cluster comprises one price change. The specification limits the sample to the clusters with price decreases, and runs the non-experimental regressions with price-cluster dummy variables to estimate elasticities off of the within-price-cluster variation—i.e., off of the base price decreases. Note, this specification's clustering technique effectively doubles the size of the sample, as all but the end-most observations belong to two clusters. Our clustered standard errors account for the sample duplication. The fifth specification is the same as the fourth, except it uses the groups with price increases for the regressions. The sixth specification is the same as the fifth, except it drops the products without pre- and post-period estimates. In other words, this specification fixes the product mix over time. The filters discussed in Figure 2's captions still apply.

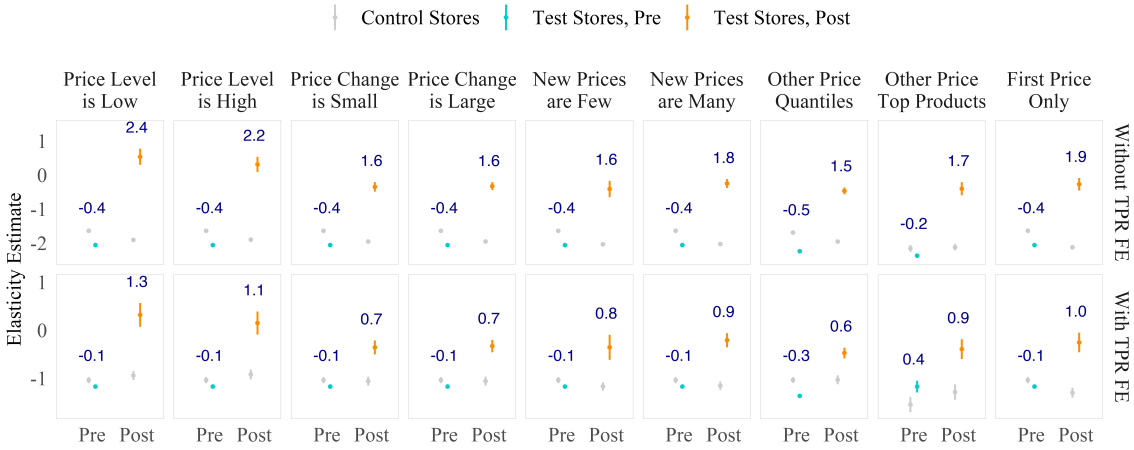


Figure 7

Referenced on page(s) 13, 13, 13, 13. This figure recreates the "Aggregated" elasticity estimates of Figure 2 with nine panels corresponding to six specifications. All specifications change only the experimental estimates, except for "Other Price Quantiles" and "Other Price Top Products," which modify the experimental and observational estimates. The "Price Level is Low" and "Price Level is High" panels report the first specification's estimates. This specification interacts the experimental prices and instruments with a set of dummy variables that indicate whether the current experimental price is lower or higher than the price on the day preceding the start of the experiment. In other words, this specification uses dummy variables to split the experimental prices into two series: one comprising the prices that are larger than usual, and the other comprising the prices that are smaller than usual. The specification then estimates a separate experimental elasticity for each series. The "Price Change is Small" and "Price Change is Large" panels report the second specification's estimates. This specification is the same as the first specification, except its dummy variables indicate whether or not the most recent price change exceeds the magnitude of the median control-store price change. The "New Prices are Few" and "New Prices are Many" panels report the third specification's estimates. This specification is the same as the first specification, except its dummy variables indicate whether or not the "synchronization level" of the most recent price change exceeds that of the median product in the given category at the control stores. We define the synchronization level of a price change as the number of other price changes for the given category at the given store, within a week of the given date. The "Other Price Quantiles" and "Other Price Top Products" specifications control for the prices of the competing products: the former includes as regression control variables the highest price, lowest price, and price quantiles, for the category at the given store, and the latter includes as regression control variables the prices of the five highest revenue products, for the given category at the given store. Finally, the "First Price Only" specification limits the experimental period of each store-product to the duration of its first test price. The filters discussed in Figure 2's captions still apply.



Figure 8

Referenced on page(s) 13, 13. This figure recreates the "Aggregated" elasticity estimates of Figure 2, but with either six or twelve additional control variables that account for the given store-product's previous prices: we include the log-prices from the preceding six weeks in the "Without TPR FE" regressions, and we include the log-prices and promotion dummies from the preceding six weeks in the "With TPR FE" regressions. We instrument for the lagged prices with lagged experimental instruments, for the "Test Stores, Post" regressions. We plot the seven price coefficient estimates—i.e., the elasticity of current demand to the price 0, 1, 2, 3, 4, 5, and 6 weeks ago. The filters discussed in Figure 2's captions still apply.

