Endogenous Elasticities: Price Multipliers Are Smaller for Larger Demand Shocks *

Aditya Chaudhry and Jiacui Li

August 29, 2025

ABSTRACT

A growing literature estimates price multipliers — per-unit impacts of demand shocks on asset prices — using plausibly exogenous but small shocks. However, it is theoretically ambiguous how these estimates extrapolate to the large shocks that are often of interest for policy and economic questions. This paper documents a new stylized fact: stock-level price multipliers decrease as the magnitude of demand shocks increases. We study three widely-used demand shocks unrelated to cash flow news, and find that multipliers are smaller for both larger contemporaneous and larger cumulative past shocks. Using investor holdings data, we also find higher demand elasticities in securities with larger price dislocations. Our findings shed light on the microfoundations of inelastic demand. While these findings cannot be explained by several prominent mechanisms, they are consistent with models in which investors endogenously allocate more attention and capital toward greater profit opportunities.

JEL classification: G11, G12

Keywords: Demand-System Asset Pricing, Inelastic Demand, Price Multipliers

*Chaudhry (chaudhry.127@osu.edu) is at the Fisher College of Business at The Ohio State University. Li (jiacui.li@eccles.utah.edu) is at the David Eccles School of Business, University of Utah. We are grateful for comments from Ehsan Azarmsa, Carter Davis, Espen Eckbo, Zhiyu Fu, Michael Gallmeyer, Niels Gormsen, Valentin Haddad, Lars Hansen, Paymon Khorrami, Ralph Koijen, Erik Loualiche, Stefan Nagel, Simon Oh, Anna Pavlova, Kaushik Vasudevan, as well as participants at Hong Kong Poly U, Hong Kong University, MARC, Stanford SITE, The Ohio State University, Utah, and UVA McIntire.

1 Introduction

Many questions in asset pricing involve quantifying the impact of *large* shifts in quantities on asset prices. For instance, how do central bank asset purchases move prices (Koijen et al., 2021)? How has the shift to passive investing impacted price efficiency (Haddad et al., 2024), and how has the rise of green investing impacted stock prices and firms' cost of capital (Van der Beck, 2021; Pástor et al., 2022)? These questions motivate a growing literature that studies demand effects in asset markets, where estimating "price multipliers" — the per-unit impact of investor demand shocks on prices — is crucial (Koijen and Yogo (2019); Gabaix and Koijen (2022); Haddad et al. (2024)). Given the difficulty in finding natural experiments involving large demand shocks, the literature has focused on plausibly exogenous but relatively small shocks. These studies consistently find large price multipliers.

Yet a fundamental question remains unresolved: How do these price multipliers measured from small shocks relate to those that would arise for the large shocks of interest in many asset pricing questions? The answer is theoretically ambiguous, as different models offer conflicting predictions about how multipliers vary with shock size. For instance, some models suggest that multipliers increase with shock size, while others predict they remain constant or even decrease. Thus, new empirical evidence is required to answer this question.

This paper documents a new stylized fact: stock-level price multipliers decrease as the magnitude of demand shocks increases. Using three demand shocks from the literature, we find that price multipliers monotonically decrease as the magnitude of both contemporaneous and accumulated past demand shocks grows. We interpret this result as evidence that investors' asset demand endogenously becomes more price-elastic in response to larger price dislocations. This finding helps distinguish competing theories of inelastic demand. Several prominent mechanisms—such as those based on arbitrageur financial constraints—cannot account for this empirical regularity. However, this finding is consistent with models in which investors allocate more attention and capital to securities with greater profit opportunities.

We begin by providing empirical evidence that larger uninformed demand shocks have smaller price multipliers (in Section 3). To estimate multipliers, we use three demand shocks from the literature: Russell index reconstitution-induced changes in benchmarking intensity (BMI) from Pavlova and Sikorskaya (2023), fund flow-induced trading (FIT) from Lou (2012), and order flow imbalance (OFI) from Li and Lin (2023). We acknowledge that any such measure is imperfect because identifying purely exogenous variation in demand is difficult, especially given our focus on large shocks. Nevertheless, we employ these measures because previous literature have studied them extensively and showed them to be largely unrelated to cash flow news.

Across all three measures, our cross-sectional contemporaneous regressions reveal concave price impact curves: price multipliers decline as demand shock size increases. We analyze these price effects at horizons relevant for asset pricing, studying FIT and OFI at quarterly frequencies and BMI at a monthly frequency (as in Pavlova and Sikorskaya (2023)). This concavity is economically significant. For example, using the FIT measure, small shocks have a price multiplier of approximately 4: a shock of size $X \approx 0\%$ raises price by about $4X\%$. However, this multiplier falls by roughly 0.7 for each one percentage point increase in shock size. We find qualitatively similar patterns across all three demand measures.

Our findings are inconsistent with a key alternative hypothesis: that multipliers do not vary with shock size *per se*, but rather with omitted variables that are correlated with shock size. We address this potential concern in several ways. First, by estimating all regressions within each time period, we mitigate the concern that large demand shocks might cluster in periods when price multipliers are systematically lower. Second, our results are robust to alternative specifications that allow price multipliers to vary cross-sectionally with various liquidity measures. This approach assuages the potential concern that large shocks are concentrated in more liquid stocks, which may generally have smaller multipliers.

We next show that this endogenous price multiplier behavior is time-consistent (in Section

4). Using the FIT and OFI shocks, we find that multipliers decrease not only with contemporaneous demand shock magnitude, but also with the magnitude of accumulated demand shocks over the past several quarters. Thus, several consecutive small shocks (e.g. three 1% shocks) have the same price impact as one large shock (e.g. one 3% shock). These results suggest price multipliers decrease in the face of larger accumulated price dislocations.

In the final empirical section, we examine holdings data to shed light on the origins of the endogenous price multiplier behavior we document (in Section 5). We estimate demand elasticities — which are inversely related to price multipliers — for institutional investors in the U.S. stock market using holdings data from the SEC Form 13F. Motivated by our preceding findings, we allow demand elasticity to depend on the magnitude of cumulative past price dislocations. Using the optimal granular instrumental variable (GIV) approach of Chaudhary et al. (2024), which extends the approach of Gabaix and Koijen (2024), we find that most institutions become more price-elastic as past price dislocations grow larger — consistent with our price impact findings. This effect is particularly pronounced for more active institutions. Moreover, we document this variation in elasticities on the intensive margin of holdings, within each investor and quarter, suggesting that our findings are not driven solely by investors entering or exiting positions or reallocation of capital among investors.

Lastly, we use these empirical results to shed light on microfoundations of inelastic demand (in Section 6). We find that many existing mechanisms cannot explain our results. Models with position constraints on arbitrageurs or intermediaries (e.g., Diamond and Verrecchia, 1987; Shleifer and Vishny, 1997; Duffie et al., 2002; He and Krishnamurthy, 2013) cannot explain our static result that multipliers decrease with shock size. In these models, multipliers actually *increase* with shock size, as larger shocks tighten these constraints. Models with benchmarking and investment mandates (e.g., Petajisto, 2009; Gabaix and Koijen, 2022; Pavlova and Sikorskaya, 2023), as well as those using logit demand specifications (Koijen and Yogo, 2019), predict that multipliers are unrelated to shock size. Models with convex position adjustment costs (e.g., Gârleanu and Pedersen, 2013; Van der Beck, 2025) cannot explain our dynamic result that multipliers decrease with the magnitude of accumulated *past* demand shocks. While models with participation costs (e.g., Vissing-Jorgensen, 2002; Gomes and Michaelides, 2005; Alan, 2006) or endogenous reallocation of capital across investors (e.g., Duffie and Strulovici, 2012) can explain our price multiplier results, they cannot explain our results from holdings data that demand elasticities vary on the intensive margin of holdings within an investor-quarter.

Our findings are consistent with mechanisms in which investors in general face frictions, but can pursue costly actions to overcome these frictions and become more elastic when price dislocations (and so expected returns) are large. This behavior arises in models with fixed adjustment costs Constantinides (1986); Grossman and Laroque (1987)). Fixed-adjustment costs gives rise to (S, s)-style adjustment behavior (Scarf (1960)). For small shocks, an investor's initial position is close to the optimum, so the benefit of adjustment does not exceed the fixed cost. Large shocks, however, significantly alter the optimal quantity (since they alter expected returns), making adjustment worthwhile. Thus, for sufficiently large shocks, investors become more price-elastic. This behavior also arises in models with costly information acquisition (Grossman and Stiglitz (1980); Kyle (1989); Van Nieuwerburgh and Veldkamp (2009, 2010); Kacperczyk et al. (2016); Han (2018)). Uncertainty about cash flows renders demand inelastic. However, if investors can pay to acquire information to reduce this uncertainty, they will be more willing to do so when they believe expected returns are higher. Lower uncertainty allows investors to take larger positions to more aggressively exploit the high expected returns without exposing themselves to a lot of perceived risk. Thus, for large shocks that create large price dislocations and high expected returns, investors acquire more information, and so become more price-elastic.

