Outlet Substitution Bias Estimates for Ride Sharing and Taxi Rides in New York City[1]

Authors Ana Aizcorbe, U.S. Bureau of Economic Analysis
Jeff Chen, Bennett Institute for Public Policy at the University of Cambridge

Author of Contact Ana.Aizcorbe@bea.gov

Date February 2022

Abstract The arrival of new merchants poses problems for measuring inflation, and many think the resulting biases in the official statistics are nontrivial. The BLS methods treat identical commodities sold by different merchants as distinct, different goods but to the extent the goods are close substitutes then the CPI will be biased upward by an estimated 0.08 percentage point per year (Moulton 2017).

There have not been many empirical studies to inform these estimates, owing to the paucity of the highly granular merchant-level data required. Studies based on external non-BLS sources have typically used a unit value index that essentially treats goods sold at different merchants as perfect substitutes, a controversial assumption. We also use a unit value index but with a different interpretation: We view a quality adjusted price index as the target and demonstrate that, in our context, the unit value index we calculate may be viewed as an upper bound to this unobserved target.

Using detailed data from email receipts, we find that the arrival and growth of ride-sharing services in New York City likely imparted a nontrivial bias in the official price indexes for that city: a lower bound of 0.5 percentage point per year over the period 2015–2017. We attribute the magnitude of the bias to the sustained growth of ride sharing over this period, from 40 percent of the market in 2015 to 70 percent by 2017.

Keywords Price indexes, inflation measurement

JEL codes C43, E31

[1] We thank Jan deHaan, Abe Dunn, Marshall Reinsdorf, and Dominic Smith for helpful comments.

I. Introduction

Outlet substitution bias is one of the biases thought to overstate price growth as measured in the Consumer Price Index (CPI). This problem was originally studied in the context of the arrival of discount stores, which is thought to have lowered the cost of food purchases, but those declines are not reflected in the CPI for food at home (Reinsdorf 1993). The rise of digital platforms has also spawned new ways of providing lower-cost substitutes for traditional services. Consumers are increasingly buying goods and services online, which has raised questions about potential biases there (Hatsius 2017). Other CPI categories with potential outlet substitution bias problems are accommodations (7211)—where organizations like Airbnb have provided a way for consumers to arrange overnight stays in private homes—and taxi and limo services (4853)—where ride-sharing platforms like Uber potentially offer a lower-cost alternative to traditional taxis.

This problem has been studied from one of two perspectives. In a traditional cost of living index (COLI) interpretation to price measurement, the problem is very similar to that associated with the arrival of new goods. In this view, inflation is measured as the amount of money you would have to give consumers to keep them indifferent between two choice sets. The challenge in this context is to account for any welfare gains associated with the introduction of new goods. The conceptual solution draws on reservation prices for the good before entry to control for any quality improvements associated with the arrival of the good (Fisher and Shell 1972). Empirically, this approach uses assumptions about the underlying utility function to define the target index and has been implemented in the context of the entry of generic drugs (Griliches and Cochburn 1994; Feenstra 1997; Berndt et al. 1996), and offshoring (Byrne et al. 2017).

In the alterative perspective, the goal is to construct a constant-quality price index without a COLI interpretation to measure inflation. Here, the challenge is defining the good properly. For example, are bananas bought at Costco identical to those bought at Whole Foods? If so, the price index for bananas should treat them as perfect substitutes and measure the price as an average price of bananas no matter where sold (a *unit value*). The unit value index is often used to study the outlet substitution bias problem, both in analytical studies (Diewert 1998; Nakamura et al. 2015) and empirical work to quantify the bias (Reinsdorf 1993; Leibtag and Hausman 2009; Ivancik and Fox 2014).

Of course, goods are rarely identical and so the notion of a *quality-adjusted* unit value (QAUV) index, where consumers view the two goods as broadly comparable, has been suggested as the relevant target index (Dalen 2001; deHaan 2002; Silver 2010). Empirically, this approach requires that one estimate quality parameters, typically using hedonic methods (deHaan 2004; Silver 2010 and 2011). In our approach, we derive conditions under which the unit value index that is typically used may be viewed as an upper bound to the target index, thus providing an alternative strategy when additional data to estimate the quality parameters are not available, as is the case for us.

We then use highly detailed data from email receipts to assess outlet substitution bias in the market for ride sharing and taxi service in New York City. We find that the type of indexes typically estimated by statistical agencies would overstate this bound for taxi rides and ride-sharing services in New York City, on average, by about one-half a percentage point per year over 2015 to 2017.

The paper is laid out as follows. The next section describes our empirical strategy for examining outlet bias. Section 3 describes the ride-level data that we use and section 4 reports out findings.

II. Quantifying Outlet Substitution Bias

Comparison of Unit Value and Noncomparable Indexes

The outlet substitution problem has to do with how one defines the good or the basic commodity to be priced in a price index. For our empirical application, the basic service provided by merchants is a ride from point A to point B, and the question is, "To what extent do riders view rides on this route with ride-sharing vehicles and taxis as perfect substitutes?" If consumers view the two types of rides as identical, then the "good" is a ride on this route regardless of who provides it and the price is an average price that divides all spending (ride sharing and taxi) by the total number of rides, called a unit value. *The unit value index* is formed as a ratio of these prices:

$$I_t^{UV} = \frac{P_t^T Q_t^T + P_t^R Q_t^R}{Q_t^T + Q_t^R} \bigg/ \frac{P_o^T Q_o^T + P_o^R Q_o^R}{Q_o^T + Q_o^R} \qquad (1)$$

where P_t^j and Q_t^j are the average price and number of rides provided by taxis (T) and ride sharing (R) vehicles.

On the other hand, if consumers view the rides as entirely different goods—or what the U.S. Bureau of Labor Statistics (BLS) calls noncomparable items—then changes in the prices of taxi and ride-sharing rides are tracked separately, using, ideally, a superlative index.

Our measure for outlet substitution bias is the difference between a target index that treats taxi and ride-sharing rides as broadly comparable to a BLS-type index that treats taxi and ride-sharing rides as noncomparable.[2] For the target index, we would like to calculate a quality-adjusted unit value index that treats the two types of rides as broadly comparable but do not have the data to calculate that index directly. Instead, we use the unit value index in (1) and argue below that it may be used as an upper bound to the quality adjusted unit value index. That is, we compare price change from a unit value index to that from a BLS-type index to obtain a lower bound to outlet substitution bias.

[2] Von Auer (2014): "In all of these publications the price indices derived were concerned with the problem of aggregating the price changes of similar products into some average price change. …. Similar products are defined as having innate differences that are observable and measurable. Such product differences occur frequently and stem from such things as quality levels, operating features, or simply the size of the packaging. These products have dissimilar product-identifying units and, consequently, they are unsuitable for the quantity summations in the UV index."

For the BLS-type index, one would ideally want to do a precise job of mimicking the CPI indexes. However, this is very difficult to do without access to the underlying source data.[3] Instead, we use a Laspeyres index of taxi and ride-sharing rides, similar to the fixed-weight indexes that the BLS uses, which we call the noncomparable index, $I_{t,o}^{NC}$:[4]

$$I_{t,o}^{NC} \quad = \quad w_o^T \ (P_t^T / P_o^T) + \ w_o^R (P_t^R / P_o^R) \qquad (2)$$

where the w's are expenditure shares from an earlier period, $w_o^j = P_o^j Q_o^j / (P_o^j Q_o^j + P_o^k Q_o^k)$.