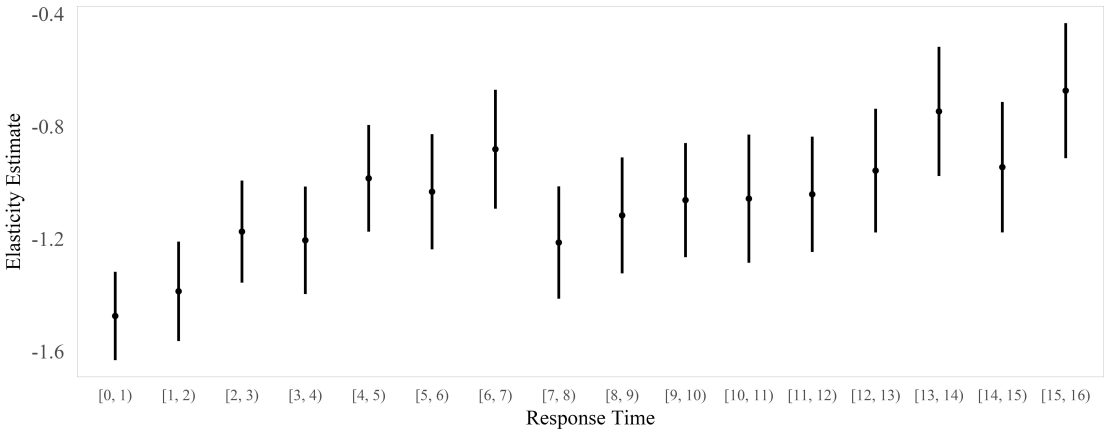


Figure 9

Referenced on page(s) 14, 14. This figure illustrates how the sensitivity of demand to a non-experimental price change varies with the response time. To create this figure, we remove observations that are within a week of a promotion and then cluster the remaining data by base price, in the fashion of Figure 6: a store–product's ith cluster comprises the observations that lie between its $(i-1)$th and ith base-price change. By design, each price cluster comprises one price change and hence comprises a pre-change and a post-change subsample. We limit the post-change subsamples to the observations between τ and $\tau+1$ weeks after the corresponding price change. We use the resulting sample to run our elasticity regressions. We include store–product–cluster dummy variables to estimate elasticities off of the within-cluster price variation—i.e., off of the demands that occur τ weeks after a base-price change. We use the full non-experimental sample, pooling the control-store, pre-COVID data (time windows -6 to 1) with the test-store, pre-experimental data (time windows -6 to -1). Also, for additional power, we use all products from our nine test categories, even those for which we do not have experimental prices. We run the regressions separately for all values $\tau \in \{0, \cdots, 15\}$. The filters discussed in Figure 2's captions still apply.

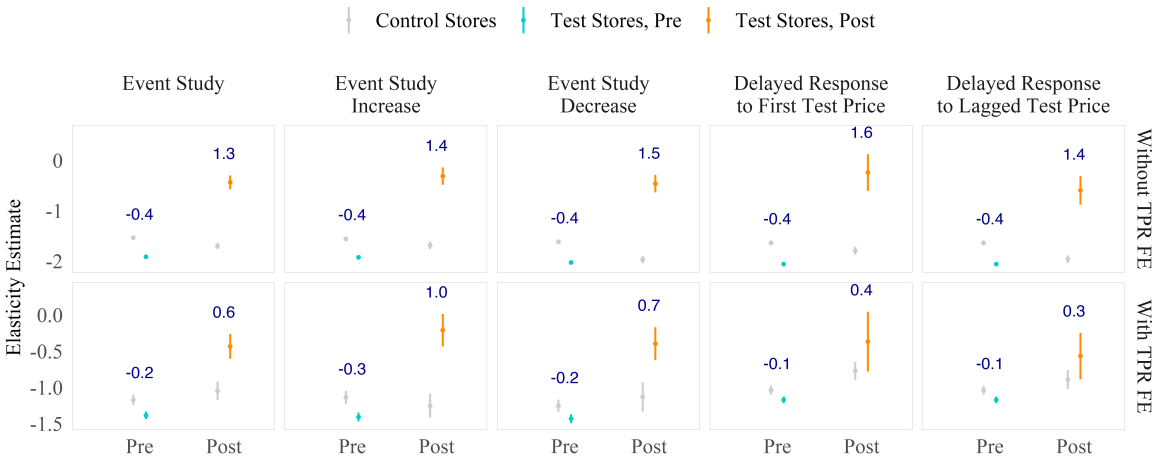


Figure 10

Referenced on page(s) 14, 14, 14, 14, 14. This figure recreates the "Aggregated" elasticity estimates of Figure 2 with five alternative specifications. Both the experimental and observational estimates are new for the first three specifications, but only the experimental estimates are new for the last two. For the first specification, we limit our sample to the one-week window surrounding each price change. We then run our elasticity regressions with price-change-window dummy variables to estimate elasticities using within-window price variation (i.e., off of the one-week demand responses). The second and third specifications are the same, except we use only the price-change windows with price increases for the former and those with price decreases for the latter. The fourth specification is the same as our primary specification, except we replace the contemporaneous experimental instrument, $\log \hat{p}_{jst} - \log p_{js0}$, with the first experimental instrument, $\log \hat{p}_{js1} - \log p_{js0}$, where $t = 1$ denotes the first day of the experiment. The fifth specification is the same as our primary specification, except we replace the contemporaneous experimental instrument, $\log \hat{p}_{jst} - \log p_{js0}$, with the 28-day-lagged experimental instrument, $\log \hat{p}_{js(t-28)} - \log p_{js0}$. The filters discussed in Figure 2's captions still apply.



Figure 11

Referenced on page(s) 14, 15. This figure recreates the "Aggregated" elasticity estimates of Figure 2 with twelve alternative specifications. The observational estimates are the same as before, but the experimental estimates are new. The control-store observational estimates vary across specifications, however, because we drop the post-period estimates for the control stores that do not have corresponding experimental estimates. The ith alternative specification is the same as our primary specification, except its experimental sample includes only the "price regimes" that last at least i weeks long, where a "price regime" is a span of a random price that is sufficiently stable so that the maximum price does not exceed the minimum price by more than 5%. For example, in the post-experimental period of the final panel labeled "12," we include only store–products with experimental prices that last at least 12 weeks. The filters discussed in Figure 2's captions still apply.



Figure 12

Referenced on page(s) 15, 15, 15, 15, 15. This figure recreates the "Aggregated" elasticity estimates of Figure 2, with four alternative specifications. The experimental estimates are the same as before ("Test Stores, Post"), but the observational estimates are new. The first specification instruments for the observational log prices with the modal log price of the given product at the stores at least 20 miles away from the focal store. The second specification uses these modal prices not as instrumental variables but as control variables in the non-experimental regressions. The third and fourth specifications are the same as the first and second, except with the prior week's price replacing the modal price in the other markets. The filters discussed in Figure 2's captions still apply.

