Received: 1 August 2018 | Revised: 13 March 2020 | Accepted: 20 March 2020

DOI: 10.1002/joom.1090

WILEY

RESEARCH ARTICLE

On empirically estimating bullwhip effects: Measurement, aggregation, and impact

Yuliang Yao[1] | **Yongrui Duan**[2] | **Jiazhen Huo**[2]

[1]College of Business, Lehigh University, Bethlehem, Pennsylvania

[2]School of Economics and Management, Tongji University, Shanghai, China

Correspondence
Yongrui Duan, School of Economics and Management, Tongji University, Shanghai 200092, China.
Email: yrduan@tongji.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 71771179, 71532015

Handling Editor: Gregory Heim

Abstract

Empirical estimation of the bullwhip effect poses several challenges, although the bullwhip effect has been well studied in modeling papers. Using a dataset from a large supermarket chain, we estimate the product level bullwhip effect using various methods, analyze consequences of its different measurements and aggregations, and examine its impact on supply chain performance in terms of inventory ratio and stockouts. We have three major findings. (a) Bullwhip effect estimates exhibit different magnitudes dependent on how they are measured. The material bullwhip effect is greater in magnitude than the information bullwhip effect in our data, where demand correlations are sufficiently low. (b) The aggregated bullwhip effect ratios by store and by time are lower than the disaggregated bullwhip effect ratios, indicating that the aggregated bullwhip effect ratios underestimate the bullwhip effect. The aggregated bullwhip effect ratios by product are lower than the disaggregated bullwhip effect ratios, indicating the bullwhip effect is not as strong as theory predicts due to order pooling. (c) The bullwhip effect is associated with poor supply chain performance, as measured by elevated inventory ratio and stockouts. However, if the bullwhip effect is measured inaccurately, these benefits can be underestimated as much as 75% for inventory and 25% for stockouts.

KEYWORDS

bullwhip effect, empirical analysis, supply chain management

1 | INTRODUCTION

The rapid development of information technology has made collecting, storing, and analyzing large-scale data feasible. As a result, companies today are competing against each other on data analytics, an important source of competitive advantages. According to an industry survey, 53% of firms used data analytics in 2017, up from 15% in 2015 (Forbes, 2017). Trade press has touted some success stories from using data analytics. For example, Wal-Mart and Target have been using data analytics to direct their retail business strategies (Forbes, 2018). However, in supply chain management (SCM), managers have not applied data analytics that could radically transform SCM practices to the same extent (McKinsey, 2016), perhaps in part due to the limited research in SCM analytics.

Research shows an increasing trend of using data analytics in operations management (OM) and SCM, for example, to detect on-shelf stockouts using point-of-sale (POS) data (Montoya & Gonzalez, 2019), support hurricane inventory management decisions using consumer demand data (Morrice, Cronin, Tanrisever, & Butler, 2016), and quantify the value of information sharing using sales information (Cui, Allon, Bassamboo, &

Van Mieghem, 2015). A recent trend in supply chain analytics research is to empirically examine the bullwhip effect (e.g., Bray & Mendelson, 2012; Cachon, Randall, & Schmidt, 2007; Chen & Lee, 2012), a topic well studied by analytical models. The bullwhip effect is defined as "the phenomenon where orders to the supplier tend to have larger variance than sales to the buyer (i.e., demand distortion), and the distortion propagates upstream in an amplified form" (Lee, Padmanabhan, & Whang, 1997b, p. 546). Scholarly and trade press publications have documented the bullwhip effect at Merloni Elettrodomestici, PSS/World Medical, Sara Lee Household & Body Care (Wheatley, 2004), Philips Semiconductors (de Kok et al., 2005), European grocery supply chains (Fransoo & Wouters, 2000; Holmström, 1997), car manufacturing (Klug, 2013),

TABLE 1 Summary of key empirical literature on bullwhip effect

Paper	Data	Measure	Findings	BW range
Fransoo and Wouters (2000)	Daily data of four convenience food companies in Netherlands from Mar 23 to Jun 5, 1998	Ratio of the coefficient of variation (CV) of demand generated by an echelon to the CV of demand received by this echelon	The bullwhip effect exists at different supply chain echelons	2.449 ~ 4.796 Mean: 4.495
Lai (2005)	Monthly data on 3,754 stock keeping units (SKUs) from a Spain supermarket chain from Jan 1990 to May 1992	Ratio of sales variance to supply variance	There is a significant amount of bullwhip effect and that order batching is a major driver for the effect	0 ~ 4,356.00
Cachon et al. (2007)	Monthly, industry level data in the United States during Jan 1992 to Feb 2006	Ratio and difference of production variance to sales variance	Wholesale industries exhibit the bullwhip effect, whereas retail industries do not, and that seasonality contributes a large portion to the bullwhip effect	Raw ratio: 0.15 ~ 4.15 Raw difference: −0.3519 ~ 0.0839 Seasonally adjusted ratio: 0.21 ~ 9.18 Seasonally adjusted difference: −0.1602 ~ 0.0657
Bray and Mendelson (2012)	Quarterly data of 4,689 U.S. public firms during 1974–2008	Percent of total demand variance (difference of order variance and demand variance decided by demand variance)	About two-thirds of firms experience the bullwhip effect, and that demand signals contribute to the bullwhip effect	−26.55 ~ 29.73% Mean: 15.81%
Chen and Lee (2012)	Weekly, SKU level data from a European retail store during a 1-year period	Ratio of shipment variance to sales variance	The bullwhip effect ratios decrease with the estimation time window	Weekly: 1.31 ~ 3.04 Biweekly: 1.25 ~ 2.68 Four-weeks: 1.08 ~ 2.03
Shan, Yang, Yang, and Zhang (2014)	Quarterly data of 1,200 public firms in China from 2002 Q1 to 2009 Q2	Ratio of production variance to sales variance	More than two-thirds of the companies exhibit the bullwhip effect	Mean: 1.26 *SD*: 0.60
Bray and Mendelson (2015)	Monthly data of 162 car models produced by 20 auto manufacturers from 1985 to 2013	Ratio of production variance to sales variance	Among 162 car models, on average, production is 220% as variable as sales	Mean: 2.20
Duan, Yao, and Huo (2015)	Daily data of a large Chinese super market chain from April to October 2011	Ratio of order variance to sales variance	Bullwhip effect is prevalent and intensive, and is not only affected by their own product's but also by substitute product's price changes and stockouts	Raw ratio: Mean: 20.64 *SD*: 37.98 First differenced ratio: Mean: 33.6 *SD*: 71.17

and various U.S. industries (Cachon et al., 2007; Dooley, Yan, Mohan, & Gopalakrishnan, 2010).

The prevalence of the bullwhip effect is a general consensus among researchers. Yet, the wide range of bullwhip effect magnitudes reported in empirical studies is puzzling (see Table 1 for details). These limited empirical findings do not always agree with theoretical predictions that the bullwhip effect should be prevalent and strong (e.g., Lee et al., 1997b). Moreover, the wide range of bullwhip effect estimates implies that the bullwhip effects may be estimated in different ways and some estimation approaches do not accurately capture the actual effects. Indeed, during our data collection at a large supermarket chain in China, when we posed to the distribution center's supply chain manager the question of whether the firm suffers from the bullwhip effect, the manager answered: "*I know what bullwhip effect is, but we have never worried about it because we don't think it is as bad as one would think in our supply chain*."

This seemingly counter-intuitive observation from this manager, together with the disparities between theoretical predictions and the empirically documented wide ranges of bullwhip effect estimates, are worrisome in several ways. First, from a theoretical perspective, the wide range of bullwhip effect estimates is likely due to the challenges in empirically measuring and estimating the bullwhip effect and the fact that some assumptions in analytical models do not reflect the actual reality in businesses. For example, in analytical models, a single product in a single firm is often assumed, whereas in reality, multiple products in multiple firms or stores, which may mask the bullwhip effect, are common. Second, from a practitioner perspective, when data analysts prepare a firm's data in one way and conclude that there is a bullwhip, and then prepare the data in another way and conclude that there is no bullwhip, the data analytics become dubious. As suggested by our findings, it is clear that a data analyst may infer anything between a strong bullwhip effect, to an insignificant bullwhip effect, simply by analyzing the exact same data that has been prepared in different ways. When firms rely on doubtful data analytics to make key business decisions, it may lead to catastrophic results. Therefore, it is critical for both researchers and practitioners to understand the potential impacts and implications of data analytics in terms of estimating the bullwhip effect.

Empirical estimation of the bullwhip effect poses several unique challenges. First, the bullwhip effect is theoretically measured by comparing order variance to demand variance (Lee et al., 1997b). However, the necessary demand information and order information are typically not readily available to supply chain data analysts, as the ideal data would require the firm to track the information flow from their customers and to their suppliers. As a result, proxy information is often used to estimate bullwhip effects (see Table 1 for details). Second, the bullwhip effect is theoretically defined at the product level in a single firm. Yet, prior studies have used both disaggregated, product level data (Duan et al., 2015; Fransoo & Wouters, 2000; Lai, 2005) and aggregated data (e.g., Cachon et al., 2007 at the industry level; Bray and Mendelson, 2012 and Shan et al., 2014 at firm level) to estimate the bullwhip effect. Aggregated data may result in over- or under-estimations of the bullwhip effect (Fransoo & Wouters, 2000), leading to different conclusions (Chen & Lee, 2012). Therefore, it is of great theoretical and managerial importance for researchers and practitioners to demonstrate the aggregation at work and to illustrate how different forms of aggregation may affect the bullwhip effect estimation.

In addition to the challenges in estimating the bullwhip effect, the impact of the bullwhip effect has also been understudied in empirical literature. Prior modeling literature is generally in consensus that the bullwhip effect is associated with worse supply chain performance, such as increased inventory and increased stockouts (e.g., Lee et al., 1997b). However, the conclusion is not without question. Chen and Samroengraja (2004), using analytical modeling and numerical examples, show that supply chain costs are not necessarily reduced by strategies for dampening a bullwhip effect, suggesting that a higher bullwhip effect may not necessarily be associated with worse supply chain performance. Few papers have empirically studied the consequences, especially when the consequences may be nuanced. An exception is Mackelprang and Malhotra (2015) who show that, surprisingly, the bullwhip effect has no relationship with a firm's operating margin. They further note, "...the relationship between bullwhip and firm performance is far more nuanced and complicated than previously believed" (p. 15).

Clearly, addressing the empirical challenges of estimating the bullwhip effect and examining its impact on supply chain performance have important theoretical and practical implications. Doing so can "improve our perspective on the phenomenon" (Bray & Mendelson, 2012), thereby contributing to the literature and shedding light on improving firms' business practices in supply chain operations. The objective of our study is to take up these challenges in measuring and estimating the bullwhip effect and its performance impact through rigorous analyses using a large-scale, granular dataset. Our research question centers on what the best way is for OM/SCM researchers and data analysts to empirically estimate the bullwhip effect. Comparing all measures used in prior literature, do they accurately capture the bullwhip effect, or do they over- or underestimate it? What is the impact of the bullwhip effect and what is the

impact when it is over- or underestimated? We replicate the measures used in the literature and examine where the disparities emanate from, and also quantify the performance degradation due to the bullwhip effect and the potential impact of inaccurate estimation of the bullwhip effect on estimates of this performance degradation.

Our key findings are that: (a) bullwhip effect estimates exhibit different magnitudes dependent on how they are measured. The material bullwhip effect is greater in magnitude than the information bullwhip effect in our data, where demand correlations are sufficiently low. (b) The aggregated bullwhip effect ratios by store and by time are lower than the disaggregated bullwhip effect ratios, indicating that the aggregated bullwhip effect ratios underestimate the bullwhip effect. The aggregated bullwhip effect ratios by product across stores is lower than the disaggregated bullwhip effect ratios, indicating the bullwhip effect is not as strong as theory predicts due to order pooling. (c) No matter how a bullwhip effect is measured, the bullwhip effect is associated with poor supply chain performance, as measured by elevated inventory ratio (IR) and stockouts. A unit decrease in the bullwhip effect ratio can save a firm's inventory by $26.03 and stockouts by 0.15 days for a product during a year on average. These results suggest that, if a firm can mitigate the bullwhip effect, the firm can expect to have lower inventory costs and better service level. However, if the bullwhip effect is measured inaccurately, the benefits can be underestimated by as much as 75% for inventory and 25% for stockouts, which may mislead data analysts and managers, providing less incentive to implement measures to mitigate the bullwhip effect.