When will a comparison of the noncomparable index in (2) and the unit value in (1) reveal bias? Analytical expressions for bias are most recently provided in Nakamura et al. (2015). In their analysis, outlet substitution bias is defined as the difference between an index for taxis only and a unit value index. Their focus on an index for taxis only allows them to assess any biases over the period before the new merchants are folded into the CPI sample. We also provide estimates for this bias in our empirical work but focus here on the pure substitution bias piece (bias that arises from how one defines the good).[5]

Comparing the two indexes in (1) and (2), there is one limiting, and perhaps obvious, case where the two indexes coincide and that is when taxi and ride-sharing prices are the same in both periods. In that case, $I_{t,o}^{UV} = I_{t,o}^{NC} = P_t^T / P_o^T = P_t^R / P_o^R$.

---

[3] Two studies that were successful in this regard are Reinsdorf (1993) and Greenlees and McClelland (2005). There, the issue arises when a new outlet (ride sharing in our case) is included in the sample replacing a taxi ride, at which point one must make an assumption about how much of the gap in the new and old prices can be attributed to quality. In this context, applying the polar assumptions provides bounds on where "truth" lies. However, in other contexts, such as matched model indexes, that strategy does not work: knowing that quality component of the price gap is between zero and one does not provide bounds on those indexes.

[4] This is different from the usual comparison. In both Nakamura et al. (2015) and Hausman and Leibtag (2009), for example, the BLS-type index they would use here is simply the price of taxis. This makes sense in their context. Hausman and Leibtag (2009) calculated all biases that would arise if the BLS left Walmart out of its sample entirely. However, as they note, their estimate reflects more than outlet substitution bias which is our interest.

[5] Another difference is that the expressions they develop here are at the route-level (where the bias occurs), before aggregating into a city-wide index. As seen below, doing so provides relatively simple expressions that are much simpler and intuitive.

But difference in prices is only one condition that must hold for bias to exist: bias also requires that there be shifts across merchants. This is easier to see in a restatement of the unit value index in (1) from Aizcorbe and Nestoriak (2011):

$$I_{t,o}^{UV} \quad = \quad w_o^T (P_t^T / P_o^T)\, \Delta S_{t,o}^T + \; w_o^R (\, P_t^R / P_o^R)\, \Delta S_{t,o}^R \qquad (3)$$

where the ΔS terms measure changes in the unit shares for each type of ride. For taxis, for example, $\Delta S_{t,o}^T = \frac{Q_t^T}{Q_t^T + Q_t^R} / \frac{Q_o^T}{Q_o^T + Q_o^R}$ measures the change in the taxi share of total rides from time o to time t. Comparing (3) and (2), the indexes coincide when there is no shifting across types of rides, $\Delta S_{t,o}^T = \Delta S_{t,o}^R = 1$.

To sum up, there is potential for bias when taxi and ride-sharing prices differ *and* consumers are shifting across merchants. This is interesting because it raises questions about how long this bias can persist.

Finally, we note that while our priors are that a noncomparable index will show faster price growth than the unit value index, comparing the indexes shows that this is not necessarily true. The difference in the noncomparable and unit value indexes can be written:

$$I_{t,o}^{NC} \; - \; I_{t,o}^{UV} \; = \; w_o^T (P_t^T / P_o^T)\, (1 - \; \Delta S_{t,o}^T) + \; w_o^R (\, P_t^R / P_o^R)\, (1 - \Delta S_{t,o}^R) \;\; (4)$$

Here, shifts across merchants will always cause one of the terms to be positive and the other negative so that the net effect on depends on the relative magnitudes of the terms. For example, as ride sharing diffuses, the unit share for taxis falls, $dS_{t,o}^T < \; 1$ —causing the first term to be positive—and the unit share for ride sharing rises, $dS_{t,o}^R > 1$ —causing second term to be negative. The noncomparable index will only show faster price growth if the positive effect from the first term exceeds any negative effect from the second term.

Quality-Adjusted Unit Value Index

Studies that use unit value indexes to assess outlet substitution bias (essentially treating it as the target index) have helped establish potential problems with CPI methods. But as noted in Diewert (1998) and Nakamura et al. (2015), the unit value assumption that consumers view the two services as identical is a controversial one.

A better target index would be a quality adjusted unit value that allows for quality differences across merchants (Dalen 2001; deHaan 2002; Silver 2010). A quality adjusted unit value index is very similar to the unit value index in (1) except that it allows for quality differences across merchants:[6, 7]

$$I_{o,t}^{QAUV} = \quad QAUV_t / QAUV_o \qquad (5)$$

with:

$$QAUV_t = \frac{P_t^T Q_t^T + P_t^R Q_t^R}{\lambda^T Q_t^T + \lambda^R Q_t^R}$$

Just as with unit values, the numerator for the quality adjusted unit value measures spending on both types of rides. The difference is in the denominator, where the quantities are quality-adjusted: the λ's are quality parameters that represent differences in the quality of ride sharing vs. taxi rides and are normally assumed constant, an assumption that we relax below.

Empirically implementing this approach requires that one estimate the quality parameters. Often, however, data on characteristics are not readily available, and one cannot estimate the quality parameters directly. For our empirical work, for example, we do not have data on the attributes of rides that might matter to riders: waiting time, quality of vehicle, safety issues, experience of the driver, and so on).[8]

Unit Value Index as an Upper Bound to the Quality Adjusted Unit Value Index

In these cases, the commonly used unit value index may be the only option. In this section, we derive conditions under which the unit value index provides an upper bound to the target index. First, we show that the difference in the two is an unobserved quality index so the unit value provides a bound if

[6] The resulting price index can provide measures very similar to those obtained directly from a hedonic regression. In particular, deHaan and Krsinich (2014) have shown that measuring price change with the $I_{o,t}^{QAUV}$ using quality parameters from a hedonic regression can give price indexes similar to those obtained directly from a hedonic approach. Specifically, if one uses a hedonic regression weighted with expenditure shares to estimate the quality parameters, the quality adjusted unit value index in (5) approximates a time dummy price index from a hedonic regression.

[7] This target index was first used in the outlet substitution bias context by Byrne, Kovak, and Michaels (2017), where they used an equilibrium condition to estimate the quality parameters. More often, the approach has been used to account for quality change in sectors where rapid product innovation presents measurement difficulties and, there, the λ's have been measured using predicted quality from hedonic regressions.

[8] See Shapiro (2021).

quality is increasing. Second, we argue that the diffusion of new outlets likely involves changes in consumers' assessments of quality that pushes up the quality index over time.

The difference between (1) and (2) is a quality index that tracks changes in average quality over time. Letting $s_o^{\,j}$ be merchant j's share of the rides taken at time t, $s_o^{\,j} = Q_o^j/(\,Q_o^{\,j} + Q_o^{\,k}\,)$, and λ^j be the average quality of a ride with them, then the average quality of a ride at time t is written $\bar{\lambda}_t = (s_t^{\,R}\lambda^R + s_t^{\,T}\lambda^T)$, a weighted average of the quality parameters. The quality index for ride-sharing and taxi rides is then written:[9]

$$I_{t,o}^{\,UV}/I_{t,o}{}^{QAUV} = \frac{s_t^{\,T}\lambda_t^{\,T} + s_t^{\,R}\lambda_t^{\,R}}{s_o^{\,T}\lambda_o^{T} + s_o^{R}\lambda_o^{\,R}} = \frac{\bar{\lambda}_t}{\bar{\lambda}_o} \qquad (6)$$

where unlike the usual specification for $I_{t,o}{}^{QAUV}$, we allow the quality of each type of ride to change over time.