A Additional Exhibits

Table A: Summary statistics: pre-experiment May 2, 2018 – July 9, 2019 / post-experiment July 10, 2019 – December 31, 2019

	Butter Spreads		Candy		Cheese, Other		Cheese, Shredded		Cheese, Sliced		Cheese, Snacking		Facial Tissues		Pasta		Pasta Sauce	
Control stores																		
Observations																		
products	29 /	29	58 /	60	36 /	37	31 /	32	27 /	28	31 /	36	20 /	21	83 /	85	79 /	79
product–stores	880 /	829	1655 /	1469	1117 /	965	958 /	844	857 /	796	928 /	995	636 /	599	2547 /	2293	2432 /	2217
product–store–dates	331,695 /	121,034	489,373 /	162,028	358,160 /	125,414	292,492 /	111,792	292,926 /	100,806	256,741 /	102,251	223,758 /	75,314	591,604 /	235,346	562,512 /	214,914
variant–store–dates	641,386 /	238,944	1,130,014 /	425,643	502,195 /	178,702	697,588 /	234,765	604,845 /	215,480	428,846 /	165,276	292,214 /	98,798	1,011,773 /	387,671	714,362 /	271,485
Median (across variant–stores)																		
median price duration in days	26 /	21	27 /	80	14 /	14	7 /	7	12 /	12	20 /	21	14 /	14	14 /	7	14 /	16
% dates at market price	100 /	99	100 /	100	100 /	100	100 /	96	100 /	100	99 /	100	100 /	100	100 /	99	100 /	100
median *base*-price duration in days	431 /	134	308 /	173	210 /	144	140 /	88	414 /	143	217 /	155	384 /	117	149 /	86	154 /	135
% dates at market *base* price	100 /	100	100 /	100	100 /	100	100 /	100	100 /	100	100 /	100	100 /	100	100 /	100	100 /	100
Test stores																		
Observations																		
products	29 /	29	59 /	62	36 /	37	31 /	32	27 /	28	29 /	36	20 /	21	83 /	85	79 /	79
product–stores	2147 /	2233	4110 /	3655	2780 /	2846	2495 /	2613	2191 /	2284	2273 /	2771	1563 /	1661	6211 /	6184	6210 /	5948
product–store–dates	765,008 /	287,426	1,177,169 /	377,413	824,254 /	336,434	720,743 /	321,037	742,793 /	267,511	602,310 /	241,825	470,188 /	166,456	1,399,253 /	605,718	1,395,291 /	518,371
variant–store–dates	1,492,673 /	565,544	2,713,401 /	1,029,227	1,193,966 /	513,276	1,699,573 /	715,398	1,479,847 /	589,096	996,986 /	392,131	608,486 /	219,594	2,394,671 /	1,038,391	1,824,536 /	682,027
Median (across variant–stores)																		
median price duration in days	18 /	7	28 /	11	14 /	7	7 /	7	12 /	7	18 /	7	13 /	7	14 /	7	11 /	6
% dates at market price	100 /	59	100 /	58	100 /	77	100 /	78	100 /	73	100 /	62	100 /	73	100 /	77	100 /	74
median *base*-price duration in days	402 /	7	265 /	8	370 /	7	140 /	10	427 /	7	263 /	7	206 /	7	143 /	32	134 /	7
% dates at market *base* price	100 /	53	100 /	53	100 /	60	100 /	58	100 /	58	100 /	56	100 /	61	100 /	67	100 /	64

Referenced on page(s) 6. We use our core sample to construct this table, restricting ourselves to the test products. The unit of observation is a variant–store–date. For simplicity, all statistics reported in this table are paired: the statistics on the left correspond to the pre-experimental period, and those on the right correspond to the post-experimental period. We also analyze price variation, reporting summary statistics about base-price variation separately from price variation that includes promotion. "Median price duration in days" first calculates, for each variant–store, the median price duration in days, and then gives the median of this statistic across store–products. Note that this duration is censored for very long spells, which is particularly apparent in the final two rows of the table where we calculate the post-experimental duration of base prices in control stores. "% dates at market price" takes, for each variant–store, the fraction of dates where the focal price equals the modal price for that variant in that market (i.e., test versus control stores). Then, we take the median of this statistic across variant–stores.

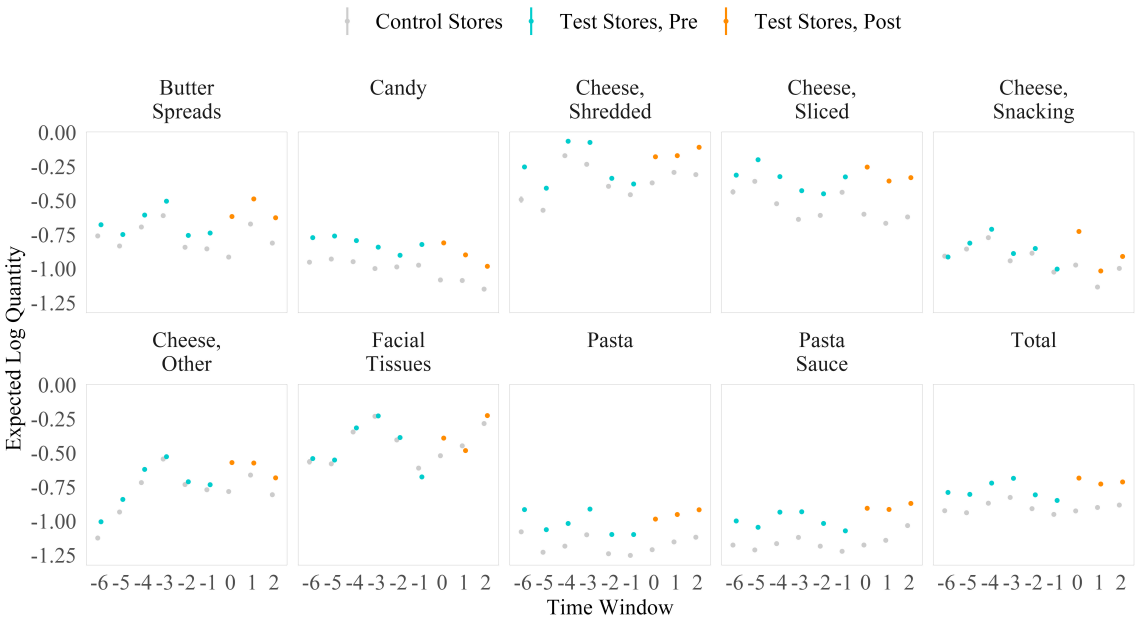


Figure A

This figure reports the mean values of $\log(q_{jst}+0.1)$, where q_{jst} is the sales quantity of product variant j, at store s, on day t. These means are calculated by time window and market, in the fashion of Figure 2. (i.e., for "Test Stores" and "Control Stores" separately). And like in Figure 2, we plot the corresponding 95% confidence intervals, but they are too small to see.