Our article makes several important contributions to the OM/SCM fields. First, our article extends the work in the OM/SCM fields by taking up the empirical challenges in estimating the magnitude and impact of the bullwhip effect. Previous studies document a wide range of bullwhip effects, likely due to the different ways in which they measure the bullwhip effect. Our findings reconcile the disparities and demonstrate that there are two scenarios when such disparities may occur. The first scenario is that, although the bullwhip effect is prevalent and intensive, when measured at different levels, it may be underestimated, even to the level that it incorrectly seems nonexistent. The second scenario is that, while theory predicts that order variance oscillates going up the supply chain (i.e., the bullwhip effect), it may not be happening in reality when order pooling from multiple firms is available, because bullwhip effect theory assumes a single product in a single firm and does not consider order pooling. Second, although there is a general consensus in bullwhip modeling literature that the bullwhip effect results in poor supply chain performance, few studies have empirically shown and quantified the consequent degradation of supply chain performance. We are among the first few attempts to empirically verify the theoretical prediction and estimate the performance impact of the bullwhip effect. Third, in addition to the contributions to the research community, our study makes important contributions to managers. We outline an implementable data analytics methodology to measure and estimate the bullwhip effect that practitioners can use in analyzing their supply chain to make informed decisions.

2 | LITERATURE REVIEW

Prior literature using analytical modeling studies the bullwhip effect (e.g., Cachon, 1999; Chatfield, Kim, Harrison, & Hayya, 2004; Lee et al., 1997b; Metters, 1997; Raghunathan, Tang, & Yue, 2017; Warburton, 2004). The general consensus is that the bullwhip effect exists and may be caused by rationing games, demand signal processing, price variation, and order batching (Lee et al., 1997b). Several ways to mitigate the bullwhip effect such as information sharing are proposed (Chen & Lee, 2009; Wu & Katok, 2006).

A few empirical studies have estimated the bullwhip effect and documented a wide range of bullwhip effect estimates. Cachon et al. (2007), using industry data from the United States, show that wholesalers exhibit an industry level bullwhip effect, whereas retailers do not. At the firm level, Bray and Mendelson (2012), using data from 4,689 public firms during the period 1974–2008, find that a majority of firms have the bullwhip effect, and that demand signals during short, midrange, and long lead times all contribute to the bullwhip effect. Using firm level data on more than 1,200 companies in China during the period 2002–2009, Shan et al. (2014) also show that a majority of the companies exhibit the bullwhip effect. At the product level, Fransoo and Wouters (2000) use data from two European supply chains and demonstrate that different supply chain echelons can have different levels of bullwhip effect. Bray and Mendelson (2015) analyze product-level data from 162 car models and find a strong bullwhip effect measured by production smoothing. Duan et al. (2015) use product-level data from a large Chinese supermarket chain and find a prevalent bullwhip effect, and that the effect is affected by a substitute product's characteristics such as price changes and stockouts. Bray et al. (2018) study one of the major driving factors of the bullwhip effect (i.e., rationing games) and, using product-level data from a supermarket chain via a structural econometric model, show the long-standing hypothesis that

rationing gaming causes a bullwhip effect. Most previous studies estimate the bullwhip effect either at the monthly or quarterly level (Bray & Mendelson, 2012, 2015; Cachon et al., 2007; Lai, 2005; Shan et al., 2014). Few studies use product-level daily data; Fransoo and Wouters (2000) analyze a limited number of products, while Duan et al. (2015) focus only on estimating drivers of the bullwhip effect.

Our article differs from empirical studies in the literature. First, we use product level and daily level data, while most prior studies use firm level or industry level data, and at monthly (or longer time period) levels. Furthermore, we are able to collect order data that enable us to test the bullwhip effect, as measured by its definition in theoretical models, and compare it against other proxy measures. Second, our context allows us to collect data from multiple retail stores and for multiple products, enabling us to aggregate the data in many different ways to measure and compare the measurement and aggregation effect on estimates of the bullwhip effect. The comparisons help us to gain a deeper understanding about why the extant literature has documented a wide range of bullwhip effect estimates and why some managers do not observe the bullwhip as strongly as theory predicts. Third, we examine a supply chain that consists of a distribution center and many stores, a setting similar to the analytical models in the literature, whereas most prior studies examine a firm setting without information about the firm's suppliers or customers. Fourth, we use a large-scale dataset that encompasses over 700,000 observations on 487 consumer products collected from a supermarket chain, which is much more than in previous empirical studies. The scale of the data and the variety of types of products increase the generalizability of our findings. Fifth, we empirically quantify the impact of the bullwhip effect on supply chain performance, which has not been done in prior empirical literature. Therefore, our study contributes to the bullwhip effect literature by narrowing the literature gaps in estimating the bullwhip effect and by quantifying the impact of the bullwhip effect at the product level. Table 1 summarizes the key relevant studies.

3 | MEASUREMENTS OF BULLWHIP EFFECT

3.1 | Material bullwhip effect versus information bullwhip effect

There are two primary ways to measure the bullwhip effect: one using shipping and sales variance, termed the *material bullwhip effect*, and the other using order and demand variance, termed the *information bullwhip effect*. To better illustrate these bullwhip effect measures, we adapt a figure from Chen and Lee (2015). As shown in Figure 1, a firm deals with both upstream and downstream parties in the supply chain. On the downstream side, the firm faces consumer demand and generates sales when the demand is realized; on the upstream side, the firm places orders to and receives shipments from its suppliers. Most modeling papers use the information bullwhip effect (i.e., the ratio of order variance to the demand variance) to measure the bullwhip effect (e.g., Cachon, 1999; Chen, Drezner, Ryan, & Simchi-Levi, 2000; Chen & Lee, 2012; Lee, Padmanabhan, & Whang, 1997a). For empirical studies, however, it is challenging to obtain data on orders and demand (Cachon et al., 2007). As a result, previous studies often use shipments to proxy orders, and use sales to proxy demand; that is, they measure the material bullwhip effect (e.g., Bray & Mendelson, 2012; Cachon et al., 2007; Lai, 2005).

The material flow and the information flow account for different levels of supply chain uncertainty. The uncertainty may come from supply shortages. If a supplier can supply a firm's orders perfectly, and the firm can satisfy the demand from its consumers perfectly, the material flow and the information flow should have the same level of uncertainty (Chen & Lee, 2015). However, this scenario rarely happens due to stockouts. It is not straightforward as to which bullwhip effect contains more uncertainty. On one hand, the information bullwhip effect takes account of both demand uncertainty

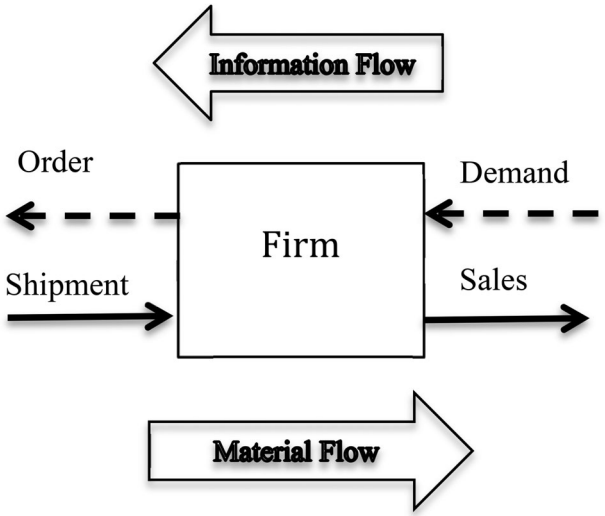

FIGURE 1 Information and material flows for a firm in a supply chain. Note: The figure is adapted from Chen and Lee (2015)

and supply uncertainty because, when a firm makes order decisions, both demand uncertainty (e.g., demand signal processing) and supply uncertainty (e.g., shortage gaming) are considered. The information bullwhip effect drives the material bullwhip effect, thus accounting for more uncertainty (Chen & Lee, 2012). On the other hand, the information bullwhip effect involves only one decision maker (i.e., the firm) who observes the demand and decides on orders. In contrast, the material bullwhip effect involves three decision effects, where sales are determined by the consumer demand and inventory, orders are determined by the firm's decisions, and shipments are determined by the suppliers' product and inventory decisions. Thus, the material bullwhip effect may account for more uncertainty as more decision effects suggest a greater level of uncertainty.

In a recent paper, Chen, Luo, and Shang (2017) specifically examine the relationship by studying discrepancies between the information bullwhip and material bullwhip effects. They find that the material bullwhip effect is always greater than the information bullwhip effect under stationary demand and ample supply. For systems with correlated (i.e., Auto regression or AR [1]) demand and with supply shortages, the relationship between the material bullwhip and the information bullwhip effect depends on the magnitude of demand correlation between the two AR(1) periods. They analytically prove that when the demand correlation is sufficiently low, the material bullwhip effect is greater than the information bullwhip effect; when the demand correlation is sufficiently high, the material bullwhip effect is smaller than the information bullwhip effect.

3.2 | Temporal aggregation of bullwhip effect

In analytical models, the variances of demand and orders are typically given through their respective distribution function assumptions. However, in empirical estimations, the sales and order data may be at different levels dependent on the level of temporal aggregation; for example, daily sales, weekly sales, or monthly sales. Chen and Lee (2012) develop a model and show that the aggregated bullwhip effect ratio temporally approaches one in the limit when the aggregated demand variance becomes sufficiently large and the aggregation period increases. They also analytically show the bullwhip effect ratio decreases monotonically under time aggregation when the ratio is greater than one at the order decision point. These theoretical results suggest that longer temporal aggregation of data generates lower bullwhip effect ratios.

3.3 | Order aggregation of bullwhip effect

Prior economics literature also has explored implications of order aggregation across products on the bullwhip effect. For example, Caplin (1985) demonstrates that, no matter what the demand correlation assumption is, aggregation still will likely preserve a bullwhip effect under (s, S) policies. Relatedly, Cachon et al. (2007) state that "whether aggregation preserves or masks the bullwhip effect or production smoothing depends on the correlation of production and demand across the unit being aggregated and on the particular causes of amplification in place." Chen and Lee (2012), using a modeling approach, investigate the aggregation over products and locations and find that product aggregation and location aggregation both can mask the bullwhip effect.

One reason that aggregation may mask or dampen the bullwhip effect is because of ordering processes among multiple firms or across multiple stages within a firm (e.g., multiple retail stores to the distribution center [DC]). The aggregated order variance from multiple firms with correlated ordering (i.e., all retailers order at the same time) is the largest, followed by that from random ordering (i.e., retailers order randomly over time), and that from balanced ordering (i.e., same retailers order each period) (Cachon, 1999; Lee et al., 1997b). There are two approaches of order aggregation for multiple products carried by multiple stores. One is aggregation over the same product across different stores, and the other is aggregation over the same store across different products. The former approach calculates the bullwhip effect faced by the distribution center (or upstream firms) in practice as the distribution center pools all orders from its retail stores. The latter approach resembles how orders are placed and shipped in practice as retail stores always place orders with multiple products for improved economy of scale in order and transportation costs. For both approaches, random ordering averages out the data peaks and bottoms, resulting in aggregated order data with smaller variation.

Through conversations with our data provider, we understand that the order processes between stores in our research setting are neither coordinated nor balanced. That is, each store makes its own ordering decisions without coordinating with any other stores, suggesting orders across stores are not correlated but rather are randomly placed, which will result in lower order variance after aggregation. A quick examination of our data shows that this is indeed the case. The average coefficients of variation of the aggregated order data series by the same product across stores and by the same store across products are 2.55 and 1.56, respectively,

whereas the average coefficients of variation of the disaggregated order series are 3.54 and 3.45, respectively. *T*-tests show that the average coefficients of variation of the aggregate order series are significantly smaller than those of the disaggregate order series. Furthermore, Figure 2 graphs the disaggregated order data for a sample product at a sample store, the aggregated order data over products in the sample store, and the aggregated order data for the sample product over all stores by the days of week (e.g., Monday, Tuesday, etc.). The aggregated series show smaller variation (i.e., smoother) than the disaggregate data series, indicating order aggregation is associated with lower order variance in our case. Therefore, the bullwhip effect is likely to be masked due to data aggregation in our research context.

4 | IMPACT OF BULLWHIP EFFECT

Modeling literature demonstrates that the bullwhip effect is detrimental to supply chain performance since it may lead to supply chain inefficiencies such as "excessive inventory investment, poor customer service, lost revenues, misguided capacity plans, ineffective transportation, and missed production schedules" (Lee et al., 1997a, p. 93). In modeling literature, supply chain performance related to the bullwhip effect is modeled mainly using inventory costs and stockouts, both of which are widely used supply chain performance metrics in empirical studies (e.g., Dong, Dresner, & Yao, 2014). In theory, when a firm faces demand with larger order variation (assuming the same mean), the firm will need to stock a greater level of inventory to maintain the same level of service, resulting in higher inventory cost (Tsay & Lovejoy, 1999). Chen and Lee (2012), through modeling, show that a firm's inventory cost is proportional to the square root of the bullwhip effect; that is, when the bullwhip effect ratio is greater than 1, the inventory cost increases.