How does diffusion occur in this market and how does that affect this quality index? The arrival of ride sharing presents a new choice to consumers. Some view the choice as superior while other continue to take taxis. Much as in a hedonic approach, we define "quality" as consumers' valuations of the different rides and "average quality" as a weighted average of these valuations over all consumers.[10] Specifically, suppose that there is one only characteristic, or attribute, *X*, that defines the "quality" of a ride and that all potential riders agree on the value of that attribute, b. But, because ride sharing is new, consumers' perceptions of the attribute, X, differ.

There are two types of riders (1 and 2). $Q_{1,t}$ of riders at time t are Type 1 riders, who believe that ride-sharing is of higher-quality than taxis; the remaining $Q_{2,t}$ riders are Type 2 buyers that believe the opposite:

Type 1: $X_1^R > X_1^T$

Type 2: $X_2^R < X_2^T$

[9] Dividing (1) by (5), the spending terms cancel and, after simplifying, one obtains (6).

[10] Although one can take a utility interpretation of a hedonic regression (e.g., deHaan and Diewert, 2017), hedonic regressions also have a constant-quality interpretation (Aizcorbe, Corrado, and Doms 2003)

Riders choose whether to take a taxi or use ride sharing by comparing quality adjusted prices. For example, a Type 1 rider chooses the ride-sharing service when their perceived quality-adjusted price of ride sharing is less than that of taxis. Dropping time subscripts for now,

$$\textit{Ride with a ride-sharing service if} \quad P^R - (X_1^R b) < P^T - (X_1^T b)$$

And we define the probability that they do so as: Pr_1^R = Pr($P^R - (X_1^R b) < P^T - (X_1^T b)$). Though the assessments of type 1 and type 2 riders are assumed fixed, prices could change and prompt them to change their choice of merchant. For a given route, the number of rides taken with ride sharing at some time period is Q^R=($Pr_1^R Q_1$+ $Pr_2^R Q_2$) and with taxis is Q^T= (($1 - Pr_1^R)Q_1$+ $(1 - Pr_2^R)Q_2$).

In our data, the ride-sharing market share increases throughout the period. With regard to prices, there are time spans where ride-sharing merchants implemented sharp price declines and that in and of itself could have generated the increases in market share. There, drops in the ride-sharing price relative to taxi prices increase the probability that consumers take ride share, reduce the probability that they would take a cab which increases the number and unit share of ride-sharing rides.

How does one reconcile increases in market share over periods where ride-sharing prices are rising? Those patterns can be explained if one allows consumers to learn about the new service over time and update their assessments of the quality of the ride. We do so by assuming that there is some true value for the ride-sharing attribute, X_*^R, that type 1 riders see right away (X_1^R= X_*^R) but type 2 riders don't (X_2^R< X_*^R)). Over time, type 2 riders revisit their evaluation periodically as they learn more about ride sharing. At that point, they either keep their previous evaluation and take a taxi, or change their evaluation to that of type 1 riders (X_1^R= X_*^R) and use ride sharing. As some of the type 2 riders switch from taxis to ride sharing, the market share for ride-sharing merchants increases despite increases in prices, consistent with what we see in the data.

This evolution also translates into increases in the perceived quality of rides. We assume that the perceived quality of taxi rides in (6) is the same for all riders and does not change: $\lambda_t^T = X_1^T$= X_2^T = X_*^T. But the perceived quality of ride-sharing rides depends on the composition of type 1 and type 2 riders:

$$\lambda_t^R = \frac{Q_{1,t} X_1^R + Q_{2,t} X_2^R}{Q_{1,t} + Q_{2,t}}$$

Because $X_1^R > X_2^R$, as type 2 riders convert to type 1 riders, the associated increases in Q_1^R and declines in Q_2^R raise the perceived quality of ride sharing ($\lambda^R \uparrow$). Moreover, as type 2 riders switch from taxi rides to ride sharing, the share of ride-sharing rides in (6) also increases: s_t^R. These increases in the perceived quality of ride sharing along with increases in the share of ride-sharing rides ensures that the overall average quality in (6) increases, thus ensuring that the unit value index is an upper bound to the quality-adjusted unit value index: $I_{t,o}^{UV} > I_{t,o}^{QAUV}$. [11]

We then define outlet substitution bias as the difference between the noncomparable index and the target index. When the unit value index is an upper bound to the target index, we can obtain a lower bound to the bias by comparing the noncomparable and unit value indexes: [12]

Toward Implementation: Defining the Good

We define the basic service provided by taxis and ride sharing as a trip from a pickup location to a drop-off location. Whether consumers view ride sharing and taxi rides as comparable depends in part on the wait times involved. Under normal circumstances, we assume the waiting times are on average roughly comparable across merchants.

However, a key feature of ride-sharing pricing requires special attention. In times of peak demand, (e.g., sporting events or concerts), there is evidence that waiting times for taxis are much longer than for ride sharing. Despite the rise in demand, taxi prices are regulated and thus are held down, leaving excess demand on the market. But ride-sharing merchants can and do adjust prices in periods of high demand. In particular, they apply dynamic pricing strategies that increase prices when demand increases. In periods and locations where demand for rides outstrips the supply of available ride-share drivers, the services apply a *surge multiplier* (increases the price of the ride) to incentivize drivers to take rides, which in turn increases the supply of drivers.

---

[11] To see this, we show that the following expression is positive when s^R and λ^R increase from time o to time t: $s_t^T\lambda_t^T + s_t^R\lambda_t^R$ - $s_o^T\lambda_o^T - s_o^R\lambda_o^R$. Using $s_t^T = 1 - s_t^R$ to substitute out s_t^T and simplifying yields: $s_t^R(\lambda_t^R - \lambda^T) - s_o^R(\lambda_o^R - \lambda^T)$, which is positive when $s_t^R > s_o^R$ and $\lambda_t^R > \lambda_o^R$.

[12] : $I_{t,o}^{NC} - I_{t,o}^{UV}$ = Y but the unit value index is an upper bound to the target index: $I_{t,o}^{UV} = I_{t,o}^{QAUV} + X$ with X>0.

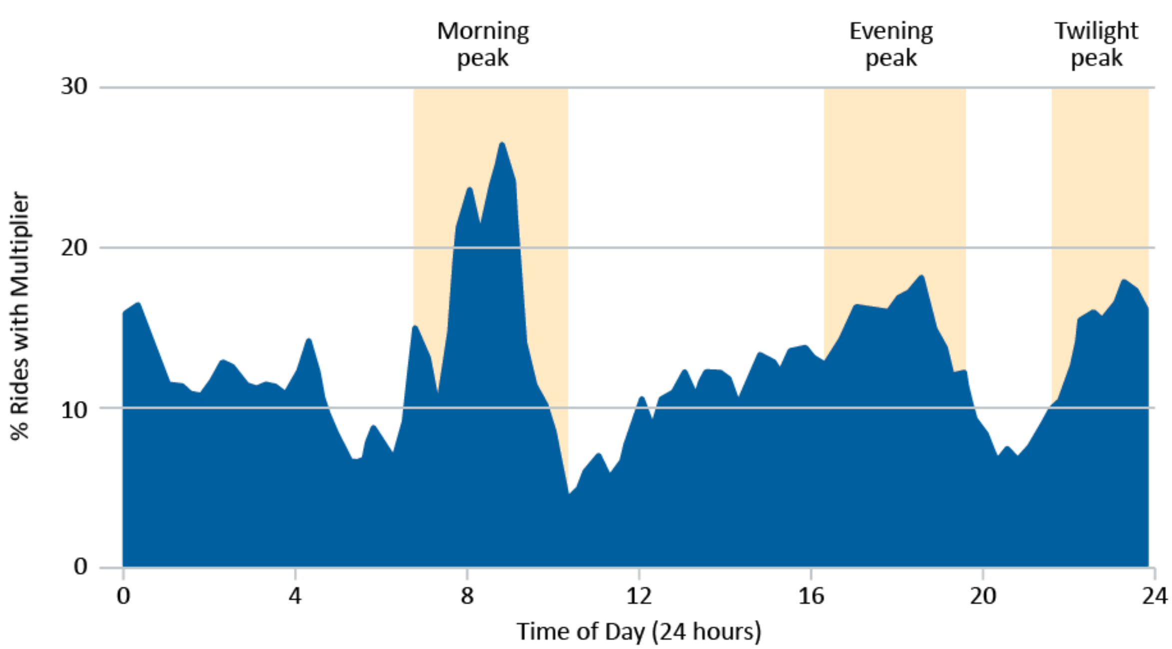
So $I_{t,o}^{NC} - [I_{t,o}^{QAUV}$ + X] = Y, or $I_{t,o}^{NC} - I_{t,o}^{QAUV}$ - X= Y or $I_{t,o}^{NC} - I_{t,o}^{QAUV}$ = Y+X . So, the observed Y is a lower bound for (smaller than) the true bias.