Table B: Initial test-price randomization checks: balanced revenues and elasticities

	Revenue per day			Elasticity		
Category	Low price	High price	p-value	Low price	High price	p-value
Butter Spreads	4.59 (0.22)	4.70 (0.23)	0.74	−2.97 (0.24)	−2.96 (0.23)	0.86
Candy	1.79 (0.10)	1.87 (0.10)	0.58	−1.66 (0.12)	−1.66 (0.12)	0.83
Cheese, Other	5.41 (0.31)	5.48 (0.34)	0.88	−2.68 (0.15)	−2.66 (0.14)	0.4
Cheese, Shredded	5.68 (0.31)	5.65 (0.31)	0.95	−2.64 (0.13)	−2.65 (0.13)	0.87
Cheese, Sliced	5.97 (0.33)	5.72 (0.33)	0.6	−2.30 (0.10)	−2.40 (0.10)	0.27
Cheese, Snacking	3.41 (0.26)	3.61 (0.27)	0.61	−2.50 (0.10)	−2.46 (0.10)	0.23
Facial Tissues	3.91 (0.20)	3.76 (0.19)	0.6	−3.12 (0.18)	−3.15 (0.18)	0.42
Pasta	1.15 (0.13)	1.18 (0.15)	0.87	−2.19 (0.16)	−2.17 (0.16)	0.51
Pasta Sauce	2.05 (0.31)	2.00 (0.28)	0.92	−2.89 (0.18)	−2.87 (0.18)	0.55
All	3.04 (0.24)	3.07 (0.24)	0.92	−2.32 (0.08)	−2.32 (0.08)	0.87

Referenced on page(s) 7. This table uses our core sample and presents two randomization checks for the initial test price: one for mean daily revenue per product and one for average pre-test elasticity. Recall that for each week, the unit of randomization is a product–store. "Low price" indicates a product–store was given a below-median experimental price on the initial randomization date. "High price" means a product–store was given an above-(or-equal-to)-median experimental price on the initial randomization date. The first two columns (after category) report "Revenue per day," or the average revenue (first across days, then across products) for each category, for the product–stores with low and high prices, respectively. The fourth column reports the p-values for a pooled-variance, two-sample t-test with the null hypothesis that the means in the two columns are equal. The fifth and sixth columns report average product-level elasticities before the first wave of the experiment for the product–stores in the "low-price" and "high-price" conditions. Category-level elasticities are estimated from regressing log(quantity + 0.1) on log(price) + product fixed effects, and the slope on log(price) is pooled in a single (OLS-weighted) regression. Clustered standard errors (store + product) are in parentheses. The seventh column reports p-values from t-test with the null of equal average elasticity between the previous two columns. (For these last three columns, we only include products with test prices that yielded a strong instrument—an F-stat greater than 10 in the first stage—in the first wave.)

Table C: Exogeneity of test price multipliers: previous week's revenue does not predict this week's multiplier

	Multiplier$_t$										
	All	All	Butter	Candy	Cheese, Other	Ch. Snacking	Ch. Shredded	Ch. Sliced	Facial Tissues	Pasta	Pasta Sauce
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Intercept	1.21***										
	(0.03)										
Multiplier$_{t-1}$	−0.21***	−0.21***	−0.2***	−0.13	−0.2***	−0.24***	−0.31***	−0.19***	−0.24***	−0.21***	−0.31***
	(0.03)	(0.03)	(0.02)	(0.11)	(0.03)	(0.04)	(0.04)	(0.02)	(0.03)	(0.02)	(0.03)
Revenue$_{t-1}$	−5.05e-06**	−5.14e-06	−3.06e-06	−1.08e-04*	−2.02e-06	−1.32e-07	−5.85e-06	−2.34e-05	−3.49e-06	−7.85e-05***	−3.48e-06
	(2.24e-06)	(3.28e-06)	(4.54e-06)	(6.31e-05)	(3.11e-06)	(5.95e-06)	(5.06e-06)	(1.63e-05)	(2.47e-05)	(2.37e-05)	(3.32e-05)
Product FE:	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	181592	181592	21357	23809	18932	18161	17011	16583	8565	26968	30206
R^2	0.05	0.05	0.04	0.02	0.05	0.09	0.11	0.04	0.07	0.05	0.11

Note: *p<0.1; **p<0.05; ***p<0.01

Referenced on page(s) 7. We use our core sample to construct this table and focus only on the test products during the first and second waves (up to December 31, 2019). The unit of observation is a product–store–test-week. Standard errors are reported in parentheses and clustered at the product–store + week level.