We wish to clarify several points regarding the interpretation of our paper. First, while many mechanisms cannot explain all of our findings, we do not claim these frictions are unimportant. Rather, our paper should be interpreted as arguing that these frictions are not, by themselves, sufficient to *fully* explain inelastic demand. Second, although price multipliers decrease with shock size, they remain substantially larger than predicted by frictionless models.[1] Thus, our findings shed light on endogenous variation in price multipliers, rather than suggest that multipliers are small in an absolute sense. Finally, our findings are distinct from the "square root" price impact curves at the trade or order level documented in the microstructure literature (e.g., Frazzini et al., 2018). Those studies focus on high-frequency, intraday time horizons, whereas our analysis is at much lower frequencies of interest in asset pricing, at which most microstructure effects have dissipated.

The paper is structured as follows. Section 2 describes our data and the construction of demand shock measures. Sections 3 and 4 present empirical evidence that price multipliers decrease with the magnitude of both current and cumulative past demand shocks. Section 5 analyzes institutional holdings data and shows that demand elasticities increase with the magnitude of past price dislocations. Section 6 uses our empirical findings to differentiate among competing theories of inelastic demand. Section 7 concludes.

1.1 Related Literature

This paper contributes to a growing literature that estimates the effects of demand on asset prices. This literature, using both plausibly exogenous demand shocks (Shleifer (1986); Harris and Gurel (1986); Lou (2012); Chang et al. (2015); Hartzmark and Solomon (2022); Schmickler and Tremacoldi-Rossi (2022); Pavlova and Sikorskaya (2023)) and structural asset demand systems (Koijen and Yogo (2019, 2020); Huebner (2023); Haddad et al. (2024); Koijen et al. (2024)), finds that empirical demand elasticities are much smaller — and so price multipliers much higher — than predicted by classical asset pricing models. Our paper contributes to this literature by documenting that price multipliers decline with the magni-

[1] Frictionless asset pricing models predict that uninformed demand shocks have stock-level price multipliers on the order of 1/6000 (e.g., Petajisto, 2009).

tude of current and past demand shocks, and by establishing the theoretical conditions under which this pattern can arise. While the literature has generally not focused on endogenous variation in multipliers, a notable exception is Haddad et al. (2024), who use a structural model to show that demand elasticities vary *across stocks* depending on strategic investor interactions. In contrast, we show that demand elasticities vary with the *magnitude of the shocks*, representing a different source of multiplier variation.

Prior research has proposed a myriad of microfoundations for inelastic demand. These include leverage constraints (e.g., Shleifer and Vishny, 1997; He and Krishnamurthy, 2013), short-selling constraints (e.g., Diamond and Verrecchia, 1987; Duffie et al., 2002), investment mandates and benchmarking (e.g., Petajisto, 2009; Gabaix and Koijen, 2022; Pavlova and Sikorskaya, 2023), and convex adjustment costs (e.g., Gârleanu and Pedersen, 2013; Van der Beck, 2025). To generate heterogeneity in price multipliers, researchers have also explored mechanisms involving capital reallocation between investors (e.g., Duffie and Strulovici, 2012) or investor entry and exit (e.g., Vissing-Jorgensen, 2002; Gomes and Michaelides, 2005; Alan, 2006). As we argue in detail in Section 6, these mechanisms cannot fully explain our findings. Instead, our evidence is most consistent with two classes of models: those featuring endogenous information or attention (e.g., Van Nieuwerburgh and Veldkamp, 2010; Gabaix, 2019), and those with fixed adjustment costs (e.g., Constantinides, 1986; Grossman and Laroque, 1987), which give rise to "(S,s)"-style adjustment behavior (Scarf, 1960).

Additionally, the industry-oriented market microstructure literature documents that price impacts appear to follow a "square-root law" at high frequencies, such as at the trade or order level (e.g., Tóth et al., 2011). In the econophysics literature, researchers explain this effect using microstructure arguments (e.g., Alfonsi et al., 2010; Gatheral, 2010; Donier et al., 2015). For instance, Donier et al. (2015) derives square-root price impact under the assumption that the density of limit orders in the order book is locally linear. Bouchaud et al. (2018) (chapter 18) links the shape of price impact curves to another microstructure consideration:

the frequency of trading. Specifically, he argues that price impact tends to be concave in frequently traded markets, while infrequently traded markets should exhibit patterns closer to linear price impact. Our work is distinct, as we focus on a much lower frequency: monthly to quarterly time horizons. To our knowledge, no previous work has established whether the nonlinearity documented in microstructure settings extends to these lower frequencies and the much larger demand shocks we analyze. Crucially, because we study a frequency at which many microstructure effects have likely dissipated, the potential explanations for our findings differ from those in the microstructure literature. Furthermore, microstructure explanations cannot readily explain our results derived from quarterly institutional holdings.

2 Data and Demand Shocks

To study how price multipliers vary with demand shock size, we use three demand shocks from previous work that researchers argue are largely devoid of cash flow information, and thus are useful for studying demand-induced price effects. Admittedly, finding cash flow-unrelated demand shocks is a difficult task. The literature on demand-based price effects has yet to reach a consensus on the "optimal" demand measure, as each measure has its own strengths and limitations. To ensure robustness, we analyze all three measures and highlight findings that are consistent across them.

2.1 Data

We download monthly stock returns and market capitalization from CRSP, and we aggregate monthly returns to the quarterly frequency when needed. To control for known return predictors, all regressions control for 13 commonly used stock characteristics from the website of Chen and Zimmermann (2022).[2] To capture stock liquidity, we also include size, effec-

[2]The stock characteristics we use include accruals, asset growth, beta, book-to-market, gross profitability, industry momentum, intermediate momentum, 1 year issuance, 5 year issuance, momentum, seasonal momentum, net operational assets, and short-term reversal.

tive bid-ask spread, quoted bid-ask spread, realized volatility, turnover, and dollar trading volume. These liquidity variables are obtained from CRSP and WRDS Intraday Indicators. Following common practice in the literature on modeling the cross-section of returns, for all these characteristics, we transform each into uniform distributions over [-0.5, 0.5] in each cross-section (e.g. Kelly et al., 2019).

In constructing the demand measures, we download the data for the benchmarking intensity (BMI) measure from Pavlova and Sikorskaya (2023). For the flow-induced-trading (FIT) measure, we use mutual fund flows from CRSP, holdings data from Thomson Reuters, and link the two using MFLINKS from Russ Wermers. For the order flow-imbalance (OFI) measure, we download daily Lee-Ready classified order flows from WRDS intraday indicators.

For the holdings-based estimation in Section 5, we use institutional holdings data from SEC Form 13F (from 1980 to 2021), which are provided by Thomson Reuters through WRDS. The SEC requires all institutional investors with at least $100 million in assets under management (AUM) to report stock-level long positions each quarter. Following Koijen and Yogo (2019), we allocate all stock holdings not covered by 13F institutions to a residual "household" sector, which includes both direct stock holdings by households and those by non-13F institutions (i.e. institutions with less than $100 million AUM).

2.2 Demand measures

We briefly explain the construction and logic behind the three demand measures we use. For more details of these measures, as well why researchers argue they are unrelated to cash flow news, we refer the reader to the original papers.

Benchmarking Intensity (BMI) Our first demand measure is based on index inclusion, which captures changes in benchmarked investors' demand for a stock when it enters or exits an index (Shleifer (1986); Harris and Gurel (1986); Chang et al. (2015)). In particular, we use the benchmarking intensity (BMI) measure from Pavlova and Sikorskaya (2023), which provides a continuous measure of demand changes driven by Russell index reconstitutions. While there are also other index-based demand shocks, such as S&P 500-based measures (e.g. Shleifer, 1986), BMI stands out for its rule-based inclusion criterion and the fact that it exhibits cross-sectional variation in demand shock magnitudes (as a fraction of shares outstanding).

Each May, Russell ranks eligible stocks by market capitalization to determine index membership. The inclusion criterion is solely based on market capitalization on the cutoff date, and there is no discretionary judgment on inclusion. Stocks above a specified rank enter the Russell 1000, while those below join the Russell 2000. The Russell 2000 historically attracts more benchmarked institutional capital than does the Russell 1000. When stocks cross the cutoff during the annual reconstitution in June, they experience institutional flows: stocks moving down to the Russell 2000 see inflows and positive returns, while those moving up to the Russell 1000 face outflows and negative returns. Conditional on the market cap as of the May ranking date, Russell index membership in June is exogenous to June cash-flow news, and so these reconstitution-driven flows are an uninformed demand shock (Chang et al. (2015); Crane et al. (2016); Glossner (2019)).