> "In the event that there are relatively more riders than driver partners such that the availability of driver partners is limited and the wait time for a ride is high or no rides are available, Uber employs a "surge pricing" algorithm to equilibrate supply and demand. The algorithm assigns a simple "multiplier" that multiplies the standard fare in order to derive the "surged" fare. The surge multiplier is presented to a rider in the app, and the rider must acknowledge the higher price before a request is sent to nearby drivers." (Cohen et al. 2016)

These multipliers can be quite large, often around 1.5x the normal fare but occasionally rising to 5x the normal fare. And, as shown in figure 1, periods where ride-sharing services tend to use surge multipliers are not confined to special events and also occur during other periods of high demand, such as rush hour periods.

We assume that the resulting increase in waiting times for taxis relative to ride sharing during these periods leads riders to view ride sharing and taxi rides as different goods (noncomparable) and use a superlative index (not a unit value) to measure price change. For non-surge periods, we assume waiting times are roughly comparable, and thus rides are more like broadly comparable services and the QAUV is the relevant target index.

Figure 1. Percent of Rides with Multiplier, by Time of Day



III. Data

This section describes the data that we used to represent routes and to identify periods of peak demand (i.e, where ride sharing applied surge multipliers). We combine near-census data from the New York City Taxi Limousine Commission (TLC), supplement with a sample of consumer email receipts obtained from Rakuten Intelligence ("email receipts").[13] As explained below, the TLC data allow us to construct post-stratification weights for the sample to ensure representativeness.

The TLC publishes full market, ride-level data containing high-resolution detail on pickup time and location for all rides of all types. While the time dimension provides detail to the second, location is recorded using NYC taxi zones—an aggregate unit of geography—to anonymize riders' identifiable information. Over the entire study period (2015 through 2017), approximately 309.1 million rides with the traditional yellow and green taxis sample ("TLC taxi data")[14] and an additional 213.1 million rides with ride-sharing companies (e.g. Uber and Lyft) are seen in the for-hire vehicles data ("TLC FHV data").

Table 1. Qualities of Three Taxi Transportation Market Data Sets for NYC

Data set	Source	Coverage	Price data	Geography	Dynamic pricing status	Type of ride
Yellow and Green Taxis (Taxis)	NYC TLC	All traditional	Yes	Taxi zones	No	Operator only
For-Hire Vehicles (FHV)	NYC TLC	Ride share and other FHV	No	Taxi zones	No	Operator only
Email receipts	Rakuten Intelligence	Approx. 3% of ride share	Yes	Zip code + 4	Yes	Operator, service, etc.

While the Taxi data also contain dropoff location and prices, the FHV data do not. We thus supplement the full market data with a sample of consumer email receipts obtained from Rakuten Intelligence ("email receipts"). Information on routes for taxi rides is provided directly from the TLC data. For rides

[13] www.rakutenintelligence.com

[14] Other for-hire vehicle services also operate, including Juno and Via as well as locally owned car services. Uber and Lyft comprise most of the ride volume.

taken with ride-sharing services, geo-processing techniques are applied to information on pickup and dropoff locations in the Rakuten sample to assign those rides to NYC taxi zones.

Our email receipt data provide information to identify when dynamic pricing periods are effect. The 15 million email receipts account for 3.2 percent of all ride-share trips (based on 2017). While this is a large sample, the sample may become thin when subdivided across every minute of every day over the course of a year and across taxi zones. Thus, the time unit of analysis needs to be small enough so the effect of price changes can be measured (temporal specificity), but large enough to overlap with other FHV rides (coverage). We then tested time intervals of 5, 10, 15, 20, 30, 60, and 120 minutes, settling on an interval of 30 minutes that maximized coverage with the FHV universe while preserving temporal specificity. Based on the half-hour increment, every trip (both ride sharing and taxi) is assigned one of these dynamic pricing flags:

- *Dynamic pricing in effect*. For a given 30-minute period in a taxi zone, at least one email receipt was priced with a dynamic pricing multiplier greater than one.
- *Dynamic pricing not in effect*. For a given 30-minute period in a taxi zone, all receipts were priced with a dynamic pricing multiplier of one.
- *No information observed*. For a given 30-minute period in a taxi zone, no email receipts were available.

These pricing flags are then applied to the full taxi ride population by taxi zone and 30-minute interval, finding that approximately 69 percent of all FHV rides and 81 percent of taxi rides, making up over 90 percent of spending, coincide with periods when pricing status is observed. As detailed below, we develop a set of benchmark price indexes using cells for which we observe surge status and then develop a reweighting strategy to assess the robustness of these indexes to the exclusion of the other rides.

We can exploit the full FHV population to construct post-stratified weights for the Rakuten data. While the TLC taxi data are self-weighted, the email receipts can be weighted by the FHV data that are aggregated along the same dimensions. Specifically, for ride-share rides, we obtain monthly population counts, N_{ijk}, from the NYC data for the following strata: pickup location, *i*; surge status, *j;* and merchant type, *k* (for Uber vs. Lyft). We then tabulate counts from the Rakuten data for the same strata, n_{ijk} and form the weights $w_{ijk} = N_{ijk}/n_{ijk}$. Each weight is interpreted as the number of rides in the population represented by each ride in the Rakuten sample and is used to scale up counts in the sample to those in the population. For example, if revenue for a particular stratum in the sample is r_{ijk}, we estimate the attendant population revenue, R_{ijk}, as: $R_{ijk} = (N_{ijk}/n_{ijk})r_{ijk}$.

We construct a set of benchmark weights whose target population is the 139.8 million ride-sharing rides and 263.4 traditional rides that occurred in time periods where we observe dynamic pricing, namely surge and non-surge status.

We use two alternative sets of weights to assess the potential impact of the strata for which we do not observe whether dynamic pricing is in effect.