In our empirical approach, we use IR and stockouts at the downstream stores to measure the supply chain performance. Based on theory, the bullwhip effect is upward facing; that is, the bullwhip effect is generated first at the stores, but is observed by the distribution center (in our case). There are several ways to study the performance implications of the bullwhip effect. The first way is to study the performance of the distribution center and the second way is to study the performance of the stores. We chose to focus on retail stores because the retail store's performance is more important, since the ultimate goal of a supply chain is to satisfy its end-consumer demand. For the downstream stores, the impact of the bullwhip effect is not straightforward. We posit the bullwhip effect the downstream store transmits to the distribution center upstream also hurts its own supply chain performance. The logic is that the bullwhip effect affects the upstream distribution center's supply chain planning and performance because of increased order variations, which in return will affect the downstream store's supply chain planning and performance because of worsened upstream performance. For example, due to increased order variability, the upstream distribution center will not be able to plan its inventory and production as well as before. As a result, the poor planning will affect its fulfillment performance to the

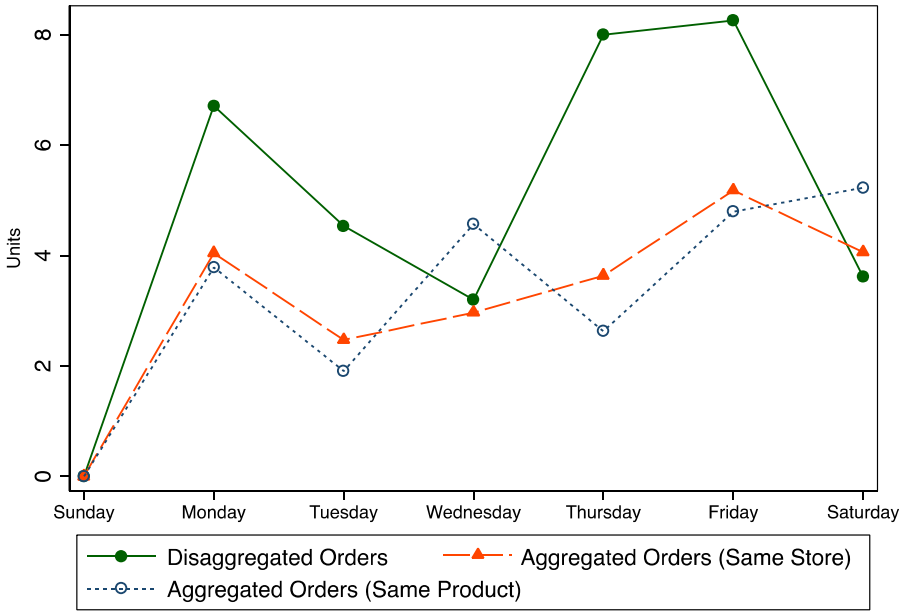

FIGURE 2 Aggregated versus disaggregated orders [Color figure can be viewed at wileyonlinelibrary.com]

downstream store's orders. When the downstream store's orders are not fulfilled completely, the downstream stores' stockouts will increase. Hence, the downstream stores will have to place orders in a greater quantity by factoring in the supply uncertainty, which leads to higher inventory. Therefore, we expect that both inventory and stockouts are positively associated with the bullwhip effect; that is, greater inventory and stockouts (poorer supply chain performance) are associated with a higher bullwhip effect.

5 | ANALYSIS AND RESULTS

5.1 | Our context and raw data

We collected the data from a large supermarket chain in China. At the time of our data collection (i.e., 2011), the company had $4.37 billion in sales. By the end of 2011, the firm operated 5,150 stores mainly in Shanghai and surrounding areas in China. The stores can be categorized into large, medium, and small sized stores. The large stores combine department stores and grocery stores; the medium stores have only grocery stores; and the small stores are convenience stores that sell only limited grocery items. The retailer directly operates all of the large stores, but relies on franchise operations for a considerable number of small to medium stores. Our data were collected only from the large stores, which are similar to the standard supermarket stores in the United States in terms of products and store size.

In order for the orders and demand to be comparable, we chose a supply chain setting that consists of 71 supermarket stores and a distribution center, all of which are owned by the supermarket chain. Although both supermarket stores and the distribution center are owned by the supermarket chain, they operate independently and make their own operations decisions. The distribution center serves these supermarket stores for the products in this study. The supermarket stores observe and fulfill the demands from their consumers and make determination on the timing and quantity of the orders to the distribution center. The orders from the supermarket stores become the demands to the distribution center. Based on the demands, the distribution center decides the timing and quantity of the orders to its suppliers. Each supermarket store has its own operational objectives, and operates independently. The stores do not coordinate the timing and quantity of their orders among themselves. Hence, this supply chain setting between a supermarket store and the distribution center is similar to the supplier–buyer dyad setting that is commonly used in the bullwhip effect models in prior studies (e.g., Chen & Lee, 2009; Lee et al., 1997b). Figure 3 describes our research setting.

We selected 15 product categories: potato chips, juices, tea drinks, toothpastes, kids' toothpastes, toothbrushes, facial tissues, mouthwash, wipes, toilet paper, baby wipes, detergents, shampoos, vinegar, and cooking oil. The reason to select these product categories is that they are common household products that are in supply over a long term and year-round (i.e., not products in the market for a few months and then discontinued, or in the market only for a few months per year), and are popular products so that their data are accumulated quickly and sufficiently for consistent bullwhip effect estimation (Lai, 2005) (i.e., estimating the bullwhip effect requires many occurrences of sales and orders). The selection of these products is in line with those product categories used in the prior literature using product level data; for example, supermarket products in Lai (2005) and Duan et al. (2015). The 15 product categories yielded 1,656 products. We further filtered these data by selecting 487 fast moving products. The fast-moving products are defined as the products that sell at least three units in a week and place at least one order every other week, on average. Table 2 shows the products and observations by product categories. Detergents and shampoos have the largest number of products and observations.

We collected the data for the most recent 7 months at the time of our data collection (i.e., April–October 2011). In particular, we gathered POS data and daily inventory data for each product-store combination. We also collected data on the store orders. The sales data and order data are aggregated to merge with the inventory data. Consequently, our data set consists of 703,632 observations (3,288 x 214). Figure 4 shows the operations activities including sales, shipments, orders, and inventory for a sample product at a sample store.

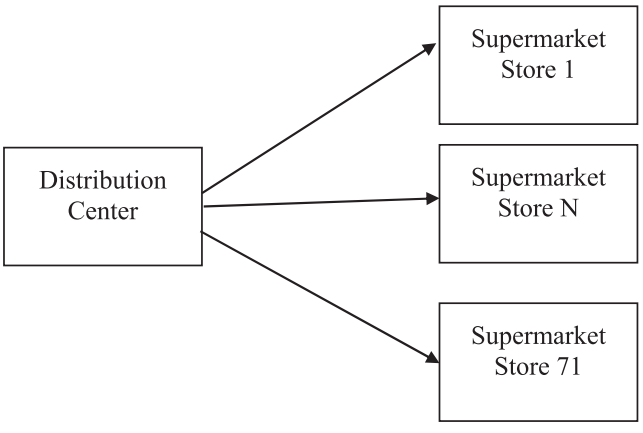

FIGURE 3 Supply chain dyads in our empirical context

5.2 | Measurement of bullwhip effect

We construct three bullwhip effect measures:

1 Information bullwhip effect (BW-INFO). As discussed above, the challenge of computing the information bullwhip effect is to obtain the data on orders and demand. We were able to collect order information from the stores. With respect to the demand information, it is not available as the stores do not track lost sales when stockouts occur. A customer may simply walk away or buy a substitute product without telling the store when the product she wants to buy is out of stock. As a result, we use the method developed in a recent paper by Chen et al. (2017) through which the variance of demand can be derived using the variance of sales series. BW-INFO is the ratio of order variance to the imputed demand variance.
2 Material bullwhip effect (BW-MAT). BW-MAT is the ratio of shipments variance to sales variance as defined in Chen and Lee (2012). As shown in Figure 1, the focal firm (retail store in our case) receives shipments from their supplier (distribution center in our case), and sells the products to the customers. The ratio of shipments variance to sales variance is derived from the actual product or material movements.
3 Hybrid bullwhip effect (BW-HYD). BW-HYD is the ratio of order variance to sales variance. Note that BW-HYD has the numerator from the information bullwhip effect and the denominator from the material bullwhip effect. The reason we construct BW-HYD is to provide a measure of bullwhip effect using raw sales data (as compared with BW-INFO, which uses derived demand data). In addition, since BW-HYD has also been used in prior literature (i.e., Duan et al., 2015), this measure provides continuity and comparison.

We also compute the first differenced ratios for both the hybrid bullwhip effect (BW-HYD-FD) and the

TABLE 2 Distribution of products, product categories, and observations

Product category	Number of products	Store-product combinations	Number of observations
Juices	66	706	151,084
Potato chips	18	93	19,902
Tea drinks	15	111	23,754
Kids' tooth pastes	4	38	8,132
Toothpastes	51	175	37,450
Toothbrushes	26	113	24,182
Mouthwash	3	15	3,210
Facial tissues	8	95	20,330
Wipes	3	7	1,498
Toilet paper	14	190	40,660
Baby wipes	8	43	9,202
Shampoos	116	502	107,428
Detergents	132	848	181,472
Vinegar	18	333	71,262
Cooking oil	5	19	4,066
Total	487	3,288	703,632

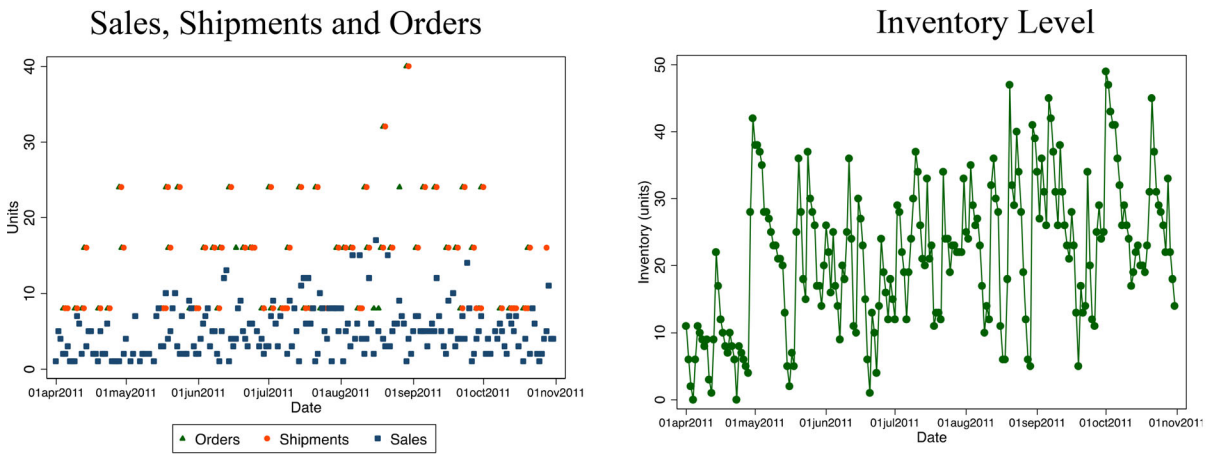

FIGURE 4 Inventory, sales, shipments, and orders for the sample product at a store. Note: The sample product is a brand of drinks packaged in 145 ml × 4 bottles [Color figure can be viewed at wileyonlinelibrary.com]

material bullwhip effect (BW-MAT-FD) (Cachon et al., 2007). (Note, we did not compute the first differenced BW-INFO because the method in Chen et al., 2017 only applies to raw data.) Since our data are at a granular daily level, we can aggregate the data to different, higher levels and calculate BW effect ratios using the aggregated data. In particular, we start by computing the bullwhip effect ratios using the daily data (without aggregation). We then calculate bullwhip effect ratios using aggregated data series by (a) 1-week period, (b) 2-week period, and (c) 4-week period. We use the same data collection periods to compute sales variances, shipment variances, and order variances for the bullwhip effect metrics.