Even within one reconstitution window, Russell reconstitution provides cross-sectional variation in demand shock magnitudes (normalized by shares outstanding) because the construction in Pavlova and Sikorskaya (2023) considers multiple indices. Every stock in the Russell 2000 Blend index is also in the Russell 2000 Value or Growth indices, which have different levels of benchmarked capital. Every stock in the Russell 1000 Blend index is also in the Russell 1000 Value or Growth indices, and some — those with market cap rank between 1000 and 2000 — are in the Russell Midcap Blend, Value, and Growth indices. This is a desirable feature as it allows us to estimate the relationship between price impact and demand shock size within time periods (Section 3).

Concretely, the Pavlova and Sikorskaya (2023) benchmarking intensity (BMI) measure captures the benchmarking-induced variation:

$$BMI_{n,t} = \sum_{\text{Index } j} \frac{\substack{\text{Institutional AUM Benchmarked to Index } j \text{ in Month } t \\ \times \text{ Weight of Stock } n \text{ in Index } j \text{ in Month } t}}{\text{Stock } n \text{ Market Value in Month } t}.$$

This measure quantifies the index-induced demand for each stock from benchmarked funds. Variation in this measure depends on which indices a stock is part of and the proportion of the total market value of stock n that is held by benchmarked investors. BMI is calculated from thirty-four indices that cover approximately 90% of mutual fund and ETF assets.

To exploit exogenous changes in BMI, we focus on June BMI changes in each year, denoted as ΔBMI, for stocks in a narrow window (150 stocks on both sides of the threshold in the baseline specification) around Russell 1000/2000 reconstitution thresholds (following Pavlova and Sikorskaya (2023)). Stocks with positive (negative) ΔBMI experience benchmarking inflows (outflows). While the *level* of BMI is generally correlated with stock profitability because index membership is based on market capitalization, the *change* ΔBMI for stocks in this window are driven by Russell index membership changes, which are exogenous to June cash flow news conditional on the May rank-date market cap.[3]

We use the BMI and Russell index constituents data provided by Pavlova and Sikorskaya (2023). All of our specifications include the stock-level controls used by Pavlova and Sikorskaya (2023): May rank-date log market cap, one-year monthly average bid-ask percentage spread, and the banding controls from Appel et al. (2019) (an indicator for having rank-date

[3]More specifically, prior to 2007 the rank cutoff was the 1,000th stock. To reduce turnover, since 2007 Russell has used a "banding policy" under which there are two separate cutoffs for stocks starting in the Russell 1000 and 2000 pre-reconstitution, both of which are still mechanical functions of the firm size distribution. Thus, there is a "band" of market caps including stocks from the Russell 1000 and 2000. Appendix A.1 explains the Russell methodology we use to calculate these cutoffs. Since Russell ranks stocks using a proprietary market cap that we lack access to, we use the method of Ben-David et al. (2019) to approximate this proprietary market cap using standard databases. Doing so predicts assignment to the Russell 1000 and 2000 with high accuracy. Following previous work, we use May — not June — market caps to calculate the Russell reconstitution thresholds to avoid selection bias (e.g. Chang et al. (2015); Appel et al. (2021); Wei and Young (2021)).

market cap in the "band", an indicator for being in the Russell 2000 in May, and the interaction of these indicators). Whereas Pavlova and Sikorskaya (2023) use the proprietary Russell market cap, we calculate market cap from standard databases using the method of Ben-David et al. (2019).

Fund Flow-Induced Trading (FIT) Our second demand measure is flow-induced trading (FIT) from Lou (2012), which captures the non-discretionary stock-level trading by mutual funds and ETFs in response to fund flows. While not initially intended as a demand instrument, it has been used as such by subsequent work (e.g. Li (2022); Chaudhry (2023); Van der Beck (2025)).

In response to inflows (outflows), funds tend to scale up (down) their pre-existing holdings in a non-discretionary fashion, a behavior that is documented in Coval and Stafford (2007) and Lou (2012), among others. For example, if Apple's existing weight is 5% in a fund's portfolio, a $1 inflow (outflow) induces the fund to increase (decrease) capital allocation of about five cents to Apple. This behavior is not only true for index funds, but also on average true for active mutual funds and exchange-traded funds (Figure A4 in Li (2022)). Frazzini and Lamont (2008) and Lou (2012) show that the fund flow-induced trades appear to have price impacts that revert over time, which is consistent with the idea that such trades are not motivated by superior information about cash flows.

While funds tend to trade one-for-one in stocks in response to fund flows, the relationship can exhibit heterogeneity as a function of realized fund flows and pre-existing fund portfolio positions. Therefore, we construct a refined version of the Lou (2012) FIT measure to account for this heterogeneity and avoid any potential mechanical bias in our results (Appendix A.2). We first calculate the quarterly percentage flow to fund i as

$$f_{i,t} = \frac{TNA_{n,t} - TNA_{n,t-1} \cdot (1 + \text{Ret}_{n,t})}{\text{TNA}_{n,t-1}}.$$

where $TNA_{n,t}$ and $\text{Ret}_{n,t}$ are fund i's end-of-quarter-t total net assets and return from end of quarter $t-1$ to t, respectively. The predicted non-discretionary trading by fund i in stock n due to this flow is proportional to $\text{SharesHeld}_{i,n,t-1} \cdot f_{i,t}$. To allow these predicted trades to vary flexibly with the realized fund flow $f_{i,t}$ and the pre-existing portfolio weight $w_{i,n,t-1}$, we estimate a scaling factor $b(\cdot)$, which is empirically approximately equal to one. Appendix A.2 provides further details for the estimation of this scaling factor. Aggregating across all funds and scaling by shares outstanding yields:

$$\text{FIT}_{n,t} = \sum_{\text{fund } i} b(w_{i,n,t-1}, f_{i,t}) \cdot \underbrace{\frac{\text{SharesHeld}_{i,n,t-1}}{\text{Shares Outstanding}_{n,t-1}}}_{\equiv S_{i,n,t-1}} f_{i,t}. \tag{1}$$

where $S_{i,n,t-1}$ is the proportion of stock n shares owned by mutual fund i in quarter $t-1$.

Why does FIT constitute a cash flow-unrelated demand shock? The original Lou (2012) paper, as well as follow up papers of Li (2022) and Huang et al. (2024), argue that a large fraction of fund flows are directed by less sophistiated retail investors. They also show that FIT-induced price effects revert over time, which is consistent with it being unrelated to genuine cash flow news. Chaudhry (2023) further argues that, because FIT is a shift-shares instrument (Goldsmith-Pinkham et al., 2020), a weaker requirement is that the lagged ownership shares ($S_{i,n,t-1}$) of a given fund i is uncorrelated with stock cash flows in the cross-section of stocks, after controlling for various cross-sectional stock characteristics (detailed in Section 3.1).

Order flow imbalance (OFI). Our third demand measure is the Lee-Ready signed order flow imbalance (OFI). Specifically, this measure takes all trades in the U.S. stock market, and uses the Lee and Ready (1991) algorithm to classify trades as buyer- or seller-initiated. We download daily Lee-Ready signed OFI data from WRDS intraday indicators, aggregate to quarterly frequencies, and then normalize it by lagged shares outstanding. We should clarify that OFI captures all trades that are executed *aggressively*, but not all trading flows. Many sophisticated institutional investors tend to execute trades slowly and passively to reduce price impact, and those flows will not be captured by OFI. The OFI measure has been used heavily in the microstructure literature to study price effects at daily or higher frequencies. Recently, Li and Lin (2023) argue that it can also be useful for studying demand effects at "asset pricing frequencies" (i.e. monthly or lower).

Relative to the previous two measures, OFI has the benefit of having a lot more variation. This can be seen in the summary statistics of Table 1. One standard deviation is equal to 0.54% of shares outstanding for FIT but 1.59% for OFI. In the tails, the 1% and 99% percentiles of OFI are -4.66% and 4.48%, respectively. This is particularly helpful for this paper because we are interested in price impacts of large demand shocks.

At the same time, a drawback of OFI is the concern about its information content. By its nature, OFI captures trading behavior by many investors, as opposed to FIT and BMI which zoom in on the trades by specific investors in specific circumstances. Li and Lin (2023) run extensive tests and do not find evidence that OFI is related to various measures of cash-flow news. However, they also admit that, due to the difficulty of measuring news, one cannot be fully certain about the information content of OFI.

Overall, our three demand shocks have their own respective strengths and weaknesses. Successful instruments need to be exogenous and relevant. In our context, exogeneity means not being correlated with cash flow-relevant news, and relevance refers to the amount of return variation it can explain. BMI is arguably the most exogenous, but index changes happen relatively rarely, so it has the lowest explanatory power over returns. Some researchers may be somewhat concerned about the exogeneity of FIT, but it has more variation as fund flow shocks impact all stocks over all periods. Finally, OFI probably has even more concern regarding its exogenenity, but it also has the highest amount of variation. Importantly, the findings we present in this paper are robust across these three measures, which we see as a strength: we do not rely on any single demand shock to draw our conclusions.