1. *All non-surge*: We construct weights under the assumption that all the rides with missing surge status occurred in non-surge periods. Let X_{ik} be the number of rides in the population for which we do not observe surge status in pickup location i for merchant type k. Then the alternative weight for when i = non-surge is $w_{ijk} = (N_{ijk} + X_{ijk}/n_{ijk})$ and that for when i = surge is $w_{ijk} = N_{ijk}/n_{ijk}$. Note that weights are constructed not just for the ride share rides (k = FHV) but also rides in traditional taxis (k=taxi). Reweighting the data in this way places a larger weight on the (positive) bias from surge strata.
2. *All surge*: The polar assumption treats the rides for which we do not observe surge status as if they took place during surge periods. The weights in this case are $w_{ijk} = (N_{ijk} + X_{ijk}/n_{ijk})$ for i = surge and $w_{ijk} = N_{ijk}/n_{ijk}$ for i = non-surge. Reweighting the data in this way places a larger weight on the (negative) bias from non-surge strata.

Because using the non-surge assumption increases the relative importance of strata with positive bias and the surge assumption increases that of strata with negative bias, constructing price indexes using these weights provides bounds on the true bias.

With the geographic and time units of analysis harmonized, we apply outlier filters to eliminate illogical and anomalous records from the email receipt and TLC taxi data sets. These filters include restrictions on total cost (less than $300 per ride), time duration (less than three hours), distance traveled (less than 100 miles), speed (less than 50 mph). The trip-level records are then aggregated by month, year, dynamic pricing status, pickup taxi zone, merchant (e.g., yellow, green, ride-share company) and service type (e.g., standard, premium, group).

Finally, we make the obvious point that there can only be bias in routes that are serviced by both merchants. The data show that 90 percent of spending in our sample occurs in routes serviced by both.

IV. Results

We begin our analysis using only observations for which we observe surge status, which we call our sample, first focusing on periods non-surge periods—where we view the quality adjusted unit value index as the target—and then folding in rides known to have occurred in surge periods—where a superlative index like the Fisher is the relevant target. We, then show that our results are robust to the treatment of observations where we do not observe surge status.

Observed Non-Surge Observations

Recall that there is potential for outlet substitution bias when two conditions are met: taxi and ride-sharing prices differ and consumers are shifting across merchants. Figure 2 below checks the first condition and shows riders typically pay less for ride-sharing rides than for taxis. The plot compares prices on taxis vs. ride sharing on the same route (each point in the plot is a route). Most of the points lie below and to the right of the line of equality (diagonal) showing that ride-sharing prices are typically lower than taxi prices in non-surge periods.

Figure 2. Comparison of Taxi and Ride-Sharing Prices During Non-Surge

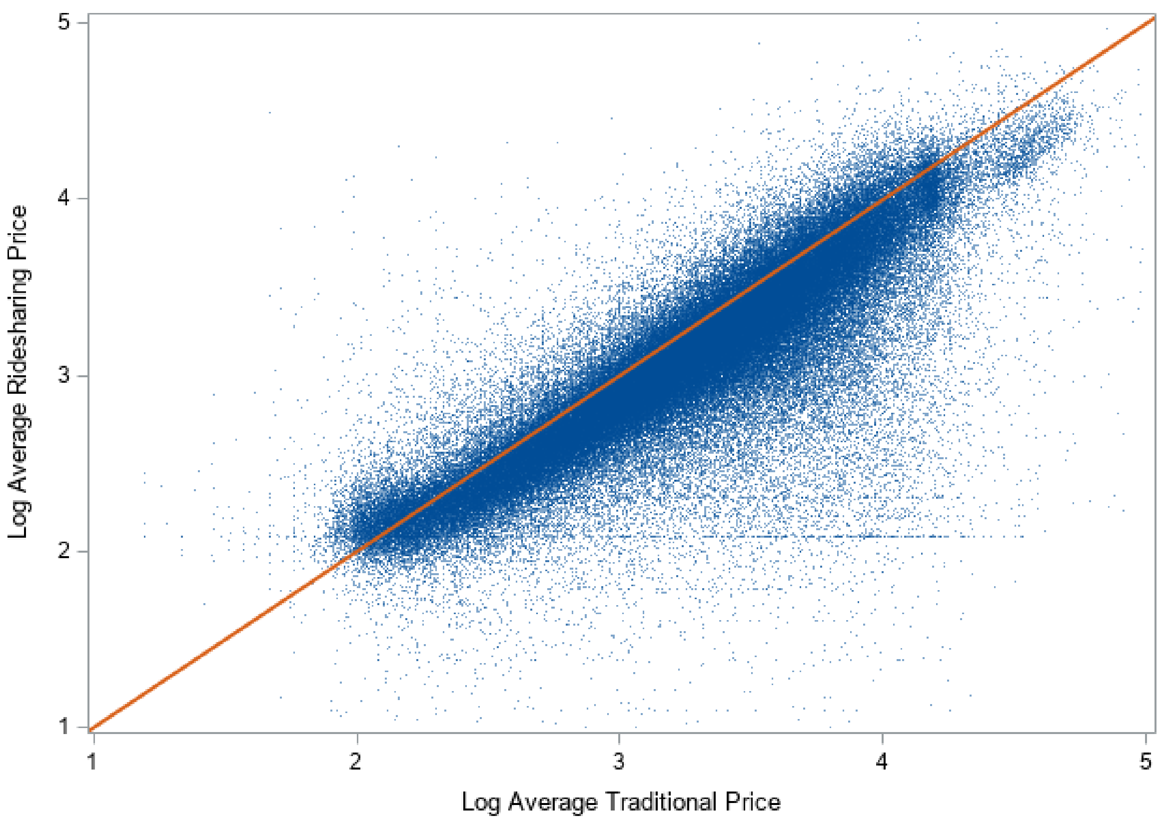
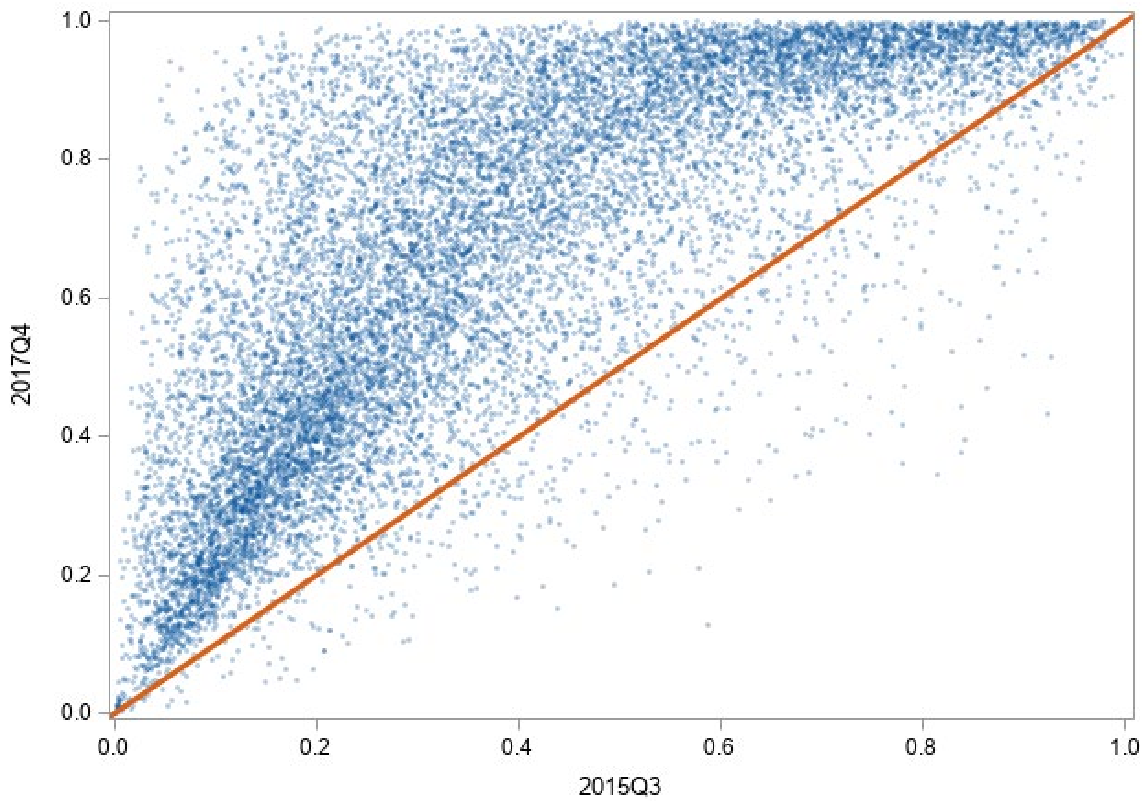