Table D: Initial store-assignment randomization check between optimized and test conditions (in the second wave): balanced store revenues in pre-experiment and first wave

	Pre-experiment			First wave		
Category	Test	Optimized	p-value	Test	Optimized	p-value
Butter Spreads	167,060 (14,298)	136,734 (8,189)	0.07*	25,844 (1,896)	20,873 (1,078)	0.03**
Candy	88,592 (8,252)	79,455 (3,928)	0.32	15,870 (1,298)	14,453 (554)	0.32
Cheese, Other	51,061 (3,158)	51,326 (2,749)	0.95	8,512 (491)	8,565 (414)	0.93
Cheese, Snacking	64,338 (5,840)	70,281 (6,213)	0.49	10,016 (865)	11,258 (845)	0.31
Cheese, Shredded	191,494 (15,340)	202,401 (10,855)	0.56	37,153 (2,503)	37,512 (1,755)	0.91
Cheese, Sliced	141,869 (9,349)	154,421 (8,501)	0.32	24,696 (1,233)	25,715 (1,291)	0.57
Facial Tissues	36,526 (3,363)	30,537 (2,542)	0.16	5,705 (486)	4,793 (356)	0.13
Pasta	70,093 (5,810)	70,484 (4,102)	0.96	13,105 (921)	12,912 (660)	0.86
Pasta Sauce	82,991 (6,637)	91,199 (4,889)	0.32	16,008 (1,071)	17,045 (780)	0.44
All	94,894 (7,885)	93,360 (7,308)	0.89	16,587 (1,336)	16,117 (1,228)	0.8

Referenced on page(s) 7. We use our core sample to construct this table. In the second wave, each store–category is randomly assigned to either the test or optimal condition. The first two columns (after category) report the average total-category revenue (across stores, in dollars) before the experiment, for stores that subsequently are randomized to stay in the test condition in the second wave of the experiment (column two) and for stores that are randomized into the optimal condition (column three). Standard errors are reported in parentheses. The fourth column reports the p-values from pooled-variance, two-sample t-test with the null hypothesis that the previous two columns have the same average. * p-value < 0.10, ** p-value < 0.05. The fifth, sixth, and seventh columns are analogous to the preceding three columns, except they measure the average total-category revenue (across stores) during the first wave of the experiment.

Table E: Initial store-assignment randomization check between optimized and test conditions (in the second wave): balanced category elasticities in pre-experiment and first wave

	Pre-experiment			First wave		
Category	Test	Optimized	p-value	Test	Optimized	p-value
Butter Spreads	−2.64	−2.73	0.22	−0.54	−0.48	0.39
	(0.21)	(0.20)		(0.11)	(0.08)	
Candy	−1.26	−1.33	0.34	−0.26	−0.26	0.91
	(0.13)	(0.12)		(0.10)	(0.08)	
Cheese, Other	−1.74	−1.80	0.3	−0.58	−0.55	0.54
	(0.18)	(0.17)		(0.13)	(0.14)	
Cheese, Shredded	−2.23	−2.32	0.31	−0.61	−0.57	0.62
	(0.12)	(0.09)		(0.31)	(0.28)	
Cheese, Sliced	−2.10	−2.18	0.37	0.20	0.33	0.11
	(0.13)	(0.11)		(0.30)	(0.31)	
Cheese, Snacking	−2.32	−2.45	0.09*	−0.31	−0.25	0.42
	(0.12)	(0.10)		(0.14)	(0.14)	
Facial Tissues	−2.64	−2.71	0.33	−0.32	−0.25	0.24
	(0.19)	(0.18)		(0.14)	(0.13)	
Pasta	−2.38	−2.50	0.21	−0.04	0.06	0.23
	(0.18)	(0.17)		(0.13)	(0.11)	
Pasta Sauce	−2.79	−2.83	0.44	−0.16	−0.14	0.56
	(0.11)	(0.10)		(0.23)	(0.23)	
All	−2.10	−2.18	0.24	−0.32	−0.27	0.35
	(0.08)	(0.07)		(0.06)	(0.05)	

Referenced on page(s) 7. We use our core sample to construct this table. In the second wave, each store–category is randomly assigned to either the test or optimal condition. The first two columns (after category) report the average product-level elasticities before the experiment for stores that subsequently are randomized to stay in the test condition in the second wave of the experiment (column two) and for stores that are randomized into the optimal condition (column three). Elasticities are estimated from regressing log(quantity + 0.1) on log(price)+ log(price)$\cdot I_{\text{optimized}}$ + product fixed effects, and the slopes are pooled in a single (OLS-weighted) regression. Clustered standard errors (store + product) are in parentheses. The fourth column reports p-values from t-test with the null of equal average elasticity between the previous two columns. The fifth, sixth, and seventh columns are analogous to the preceding three columns, except they measure the average elasticity during the first wave of the experiment, and we instrument the respective price variables with first-wave experimental prices.

Table F: Store reassignment randomization check between optimized and test conditions (in the second wave): balanced store revenues in previous week

	Test			Optimized		
Category	No change	Change	p-value	No change	Change	p-value
Butter Spreads	3,456 (113)	3,268 (306)	0.56	3,250 (93)	3,040 (261)	0.45
Candy	4,913 (305)	4,251 (529)	0.28	4,185 (98)	4,966 (446)	0.09*
Cheese, Other	1,055 (29)	1,071 (72)	0.84	1,075 (21)	1,105 (74)	0.7
Cheese, Shredded	3,308 (75)	3,478 (276)	0.55	3,736 (74)	3,402 (235)	0.18
Cheese, Sliced	2,714 (60)	2,799 (163)	0.62	2,849 (51)	3,046 (251)	0.44
Cheese, Snacking	1,023 (33)	1,146 (115)	0.3	1,175 (31)	1,228 (135)	0.7
Facial Tissues	664 (20)	696 (80)	0.71	637 (20)	708 (69)	0.32
Pasta	1,385 (31)	1,378 (102)	0.95	1,504 (30)	1,390 (120)	0.36
Pasta Sauce	1,653 (38)	1,592 (125)	0.64	1,626 (29)	1,742 (140)	0.42
All	1,858 (28)	1,896 (83)	0.66	1,971 (23)	1,953 (84)	0.84

Referenced on page(s) 7. We use our core sample to construct this table. In the second wave, after the first week, in each category, three stores from each condition are randomly reassigned to from their current condition: test or optimal. The first two columns (after "Category") report the average total-category revenue (across stores, in dollars) before the experiment, for stores that are not reassigned from the test condition (column two, "No change") and for those that are reassigned (column three, "Change"). Standard errors are reported in parentheses. The fourth column reports p-values from a pooled-variance, two-sample t-test with the null hypothesis that the averages in the two columns are equal. * p-value < 0.10, ** p-value < 0.05. The fifth, sixth, and seventh columns report the corresponding statistics for the optimal condition.