Table 3 presents the results. As shown in Table 3 (top section), the bullwhip effect ratio as measured by the ratio of order variance to sales variance (BW-HYD) is 20.64, 13.31, 12.97, and 9.95 for the daily, 1-week aggregation, 2-week aggregation, and 4-week aggregation data series, respectively. The material bullwhip effect ratio (BW-MAT) is 13.60, 6.00, 4.43, and 2.73 for the daily, 1-week, 2-week, and 4-week aggregation data series, respectively. Similarly, the information bullwhip effect ratio (BW-INFO) is 7.70, 1.55, 1.09, and 0.66 for the daily, 1-week, 2-week, and 4-week aggregation data series, respectively. The outcomes of *t*-tests between BW-HYD and BW-MAT, and those between BW-MAT and BW-INFO, are significant for all four data series, indicating that BW-HYD is greater than BW-MAT, and BW-MAT is greater than BW-INFO.

5.2.1 | Information BW versus material BW

In Section 3.1, we discussed that the material bullwhip effect is expected to be greater than the information bullwhip effect when the AR(1) demand correlation for a product is sufficiently low, and is expected to be smaller than the information bullwhip effect when the AR(1) demand correlation is sufficiently high. A threshold value exists in the demand correlation where the material bullwhip effect is changed to be smaller than the information bullwhip effect. Our results show that the information bullwhip effect ratio is smaller than the material bullwhip effect. This result is drawn from the analysis using the pooled data, thus the analysis does not reveal if the material bullwhip effect has changed to be smaller than the information bullwhip effect at some threshold value of the demand correlation. Chen et al. (2017) do not provide guidance on the threshold value in demand correlation.

To explore the threshold value of the AR(1) demand correlation (i.e., RHO (ρ)), we divide products having different AR(1) RHOs into bins of size 0.10, and average the bullwhip effect ratios for each product bin. Figure 5 shows that the material bullwhip effect is greater than the information bullwhip effect for most product bins, but not for the last two bins (0.7–0.8 and 0.8–0.9). In the last two bins, the information bullwhip effect becomes larger than the material bullwhip effect for the daily and 1-week aggregation data series, but not for the 2-week and 4-week data series. We further conducted *t*-tests

TABLE 3 Measurement of bullwhip effect ($N = 3{,}288$)

	No aggregation (daily data)	One-week aggregation	Two-week aggregation	Four-week aggregation
BW-HYD	20.64	13.31 (17.69***)	12.97 (1.97*)	9.95 (13.50***)
BW-MAT	13.60	6.00 (11.17***)	4.43 (22.24***)	2.73 (15.22***)
BW-INFO	7.70	1.55 (28.65***)	1.09 (20.29***)	0.66 (18.57***)
T statistics				
BW-HYD versus BW-MAT	7.97***	21.44***	20.54***	17.25***
BW-MAT versus BW-INFO	8.47***	19.32***	11.98***	10.49***
BW-HYD-FD	33.60	17.30 (21.43***)	16.02 (2.78**)	14.61 (2.24***)
BW-MAT-FD	21.07	10.71 (9.99***)	6.84 (18.02***)	4.08 (20.95***)
T statistics	8.66***	9.79***	21.55***	18.00***

Note: The numbers in parentheses are the results of *t*-tests comparing the bullwhip effect ratio using data aggregated over its own time window with that over the previous window.

Abbreviations: BW-HYD, bullwhip effect ratio of order variance to sales variance; BW-HYD-FD, first differenced bullwhip effect ratio of order variance to sales variance; BW-INFO, bullwhip effect ratio of order variance to derived demand variance; BW-MAT, bullwhip effect ratio of shipment variance to sales variance; BW-MAT-FD, first differenced bullwhip effect ratio of shipment variance to sales variance.

*$p < .05$; **$p < .01$; ***$p < .001$.

between the two measures for each bin for the four data series, and present the results in Table 4. From Table 4, it is clear that the information bullwhip effect is smaller than the materials bullwhip effect when RHO is smaller than 0.7, but insignificantly different when RHO is greater than 0.7. The insignificant finding may be due to the fact that we have insufficient product observations with high AR(1) demand correlations. Out of the 3,288 products studied, there are only 2 products with demand correlations greater than 0.8 and 43 products greater than 0.7 and smaller than 0.8.

Table 3 (bottom section) presents the *t*-test results of a similar comparison between the BW-HYD-FD and the BW-MAT-FD. The results are consistent with those discussed above.

5.2.2 | Temporal aggregation of BW

The numbers in parentheses in Table 3 are the results of *t*-tests comparing the bullwhip effect ratio using data aggregated over its own time window with that over the previous window. For example, 21.43 is the result of a *t*-test between 33.6 and 17.30. For all of the bullwhip effect ratio series in Table 3, the *t*-test results are significant. Furthermore, when the bullwhip effect estimate for no aggregation is compared with those for the 2-week and 4-week aggregation, and when the bullwhip effect estimate for 1-week aggregation is compared with those for the 4-week aggregation, the *t*-test results are also significant (not shown in the table). These results suggest that longer temporal data aggregation in measuring the bullwhip effect leads to smaller bullwhip effect ratios. Interestingly, while the bullwhip effect is sizable in general, the information bullwhip effect ratio based on 4-week data aggregation is reduced to 0.66, a nonexistent bullwhip effect, suggesting that the temporal aggregation has completely masked the information bullwhip effect.

5.3 | Order aggregation of bullwhip effect

As discussed above, data aggregation may mask or dampen the bullwhip effect ratio because of order patterns among stores. The order patterns among the stores

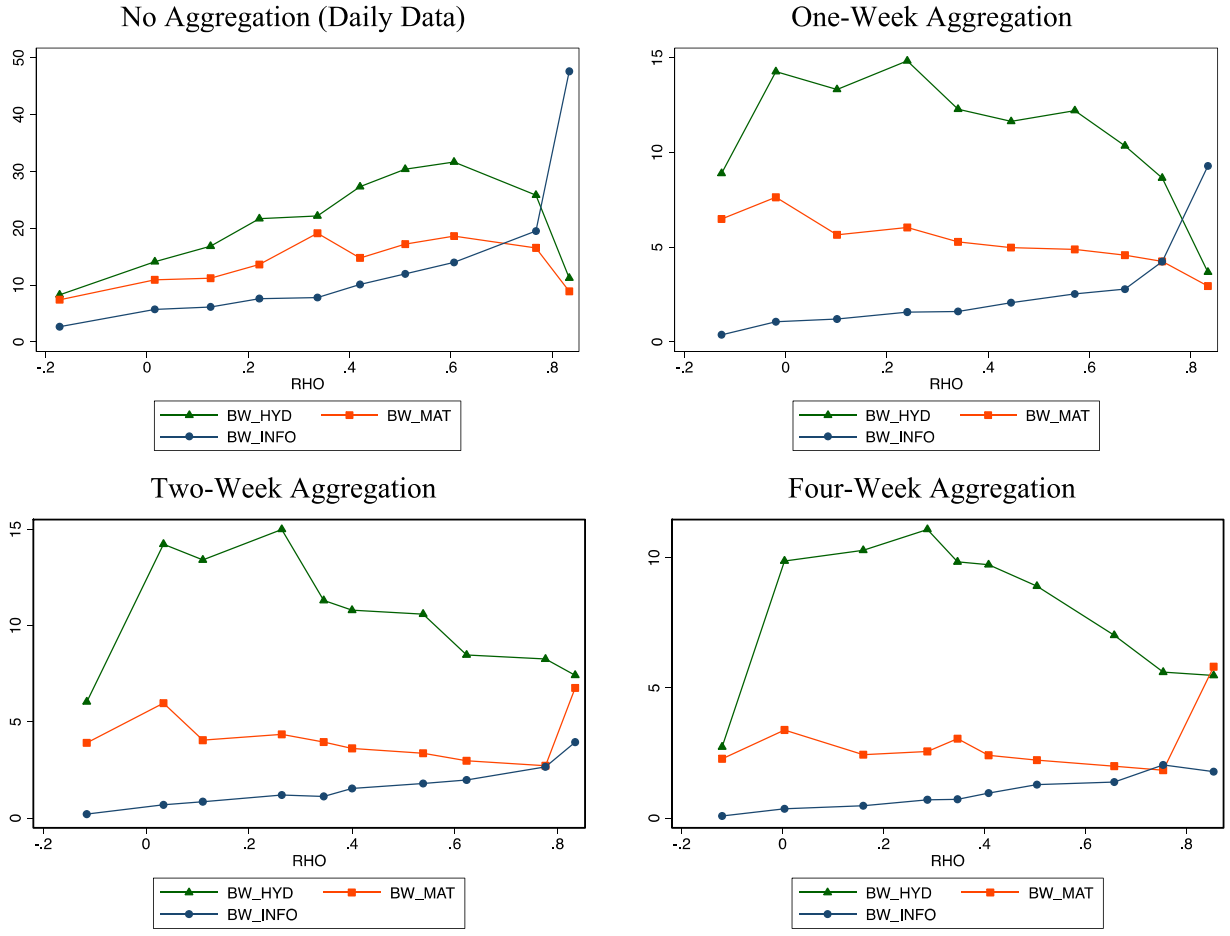

FIGURE 5 Different measures of bullwhip effect over RHO [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Differences between BW-MAT and BW-INFO by RHO

RHO	*N*	No aggregation (daily data)	One-week aggregation	Two-week aggregation	Four-week aggregation
−0.1 to 0.0	12	4.74***	6.11**	3.70***	2.20***
0.0–0.1	799	5.21***	6.57***	5.29***	3.03***
0.1–0.2	780	5.07***	4.45***	3.20***	1.97***
0.2–0.3	615	6.00***	4.47***	3.15***	1.85***
0.3–0.4	423	11.32***	3.67***	2.83***	2.33***
0.4–0.5	290	4.66***	2.91***	2.08***	1.45***
0.5–0.6	227	5.23***	2.35***	1.57***	0.94***
0.6–0.7	99	4.64**	1.80***	1.00	0.61**
0.7–0.8	41	−2.96	0.01	0.05	−0.20
0.8–0.9	2	−38.71	−6.34	2.82	4.03
Total	3,288	5.91***	4.45***	3.34***	2.07***

Note: Positive number denotes BW-MAT is greater than BW-INFO.
Abbreviations: BW-INFO, bullwhip effect ratio of order variance to derived demand variance; BW-MAT, bullwhip effect ratio of shipment variance to sales variance.
**$p < .01$; ***$p < .001$.

in our dataset are derived from random store ordering practices that result in lower order variance after aggregation. Since our data encompass multiple stores and multiple products, we construct two bullwhip effect data series aggregated over orders, one series over the same product across different stores, and the other series over the same store across different products. Both data series have important theoretical and managerial implications. The former series is the bullwhip effect faced by the distribution center in practice as it pools all orders from its retail stores. The latter series resembles how orders are placed and shipped in practice as retail stores always place orders with multiple products that may be shipped in truckloads.

We first demonstrate the random ordering pattern descriptively using the sample product carried by 47 stores, and then use *t*-tests to examine whether aggregation mitigates the bullwhip effect. We aggregate the data in two ways. First, we construct two data series: one series is the orders placed for the sample product by a single store (1 of the 47 stores) (i.e., disaggregated orders), and the other series is the average orders for the sample product placed by all 47 stores (i.e., aggregated orders for the same product across stores). Second, we construct the data series for the sample product orders at a single store (i.e., disaggregated orders), and the orders for all (154) products carried by the sample store (i.e., aggregated orders for different products over the same store). We present the descriptive graphs in Figures 6 and 7. Comparing the aggregated and disaggregated data series,

FIGURE 6 Orders for a single store versus all stores for a sample product [Color figure can be viewed at wileyonlinelibrary.com]

clearly, order aggregation across stores and across products smooth out the order variance.

Table 5 presents the *t*-test results for the aggregation of the bullwhip effect over the same product across different stores. We also consider the four aggregation periods as defined above, and all of the five bullwhip effect ratios (i.e., BW-HYD, BW-MAT, BW-INFO, BW-HYD-FD, and BW-MAT-FD). The results show that order data aggregation can reduce the bullwhip effect ratio significantly across all measures of the bullwhip effect and across all aggregation periods. For example, for the hybrid bullwhip effect ratio using a 1-week aggregation, the order aggregation reduces the bullwhip effect ratio from 13.45

to 7.42, a 44.83% reduction. Similarly, for the first differenced hybrid bullwhip effect ratio using a 1-week aggregation, the aggregation reduces the bullwhip effect ratio from 18.47 to 9.84, a 46.72% reduction. The results are consistent across all four aggregation data series and across all measures of the bullwhip effect. It is interesting to note that, after the data aggregation, some bullwhip effects disappear. For example, for BW-INFO, the estimate is reduced to 0.93, 0.68, and 0.45 for the 1-week, 2-week, and 4-week aggregation series. Since the distribution center looks at aggregated order data from all of their stores, the results demonstrate that the bullwhip effect predicted by theory may not be observable by managers when data analysts perform these data aggregations.