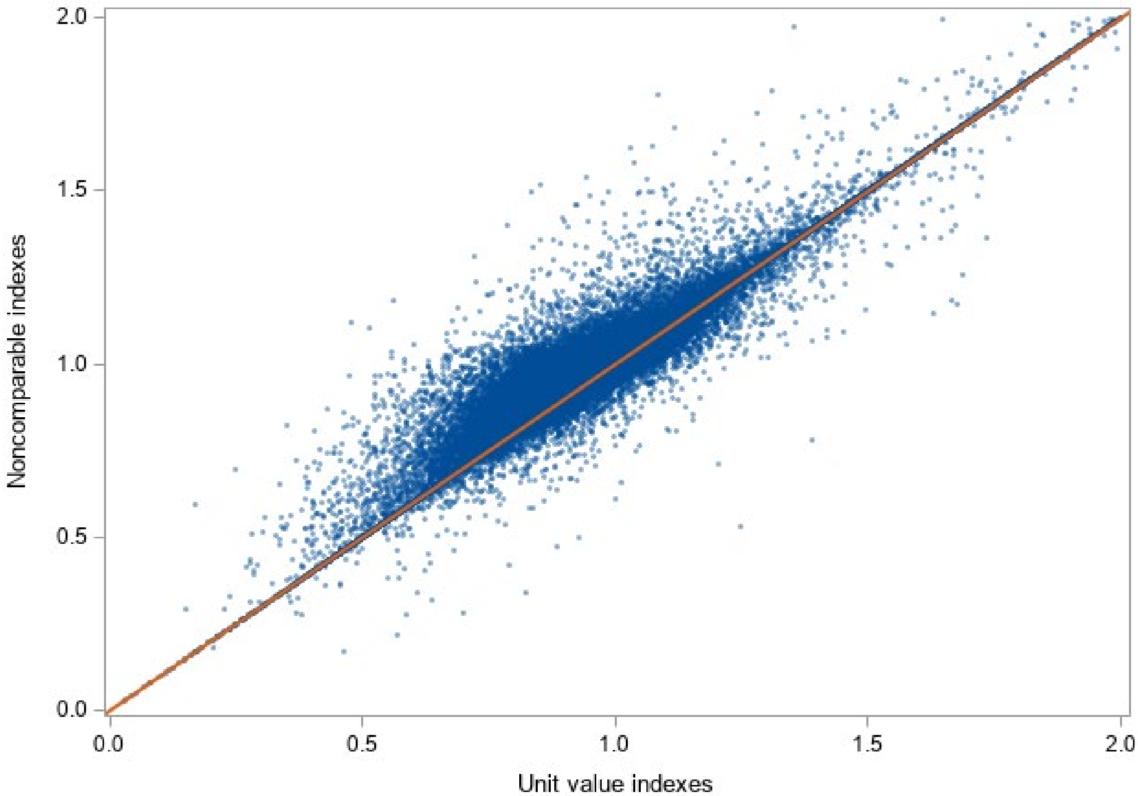
Figure 3. Comparison of Unit Market Shares for Ride Sharing in 2017Q4 (Vertical Axis) vs 2015Q3 (Horizontal Axis)



At the same time, the data are consistent with the notion that there is shifting toward ride sharing in most routes. In figure 3, each point is again a route, and unit market shares in the last quarter of our sample (measured on the vertical axis) were typically higher than the share in the first quarter of our sample (on the horizontal axis).

The resulting route-level price indexes are plotted in figure 4, where growth in the noncomparable index (calculated using equation (2) and shown on the vertical axis) typically exceeds that in the unit value index (calculated using equation (1)): most of the points lie above and to the left of the line of equality.

Figure 4. Comparison of Route-Level Price Indexes



We take averages of these route-level price changes to obtain aggregate measures for New York City. To do so, we use fixed base rather than chained indexes because chained indexes are called for in sectors marked by high turnover in order to ensure that new and existing goods are properly accounted for. That is not the case here: ride sharing entered the market in 2011 and was serving most routes to some extent by the beginning of our sample period.

Table 3 shows these averages using both Laspeyres and Fisher index formulas. While we would not have expected any traditional substitution bias issues (riders typically don't change routes when the price of a ride on some other route fell), the composition of rides across routes apparently changed enough that there are nontrivial differences in aggregates that use different formulas.

Using the Fisher index, the noncomparable index rises at a 1.7 percent compound annual growth rate over our sample period, 0.6 percentage point faster than the unit value index. We were surprised by the magnitude of this difference. Uber entered NYC in 2011 and one might have thought the diffusion process would have been over in a few years but, instead, ride sharing continued to gain market share (units) in our sample period: from 40 percent in 2015 to 70 percent by the end of 2017.

Table 2. Noncomparable and Unit Value Indexes for Non-Surge Observations, 2013Q3 to 2017Q4 (Compound Annual Growth Rates)

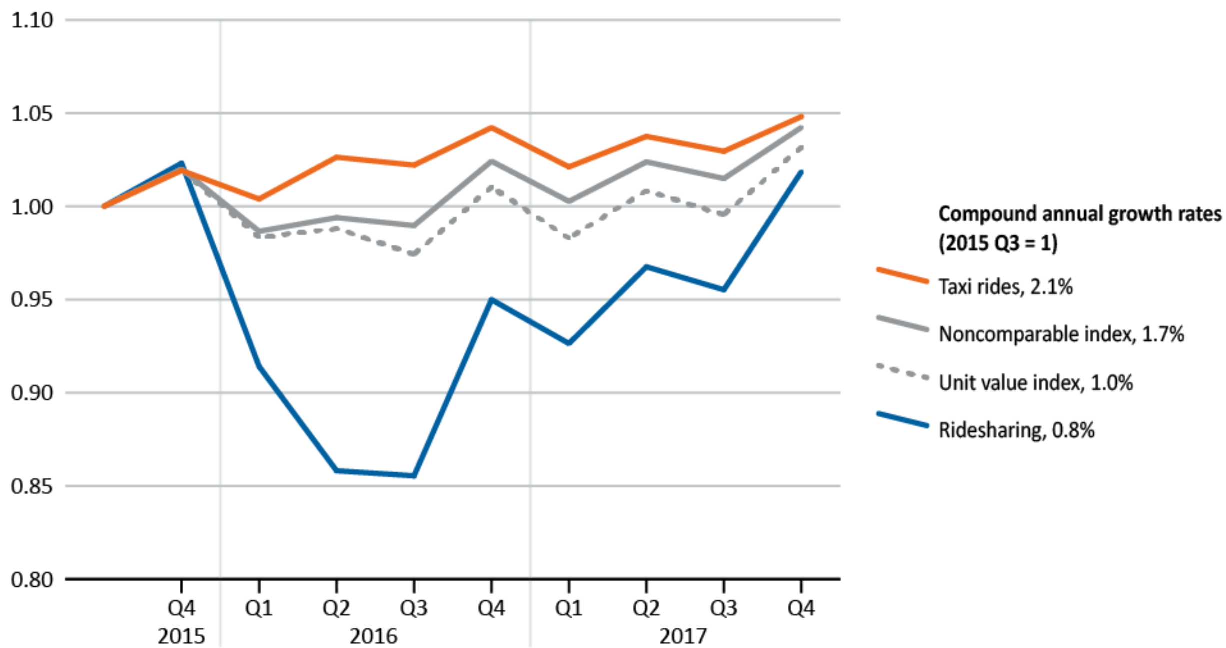
Route-level index	Formula for aggregation over routes	
	Laspeyres	Fisher
Noncomparable	1.8	1.7
Unit value	1.4	1.1
Difference	.4	.6
Memo: Taxi only	2.1	2.1