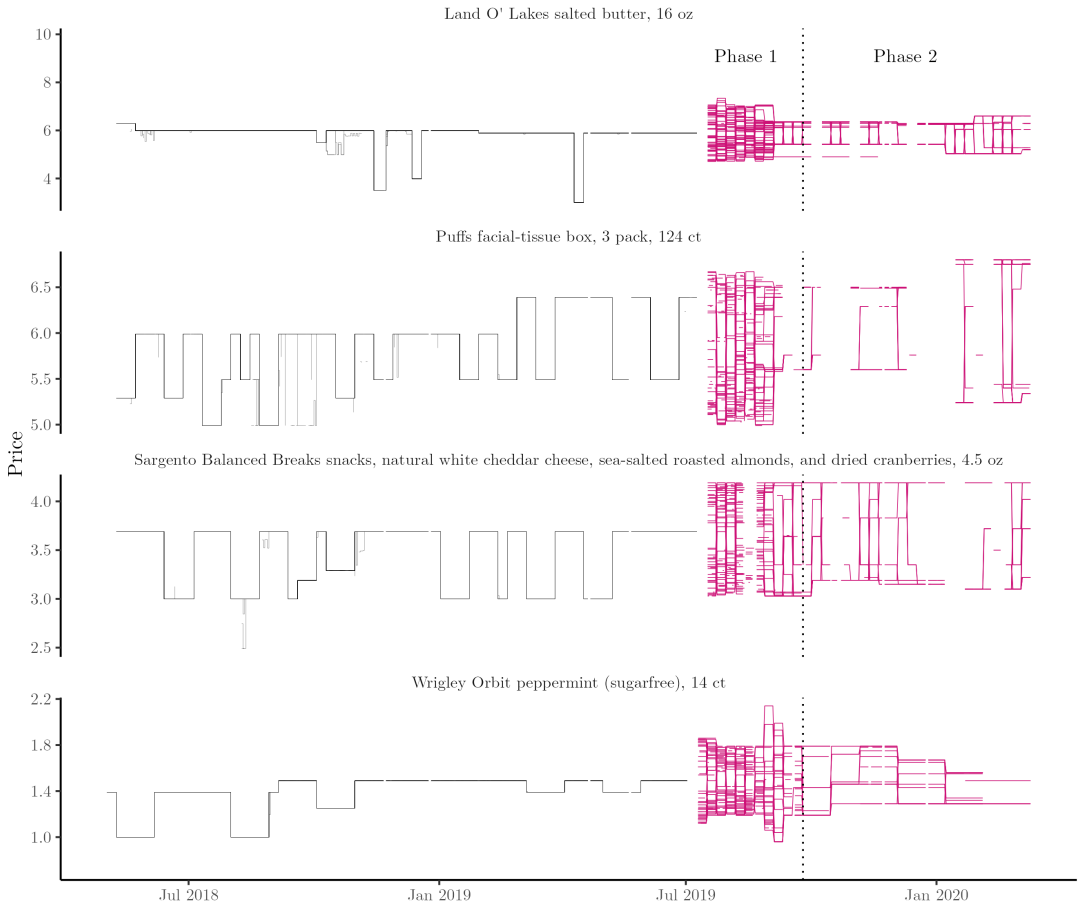


Figure B

Referenced on page(s) 8. This figure illustrates price paths before the experiment (gray) and during the experiment (pink) for four products across all stores in our data.

Table G: Cost-shifter instruments: first-stage investigation

	BLS input prices		Wholesale prices	
Subset	# products	# coefficient > 0	# products	# coefficient > 0
base prices + 0% R^2	99	37	71	57
base prices + 10% R^2	69	23	50	42
base prices + 20% R^2	39	17	45	37
base prices + 30% R^2	27	11	37	35
base prices + 40% R^2	26	10	33	32
base prices + 50% R^2	19	8	20	19
base prices + 60% R^2	7	6	16	15
base prices + 70% R^2	4	4	16	15
base prices + 80% R^2	0	0	15	14
base prices + 90% R^2	0	0	13	12

Referenced on page(s) 16, 16, 16. In this analysis, we use our core sample plus the cost-shifter instruments described in B.3. For both cost shifters, we instrument base prices with the instruments, as only base prices have been shown to respond to cost changes in Anderson et al. (2017). First, we construct the BLS PPI columns (first two columns after the "Subset") by taking the following steps:

1. Isolate base prices by dropping all observations where our inferred-promotion flag is equal to one. Define a base-price change as a change in the remaining, last non-promoted price to the next non-promoted price by an amount exceeding 2 cents. (This filter drops rounding errors, e.g., 99 cents to 98 cents, back to 99 cents, which are rare, but do happen.)
2. Drop all products that do not have at least one base-price change in every non-experimental sample, i.e., the test-pre, control-pre, control-post.
3. Drop all products that do not have at least three months of observations in each non-experimental sample. This filter is necessary to include three-month-lagged BLS PPIs. The preceding filters leave 150 products.
4. Next, we analyze the explanatory power of the BLS PPIs in explaining base prices. As in Villas-Boas and Winer (1999) (and the working paper version), we instrument base prices with contemporaneous and lagged BLS PPIs. Specifically, for each product–market–time zone (e.g., before experiment, after), we regress $\log(price) \sim PPI + PPI_{-1month} + PPI_{-2months} + PPI_{-3months}$.
5. We then save coefficients on each lag (including 0 lag), standard errors, and the R^2, and we keep only products where the coefficients are identified (i.e., sufficient variation) in every non-experimental sample. This filter leaves us with 99 products. The third column reports the number of products that have a positive average BLS coefficient and an R^2 of at least the specified amount in each row (i.e., 0%, 10%, ..., 90%.) To calculate the average coefficient (because there are four coefficients, including lags), we take a weighted average (by precision) of the coefficients.