FIGURE 7 Orders for a single product versus all products for a sample store [Color figure can be viewed at wileyonlinelibrary.com]

Table 6 presents the *t*-test results for the aggregation effect over the same store but across different products. The results show that data aggregation over the same store across different products can also reduce estimates of the bullwhip effect significantly. For example, for the hybrid bullwhip effect ratio with 1-week aggregation, the order aggregation reduces the bullwhip effect ratio from 13.89 to 1.88, an 86.47% reduction. Similarly, for the first differenced hybrid bullwhip effect ratio using 1-week aggregation, the aggregation reduces the bullwhip effect ratio from 16.99 to 2.14, an 87.40% reduction. The results are consistent across all four aggregation data series and

TABLE 5 Order aggregation of bullwhip effect (aggregation of same product across stores; $N = 487$)

	No aggregation (daily data)	One-week aggregation	Two-week aggregation	Four-week aggregation
BW-HYD				
Disaggregated	21.83	13.45	12.84	9.65
Aggregated	11.12	7.42	7.33	5.63
T statistics	12.65***	12.18***	10.08***	8.39***
BW-MAT				
Disaggregated	15.49	6.43	4.81	2.98
Aggregated	7.64	3.48	2.76	1.80
T statistics	8.21***	11.64***	7.30***	5.59***
BW-INFO				
Disaggregated	8.29	1.59	1.10	0.68
Aggregated	4.09	0.93	0.68	0.45
T statistics	11.53***	11.01***	8.80***	8.25***
BW-HYD-FD				
Disaggregated	36.15	18.47	16.67	14.25
Aggregated	22.75	9.84	10.74	10.24
T statistics	8.05***	7.47***	6.06***	5.01***
BW-MAT-FD				
Disaggregated	24.52	11.97	7.70	4.56
Aggregated	15.60	6.99	5.40	3.57
T statistics	6.84***	8.60***	4.07***	3.01***

Abbreviations: BW-HYD, bullwhip effect ratio of order variance to sales variance; BW-HYD-FD, first differenced bullwhip effect ratio of order variance to sales variance; BW-INFO, bullwhip effect ratio of order variance to derived demand variance; BW-MAT, bullwhip effect ratio of shipment variance to sales variance; BW-MAT-FD, first differenced bullwhip effect ratio of shipment variance to sales variance.
***$p < .001$.

TABLE 6 Order aggregation of bullwhip effect (aggregation of different products over the same store; $N = 71$)

	No aggregation (daily data)	One-week aggregation	Two-week aggregation	Four-week aggregation
BW-HYD				
Disaggregated	21.60	13.89	13.86	11.42
Aggregated	4.39	1.88	1.98	2.30
T statistics	16.97***	16.06***	15.00***	12.27***
BW-MAT				
Disaggregated	13.74	6.32	4.70	2.98
Aggregated	2.16	0.77	0.62	0.64
T statistics	16.10***	14.59***	11.18***	9.91***
BW-INFO				
Disaggregated	8.22	1.68	1.10	0.68
Aggregated	1.47	0.23	0.18	0.14
T statistics	16.86***	16.37***	15.47***	14.42***
BW-HYD-FD				
Disaggregated	34.13	16.99	16.10	16.60
Aggregated	10.19	2.14	2.46	3.61
T statistics	12.07***	15.14***	12.54***	8.64***
BW-MAT-FD				
Disaggregated	13.74	10.95	7.14	4.40
Aggregated	2.16	1.35	0.93	1.02
T statistics	16.10***	15.13***	13.82***	9.29***

Abbreviations: BW-HYD, bullwhip effect ratio of order variance to sales variance; BW-HYD-FD, first differenced bullwhip effect ratio of order variance to sales variance; BW-INFO, bullwhip effect ratio of order variance to derived demand variance; BW-MAT, bullwhip effect ratio of shipment variance to sales variance; BW-MAT-FD, first differenced bullwhip effect ratio of shipment variance to sales variance.
***$p < .001$.

all measures of the bullwhip effect, indicating the robustness of the results. BW-MAT and BW-INFO are reduced to below one for 1-week, 2-week, and 4-week aggregation series, suggesting a disappearing bullwhip effect. Considering that some prior studies have used BW-MAT as their measurement of the bullwhip effect, and have used aggregated data at the firm or industry levels (e.g., Bray & Mendelson, 2012; Cachon et al., 2007; Shan et al., 2014), our results advance a possible explanation for why they do not find prevalent bullwhip effects. It may be the data aggregations across firms that have masked the bullwhip effect in prior studies. For practitioners, the results demonstrate how a bullwhip effect may not be observable by managers when data analysts perform these data aggregations.

5.4 | Bullwhip effect by category

The bullwhip effect ratios vary among product categories. We compute the average bullwhip effect ratios for each product category. The results are presented in Table 7. Using the estimates from 1-week aggregation data series, Table 7 shows that the bullwhip effect ratios vary across product categories, ranging from 7.18 for baby wipes to 131.85 for Wipes, suggesting that different product categories exhibit different levels of the bullwhip effect. According to the estimation results in Duan et al. (2015) (table 6 in their paper), the heterogeneity in bullwhip effect between product categories is driven by many product level factors such as lead-time, seasonality, order intervals, and other drivers.

6 | IMPACT OF BULLWHIP EFFECT

6.1 | Estimation and results

To examine bullwhip effect impact, we construct a two-equation econometric model. The two equations estimate inventory and stockouts, respectively. Based on inventory theory, inventory and stockouts are endogenously determined (Lee, Clark, & Tam, 1999); that is, inventory

TABLE 7 Estimations of bullwhip effect by product category (BW-HYD-FD)

Product category	No aggregation (daily data)	One-week aggregation	Two-week aggregation	Four-week aggregation
Juices	48.79	26.49	27.73	28.89
Potato chips	77.42	19.25	15.71	12.45
Tea drinks	19.54	13.67	9.76	7.24
Kids' tooth pastes	45.63	31.97	31.24	14.76
Tooth pastes	66.75	37.12	30.26	23.48
Tooth brushes	70.02	27.32	18.37	14.17
Mouth wash	21.41	15.99	15.60	17.22
Facial tissues	24.41	11.83	7.41	3.81
Wipes	203.53	131.85	155.86	131.91
Toilet papers	23.41	10.54	6.45	5.30
Baby wipes	9.52	7.18	8.16	6.68
Shampoos	20.02	14.05	9.50	6.00
Detergents	19.07	9.46	11.15	11.11
Vinegar	20.82	12.42	11.94	12.13
Cooking oil	20.54	13.95	9.94	5.16

affects stockouts and stockouts also affect inventory. Empirically, they often are estimated using a system of equations (e.g., Dong et al., 2014). That is, inventory enters into the stockouts equation as an independent variable, and stockouts also enters into the inventory equation as an independent variable. Thus, the two equations form a system of equations.

For the inventory equation, the dependent variable is the IR calculated using inventory in units divided by the average sales for a product at a store. For the stockouts equation, the dependent variable is the number of stockouts (STOCKOUT), which is a count of stockouts during the data collection period. Both IR and STOCKOUT are commonly used measures for supply chain performance in the literature (e.g., Chen, Frank, & Wu, 2007; Dong et al., 2014; Lee et al., 1999). The independent variable for both equations is the bullwhip effect ratio (BW-HYD-FD).

In addition, we include several control variables. The control variables that are common for both equations are as follows. First, we add the sales in units (SALES) to control for the levels of demand, and coefficient of variation of sales (CVSALES) to control for the variations of demand (Dong et al., 2014; Wan, Evers, & Dresner, 2012). CVSALES is calculated as the ratio of the *SD* of sales to its mean. Second, we include the mean retail price (PRICE) for a product to control for the value of products since more expensive products may differ from less expensive products in service requirement and inventory planning (Abad, 1996). Third, we add the replenishment leadtime (LEADTIME). The replenishment leadtime is computed as the average days between order placement and shipment receipt by the store. Longer replenishment leadtime is associated with greater inventory and possibly a greater number of stockouts (Lee et al., 1997a). Finally, we add fixed store effects (STORE) and fixed product category effects (CATEGORY) to control for the fixed effects between stores and between product categories.

In order to have the system of equations be identified, we include order frequency (OFREQ) in the inventory equation only, and stockouts for other products of the same brands (O_STOCKOUT) in the stockouts equation only. OFREQ is calculated as the number of orders for a product placed by a store during the data collection period. O_STOCKOUT is the total number of stockouts for a product at a store during the data collection period for other products of the same brand. For example, assume Products 1, 2, and 3 are products (e.g., detergents) of the same brand. Products 1, 2, and 3 have 12, 5, and 18 stockouts, respectively. The O_STOCKOUT for Product 1 is then 23 (5 + 18). OFREQ serves as an instrumental variable for IR in the STOCKOUT equation, and O_STOCKOUT serve as an instrumental variable for STOCKOUT in the IR equation. OFREQ only directly affects inventory as it is the information between the stores and the distribution center and does not directly interact with end customers where stockouts occur. O_STOCKOUT only

affects STOCKOUT because demands for products within the same brand are often correlated (Tang & Yin, 2007), but firms do not use other products' service levels to plan the focal product's inventory (Duan et al., 2015). Therefore, they are valid instrumental variables in theory (Wooldridge, 2002). Similar models and instrumental variables have been used in the literature (e.g., Dong et al., 2014).

The first stages from both estimations show that the Cragg–Donald Wald *F*-statistics (848 for STOCKOUT and 26.79 for IR) are much higher than the Stock–Yogo critical values (16.38), rejecting that the IV is weak. The "weak-instrument-robust inference" tests are also significant, rejecting that the endogenous regressors are statistically insignificant. The Anderson's Canonical Correlation LM statistics are significant (27.33 for IR, 689 for STOCKOUT), rejecting the null hypothesis that the equation system is under identified. The Sargan–Hansen test produced an insignificant Sargan statistic (0 for both equations), hence, the null hypothesis that the instruments are valid instruments cannot be rejected. The equation system is estimated using Stata MP/14.2. Using i to denotes store and j to denote product, our model is as follows:

$$\begin{aligned}\mathrm{IR}_{ij} = {} & \beta_0 + \beta_1 \mathrm{STOCKOUT}_{ij} + \beta_2 \mathrm{BW}_{ij} + \beta_3 \mathrm{SALES}_{ij} \\ & + \beta_4 \mathrm{CVSALES}_{ij} + \beta_5 \mathrm{PRICE}_{ij} + \beta_6 \mathrm{LEADTIME}_{ij} \\ & + \beta_7 \mathrm{OFREQ}_{ij} + \gamma_i \sum_{i=1}^{N-1} \mathrm{STORE}_i \\ & + \gamma_k \sum_{k=1}^{K-1} \mathrm{CATEGORY}_k + \varepsilon_{ij},\end{aligned}$$

$$\begin{aligned}\mathrm{STOCKOUT}_{ij} = {} & \alpha_0 + \alpha_1 \mathrm{IR}_{ij} + \alpha_2 \mathrm{BW}_{ij} + \alpha_3 \mathrm{SALES}_{ij} \\ & + \alpha_4 \mathrm{CVSALES}_{ij} + \alpha_5 \mathrm{PRICE}_{ij} \\ & + \alpha_6 \mathrm{LEADTIME}_{ij} + \alpha_7 \mathrm{O\_STOCKOUTQ}_{ij} \\ & + \gamma_i \sum_{i=1}^{N-1} \mathrm{STORE}_i + \gamma_k \sum_{k=1}^{K-1} \mathrm{CATEGORY}_k + \varepsilon_{ij}.\end{aligned}$$

Table 8 shows the descriptive statistics and Table 9 shows the correlation matrix. We calculate variance inflation factor scores for both equations. The variance inflation factor scores are between 1.03 and 6.76 for the IR equation, and between 1.03 and 6.89 for the STOCKOUT equation. The fact that variance inflation factor (VIF) scores are lower than 10 indicates multicollinearity is not a concern (Kennedy, 2003).