For completeness, we also look at another comparison that has been done in the literature: an index of taxi prices vs the unit value index. This comparison gives an estimate of distortions in the CPI when the new merchant is omitted from the index entirely, as happens in official statistics when there are lags in bringing in the new merchant into the sample. The index of taxi prices rises 2.1 percent over this period, 1 percentage point faster than the unit value. We attribute 0.6 percentage point of that gap to outlet substitution bias and the remaining 0.4 percentage point to the effect of omitting ride sharing from the index altogether. This suggest that lags in bringing in new merchants could have important numerical implications for price measurement.

The differences in the taxi and noncomparable indexes arise because the taxi vs. ride-sharing price patterns are very different in our sample. Taxi prices are regulated and, as seen in the gold line in figure 5, grew about 2 percent per year over this period. In contrast, ride-sharing prices (the dotted blue line) fell sharply early in the period and show substantial increases in subsequent quarters.

The patterns in the ride-sharing prices reflect well-publicized changes in their pricing strategies over this period. For example, the large drop in Uber prices in the early half of 2016 followed the service's announcement that it would cut prices 15 percent to increase its market share. The large jump in the end of 2016 coincides with a shift to an "upfront pricing" strategy announced mid-year. Our data are consistent with allegations in the news media that this shift in strategy allowed ride-sharing companies to mask price increases in the latter half of 2016. And, finally, another increase in 2017Q4 followed an announced shift in pricing that would allow ride-sharing companies to charge premiums for higher-demand routes. Previously, ride-sharing pricing reflected only distance and duration of the trip.

Figure 5. Alternative Fixed-Base Fisher Price Indexes During Non-Surge Periods, 2015Q3 to 2017Q4



Folding in Surge Observations and the Preferred Index

While rides taken over some route during non-surge periods are arguably very similar, we argued that taxi vs. ride-sharing rides are very different during surge periods and, thus, require a different treatment. In particular, we argue that, during surge periods, consumers view them as noncomparable goods because of the long waiting times associated with taxi rides. In that case, there is no bias associated with rides during surge periods—both indexes will use a noncomparable index for those rides. We combine that index with our unit value index from the non-surge observations to obtain an upper bound on overall price change in this sector and a lower bound for outlet substitution bias.

The table below shows these calculations. A noncomparable index over both surge and non-surge periods grows at a 1.6 percent compound annual growth rate over our sample period, an average of the 1.7 percent and 1.3 percent rates during non-surge and surge periods, respectively. The preferred index shows slower growth at a 1.1 percent annual rate, held down by the slow growth of the unit value index used over non-surge observations. The difference in the two indexes is 0.5 percentage point, which we interpret as a lower bound to the "true" underlying outlet substitution bias.

Table 3. Noncomparable and Preferred Indexes, 2013Q3 to 2017Q4 (Compound Annual Growth Rates)

Surge status	Noncomparable index (1)		Preferred index (2)		Difference (1) – (2)
Non-surge	Fisher	1.7	Unit value	1.1	.6
Surge	Fisher	1.3	Fisher	1.3	.0
Fisher Aggregate		1.6		1.1	.5

Robustness

The sample we used for these results exclude observations for rides where we cannot observe surge status because the Rakuten sample did not contain prices for that route in the relevant time period. As discussed in the data section, we can assess the potential importance of excluding these observations by exploiting the fact surge prices are substantially higher than taxi prices and yield bias estimates at the route level that are lower than those when using non-surge observations. So, we calculate two alternative indexes. One assumes all the excluded observations were, in fact, rides that occurred during surge periods and recalculate the sample weights to increase the importance of surge observations. This will yield bias estimates that are lower than those reported in table 3, and comparing the magnitude of the biases provides a way to check the robustness of excluding those observations from our sample. Similarly, assuming that all the excluded observations instead occurred during non-surge periods will yield bias estimates that could be greater than those reported in table 3.

Table 4 shows that the calculated bias changes very little when we add the excluded observations under these polar assumptions. Growth rates are within 0.1 percentage point of each other, regardless of the treatment of observations with missing prices.

Table 4. Noncomparable and Preferred Indexes, 2013Q3 to 2017Q4 (Compound Annual Growth Rates)

Treatment of observations with missing prices	Noncomparable index (1)	Preferred index (2)	Difference (1) – (2)
Assumed all surge	1.5%	1.1%	.4%
Excluded from indexes	1.6%	1.1%	.5%
Assumed all non-surge	1.6%	1.0%	.6%

V. Discussion and Conclusions

We have provided new expressions that emphasize the important role of diffusion in generating outlet substitution bias. Our empirical application shows that the diffusion process associated with the entry of new merchants can be quite long—in our case, ride sharing continued to gain market share through 2017, long after Uber's original entry in 2011. This doesn't seem related to supply constraints—bringing in new ride-sharing drivers is certainly more flexible than opening new stores, for example. Instead, the slow rate of diffusion might have more to do with how long it took some potential riders to warm up to ride sharing. Following the common practice of comparing BLS-type indexes to a unit value index shows that the associated bias can be large, 0.5 percentage point per year.

We also provide a simple model of diffusion and find conditions under which the bias calculated using a unit value index as a target may be viewed as a lower bound for the bias relative to a quality adjusted unit value index. Specifically, this occurs when diffusion is driven by heterogeneous consumers whose assessment of the quality of the new service increases over time, something consistent with increases in ride-sharing market shares even in periods of price increases. We, thus, conclude that outlet substitution bias for this market is at least 0.5 percentage point per year.

References

Aizcorbe, Ana. 2014. *A Practical Guide to Price Indexes and Hedonic Techniques,* Oxford University Press.

Aizcorbe, Ana. 2006. "Why Did Semiconductor Price Indexes Fall So Fast in the 1990s? A Decomposition," *Economic Inquiry*, 44(3):485–496, July.

Aizcorbe, Ana, Carol Corrado, and Mark Doms. 2003. "When Do Matched-Model and Hedonic Techniques Yield Similar Measures?" In Working Papers in Applied Economic Theory: Federal Reserve Bank of San Francisco.

Aizcorbe, Ana and Tina Highfill. 2018. "Price Indexes for US Medical Care Spending, 1980–2006," *Review of Income and Wealth*, Review of Income and Wealth - Wiley Online Library.

Aizcorbe, Ana and Nicole Nestoriak. 2011. "Changing mix of medical care services: Stylized facts and implications for price indexes," *Journal of Health Economics* 30(3): 568–574.

Berndt, Ernst R. 1996. "The Measurement of Quality Change." In *The Practice of Econometrics*. Reading: Addison-Wesley Publishing Company.