Table 1 reports summary statistics. Because we are interested in how the price multiplier varies by demand size, we show the percentile distributions in the right-most columns. All demand measures are reported as a percentage of lagged market capitalization. ΔBMI and FIT have similar ranges, with standard deviations of 0.72% and 0.54%, respectively.[4] OFI, which has the largest range, has a standard deviation of 1.59% and varies from -4.66% to 4.48% from the 1^{th} to the 99^{th} percentile.

Our analysis spans 1998 to 2018 for the BMI demand shock, and 1993 to 2022 for FIT and OFI. The BMI sample is limited by the period for which we observe Russell index constituents. The OFI sample is limited by the availability of the signed OFI measure from WRDS. We constrain the FIT sample to match that of OFI.

				Percentiles						
	Obs	Mean	StDev	1%	5%	25%	50%	75%	95%	99%
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Quarterly return (%)	4,712	2.82	30.36	-63.81	-41.00	-12.40	1.09	14.42	50.00	108.70
Demand (ΔBMI (scaled), %)	472	0.05	0.72	-2.10	-1.14	-0.16	0.00	0.24	1.32	2.32
Demand (FIT, %)	4,564	0.07	0.54	-1.22	-0.65	-0.12	0.01	0.21	0.98	1.92
Demand (OFI, %)	2,877	0.09	1.59	-4.66	-2.29	-0.57	0.12	0.79	2.42	4.48
Market cap ($m)	4,712	4,082	26,608	6	15	76	311	1,471	14,915	68,551

Table 1. Summary Statistics

Column (1) reports the average number of stocks per period. For FIT and OFI, one period is one quarter. For BMI, one period is the month of June — which is when Russell index reconstitution occurs — in a specific year. The sample for all variables consists of quarterly data from 1993 to 2022, except the BMI sample which consists of data for each June from 1998 through 2018. For the demand measures, ΔBMI refers to changes in benchmarking intensity in Pavlova and Sikorskaya (2023), FIT refers flow-induced trading in Lou (2012), and OFI refers to order flow imbalance in Li and Lin (2023). All demand measures (in percent) are expressed as a fraction of shares outstanding; for this purpose, we multiply ΔBMI by 0.2, following Pavlova and Sikorskaya (2023).

[4] Since ΔBMI measures the Russell index reconstitution-induced change in inelastic demand for a stock by only benchmarked mutual funds and ETFs — not the total change in demand by all institutional investors — we scale it to have the same units as OFI and FIT. Pavlova and Sikorskaya (2023) find that a one percentage point reconstitution-induced change in BMI raises total institutional ownership by 0.20 percentage points. See Column 3 of Table 3 in Pavlova and Sikorskaya (2023) for justification for this scaling.

3 Static Evidence of Endogenous Elasticities

In this section, we estimate the contemporaneous relationship between demand shocks and returns, and find evidence of concave price impact: price multipliers shrink with demand shock size. Section 3.1 provides evidence of this behavior using the three demand shocks discussed in Section 2. Section 3.2 examines alternative hypotheses.

3.1 Testing for Nonlinear Price Impact

To test for nonlinear demand effects, we estimate cross-sectional regressions of stock-level returns $r_{n,t}$ on demand shocks $d_{n,t}$ (ΔBMI, FIT, or OFI) and their interaction shock magnitudes $|d_{n,t}|$ in each time period t:

$$\forall t : r_{n,t} = b_{1,t} \cdot d_{n,t} + b_{2,t} \cdot |d_{n,t}| \cdot d_{n,t} + \boldsymbol{c}_t^{'} \boldsymbol{x}_{n,t-1} + \tau_t + \epsilon_{n,t}. \tag{2}$$

where the controls $\boldsymbol{x}_{n,t-1}$ include the stock characteristics discussed in Section 2, as well as the BMI-specific controls discussed in Section 2 when studying the BMI demand measure. We use the quarterly frequency for FIT and OFI because we are interested in price impacts that are sufficiently long-lasting to be of interest to asset pricing. For BMI, to be consistent with the prior literature (Pavlova and Sikorskaya, 2023), we use the monthly frequency and focus on the effects in the month of June when Russell index reconstitution occurs.

The implied shock-size dependent price multiplier in regression (2) is $M = b_1 + b_2 \times |d_{n,t}|$. The key coefficient of interest is b_2. If price impacts are smaller for larger demand shocks, then b_2 would be negative.

We estimate regression (2) in each cross section and report the average regression coefficients (in the style of Fama and MacBeth (1973)). Doing so ensures we use only cross-sectional variation to identify the regression coefficients, and thus avoids concerns about omitted variables driving time variation in both price multipliers and demand shock sizes. In Appendix Table B.3, we find that estimating regression (2) using panel regressions with time fixed effects yield similar results.

	Dependent variable: stock return $r_{n,t}$								
$d_{n,t}$ =	Δ BMI			FIT			OFI		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$d_{n,t}$	2.60***	2.06***	1.96***	4.48***	3.85***	3.97***	4.94***	4.67***	4.65***
	(0.69)	(0.58)	(0.57)	(0.62)	(0.45)	(0.37)	(0.21)	(0.18)	(0.18)
$d_{n,t} \times \lvert d_{n,t} \rvert$	−0.72***	−0.60***	−0.54**	−0.70**	−0.63***	−0.71***	−0.31***	−0.28***	−0.28***
	(0.25)	(0.22)	(0.21)	(0.33)	(0.24)	(0.21)	(0.02)	(0.02)	(0.02)
Predictor controls	N	Y	Y	N	Y	Y	N	Y	Y
Liquidity controls	N	N	Y	N	N	Y	N	N	Y
Obs	9,914	9,914	9,914	561,405	561,405	561,405	333,772	333,772	333,772
R^2	0.048	0.139	0.170	0.009	0.064	0.080	0.063	0.127	0.144
Marginal $R^2(d_{n,t})$	0.016	0.012	0.011	0.009	0.005	0.005	0.063	0.056	0.054

Table 2. Static Price Impact Regressions
We estimate cross-sectional regression (2) of stock returns on demand shocks and demand shocks interacted with their absolute values in each time period and report time-series average coefficients. Columns (1) through (3) report results using the ΔBMI demand shock and monthly returns. Columns (4) through (6) and (7) through (9) report results using quarterly returns and the FIT and OFI demand measures, respectively. In columns (2), (5), and (8), we also control for return-predicting stock characteristics. In columns (3), (6), and (9), we further control for liquidity-related proxies. Those controls are described in Section 2. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

Columns (1), (4), and (7) of Table 2 report the results of regression (2). The coefficient on $d_{n,t}$ (b_1) implies that price multipliers for small demand shocks (i.e. $|d_{n,t}| \approx 0$) lie in a range of 2.6 to 4.9. That is, for small X, buying $X\%$ of all shares outstanding of a stock raises price by around $2.6X\%$ to $4.9X\%$.[5]

The coefficient on $d_{n,t} \times |d_{n,t}|$ (b_2) is negative, indicating that larger shocks are associated with smaller price multipliers. The coefficient magnitudes represent how much the multiplier falls by when demand shock magnitude rises by 1% (i.e. $|d_{n,t}| = 1\%$ versus 2%). This interaction coefficient is statistically significantly negative at the 5% level for FIT and at the 1% level for BMI and OFI. The magnitude of b_2 is economically significant: assuming a constant price

[5]There are multiple reasons why price multiplier may between demand shocks beyond just shock sizes. For instance, ΔBMI primarily impacts stocks with smaller market capitalizations than FIT and OFI. Therefore, examining differences across demand shock types is not a "ceteris paribus" exercise and so we focus on variations within demand shock type.

multiplier significantly overestimates the actual price impact of large demand shocks.[6]

The point estimates for the three demand shocks suggest different degrees of nonlinearity, but the result are not statistically significantly different. The b_2 estimates in columns (1), (4), and (7) in Table 2 indicate that a 1% larger shock lowers the multiplier by about 0.72, 0.70, and 0.31 for the BMI, FIT, and OFI shocks, respectively. Due to the sizeable standard errors for BMI and FIT, however, the differences are not statistically significant.

In the other columns of Table 2, we add additional control variables and find qualitatively similar results. Specifically, columns (2), (5), and (8) control for the 13 return predictors discussed in Section 2.1. Columns (3), (6), and (9) further control for the 6 liquidity proxies discussed in Section 2.1.

Note that while we find price multipliers decline with shock size, we still find that multipliers — even for large shocks — are far larger than suggested by frictionless asset pricing models, which imply multipliers on the order of 1/6000 (e.g. Petajisto, 2009). Therefore, our findings should be interpreted as shedding light on the endogenous variation of price multipliers, rather than arguing that multipliers are as small as in frictionless asset pricing models.