Next, we conduct a similar exercise for the wholesale-price instruments, which we only observe around when the first wave of the experiment begins. Specifically, in our core sample, after dropping promotions and focusing on base prices, we observe wholesale prices with non-zero variance for 92 products during the non-experimental part of the sample at control stores (which did not get experimental prices).

1. We isolate base prices and their changes, as in step (1) with the BLS-PPI instrument. Applying these filters leaves 71 products with: at least one base-price change in the sample and positive variance in the wholesale-price instrument.
2. Next, we analyze the explanatory power of the wholesale prices in explaining base prices. For each product, in its control stores, from the start of the first wave to the cutoff date during the second wave, we regress $\log(price) \sim \log(wholesaleprice)$.
3. We then save coefficients and the R^2. The fifth column reports the number of products with a positive $\log(wholesaleprice)$ coefficient and an R^2 of at least the specified amount in each row (i.e., 0%, 10%, ..., 90%.)

B Data Appendix

B.1 Sample Construction and Variable Definitions

Continuing from page 6. To construct our core sample, we apply the following filters and transformations to the raw data:

1. We keep only the retailer's 11 experimental product categories, which are: "Butter Spreads;" "Candy;" "Parmesan and Velveeta Cheese;" Cheese, Shredded;" "Cheese, Sliced, "Chunk Cheese, "Facial Tissues;" "Pasta;" "Pasta Sauce;" "Cheese, Snacking;" and "Halloween Candy." We drop "Halloween Candy," which is highly seasonal, and combine "Parmesan and Velveeta Cheese" with "Chunk Cheese" into a single category, "Cheese, Other," leaving 9 test categories.

2. We drop transaction lines with: (i) negative revenue and negative quantity (i.e., returns), (ii) negative revenue and positive quantity, (iii) zero revenue and zero quantity, and (iv) zero revenue and negative quantity (all very rare and likely errors).

3. We flag each variant–store–date in which at least one transaction involved "Buy-X-get-Y-free" promotions (henceforth, BOGO), represented in the data as zero-revenue-positive-quantity transaction lines. We use this flag later to filter our sample. After creating this flag, we drop transaction lines that involve BOGOs, coupons, and employee discounts.

4. We aggregate revenue, quantity, and the BOGO flag to the variant–store–date level. We calculate each observation's "effective price" as total revenue / total sales quantity.

5. We impute missing observations with variant–store–dates with zero sales. We then drop zero-sales strings at the variant–store level exceeding ten days (likely stockout/not carried periods). We also impute missing prices for the remaining zero-sales days in two ways: by filling the variant's price backward or forward between transactions.

6. We winsorize the bottom and top 1% of prices and the top 1% of quantities at the variant–store level.

7. We create a promotion flag, at the variant–store–date level, in the manner described in Appendix B.2 (similarly to Butters et al. 2020 and Hitsch et al. 2021).

8. We create products by splitting line-groups (groups of variants meant to have the same price) using price variation in the data. First, we define variants with missing line-groups as their own products. Second, we separate as their own products those variants whose average (across store–dates) price deviation from its corresponding median (across variants) line-group price exceeds 1% of the average (across store–dates) median (across variants) price. We define the remaining variants left in each line-group as a single product (after separating out the aforementioned variants). Third, within product–store–date, we set every variant's price equal to the median price across variants with non-zero sales (if all variants have zero sales, we set all prices equal to the median *imputed* price).

9. We drop observations where forward- and backward-imputed prices disagree.

10. We drop observations where forward- and backward-imputed promotion flags disagree.

11. We drop variant–store–dates with a positive BOGO flag as defined in step 3, and, after June 25, 2019—when we observe richer promotion data, see Appendix B.2—we drop observations where there is a promotion type of "Buy X Get Y Free."

12. We drop all observations after December 31, 2019, to mitigate the effect of the COVID-19 pandemic on our analysis.

B.2 Promotions and Base Prices

Although we observe richer promotion data after June 25, 2019 (around the time the price experiments began), we use the imputed promotion variable for experimental observations, for consistency in our analyses. This promotion-classifying algorithm was designed with the classic "saw-tooth" pattern in observational price variation in mind. Experimentally induced price variation could thus "trick" the algorithm into thinking some experimental prices are promoted prices, so we are careful to use non test stores' promotion statuses to infer promotion status during price experiments.

First, we describe how we impute promotion status to distinguish base prices from promoted prices. Second, we describe our richer promotion dataset that we observe for the latter part of our sample window.

Inferring Promotions and Base Prices. Continuing from pages 5 and 44. We use an algorithm to impute promotion status in a manner similar to Butters et al. 2020 and Hitsch et al. 2021. Specifically, we first create the promotion flag for non-experimental observations. We create two flags, "promo backward" and "promo forward," one for forwardly imputed prices and one for backwardly imputed prices. To construct these flags, we apply the following algorithm to each variant–store:

1. If a price is lower than its preceding price, we label it as a potential promotion.

2. If a price is labeled a potential promotion, then if next price is not within 5% of the last non-potential-promotion price, then this next price is also labeled a potential promotion.

3. If the first price in the time series lasts for fewer than 3 weeks and is lower than the second price, we label the first price a potential promotion.

4. If a price is not labeled a potential promotion, but that identical price level is labeled a potential promotion on more than 50% of other dates (for that variant–store), then we call it a potential promotion. Similarly, if a price is labeled a potential promotion, but that identical price level is labeled a potential promotion on less than than 80% of other dates (for that variant–store), we remove the potential-promotion label.

5. If a price is within three cents of a another price labeled a potential promotion in the same month–year (for that product–store), then we label it a potential promotion.

6. To create the promotion flag for test products at test stores during the price experiments, we use the modal potential-promotion flag for each variant–date across control stores. This approach is sensible because promotions in our data are typically implemented chain-wide (consistent with Hitsch et al. 2021).