We estimate the model using two-stage least squares procedures (2SLS) for all four aggregation data series separately. We use the ivreg2 command in Stata MP/14.2. The results of the second stage estimations are presented in Table 10. The *R*-squared statistics are 0.38–0.44 for the IR estimation and 0.18–0.23 for the STOCKOUT estimation, demonstrating a good fit. For both equations, since the estimation results for the independent variable (i.e., the bullwhip effect variable) across all four data aggregation series are consistent, we focus our discussion on the coefficients for the 1-week aggregation estimation approach. The coefficient for BW-HYD-FD is positive and marginally significant ($\beta = 0.04$, $p < .10$). The result shows that a greater bullwhip effect ratio is associated with a higher IR. The value of the coefficient suggests that an increase in the bullwhip effect ratio by 1 or 5.78% is associated with a higher IR by 0.04, or 0.25% on

TABLE 8 Descriptive statistics ($N = 3{,}288$)

	Variables	Definition	Mean	*SD*	Min	Max
1	BW (daily)	First differenced bullwhip effect ratio	33.60	71.17	0.38	2,175
2	BW (1 week)	First differenced bullwhip effect ratio	17.30	45.38	0.08	1973.00
3	BW (2 week)	First differenced bullwhip effect ratio	16.02	36.91	0.09	1,124.25
4	BW (4 week)	First differenced bullwhip effect ratio	14.61	46.83	0.08	1,478.23
5	IR	Inventory ratio	16.05	8.17	2.79	78.98
6	STOCKOUT	Number of stockouts	0.06	0.09	0	0.54
7	SALES	Average sales in units	2.55	1.72	0.56	14.78
8	CVSALES	Coefficient of sales variation	1.16	0.31	0.47	3.21
9	PRICE	Average price	11.22	10.67	0.80	78.40
10	LEADTIME	Average days between when an order is placed and shipped for a product	2.52	5.06	1	126
11	OFREQ	Total number of orders placed by a store for a product	26.80	11.39	16	105
12	O_STOCKOUT	Number of stockouts for other brands	0.28	0.39	0	2.49

Note: BW denotes BW-HYD-FD: first differenced bullwhip effect ratio of order variance to sales variance.

TABLE 9 Correlation matrix ($N = 3,288$)

	Variables	1	2	3	4	5	6	7	8	9	10	11
1	BW (daily)	1										
2	BW (1 week)	0.82***	1									
3	BW (2 week)	0.71***	0.82***	1								
4	BW (4 week)	0.40***	0.38***	0.65***	1							
5	IR	0.34***	0.26***	0.31***	0.26***	1						
6	STOCKOUT	0.18***	0.09***	0.07***	0.04*	−0.14***	1					
7	SALES	0.11***	0.03	−0.04*	−0.07***	−0.19***	−0.08***	1				
8	CVSALES	−0.08***	−0.03	0.04*	0.09***	0.20***	0.29***	−0.46***	1			
9	PRICE	−0.07***	−0.05**	−0.10***	−0.09***	0.08	−0.05**	−0.25***	0.14***	1		
10	LEADTIME	0.39***	0.14***	0.15***	0.17***	0.25***	0.13***	−0.01	0.05**	−0.06***	1	
11	OFREQ	−0.04*	−0.05**	−0.06***	−0.06***	−0.43***	0.07***	0.39***	−0.24***	−0.06**	−0.07**	1
12	O_STOCKOUT	0.06***	0.03*	0.06***	0.09***	0.0004	0.18***	−0.04*	0.09***	−0.09***	0.003	0.07***

Note: BW denotes BW-HYD-FD: first differenced bullwhip effect ratio of order variance to sales variance.
*$p < .05$; **$p < .01$; ***$p < .001$.

average. For the stockouts equation, the coefficient for BW-HYD-FD for the 1-week aggregation estimation is positive and significant ($\beta = 0.0004$, $p < .01$). The result shows that a greater bullwhip effect ratio is associated with a higher number of stockouts. The value of the coefficient suggests that an increase in the bullwhip effect ratio by 1, or 5.78%, is associated with a greater number of stockouts by 0.0004, or 0.67% on average.

Interestingly, the coefficients for BW-HYD-FD are lower in the estimations using daily data and 4-week aggregated data. As we discussed, the average order cycle time is between 1 and 2 weeks (i.e., 7.32 days). To be conservative, since 7.32 days is closer to 1 week than to 2 weeks, we use 1-week aggregation as the "should be" estimation. Our results suggest that if daily or 4 week aggregation data were used, the impact of the bullwhip effect would have been underestimated (i.e., for IR, 0.036659 for 1-week aggregation data vs. 0.034859 and 0.025736 for daily and 4-week aggregation data estimations, respectively; for STOCKOUT, 0.000389 for 1-week aggregation data vs. 0.000353 and insignificant 0.000117 for daily and 4-week aggregation data estimations, respectively).

6.2 | Robustness checks

We perform a number of robustness checks. The robustness results using 1-week data aggregation are presented in Table 11. The estimation coefficients for the bullwhip effect in all robustness checks are positive and consistent with the main analysis, confirming that our findings are robust. In particular, we perform the following robustness checks. First, a concern is that, when estimating the impact of the bullwhip effect, the bullwhip effect ratios are computed contemporaneously with the performance variables. To allay this concern, we split the data by half based on time periods (first 15 weeks vs. second 15 weeks). We then use the first half of the data to calculate the bullwhip effect ratios, and measure the performance during the second half of the data. Second, we perform a three stage least squares estimation (3SLS). Third, although we have controlled for numerous product and store level variables, product category fixed effects, and store fixed effects, and have used robust standard errors clustered by products to deal with heteroscedasticity, we perform multilevel mixed effect models. We run this model using random product effects and using random store effects separately. Fourth, we extend the data aggregation further to 6 weeks and the coefficients of the bullwhip effect continue to weaken (Table A2 in the Appendix). The coefficients drop to 0.005031 for the IR equation (from 0.25735 for 4-week aggregation), and insignificant for the STOCKOUT

TABLE 10 Estimation results—impact of bullwhip effect (robust standard errors in parentheses; clustered by products)

	No aggregation (daily)		One-week aggregation		Two-week aggregation		Four-week aggregation	
	IR	**STOCKOUT**	**IR**	**STOCKOUT**	**IR**	**STOCKOUT**	**IR**	**STOCKOUT**
IR		−0.005426***		−0.005438***		−0.005383***		−0.005299***
		(0.000882)		(0.000906)		(0.000932)		(0.000958)
STOCKOUT	−25.366911		−20.491071		−23.083443		−27.557961	
	(20.829850)		(20.799134)		(20.656430)		(21.210662)	
BW	0.034859**	0.000353***	0.036659****	0.000355**	0.049603***	0.000389***	0.025736***	0.000117
	(0.011775)	(0.000082)	(0.019657)	(0.000131)	(0.014298)	(0.000087)	(0.007634)	(0.000073)
SALES	−0.029549	−0.003232****	0.132331	−0.001709	0.209489	−0.001013	0.183915	−0.001283
	(0.164611)	(0.001778)	(0.149867)	(0.001730)	(0.149808)	(0.001730)	(0.157724)	(0.001754)
CVSALES	3.549378	0.114584***	3.509656	0.119778***	3.787786	0.119572***	4.218538	0.119142***
	(2.561631)	(0.012498)	(2.654182)	(0.012927)	(2.640446)	(0.013112)	(2.726386)	(0.013319)
PRICE	0.039246	−0.000197	0.044209	−0.000170	0.053805	−0.000092	0.046605	−0.000162
	(0.031991)	(0.000289)	(0.033276)	(0.000300)	(0.034677)	(0.000309)	(0.036029)	(0.000311)
LEADTIME	0.202304**	0.001745*	0.324590***	0.003089***	0.320878***	0.003088***	0.340807***	0.003288***
	(0.064246)	(0.000703)	(0.047382)	(0.000621)	(0.042700)	(0.000584)	(0.045038)	(0.000630)
OFREQ	−0.299054***		−0.307399***		−0.303637***		−0.298309***	
	(0.048102)		(0.048517)		(0.047853)		(0.048867)	
O_STOCKOUT		0.019000**		0.020138**		0.019766**		0.019659**
		(0.007054)		(0.007180)		(0.007232)		(0.007402)
Store fixed effects	Included	Included	Included	Included	Included	Included	Included	Included
Product category fixed effects	Included	Included	Included	Included	Included	Included	Included	Included
N	3,288	3,288	3,288	3,288	3,288	3,288	3,288	3,288
R^2	0.440	0.226	0.418	0.199	0.420	0.193	0.382	0.176
F statistics	36***	89***	33***	143***	35***	93***	32***	61***

*$p < .05$; **$p < .01$; ***$p < .001$; ****$p < .10$.

TABLE 11 Estimation results—robustness checks (standard errors in parentheses)

	Split sample (50 vs. 50%)		Three stage least squares		Mixed model on random product effects		Mixed model on random store effects	
	IR	**STOCKOUT**	**IR**	**STOCKOUT**	**IR**	**STOCKOUT**	**IR**	**STOCKOUT**
IR		−0.0059379***		−0.005438***		−0.002169***		−0.003295***
		(0.000975)		(0.000445)		(0.000158)		(0.000200)
STOCKOUT	−134.7847		−20.491071		−12.877486***		−15.835914***	
	(101.8769)		(14.20924)		(1.564018)		(1.293007)	
BW	0.0283636*	0.0001779*	0.036659***	0.000355***	0.027145***	0.000083***	0.035839***	0.000279***
	(0.0138571)	(0.0000875)	(0.00345)	(0.000035)	(0.002233)	(0.000023)	(0.002379)	(0.000032)
SALES	−0.5006028	−0.0036579*	0.132331	−0.001709	0.394260***	−0.000729	0.147320****	−0.000583
	(0.4709145)	(0.0016755)	(0.091745)	(0.001048)	(0.090890)	(0.000867)	(0.079476)	(0.001010)
CVSALES	12.22575	0.08837***	3.509656*	0.119778***	1.784276***	0.047072***	2.976767***	0.111644***
	(8.594688)	(0.0134622)	(1.67978)	(0.005758)	(0.467166)	(0.004612)	(0.444012)	(0.005464)
PRICE	−0.0326839	−0.0003242	0.044209**	−0.00017	0.063700**	−0.000241	0.046801***	−0.000167
	(0.1055927)	(0.0003701)	(0.014721)	(0.000166)	(0.023815)	(0.000412)	(0.012410)	(0.000163)
LEADTIME	0.694896***	0.0044171***	0.324590***	0.003089***	0.292452***	0.001264***	0.317507***	0.002509***
	(0.1778636)	(0.0009287)	(0.031867)	(0.000336)	(0.023594)	(0.000239)	(0.023446)	(0.000313)
OFREQ	−0.1243671		−0.307399***		−0.349503***		−0.316215***	
	(0.4128294)		(0.029254)		(0.013169)		(0.011716)	
O_STOCKOUT		0.0021885		0.020138***		0.006794*		0.021452***
		(0.0085963)		(0.004324)		(0.003397)		(0.004244)
Store fixed effects	Included	Included	Included	Included	Included	Included	Included	Included
Product category fixed effects	Included	Included	Included	Included	Included	Included	Included	Included
Intercept	9.646659	0.0390125	22.541***	−0.0016773	24.117073***	0.047220****	23.198393***	−0.037231*
	(10.38895)	(0.0256665)	(2.385626)	(0.0183896)	(1.715674)	(0.026031)	(1.300597)	(0.016888)
N	3,288	3,288	3,288	3,288	3,288	3,288	3,288	3,288
R^2	0.606	0.159	0.477	0.236				
F statistics	54***	1,048***						
χ^2 statistics			2,866***	1,016***	2,456***	548***	3,025***	1,168***

*$p < .05$; **$p < .01$; ***$p < .001$; ****$p < .10$.

equation (insignificant for 4-week aggregation already), confirming that longer time aggregation is associated with a greater level of under estimation of the impact of the bullwhip effect.

6.3 | Economic impact

To demonstrate the economic impact, we compute the marginal effect of the bullwhip effect on inventory and stockouts using the regression estimates. Table 12 presents the analysis. When the bullwhip effect ratio decreases by 1, IR will decrease by 0.04 or 0.25%. If we use the average sales of 2.55 and average price of $1.78 (RMB11.22) and 365 days for a year, the change in IR translates into $26.03 in annual savings for the product. Following the same calculation, a 1 *SD* decrease in the bullwhip effect translates into $1,180 for a product in annual savings. Similarly, for stockouts, when the bullwhip effect ratio decreases by 1, the number of stockouts will decrease by 0.0004 or 0.67%. For a year of 365 days, this estimate translates into a reduction in the number of stockouts by 0.15. Following the same calculation, a 1 *SD* decrease in the bullwhip effect translates into 6.63 fewer stockouts for a product in a year. These analyses show that the effect of the bullwhip effect is not only statistically significant but also economically significant.

Next, we demonstrate how much the performance degradation can be underestimated when the bullwhip effect is underestimated using a 4-week aggregation. Table 12 also shows the economic impact calculated using regression estimates based on the 4-week data aggregation. Using the same parameters as above, a one unit reduction in the bullwhip effect translates into $19.49 in annual savings for a product, and 0.04 stockouts in a year. Comparing with the estimates using 1-week data, these numbers are 25 and 75% underestimated, respectively.