3.2 Alternative Interpretation: Variation in Multipliers

We interpret the results in Section 3.1 as evidence that *ceteris paribus*, a stock exhibits a smaller price multiplier when experiencing a larger demand shock. However, the results may also be consistent with a general class of alternative interpretations: that price multipliers do not vary with shock size per se, but do vary for other reasons that correlate with shock

[6]For example, the result in Column 1 (BMI) imply that buying 1% of shares outstanding has a multiplier of about $M = 1.88$ ($b_1 + b_2 \times 1 = 2.60 - 0.72 \times 1 \approx 1.88$) and so raises price by about 1.88% ($M \times 1\%$), not 2.60% as a constant multiplier $M = b_1$ would suggest. A 2% shock has a multiplier of about $M = 1.16$ ($b_1 + b_2 \times 1 = 2.60 - 0.72 \times 2 \approx 1.16\%$) and so raises price by about 2.32% ($M \times 2\% = 2.32\%$), not 5.2% as a constant multiplier $M = b_1$ would suggest ($b_1 \times 2\% = 5.2\%$).

size. In general, one may suspect that the true model is:

$$r_{n,t} = \underbrace{(b + \beta_{n,t})}_{\text{price multiplier}} \cdot d_{n,t} + \mathbf{c}_t' \mathbf{x}_{n,t-1} + \tau_t + \epsilon_{n,t} \tag{3}$$

where $\beta_{n,t}$ denotes unobserved price multiplier variation that happens to be correlated with $|d_{n,t}|$. We address three variants of this concern.

	Dependent variable: stock return $r_{n,t}$								
$d_{n,t}$ =	Δ BMI			FIT			OFI		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$d_{n,t}$	2.00***	2.24***	2.30***	3.98***	3.89***	3.82***	4.65***	4.50***	4.65***
	(0.59)	(0.66)	(0.70)	(0.37)	(0.42)	(0.43)	(0.18)	(0.18)	(0.18)
$d_{n,t} \times \lvert d_{n,t} \rvert$	−0.53**	−0.69**	−0.76**	−0.74***	−0.77***	−0.69***	−0.28***	−0.36***	−0.33***
	(0.23)	(0.28)	(0.32)	(0.21)	(0.23)	(0.25)	(0.02)	(0.02)	(0.03)
Direct controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Interacted: predictors	N	Y	Y	N	Y	Y	N	Y	Y
Interacted: liquidity	N	N	Y	N	N	Y	N	N	Y
Obs	9,914	9,914	9,914	561,405	561,405	561,405	333,772	333,772	333,772
R^2	0.173	0.207	0.225	0.080	0.085	0.087	0.145	0.168	0.178
Marginal $R^2(d_{n,t})$	0.012	0.011	0.011	0.005	0.004	0.004	0.054	0.040	0.034

Table 3. Static price impact regressions: controlling for demand-characteristic interactions
We estimate cross-sectional regression (4) of stock returns on demand shocks and demand shocks interacted with their absolute values in each time period and report average coefficients across all available time periods. All regressions control for stock return predictors and liquidity proxies as described in Section 2.1. Columns (1) through (3) report results using the ΔBMI demand shock and monthly returns. Columns (4) through (6) and (7) through (9) report results using quarterly returns and the FIT and OFI demand measures, respectively. In columns (2), (5), and (8), we also control for the interaction of return predictors with demand. In columns (3), (6), and (9), we further control for the interaction of liquidity proxies with demand. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

Time-varying multipliers. One potential concern is that multipliers happen to be larger in times when demand shocks are smaller. For example, it is well-documented that liquidity conditions vary over time. For instance, during the 2008 financial crisis, liquidity conditions worsen so price multipliers are larger, and there is also less trading ($|d_{n,t}|$) at the same time. Fortunately, this potential concern is already addressed by our baseline specification, as we estimate (2) only using cross-sectional variation.

Heterogeneous multipliers across stocks with different liquidity. Another potential concern is that some stocks are more liquid and have smaller price multipliers than others, and those stocks also experience larger demand shocks.

Formally, in equation (3), this concern can be modeled as $\beta_{n,t} = b_{3,t}' \cdot \boldsymbol{z}_{n,t-1}$ where $\boldsymbol{z}_{n,t-1}$ are stock-specific liquidity-related characteristics.

To address this concern, we augment our baseline specification (2) with interactions of the demand shock with stock characteristics $\boldsymbol{z}_{n,t-1}$:

$$\forall t : r_{n,t} = b_{1,t} \cdot d_{n,t} + b_{2,t} \cdot |d_{n,t}| \cdot d_{n,t} + \underbrace{\boldsymbol{b}_{3,t}^{'} \cdot \boldsymbol{z}_{n,t-1}}_{\text{added}} \cdot d_{n,t} + \boldsymbol{c}_t^{'} \boldsymbol{x}_{n,t-1} + \tau_t + \epsilon_{n,t}. \tag{4}$$

where the interaction terms $\boldsymbol{b}_{3,t}^{'} \cdot \boldsymbol{z}_{n,t-1} \cdot d_{n,t}$ absorb the $\beta_{n,t} \cdot d_{n,t}$ term in (3). As in Section 3.1, we estimate regression (4) in each cross section and report the average coefficients, in the style of Fama and MacBeth (1973).

We report results in Table 3. For comparison purposes, columns (1), (4), and (7) reproduce the specification with all controls from Table 2. Columns (2), (5), and (8) also control for interactions of demand with return predictors. Columns (3), (6), and (9) further control for interactions of demand with the liquidity variables. The results are qualitatively unchanged, and we continue to see significant evidence for concave price impact curves. Appendix B.1 further shows that our results are robust throughout the process of progressively adding each individual characteristic interaction. Specifically, we plot the evolution of the coefficient of interest (b_2) as we add one interaction at a time, and find that the estimate is stable throughout all specifications. Overall, we find no evidence that our results are explained by cross-sectional correlations between stock characteristics and demand shock sizes.

Stock-specific variation in multiplier and demand shock sizes. To the extent that our liquidity proxies can indeed capture the variation of multipliers across stocks, Table 3 shows that differential liquidity does not explain our findings. However, one may still worry that there is further variation of liquidity across stocks that cannot be controlled for. Specifically, let M_n be the stock-specific multiplier and $\sigma(d_{n,t})$ be the stock-specific demand volatility. Suppose cross-sectional variation in M_n cannot be fully captured by liquidity controls. In this case, if stocks with lower M_n happen to have higher $\sigma(d_{n,t})$, then we may still spuriously find nonlinear price impact curves.

Under this alternative hypothesis, standardizing demand shocks

$$d_{n,t}^{\text{std}} = d_{n,t}/\sigma(d_{n,t}),$$

should remove the nonlinearity. We test this in Appendix Table B.6. Specifically, for each stock n in period t, we use the previous $h = 4, 8$, or 12 quarters to estimate the stock-specific demand volatility and construct standardized demand. When using standardized demand shocks in the price impact regression, we continue to see strong evidence of nonlinear price impact, which suggests that stock-specific liquidity is unlikely to drive our results.

3.3 Additional Results and Robustness

We now summarize several robustness checks. The detailed results are in Appendix B.

Piecewise regression specification. Our evidence for concave price impact in Section 3.1 are based on a parametric specification. We also consider a non-parametric specification using piecewise linear regressions. In each time period, we sort observations by realized demand shock sizes $|d_{n,t}|$ into bins $b = 1, 2, 3$, and then estimate cross-sectional regressions:

$$r_{n,t} = \sum_{\text{bin } b} M_{b,t} \cdot I_{|d_{n,t}| \in b} \cdot d_{n,t} + \boldsymbol{c}_t' \boldsymbol{x}_{n,t-1} + \tau_t + \epsilon_{n,t} \tag{5}$$

where the $\{M_b\}_b$ coefficients represent price multipliers for different shock sizes. To sort observations into bins, we compute the cross-sectional demand shock standard deviation ($\sigma(d_{n,t})$) in each period. Bin $b = 1$ collects observations with $|d_{n,t}| < \sigma(d_{n,t})$, $b = 2$ collects those with $\sigma(d_{n,t}) \leq |d_{n,t}| \leq 2\sigma(d_{n,t})$, and $b = 3$ collects the remaining observations. We use the same stock characteristics from Table 2 as controls $\boldsymbol{x}_{n,t-1}$. We report the time-series average estimated price multipliers and Fama-MacBeth standard errors.