Berndt, Ernst R., Iain M. Cockburn, and Zvi Griliches. 1996. "Pharmaceutical Innovations and Market Dynamics: Tracking Effects on Price Indexes for Antidepressant Drugs," *Brookings Papers on Economic Activity: Microeconomics* 2: 133–88.

Byrne, D., K. Kovak, and R. Michaels. 2017. "Quality-Adjusted Price Measurement: A New Approach with Evidence from Semiconductors," *The Review of Economics and Statistics*, MIT Press, vol. 99(2), pages 330–342, May.

Cohen, P., R. Hahn, J. Hall, S. Levitt, and R. Metcalfe. 2016. "Using Big Data to Estimate Consumer Surplus: The Case of Uber," NBER Working Paper 22627 http://www.nber.org/papers/w22627

Dalén, J. 1998. "Experiments with Swedish Scanner Data," *Proceedings of the Third Meeting of the International Working Group on Price Indices*. Statistics Netherlands, Department of Statistical Methods. Voorburg, Netherlands.

Dalén, J. 2001. Statistical Targets for Price Indexes in Dynamic Universes, *Proceedings of the Sixth Meeting of the International Working Group on Price Indices*, Canberra, Australia.

deHaan, J. 2002. "Generalised Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index (CPI)," *Journal of Official Statistics*, 18(1), 61–85.

deHaan, J. 2004. "Estimating Quality-Adjusted Unit Value Indexes: Evidence from Scanner Data." Paper presented at the SSHRC International Conference on Index Theory and the Measurement of Prices and Productivity, Vancouver, Canada.

deHaan, J. 2007. “Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and Other Approaches.” Paper presented at the Seventh EMG Workshop, Sydney, Australia.

deHaan, J. 2015. “Rolling Year Time Dummy Indexes and the Choice of Splicing Method.” Research paper, Statistics Netherlands, The Hague, The Netherlands.

deHaan, J. , and W. E. Diewert. 2017. “Quality Change, Hedonic Regression and Price Index Construction,” Quality Change, Hedonic Regression and Price Index Construction (ottawagroup.org)

deHaan, J. , and F. Krsinich. 2014. “Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes,” *Journal of Business & Economic Statistics*, 32(3):341–358

Diewert, W. Erwin. 1998. “Index Number Issues in the Consumer Price Index,” *Journal of Economic Perspectives,* 12(1): 47-58.

Diewert, W. Erwin. 1995/2012. “Axiomatic and Economic Approaches to Elementary Price Indexes,” in W. Erwin Diewert, Bert M. Balk, Dennis Fixler, Kevin J. Fox, and Alice O. Nakamura (eds.), *Price and Productivity Measurement: Index Number Theory* 6, Trafford Press.

Feenstra, Robert C. 1997. "Generics and New Goods in Pharmaceutical Price Indexes: Comment," *American Economic Review* 87: 760–67.

Feenstra, Robert C. 1994. "New Product Varieties and the Measurement of International Prices," *American Economic Review* 84: 157–77.

Fisher, Franklin M. and Zvi Griliches. 1995. "Aggregate Price Indices, New Goods, and Generics," Quarterly *Journal of Economics* 110: 229–44.

Fisher, Franklin M. and Karl Shell. 1972. "The Economic Theory of Price Indices," Elsevier Monographs

Greenlees, John S., and Robert McClelland. 2011. “New Evidence on Outlet Substitution Effects in Consumer Price Index Data,” *Review of Economics and Statistics* 93 (May), 632–646. http://www.indexmeasures.ca/dc2008/papers/GreenleesMcClelland%20World%20Congress%20May%201 5.doc.

Griliches, Z. and I. Cockburn. 1994. “Generics and New Goods in Pharmaceutical Price Indexes,” *The American Economic Review,* 84(5), 1213–1232.

Hausman, J., and E. Leibtag. 2007. Consumer benefits from increased competition in shopping outlets: Measuring the effect of Wal-Mart. *Journal of Applied Econometrics*, 22(7), 1157–1177

Hatzius, J. 2017. “US Daily: The Internet and Inflation: How Big is the Amazon Effect?,” Goldman Sachs Economic Research, August 2.

Hausman, Jerry, and Ephraim Leibtag. 2009. "CPI Bias from Supercenters: Does BLS Know that Wal-Mart Exists?," *Price Index Concepts and Measurement*, Studies in Income and Wealth, 70, W. Erwin Diewert, John S. Greenlees, and Charles R. Hulten (editors). Chicago and London: University of Chicago Press, 203–231.

Houseman, Susan, Christopher Kurz, Paul Lengermann, and Benjamin Mande. 2011. "Offshoring Bias in U.S. Manufacturing," *Journal of Economic Perspectives* 25(2), Sprint 2011, 111–132.

ILO/IMF/OECD/UNECE/Eurostat/The World Bank. 2004. Consumer Price Index Manual: Theory and Practice. ILO Publications, Geneva.

Ivancic, Lorraine and Kevin J. Fox. 2014. "Understanding Price Variation Across Stores and Supermarket Chains: Some Implications for CPI Aggregation Methods," *Review of Income and Wealth* 59 (4) 629–647.

Kurz, C., P. Lengermann, and B Mandel. 2019. "New economy, same old consumption? E-commerce and Implications for Economic Measurement." Paper presented at Brookings Productivity Conference.

Moulton, B. 2018. "The measurement of output, prices, and productivity: What's changed since the Boskin Commission?" (Presented at Brookings Productivity Conference).

Nakamura, A., W.E Diewert, J.S. Greenlees, L.I. Nakamura, and M.B. Reinsdorf. 2015. "Sourcing Substitution and Related Price Index Biases," Chapter 2 (21–88) in: *Measuring Globalization: Better Trade Statistics for Better Policy 1, Biases to Price, Output, and Productivity Statistics from Trade,* Susan N. Houseman and Michael Mandel, eds. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Reinsdorf, Marshall. 1993. "The Effect of Outlet Price Differentials in the U.S. Consumer Price Index," *Price* Measurements *and their Uses.* Studies in Income and Wealth 57, Murray F. Foss, Marilyn E. Manser, Allan H. Young (editors). Chicago and London: University of Chicago Press, 227–254.

Reinsdorf, Marshall. 1994. "Price Dispersion, Seller Substitution and the US CPI," Working Paper 252, Bureau of Labor Statistics, Washington D.C., March.

Reinsdorf, Marshall. 1999a. "Formula Bias and Within-Stratum Substitution Bias in the US CPI," *Review of Economics and Statistics* 58 (2), 175–187.

Reinsdorf, Marshall and Brent R. Moulton. 1997. "The Construction of basic components of cost-of-living indexes," *The Economics of New Goods,* Bresnahan and Gordon (eds.), NBER Studies in Income and Wealth 58, University of Chicago Press, Chicago, 397–436.

Reinsdorf, Marshall and Robert E. Yuskavage. 2014. "Effects of Offshoring and Country Substitution on Measurement of Prices and Productivity," working paper, U.S. Bureau of Economic Analysis.

Shapiro, M. 2018. “Density of Demand and the Benefit of Uber,” mshapiro_jmp.pdf (shapiromh.com).

Silver, M. 2010. “The Wrong and Rights of Unit Value Indices,” *Review of Income and Wealth*, 56, S206–23.

Silver, M. 2011. “An Index Number Formula Problem: The Aggregation of Broadly Comparable Items,” *Journal of Official Statistic*s 27, 553–567.

von Auer, Ludwig. 2014. “The Generalized Unit Value Index Family,” *Review of Income and* Wealth Series 60 (4) December.

White, Alan. (1999). Measurement Biases in Consumer Price Indexes. *International Statistical Review* 67 (3), 301-325.