7 | CONCLUDING REMARKS

The objective of our study is to estimate and compare different bullwhip effect measures and aggregations through an analysis of a product level dataset, and examine the impact of a bullwhip effect on supply chain performance. In particular, using a dataset at the product level from both the distribution center and stores, we analyze the measurement and aggregation of the bullwhip effect, and examine its effects on inventory and stockouts.

7.1 | Discussion of findings

We show that the material bullwhip effect ratio (BW-MAT) is lower than that measured using order variance (BW-HYD), and that the information bullwhip effect (BW-INFO) is lower than the material bullwhip effect ratio (BW-MAT). As predicted by Chen et al. (2017), when demand is correlated, there exists a threshold value of demand correlation where the material bullwhip effect may change to be smaller than the information bullwhip effect; that is, the material bullwhip effect is larger than the information bullwhip effect when demand correlation is lower than the threshold, ceteris paribus. We find support for the scenario where the material bullwhip effect is greater than the information bullwhip effect, but do not find evidence for the scenario where the material bullwhip effect is consistently significantly smaller than the information bullwhip effect.

Chen and Lee (2012) show analytically that the bullwhip effect should be underestimated when measured using data series constructed with longer time aggregation. We provide conclusive evidence to support their argument. For example, the bullwhip effect (BW-HYD-FD) measured using daily data is 1.94, 2.10, and 2.30 times of those measured using 1-, 2-, and 4-week aggregation data series,

TABLE 12 Economic impact of bullwhip effect (predicted IR at means = 16.05; predicted stockouts at means = 0.06)

	Change in IR	% change in IR	Annual inventory savings per product ($)	% of underestimation (1 week vs. 4 week aggregation)	Change in Stockouts	% change in Stockouts	Annual reduction in Stockouts per product	% of under estimation (1 week vs. 4 week aggregation)
Based on regression estimates for 1-week aggregation in Table 10								
BW + 1	0.04	0.25%	$26.03		0.0004	0.67%	0.15	
BW + 1 SD	1.82	11.34%	$1,180		0.0182	31.25%	6.63	
Based on regression estimates for 4-week aggregation in Table 10								
BW + 1	0.03	0.23%	$19.49	25%	0.0001	0.16%	0.04	75%
BW + 1 SD	1.36	10.55%	$883.59	25%	0.0045	7.39%	1.64	75%

Based on Table 8, 1 *SD* = 45.38.

respectively. It is interesting to note that when measured using a long enough aggregation, the bullwhip effect can even disappear (e.g., BW_INFO using 4-week aggregation data is 0.66). As discussed in Chen and Lee (2012) and Chen et al. (2017), the bullwhip effect needs to be measured using an appropriate level of data aggregation over time. They suggest that the appropriate level should be similar to a product's order cycle time. In our data, the products' average order cycle time is about 7.32 days (roughly a week). Hence, comparing the bullwhip effect ratios measured using a 1-week aggregation with those measured using a 4-week aggregation, measuring the bullwhip effect using 4 week data aggregation may underestimate the bullwhip effect by as much as 61.90% (i.e., for BW-MAT-FD, 4.08 for 4 week aggregation vs. 10.71 for 1 week aggregation).

We find that the aggregated bullwhip effects, whether aggregated by products or by stores, are lower than the disaggregated bullwhip effects. Using the bullwhip effect ratio measured using a 1-week aggregation and first differenced data series as an example, the aggregated bullwhip effect ratio over the same product across different stores is 46.72% (from 18.47 to 9.84) and that over a store across different products is 87.40% (from 16.99 to 2.14) smaller than the disaggregated bullwhip effect ratio on average. For some measures of the bullwhip effect, the effect from data aggregation may be so severe that the bullwhip effect disappears completely (i.e., the bullwhip effect ratio becomes lower than 1).

The two different ways to aggregate data have different implications. First, when aggregating data by products, this type of aggregation replicates what the distribution center (or upstream firm) observes as it observes the order pooling for a product from multiple stores. In the theoretical models, the bullwhip effect is assumed to be at a single product and single firm level so it may propagate upstream in the supply chain. In reality, this may not be the case, since at the upstream level, orders for the same product across stores are pooled so that the bullwhip effect may not be as serious or amplified as expected in the theoretical model. As a result, for the upstream firm (the distribution center, in our case), which is supposed to experience worsened performance resulting from the bullwhip effect because of amplified order variation, they actually may not observe as much bullwhip as we might expect. Because the upstream managers make their decisions based on a pool of orders from multiple downstream stores, the order variances may be actually canceled out to a great extent, resulting in a much lower aggregated bullwhip effect ratio, as shown in our analysis. This may in part explain why the distribution center's supply chain manager, quoted in the first section of this article, did not think the bullwhip effect was as bad as it should be.

Second, when aggregating data by stores, this type of aggregation replicates what was used in the prior literature when the bullwhip effect was measured at the firm level. We show that aggregation at the firm level, or further at the industry level, will underestimate the individual product's bullwhip effect. This finding, in part, may explain why some prior studies using data at firm or industry levels did not find a prevailing bullwhip effect: aggregation masks the bullwhip effect to a large extent. In particular, we show that the product level bullwhip effect is prevalent and strong and substantially larger than found in other, prior empirical studies (e.g., Bray & Mendelson, 2012; Cachon et al., 2007). We directly compare our estimations with prior studies. Our industry context is the retail industry, which sells consumer products (i.e., foods, shampoos, toothbrushes, etc.). It is comparable to the "food and beverage stores" and "general merchandise stores" under the retail industries in Cachon et al. (2007), "food" and "general merchandise" in Bray and Mendelson (2012), and "consumer staples" in Shan et al. (2014). Cachon et al. (2007) estimate the bullwhip effect ratios of 1.30 and 1.41 for the "food and beverage stores" and "general merchandise stores" sectors, respectively; Bray and Mendelson (2012) estimate the bullwhip effect ratios of 0.88 and 1.85 for "food" and "general merchandise," respectively; Shan et al. (2014) estimate the bullwhip effect ratios between 1 and 1.5 for "consumer staples." For us, the estimated bullwhip effect ratio is between 1.55 and 17.30, dependent on the measures used. The bullwhip effect ratios reported in our study are much higher than others in the literature because we use product level, disaggregated data. Prior studies use data aggregated at higher levels; for example, Cachon et al. (2007) use industry level data, while Bray and Mendelson (2012) use firm level data. The bullwhip effect may be underestimated when measured using aggregated data.

We show the bullwhip effect is associated with worse supply chain performance as measured by inventory and stockouts. Because we measure the inventory and stockouts at the downstream stores, the performance deterioration may be because an increased bullwhip effect reduces performance of the upstream firm, which in return offers lower fulfillment and service levels to the downstream store. For example, the orders placed by the downstream store are not fulfilled at 100% in our data, likely due to poor inventory planning at the upstream DC resulting from the bullwhip effect transmitted from the downstream store. With stockouts at the upstream DC, the downstream store has to endure a higher level of stockout rate and/or increase its inventory to guard against supply uncertainty and maintain its own service

levels. Our findings provide empirical evidence for analytical results in the literature, for example, to Lee et al. (1997b) who conclude that the bullwhip effect leads to excessive inventory and lowered customer service levels. A unit decrease in the bullwhip effect ratio can save a firm's inventory by $26.03 and stockouts by 0.15 days for a product during a year on average. These results suggest that, if a firm can mitigate the bullwhip effect, the firm can expect to have lower inventory costs and better service level. However, if the bullwhip effect is measured inaccurately, the estimated benefits are as much as 25% smaller in terms of inventory and 75% in terms of stockouts. If a supply chain manager uses incorrect methods to estimate the bullwhip effect and its impact, she may draw the conclusion that the bullwhip is not severe or existent, and the savings from its mitigation are marginal.

7.2 | Managerial implications

Our findings have several managerial implications that are important to the managers. First, the bullwhip effect can be substantial at the retail stores. Given that the store manager's responsibility is typically within the store, they may not pay attention to the amplification of the order variance feeding into the upstream distribution center or firm, and do not realize the bullwhip effect may be intensive and causing supply chain inefficiency for the upstream firm. Or, the retail managers may think that an increased bullwhip effect is only a headache for the upstream firm to worry about, but not for their own firm. This line of thinking is problematic, as shown by our results. The retailer's performance based on inventory and service levels can also be hurt when it transmits a high bullwhip effect to suppliers. According to our findings, this can very well be the case. Hence, when the upstream and downstream parties are owned by a retailer (e.g., distribution center and stores), retailers should implement more comprehensive SCM systems through which retail managers' incentives are aligned with supply chain efficiency. When the upstream and downstream parties are two separate firms (e.g., wholesaler and retailer), they may want to consider a strategic partnership through which they work together on the bullwhip effect data analytics that guide their supply chain decision making.

Second, since both information bullwhip and material bullwhip effect ratios are valid measures, managers may use both measures if order, demand, sales, and shipment information is available. When the order and demand information is not available but sales and shipment information is available, they may use the material bullwhip effect as long as they use it consistently over time. The comparison over time should give managers a good idea how their supply chain is improving or deteriorating in terms of the bullwhip effect over time. In the meantime, managers should keep in mind that the material bullwhip effect may be higher than the information bullwhip effect, the typical measure of the bullwhip effect, when the demand correlation is sufficiently low.

Third, each supply chain has its own supply chain structure. When the supply chain structure is a 1-to-*N* kind (i.e., one supplier serving multiple buyers), the bullwhip effect may be tamed unexpectedly due to order pooling. This situation is only valid when the orders from the buyers are not highly correlated. While retail managers may coordinate with their peer managers at other stores on ordering at the same time, for the upstream firms, their managers may want to negotiate with the downstream firms and ask them not to coordinate with each other on ordering at the same time, so that the supply chain can take advantage of the order pooling as it is an effective strategy to mitigate the bullwhip effect.

Fourth, retailers should include estimating bullwhip effects in their data analytics effort. While many retailers have implemented data analytics to improve their decision making, they focus mostly on customer and demand management. As we show, estimating the bullwhip effect involves careful consideration on data aggregation. When it is not done correctly, it will lead to inaccurate estimates of the bullwhip effect, and potentially poor supply chain decision making. Therefore, retail managers and data analysts should work together to distinguish different types of data aggregation as we study in this article, all of which leads to different estimates of the bullwhip effect, but only some are correct. The data aggregation due to supply chain order pooling that leads to lowered bullwhip effects is correct, and the data aggregation due to incorrect aggregation either over time interval and unit interval may lead to underestimation of the bullwhip effects. The correct time interval should be the time interval that matches a product's order cycle time, and the correct unit interval should be at product level.

Finally, managers should be aware that the economic impact of bullwhip effect is nontrivial, and that, if they underestimate the bullwhip effect, they may miss the chances to capitalize on the savings from curbing the bullwhip effect. By our calculations, when the downstream firm can decrease the bullwhip effect, the inventory cost savings and improved service levels to its own firm is significant. However, if the data analysts underestimate the bullwhip effect, it would lead

the managers to think that the firm is not subject to the negative impact from the bullwhip effect, and thus, not to take on tactics to deal with the bullwhip effect. The potential consequence highlights the importance of correctly estimating the bullwhip effect in the first place.

7.3 | Limitations and future research

Notwithstanding the fine granularity of our data, there are several research limitations. First, we collected the data from a single, consumer product supply chain, even though the data include a wide range of products and stores. Researchers and practitioners should be cautious when generalizing our results to other supply chains, other sectors, and other products. Also, the bullwhip effect is a multiechelon phenomenon. Our article studies the bullwhip in a single echelon. Future studies can extend our research setting to multiple echelons. Second, although our data is at the product level and in large scale, the length of time (i.e., 7 month) is limited. After aggregation over a long time window, the resulting data points can be sparse. Hence, future research may consider to estimate the bullwhip effect with a longer period time. Third, our data do not have actual demand data. We use our available data and imputation methodology derived in the literature to compute order variance. Although we have used the measures that are closer to the bullwhip effect definition in theory, it still does not reflect the full demand information variations. We acknowledge it is difficult to collect actual demand data from a firm's databases and hope that future studies extend our study to consider other methodologies such as experiments, through which demand data may be full captured. Finally, we measure the performance impact of the bullwhip effect at the downstream party due to data limitation. It would be interesting to examine the impact on the upstream party's performance as well so that we would have a full picture of the impact from the bullwhip effect. Future research can collect such data and conduct the further examination.

ACKNOWLEDGMENTS

We would like to thank the editors-in-chief, the department editor, the associate editor, and the reviewers for spending their valuable time reviewing our manuscript, and for providing constructive feedback that has helped us to greatly improve the article. We are also grateful to our institutions (Lehigh University and Tongji University) for supporting our research. Yongrui Duan and Jiazhen Huo acknowledge the support of the National Natural Science Foundation of China (grant numbers 71771179, 71532015).