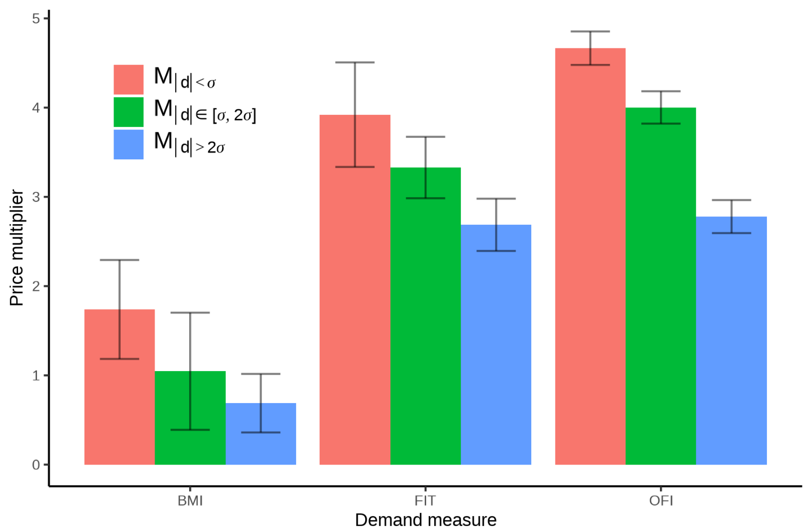


Figure 1. Multiplier by current shock size: piecewise linear specification

For each demand shock measure, we estimate cross-sectional Fama-MacBeth regressions:

$$r_{n,t} = \sum_{\text{bin } b} M_{b,t} \cdot I_{|d_{n,t}|\in b} \cdot d_{n,t} + \boldsymbol{c}_t^{'} \boldsymbol{x}_{n,t-1} + \tau_t + \epsilon_{n,t} \tag{7}$$

where $r_{n,t}$ is stock return and $d_{n,t}$ is the demand shock. In each period, we split the sample into three bins b by comparing $|d_{n,t}|$ with one or two times the cross-sectional standard deviation of $d_{n,t}$. The controls $\boldsymbol{x}_{n,t-1}$ are the same set of stock return predictors and liquidity proxies used in Table 2. The bars plot the time-series average price multiplier estimates ($\{M_b\}_b$). The whiskers represent 95% confidence intervals. Detailed regression results are in Table B.4.

The resulting estimated multipliers are plotted in Figure 1, with the error bars indicating 95% confidence intervals. The results are broadly consistent with those in Table B.4. For the BMI measure, the price multiplier is around two when considering demand shocks within one standard deviation, and it monotonically decreases to less than one for shocks larger than two standard deviations. Similar monotonic declines also appear for the multipliers based on FIT and OFI. Detailed regression results are shown in Table B.4 where we also conduct statistical tests of the differences between multiplier estimates and compute standard errors using the Delta method. For the FIT and OFI regressions, pairwise comparisons of multipliers associated with different demand shock sizes are always statistically significant at the 1% level. The regression using BMI has lower power, with around half of the multiplier differences being statistically significant, but the signs are always in the anticipated directions. Overall, we find these piecewise linear regressions to yield similar conclusions relative to the parametric specifications in Table 2.

Measurement errors. One may be concerned that our findings are driven by larger measurement error for larger demand shocks. If the shock magnitudes are overestimated for large shocks, we could spuriously find concave price impact.

Measurement error is unlikely to explain our findings. First, while measurement error is potentially a concern for the contemporaneous regressions in Section 3, it is less relevant for dynamic regressions in Section 4. We find that when the magnitude of cumulative *past* demand shocks is large in magnitude, the *contemporaneous* price multiplier is lower. As long as measurement errors are not serially correlated, this finding cannot be accounted for by attenuation biases.

Second, even for the contemporaneous results in this section, measurement error is an unlikely concern, as we explain in more details in Appendix B.2. OFI is measured directly by summing up signed trades, rather than imputed. Even though FIT is imputed, in the Appendix, we show that the results are not sensitive to filtering out large fund flows or flexibly estimating flow-to-trade pass-throughs.

Removing news days from the sample. Of the three demand measures we use, one may be most concerned about omitted variable bias in the multiplier estimates for the OFI measure. As discussed in Section 2.2, OFI captures aggressive trading from many market participants, and we do not know the exact motive behind their trading. While Li and Lin

(2023) do not find evidence that OFI is driven by information-based trading, they also admit it is difficult to rule out.

To address the concern that correlation with news releases may drive our OFI results, in Appendix B.3 we remove days with more news releases. Specifically, we remove days with earnings releases, analyst updates, or Ravenpack news releases from our sample. Our finding of nonlinear price impact is robust to removing these days.

4 Dynamic Evidence of Endogenous Elasticities

Section 3 finds that price multipliers decrease with *contemporaneous* demand shock magnitudes. If so, then price multipliers should also decrease with the magnitude of cumulative *past* shocks: if demand shocks are persistent, then price dislocations will depend on past shocks in addition to contemporaneous shocks. This section finds evidence consistent with this hypothesis.

4.1 Empirical Test

We test how multiplier vary with cumulative past shock size using FIT and OFI. We do not use the BMI measure because it only features shocks in the month of June and so does not have consecutive demand realizations. Specifically, we estimate cross-sectional regressions:

$$r_{n,t} = b_{1,t} \cdot d_{n,t} + b_{2,t} \cdot d_{n,t} \cdot \left| \sum_{l=1}^{L} d_{n,t-l} \right| + \boldsymbol{c_t}'\boldsymbol{x_{n,t-1}} + \tau_t + \epsilon_{n,t} \tag{8}$$

where the key coefficient of interest is $b_{2,t}$, which measures whether the price multiplier changes with the absolute value of cumulative demand shocks over the previous L quarters. As in Section 3, we estimate the regression by each cross-section and report the time-series average coefficients, in line with the Fama-MacBeth procedure. The vector of controls $\mathbf{x}_{n,t-1}$ is the same as in Table 2. To account for the use of lags in the regression, we compute

Newey-West standard errors with $L+1$ lags.

The regression results reported in Table 4 indicate that price multipliers decline with the magnitude of cumulative past demand shocks. Columns (1) through (4) report results based on FIT where we vary the number of lags L. For instance, column (1) shows that for each 1% increase in the magnitude of previous quarter's demand shock, the contemporaneous price multiplier declines by 1.62, and the interaction coefficient is statistically significant at the 1% level. To quantify the economic magnitude of multiplier variation, the last row reports that one standard deviation of the previous quarter's demand shock magnitude is 0.54%. Thus, one standard deviation change in this variable means that the contemporaneous price mutliplier declines by around $1.62 \times 0.54 \approx 0.87$, which is around one-fifth of the unconditional price multiplier.

Columns (2) through (4) vary the number of lags L. The interaction coefficient remains negative, indicating price impact declines with the magnitude of past cumulative demand shocks. The magnitude of coefficient b_2 declines, which is consistent with the idea that more recent demand shocks matter more.[7] Columns (5) through (8) examine OFI and find qualitatively consistent results.

Piecewise linear specification. As a robustness check, we also estimate piecewise linear specifications. Specifically, for each demand measure, we estimate:

$$r_{n,t} = \sum_{\text{bin } b} M_{b,t} \cdot I_{\left|\sum_{l=1}^{L} d_{n,t-l}\right| \in b} \cdot d_{n,t} + \boldsymbol{c}_t^{'} \boldsymbol{x}_{n,t-1} + \tau_t + \epsilon_{n,t} \tag{9}$$

where bins $b = 1, 2, 3$ are subsamples split by comparing the magnitude of past cumulative demand shock ($\left|\sum_{l=1}^{L} d_{n,t-l}\right|$) with its cross-sectional standard deviation. The resulting time-series average price multiplier estimates are plotted in Figure 2. For FIT, when the

[7] While the coefficient declines, because the standard deviation of cumulative demand shock increases with the lookback period (the last row), the resulting implied variation on price multiplier does not decrease.

Dependent variable: stock return $r_{n,t}$								
$d_{n,t}$ =	FIT				OFI			
	$L = 1$	2	3	4	$L = 1$	2	3	4
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$d_{n,t}$	4.11***	3.96***	4.00***	3.82***	3.83***	3.82***	3.74***	3.75***
	(0.36)	(0.34)	(0.32)	(0.33)	(0.26)	(0.30)	(0.34)	(0.36)
$d_{n,t} \times \lvert \sum_{l=1}^{L} d_{n,t-l} \rvert$	−1.62***	−0.84***	−0.68***	−0.44***	−0.20***	−0.14***	−0.09***	−0.10***
	(0.28)	(0.19)	(0.13)	(0.11)	(0.03)	(0.03)	(0.03)	(0.03)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Obs	538,398	517,033	497,185	478,324	314,110	296,241	279,914	264,872
R^2	0.081	0.082	0.084	0.085	0.144	0.146	0.147	0.148
$\sigma(\lvert \sum_{l=1}^{L} d_{n,t-l} \rvert)$	0.54%	0.94%	1.30%	1.63%	1.55%	2.16%	2.61%	2.95%