7. We rename the surviving potential promotions as promotion flags.

8. We drop variant–store–dates where the forward and backward promotion flags disagree.

We validate our promotion imputations by comparing them to actual promotion statuses for the part of the data where we observe richer promotion variables (Table A). We next introduce this subset of the data and describe the validation exercise.

Richer Promotion Dataset. Continuing from page 5. After June 25, 2019, we observe the retailer's actual promotional activity for the universe of its stores and variants. Specifically, we observe temporary price promotions, feature print ads, and feature digital ads. And for each promotion type, we observe its subtype: "sale," "liquidation sale," "sale with points," "points only," "clearance," "in-store markdown," and "buy X get Y free." We also observe the variant–stores for which the promotion is active, the per unit promoted price, and, for BOGOs, the number of units required at purchase to obtain the promoted price. In total, from June 25, 2019, to August 22, 2021, we observe 60,641,880 unique promotions (defined by a promotion–variant–store) for 60,024 variants at the retailer.

Promotion-Flag Validation. Table A summarizes the promotion variation for products in our nine test categories in the core sample for the test products. Note that for the major promotion types (i.e., standard "Sales"), the median (across products) share of a product's store–dates that are correctly labeled "promotion" is almost 1. Further, the median (across products) share of a product's store–dates that are correctly labeled as "base price" is 0.99. As expected, our imputations are less precise for less frequent, store- or market-specific promotions (i.e., liquidation sales, clearance sales, perishable discounts) and for promotions that do not affect posted prices (i.e., points-only promotions). These types of promotions, however, are much less common than chain-wide promotions.

Table A: Validation of promotion flag using richer promotion dataset

Promotion type	Number of products with promotion type	Median (across products)	
		Share of store–dates with promotion type	Share of correct classifications for promotion type
None	392	0.79	0.99
Sale (TPR)	314	0.23	0.94
Sale (Print Ad)	168	0.12	1.00
Sale (Digital)	117	0.07	1.00
Liquidation Sale (Print Ad)	4	0.04	0.65
Sale (TPR) with Points	26	0.04	1.00
Points only	107	0.01	0.00
Clearance (TPR)	169	0.00	0.42
In-store markdown (TPR)	95	0.00	0.24

Referenced on page(s) 46, 46. We use our core sample to construct this table, restricting ourselves to the test products. The unit of observation is a product–store–date. "Correct classifications" refers to the share of store–dates (for each product) that our imputed promotion flag is correct (i.e., 0 when there is no current promotion happening, and 1 when there is any promotion type).

B.3 Wholesale-Price and Producer-Price-Index Instruments

Continuing from pages 5, 6, and 16. First, we describe here our construction of the wholesale-price instruments for test products. Then, we describe where we obtain input-price proxies to construct our PPI instruments for test products.

Wholesale-Price Instrument. We acquired wholesale-price data through two different data pulls. In the first pull, we observe wholesale prices for all 2,276 variants in the universe of test-category variants for the test stores from July 1, 2019, to May 10, 2020. In the second pull, we observe wholesale prices for the universe of variants for all categories and stores in the retailer from May 11, 2020, to September 12, 2021. Of the 409 products in the core sample (between July 1, 2019, and December 31, 2019) that ever are treated, all 409 show up in the raw cost data. We apply the following filters and transformations to construct the wholesale-price panel, starting at the variant–store–date level and aggregating to the product–store–date level:

1. Over the 713-day period over which have wholesale-price data, we are missing wholesale prices for 24 dates. We impute these missing values by filling forward the more recent, non-missing wholesale price for that variant–store.

2. We use the chain-wide modal wholesale price for each variant–date across both test and control stores, dropping any observations still missing a wholesale price. (Wholesale prices differ across stores within product in only 0.1% of observations.)

3. We keep only products that have wholesale-price variation (for at least one variant) between July 1, 2019 and December 31, 2019. Applying this filter leaves 138 products.

4. We winsorize the bottom and top 1% of wholesale prices for each variant and reapply step 3. After this transformation, we are left with 123 products with wholesale-price variation between July 1, 2019, and December 31, 2019.

5. Finally, we aggregate wholesale prices from the variant level to the product level by taking the mean wholesale price across variants for each store–date–product. (Wholesale prices differ across variants in 9.4% of product–store–dates, and conditional on differing, the median maximum-to-minimum percent difference across variants is 10.6%.)

Producer-Price-Index Instrument. To obtain a second cost shifter we use the Producer-Price-Index (PPI) data provided by the Bureau of Labor Statistics (BLS). The BLS PPI-Commodity data are tracked monthly at the category level. Because these data are updated monthly (rather than daily), we forward fill the most recent value of the PPI to create a daily version of the PPI. For our categories, we pull the following respective datasets from the BLS website (https://www.bls.gov/ppi/):

1. Butter spreads: "PPI Commodity data for Processed foods and feeds-Butter, not seasonally adjusted."

2. Pasta: "PPI Commodity data for Processed foods and feeds-Cereal and pasta products, not seasonally adjusted."

3. Cheese categories: "PPI Commodity data for Processed foods and feeds-Process cheese, shipped in consumer packages or containers (3 lbs. or less), not seasonally adjusted."

4. Facial tissues: "PPI Commodity data for Pulp, paper, and allied products-Sanitary tissue paper products, made from purchased sanitary paper stock or wadding, not seasonally adjusted."

5. Front end products and candy: "PPI Commodity data for Processed foods and feeds–Candy and nuts, not seasonally adjusted."

6. Pasta sauce: "PPI Commodity data for Processed foods and feeds-Canned catsup and other tomato based sauces, not seasonally adjusted."

B.4 Miscellaneous Product Data Cleaning

Here, we describe additional data cleaning in the retailer's raw product files. Specifically, we:

1. Trim white space on the left and right of category strings, which create false duplicate categories.

2. All multiple white spaces are replaced by a single white space.

3. All variables are made lower case, and then the first letter of the first word is made uppercase.

4. Replace all forward slashes with underscores.

5. Remove all periods.

6. Replace all special character ampersands with "and."

7. Replace all double quotations, which signify "inches," with the word "inches."