ORCID

Yuliang Yao https://orcid.org/0000-0001-9384-6707

ENDNOTES

[1]In this article, because we measure the bullwhip effect at the retail store level, orders are those placed from the stores to the distribution center, and demand and sales are those received by stores from their customers. Figure 1 presents the graphical explanation.

[2]Converted from Chinese currency RMB. $1 = RMB 6.3 in July 2011 when the data were collected. This applies to all dollar calculations throughout the article.

[3]These estimates are the same as those in Duan et al. (2015).

[4]The bullwhip effect ratio for product category wipes is 131.85, much higher than those for other product categories. A further investigation showed that there were two cases where two stores placed orders of 840 units while their average order quantity is about 24. A discussion with the data provider indicated that these were valid orders due to anticipated demand spikes such as volume purchases. Hence, we kept the data points in our analysis.

[5]We use average sales to avoid sales value 0 in denominator.

[6]We report the estimation results using BW-HYD-FD, and note that our estimation results are consistent for all other bullwhip effect measures.

[7]The estimation results from the first stage are presented in Table A1 in the Appendix.

[8]The number is computed using one divided by the mean value of IR (17.30). We used the same calculations for all percentage calculations in this paragraph.

REFERENCES

Abad, P. L. (1996). Optimal pricing and lot-sizing under conditions of perishability and partial backordering. *Management Science*, *42*(8), 1093–1104.

Bray, R., Yao, Y., Duan, Y., & Huo, J. (2019). Ration gaming and the bullwhip effect. *Operations Research*, *67*, 295.

Bray, R. L., & Mendelson, H. (2012). Information transmission and the bullwhip effect: An empirical investigation. *Management Science*, *58*(5), 860–875.

Bray, R. L., & Mendelson, H. (2015). Production smoothing and the bullwhip effect. *Manufacturing & Service Operations Management*, *17*(2), 208–220.

Cachon, G. P. (1999). Managing supply chain demand variability with scheduled ordering policies. *Management Science*, *45*(6), 843–856.

Cachon, G. P., Randall, T., & Schmidt, G. M. (2007). In search of the bullwhip effect. *Manufacturing & Service Operations Management*, *9*(4), 457–479.

Caplin, A. S. (1985). The variability of aggregate demand with (s,s) inventory policies. *Econometrica*, *53*, 1395–1409.

Chatfield, D. C., Kim, J. G., Harrison, T. P., & Hayya, J. C. (2004). The bullwhip effect - impact of stochastic Lead time,

information quality, and information sharing: A simulation study. *Production and Operations Management*, *13*(4), 340–353.

Chen, F., Drezner, Z., Ryan, J. K., & Simchi-Levi, D. (2000). Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, Lead times, and information. *Management Science*, *46*(3), 436–443.

Chen, F., & Samroengraja, R. (2004). Order volatility and supply chain costs. *Operations Research*, *52*(5), 707–722.

Chen, H., Frank, M., & Wu, O. (2007). US retail and wholesale inventory performance from 1981 to 2003. *Manufacturing & Service Operations Management*, *9*(4), 430–456.

Chen, L., & Lee, H. (2015). Modeling and measuring the bullwhip effect. In Y. Ha Albert & C. S. Tang (Eds.), *Information exchange in supply chain management*. Berlin, Germany: Springer.

Chen, L., & Lee, H. L. (2009). Information sharing and order variability control under a generalized demand model. *Management Science*, *55*(5), 781–797.

Chen, L., & Lee, H. L. (2012). Bullwhip effect measurement and its implications. *Operations Research*, *60*(4), 771–784.

Chen, L., Luo, W., & Shang, K. (2017). Measuring the bullwhip effect: Discrepancy and alignment between information and material flows. *Manufacturing & Service Operations Management*, *19*(1), 36–51.

Cui, R., Allon, G., Bassamboo, A., & Van Mieghem, J. A. (2015). Information sharing in supply chains: An empirical and theoretical valuation. *Management Science*, *61*(11), 2803–2824.

de Kok, T., Janssen, F., van Doremalen, J., van Wachem, E., Clerkx, M., & Peeters, W. (2005). Philips electronics synchronizes its supply chain to end the bullwhip effect. *Interfaces*, *35*(1), 37–48.

Dong, Y., Dresner, M., & Yao, Y. (2014). Information sharing and beyond: An empirical analysis of vendor managed inventory. *Production and Operations Management*, *23*(5), 817–828.

Dooley, K. J., Yan, T., Mohan, S., & Gopalakrishnan, M. (2010). Inventory management and the bullwhip effect during the 2007–2009 recession: Evidence from the manufacturing sector. *Journal of Supply Chain Management*, *46*(1), 12–18.

Duan, Y., Yao, Y., & Huo, J. (2015). Bullwhip effect under substitute products. *Journal of Operations Management*, *36*, 75–89.

Forbes. (2017). 53% of companies are adopting big data analytics. Retrieved from https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/#2ddc12c639a1

Forbes. (2018). On target: Rethinking the retail website. Retrieved from https://www.forbes.com/sites/hbsworkingknowledge/2018/12/04/on-target-rethinking-the-retail-website/#10608b3816fb

Fransoo, J. C., & Wouters, M. J. F. (2000). Measuring the bullwhip effect in the supply chain. *Supply Chain Management*, *5*(2), 78–89.

Holmström, J. (1997). Product range management: A case study of supply chain operations in the European grocery industry. *Supply Chain Management*, *2*(3), 107–115.

Kennedy, P. (2003). *A guide to econometrics* (5th ed.). Cambridge, MA: MIT Press.

Klug, F. (2013). The internal bullwhip effect in car manufacturing. *International Journal of Production Research*, *51*(1), 303–322.

Lai, R. (2005). Bullwhip effect in a Spanish shop. Working Paper. Harvard Business School.

Lee, H. G., Clark, T., & Tam, K. Y. (1999). Research report. Can EDI benefit adopters? *Information System Research*, *10*(2), 186–195.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997a). The bullwhip effect in supply chains. *Sloan Management Review*, *38*(3), 93–102.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997b). Information distortion in a supply chain: The bullwhip effect. *Management Science*, *43*, 546–558.

Mackelprang, A. W., & Malhotra, M. K. (2015). The impact of bullwhip effect on supply chains: Performance pathways, control mechanisms, and managerial levers. *Journal of Operations Management*, *36*, 15–32.

McKinsey & Company. (2016). Big data and the supply chain: The big-supply-chain analytics landscape. Retrieved from https://www.mckinsey.com/business-functions/operations/our-insights/big-data-and-the-supply-chain-the-big-supply-chain-analytics-landscape-part-1

Metters, R. (1997). Quantifying the bullwhip effect in supply chains. *Journal of Operations Management*, *15*(1), 89–100.

Montoya, R., & Gonzalez, C. (2019). A hidden Markov model to detect on-shelf out-of-stocks using point-of-sale data. *Manufacturing & Service Operations Management*, *21*(4), 932–948. https://doi.org/10.2139/ssrn.3154600

Morrice, D. J., Cronin, P., Tanrisever, F., & Butler, J. C. (2016). Supporting hurricane inventory management decisions with consumer demand estimates. *Journal of Operations Management*, *45*, 86–100.

Raghunathan, S., Tang, C. S., & Yue, X. (2017). Analysis of the bullwhip effect in a multiproduct setting with interdependent demands. *Operations Research*, *65*(2), 424–432.

Shan, J., Yang, S., Yang, S., & Zhang, J. (2014). An empirical study of the bullwhip effect in China. *Production and Operations Management*, *23*(4), 537–551.

Tang, C., & Yin, S. R. (2007). Joint ordering and pricing strategies for managing substitutable products. *Production and Operations Management*, *16*(1), 138–153.

Tsay, A. A., & Lovejoy, W. S. (1999). Quantity flexibility contracts and supply chain performance. *Manufacturing & Service Operations Management*, *1*(2), 89–111.

Wan, X., Evers, P. T., & Dresner, M. E. (2012). Too much of a good thing: The impact of product variety on operations and sales performance. *Journal of Operations Management*, *30*, 316–324.

Warburton, R. D. H. (2004). An analytical investigation of the bullwhip effect. *Production and Operations Management*, *13*(2), 150–160.

Wheatley, M. (2004). The bullwhip effect lives. *MSI*, *22*(2), 36–38.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wu, D. Y., & Katok, E. (2006). Learning, communication, and the bullwhip effect. *Journal of Operations Management*, *24*, 839–850.

How to cite this article: Yao Y, Duan Y, Huo J. On empirically estimating bullwhip effects: Measurement, aggregation, and impact. *J Oper Manag*. 2021;67:5–30. https://doi.org/10.1002/joom.1090

APPENDIX

TABLE A1 Estimation results—first stage

	No aggregation (daily)		One-week aggregation		Two-week aggregation		Four-week aggregation	
	IR	**STOCKOUT**	**IR**	**STOCKOUT**	**IR**	**STOCKOUT**	**IR**	**STOCKOUT**
O_STOCKOUT	0.022032**	−0.558890	0.022663**	−0.464387	0.022571**	−0.521011	0.023020**	−0.634392
	(0.007301)	(0.479436)	(0.007346)	(0.490313)	(0.007380)	(0.481177)	(0.007424)	(0.490178)
BW	0.000190***	0.030052**	0.000175***	0.033075****	0.000139*	0.046388**	−0.000023	0.026367***
	(0.000042)	(0.010179)	(0.000044)	(0.018613)	(0.000055)	(0.014124)	(0.000037)	(0.007079)
SALES	−0.003562****	0.060799	−0.002733	0.188330	−0.002445	0.265921****	−0.002644	0.256773****
	(0.001963)	(0.148248)	(0.001954)	(0.142086)	(0.001962)	(0.143463)	(0.001966)	(0.148167)
CVSALES	0.110540***	0.745314	0.113319***	1.187618	0.113254***	1.173484	0.113339***	1.095150
	(0.013240)	(0.753406)	(0.013404)	(0.764432)	(0.013494)	(0.760323)	(0.013563)	(0.771455)
PRICE	−0.000475	0.051305****	−0.000462	0.053681****	−0.000436	0.063867*	−0.000479	0.059806****
	(0.000301)	(0.028999)	(0.000303)	(0.030096)	(0.000304)	(0.031666)	(0.000305)	(0.032216)
LEADTIME	0.000751	0.183255**	0.001490**	0.294057***	0.001554**	0.285013***	0.001735**	0.292981***
	(0.000554)	(0.061946)	(0.000566)	(0.042159)	(0.000573)	(0.036250)	(0.000616)	(0.031946)
OFREQ	0.001881***	−0.346781***	0.001881***	−0.345948***	0.001867***	−0.346723***	0.001851***	−0.349315***
	(0.000339)	(0.022134)	(0.000343)	(0.022478)	(0.000349)	(0.022047)	(0.000350)	(0.022740)
Store fixed effects	Included	Included	Included	Included	Included	Included	Included	Included
Product category fixed effects	Included	Included	Included	Included	Included	Included	Included	Included
Intercept	26.2282***	−0.13590***	25.40635***	−0.139834***	25.00182***	−0.138232***	25.72017***	−0.132522***
	(1.636855)	(0.264958)	(1.768112)	(0.0267155)	(1.720376)	(0.0270387)	(1.654165)	(0.0268117)
N	3,288	3,288	3,288	3,288	3,288	3,288	3,288	3,288
F statistics	245.46***	9.11**	236.87**	9.52**	247.33***	9.35**	235.97***	9.62**

Note: Robust standard errors in parentheses; clustered by products.
*$p < .05$; **$p < .01$; ***$p < .001$; ****$p < .10$.

TABLE A2 Estimation results—6-week aggregation

	Six-week aggregation	
	IR	**STOCKOUT**
IR		−0.005295***
		−0.000959
STOCKOUT	−20.532015	
	(21.655187)	
BW	0.005031*	0.000025
	(0.002196)	(0.000018)
SALES	0.174405	−0.001404
	(0.157877)	(0.001765)
CVSALES	3.543089	0.119700***
	(2.776329)	(0.013399)
PRICE	0.044702	−0.000185
	(0.036043)	(0.000311)
LEADTIME	0.360887***	0.003429***
	(0.046775)	(0.000666)
OFREQ	−0.311745***	
	(0.049827)	
O_STOCKOUT		0.020373**
		(0.007333)
Store fixed effects	Included	Included
Product category fixed effects	Included	Included
Intercept	23.43742***	0.0054374
	(3.478936)	(0.0278053)
N	3,288	3,288
R^2	0.381	0.174
F statistics	31***	57***

Note: Robust standard errors in parentheses; clustered by products.
*$p < .05$; **$p < .01$; ***$p < .001$.