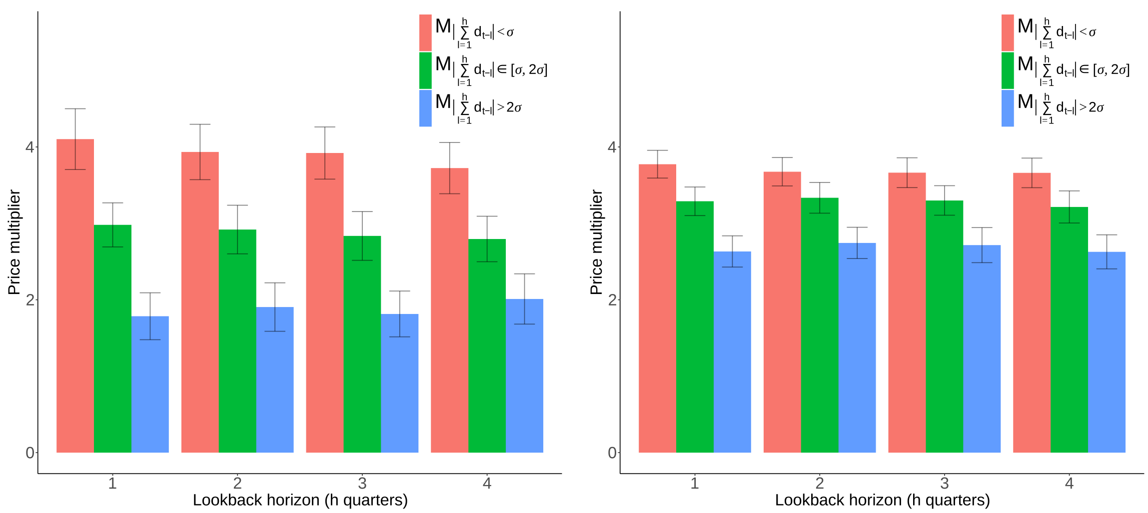
Table 4. Dynamic price impact regressions

We estimate cross-sectional regressions (8) using quarterly data and report the average time-series coefficients. Specifically, we regress stock return $r_{n,t}$ on the contemporaneous demand shock $d_{n,t}$, as well as the demand shock interacted with the absolute value of cumulative past shocks over L periods ($d_{n,t} \times \left| \sum_{l=1}^{L} d_{n,t-l} \right|$). The regressions control for the same set of characteristics as in Table 2. We compute Newey-West standard errors with $L + 1$ lags. The first four columns examine the FIT demand shock and the last four columns examine OFI. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01. The last row reports one standard deviation of the cumulative demand shock.

cumulative past demand shocks are less than one standard deviation, the price multiplier is around 4. The multiplier monotonically declines to around 2 when past cumulative demand shocks are above two standard deviations. For OFI, we see qualitatively similar patterns, though with a smaller range of multiplier variation. The detailed regression results are reported in Table B.5.

Overall, these results imply that price multipliers vary dynamically and decrease with the magnitude of past demand shocks.

Additional implications The dynamic results in this section help alleviate potential concerns about dynamic consistency. If price impact curves do not vary over time, but exhibit concavity in each period (as in Section 3), then one may be able to make profits through manipulative trades. Specifically, one can make profits by first buying 1 unit for N periods and then selling N units in period $N + 1$. While this is not a riskless arbitrage (it requires require taking risk over $N + 1$ quarters), it would nontheless be counterintuitive



(a) FIT (b) OFI

Figure 2. Multiplier by past demand shock size: piecewise linear specification

For each demand shock measure and lookback horizon h, we estimate cross-sectional regressions:

$$r_{n,t} = \sum_{\text{bin } b} M_{b,t} \cdot I_{\left|\sum_{l=1}^{h} d_{n,t-l}\right| \in b} \cdot d_{n,t} + \boldsymbol{c}_t^{'} \boldsymbol{x}_{n,t-1} + \tau_t + \epsilon_{n,t} \tag{11}$$

where $r_{n,t}$ is stock return and $d_{n,t}$ is the demand shock. In each period, we split the sample into three bins b by comparing $\left|\sum_{l=1}^{h} d_{n,t-l}\right|$ with its cross-sectional standard deviation σ. The controls $\boldsymbol{x}_{n,t-1}$ include the return-predicting and liquidity characteristics used in Table 2. The bars plot time-series average multiplier estimates $\hat{M}_b$ along with 95% confidence intervals. Panels (a) and (b) shows results for demand measures FIT and OFI, respectively. The regression results are presented with more details in Table B.5.

for such strategies to be profitable on average. In fact, Huberman and Stanzl (2004) predict that, if contemporaneous price impact curves are not linear, then price impacts should vary dynamically. Our results are consistent with their prediction.

Second, the dynamic findings also help alleviate concerns about measurement error. As explained in Section 3.3, one may worry that our contemporaneous results are due to larger measurement error for larger measured demand shocks. However, as long as measurement errors are not serially correlated over time, this is unlikely to explain how price multipliers vary with *past* demand shocks.

5 Evidence from Investor Holdings

The preceding sections have shown that price multipliers are smaller when either contemporaneous or past cumulative demand shocks are larger. Given that equilibrium price multipliers are inversely related to investor demand elasticities, we next examine institutional holdings data to directly investigate the behavior of these elasticities using the granular instrumental variable (GIV) approach. Our analysis reveals that demand elasticities increase with the magnitude of accumulated price dislocations. We further find that this effect is driven in large part by within-investor intensive-margin adjustments rather than capital shifting across investors or investors entering or exiting the asset. Together, these results help pinpoint the micro-level mechanism responsible for the endogenous variation in price multipliers.

5.1 GIV Methodology

We estimate investor-level demand elasticities using the granular instrumental variable (GIV) methodology of Gabaix and Koijen (2022, 2024). Specifically, we develop a dynamic, nonlinear extension of the optimal GIV methodology from Chaudhary et al. (2024), which is designed to maximize statistical power when estimating heterogeneous elasticities. In essence, the GIV approach constructs a price instrument for each investor by aggregating the idiosyncratic trades of all other investors under a set of structural assumptions. We adopt this methodology because it offers substantially greater statistical power for estimating demand elasticities compared to the demand measures used in the preceding sections.

5.1.1 Nonlinear and Dynamic Model for Investor Demand

We are interested in whether investors' demand becomes more elastic when accumulated past price changes are larger in magnitude. Following the results in Section 4, we model demand as:

$$\Delta q_{i,n,t} = -\underbrace{\left(\zeta_{1,i,t} + \zeta_{2,i,t}\tilde{P}_{n,t}\right)}_{\text{Price Elasticity of Demand}} \Delta p_{n,t} + \underbrace{\boldsymbol{\lambda}_{i,t}^{'}\boldsymbol{\eta}_{n,t} + u_{i,n,t}}_{\text{Demand Shock}} \tag{12}$$

$$\tilde{P}_{n,t} = \left|\sum_{l=1}^{L}\Delta p_{n,t-l}\right| - \mathbb{E}_t\left[\left|\sum_{l=1}^{L}\Delta p_{n,t-l}\right|\right],$$

where $\Delta q_{i,n,t}$ is the change in log quantity of shares demanded by investor i for stock n in quarter t. This quantity change has two components: movement along the demand curve and demand shocks (i.e. demand curve shifts). The first component is the price elasticity of demand times the percentage price change ($\Delta p_{n,t}$), which captures how demand reacts to prices after holding constant the rest of the investor's information set.

The second component captures changes in non-price determinants of demand, such as new information about future cash flows or risk. $\boldsymbol{\eta}_{n,t}$ represents stock-quarter specific characteristics such as firm-level fundamentals or news. $u_{i,n,t}$ are residual demand shocks that are assumed to be conditionally mean-independent across investors, within and across quarters:

$$\forall i \neq j, \forall t, \forall l = 0, \ldots, L : \mathbb{E}[u_{i,n,t} \mid u_{j,n,t-l}, \boldsymbol{\eta}_{n,t}] = \mathbf{0}. \tag{13}$$

Our main goal is to test if $\zeta_{2,i,t} > 0$. Here, $\zeta_{1,i,t}$ represents the average elasticity, while $\zeta_{2,i,t}$ captures how elasticity changes with the magnitude of demeaned cumulative price changes over the past L quarters ($\tilde{P}_{n,t}$). Thus, finding $\zeta_{2,i,t} > 0$ ($\zeta_{2,i,t} < 0$) indicates that demand becomes more (less) elastic when past price changes have larger magnitudes.

5.1.2 Using GIV to Identify Elasticities

A key challenge in identifying elasticities is that demand and prices are jointly determined in equilibrium. Through market clearing, investor i's demand shock correlates with the contemporaneous price change. Furthermore, if demand shocks are serially correlated, the current demand shock will also correlate with lagged price changes. Thus, a simple regression of equilibrium quantity changes on current and lagged price changes fails to identify $\zeta_{1,i,t}$ and $\zeta_{2,i,t}$ due to omitted variable bias. Instead, identification requires exogenous variation in current and lagged price changes that is uncorrelated with investor i's demand shocks.

To identify the elasticity parameters $\zeta_{1,i,t}$ and $\zeta_{2,i,t}$, we build on the optimal granular instrumental variables (GIV) methodology of Chaudhary et al. (2024). This identification strategy is motivated by market clearing. Assuming a fixed supply, the aggregate equilibrium change in quantity demanded must be zero:

$$0 = \sum_i S_{i,n,t-1} \Delta q_{i,n,t}, \tag{14}$$

where $S_{i,n,t-1}$ are lagged ownership share weights — the proportion of outstanding shares held by investor i. After plugging (12) into (14), a first-order approximation around $\tilde{P}_{n,t} = 0$ yields the equilibrium price change:

$$\Delta p_{n,t} \approx \underbrace{\left(\frac{1}{\zeta_{1,S,t}} - \frac{\zeta_{2,S,t}}{\zeta_{1,S,t}^2} \tilde{P}_{n,t} \right)}_{\text{Price Multiplier}} \left(\boldsymbol{\lambda}_{S,t}' \boldsymbol{\eta}_{n,t} + u_{S,n,t} \right), \tag{15}$$

where the S subscript denotes ownership-share weighted averages (e.g., $u_{S,n,t} = \sum_i S_{i,n,t-1} u_{i,n,t}$). Intuitively, the demand shocks of investors with larger ownership shares have a greater impact on prices. Additionally, as we find empirically in Section 4, the price multiplier depends on the magnitude of past price changes ($\tilde{P}_{n,t}$), and the relationship is governed by $-\zeta_{2,S,t}/\zeta_{1,S,t}^2$.

By the assumption in (13), residual demand shocks are uncorrelated across investors. For each investor i, we can combine the residual demand shocks of other investors $j \neq i$ to form a price instrument. Furthermore, Chaudhary et al. (2024) show that statistical power is maximized by constructing the GIV instrument as the ownership-share weighted average of other investors' residual demand shocks, $z_{i,n,t} = \sum_{j \neq i} S_{j,n,t-1} u_{j,n,t}$. Thus, the following two moment conditions identify $\zeta_{1,i,t}$ and $\zeta_{2,i,t}$:

$$0 = \mathbb{E}\left[u_{i,n,t} \cdot z_{i,n,t} \mid \boldsymbol{\eta}_{n,t}\right] \tag{16}$$

$$0 = \mathbb{E}\left[u_{i,n,t} \cdot z_{i,n,t} \cdot \tilde{Z}_{i,n,t} \mid \boldsymbol{\eta}_{n,t}\right], \tag{17}$$

$$\text{for } z_{i,n,t} = \sum_{j \neq i} S_{j,n,t-1} u_{j,n,t}, \text{ and}$$

$$\tilde{Z}_{i,n,t} = \left|\sum_{l=1}^{L} z_{i,n,t-l}\right| - \mathbb{E}_t\left[\left|\sum_{l=1}^{L} z_{i,n,t-l}\right|\right]$$

where the expectations are taken across stocks n in quarter t. The moment conditions above are formally equivalent to a two-stage least squares (2SLS) regression of $\Delta q_{i,n,t}$ on both $\Delta p_{n,t}$ and the interaction term $\Delta p_{n,t} \cdot \tilde{P}_{n,t}$, witgh $z_{i,n,t}$ and $z_{i,n,t} \cdot \tilde{Z}_{i,n,t}$ as instruments, and $\boldsymbol{\eta}_{n,t}$ included as controls. Intuitively, $z_{i,n,t}$ and $\tilde{Z}_{i,n,t}$ provide exogenous variation in the current price change and the magnitude of past price changes.

A practical challenge, however, is that constructing these instruments requires observing other investors' residual demand shocks:

$$u_{j,n,t} = \Delta q_{j,n,t} + \left(\zeta_{1,j,t} + \zeta_{2,j,t}\tilde{P}_{n,t}\right)\Delta p_{n,t} - \boldsymbol{\lambda}_{j,t}^{'}\boldsymbol{\eta}_{n,t},$$

which in turn requires knowledge of their elasticity parameters $(\zeta_{1,j,t}, \zeta_{2,j,t})$. To resolve this problem, we estimate $\zeta_{1,j,t}$ and $\zeta_{2,j,t}$ for all investors simultaneously using generalized method of moments (GMM) with moment conditions (16) and (17).

5.1.3 Empirical Specification

We estimate the nonlinear demand curve from equation (12) using institutional holdings data from the SEC Form 13F, obtained from Thomson Reuters.[8] We restrict our sample to intensive-margin quantity changes, defined as observations where investor i holds a positive quantity of stock n in both quarters $t-1$ and t.

We parsimoniously characterize heterogeneity in the elasticity parameters across investors by parameterizing $\zeta_{1,i,t}$ and $zeta_{2,i,t}$ as linear functions of each investor's active share:

$$\zeta_{k,i,t} = \zeta_{k,0,t} + \zeta_{k,\text{Active Share},t} \cdot \text{Active Share}_{i,t-1-L}.$$

Active share is the demeaned sum of the absolute deviations between an investor's portfolio weights and the market cap-weighted portfolio weights for that investor's universe (divided by two). [9] We use a four-quarter rolling average of active share, lagged to quarter $t-1-L$, to prevent correlation with demand shocks from quarters $t-L$ to t. For $k \in \{1,2\}$, $\zeta_{k,0,t}$ represents the parameter for the average investor, while $\zeta_{k,\text{Active Share},t}$ describes how the parameter varies with active share. This parametric approach preserves statistical power since we lack power to estimate the elasticity parameters separately for every investor.

[8]Following Davis and Haltiwanger (1992), we calculate $\Delta q_{i,n,t}$ as:

$$\Delta q_{i,n,t} = \frac{\hat{Q}_{i,n,t} - \hat{Q}_{i,n,t-1}}{\frac{1}{2}(\hat{Q}_{i,n,t} + \hat{Q}_{i,n,t-1})},$$

where $\hat{Q}_{i,n,t-1} = H_{i,n,t-1}$ is investor i's dollar holdings in stock n in quarter $t-1$, and $\hat{Q}_{i,n,t} = H_{i,n,t}/(1 + R^X_{n,t-1\to t})$ is the dollar holdings in quarter t adjusted for the ex-dividend return $R^X_{n,t-1\to t}$. This specification is motivated by potential measurement errors in share data (e.g., unadjusted stock splits) in 13F filings. Using dollar holdings circumvents this issue, while the denominator, which maps the expression into the range $[-2, 2]$, effectively winsorizes large percentage changes.

[9]Following Koijen and Yogo (2019), an investor's universe is all stocks held in the past 11 quarters.

As in the preceding sections, we estimate the parameters for each quarter t using the cross-section of stocks. To improve numerical stability, we perform the estimation over rolling four-quarter windows and report the average parameter estimates across all windows. We compute Fama-MacBeth standard errors with a Newey-West correction (8 lags) to account for the overlapping windows.

The stock characteristic controls, $\boldsymbol{\eta}_{n,t}$, include the observed characteristics discussed in Section 2 and five latent characteristics estimated via principal component analysis on the investor-stock ownership panel. Our results are robust to using different numbers of latent characteristics (Appendix C.2). We also include investor-quarter fixed effects to ensure that the elasticity parameters are identified solely from cross-sectional variation within each investor's portfolio. Appendix C.1 provides further estimation details.

Table 5 summarizes the holdings data we use. The mean $\Delta q_{i,n,t}$ is around zero, and one standard deviation is around 46%, which corresponds to buying or selling 46% of pre-existing holdings. The mean active share is 0.44 and one standard deviation of active share is 0.20.

				Percentiles						
	Obs	Mean	StDev	1%	5%	25%	50%	75%	95%	99%
$\Delta q_{i,n,t}$	31,544,619	-0.04	45.67	-163.13	-74.01	-7.67	0.22	7.96	72.66	161.56
$\Delta p_{n,t}$	31,544,619	3.67	18.63	-44.40	-25.61	-6.08	3.36	12.89	33.33	57.22
$\sum_{l=1}^{L} \Delta p_{n,t-l}$	31,544,542	31.55	36.11	0.40	2.03	10.41	22.39	40.87	88.00	167.66
Active Share$_{i,t}$	31,544,619	0.44	0.20	0.07	0.10	0.30	0.44	0.58	0.77	0.88

Table 5. Summary Statistics
This table provides summary statistics for the data underlying the GIV model estimation in Section 5. Variables include quarterly percentage changes in quantity of shares demanded ($\Delta q_{i,n,t}$), quarterly percentage changes in price ($\Delta p_{n,t}$), percentage absolute accumulated lag price changes over the past four quarters ($|\sum_{l=1}^{L=4} \Delta p_{n,t-l}|$), and investor-level active share (Active Share$_{i,t}$). The sample for all variables consists of quarterly data from 1982 to 2021.

5.2 Empirical Results

Table 6 displays the estimation results for the nonlinear, dynamic demand function (12). The first two columns display the time-series average of the estimates of $\zeta_{1,0}$ and $\zeta_{1,\text{Active Share}}$

	$\zeta_{1,0}$	$\zeta_{1,\text{Active Share}}$	$\zeta_{2,0}$	$\zeta_{2,\text{Active Share}}$
Coefficient	0.4580***	1.3880***	0.0072***	0.0325***
	(0.0424)	(0.1554)	(0.0028)	(0.0110)

Table 6. Nonlinear and Dynamic Demand Curve Estimation

We estimate nonlinear demand curve (12) using quarterly 13F holdings data in rolling annual window and report the average coefficient values over all time periods. The standard errors are computed using the Newey-West method with 8 quarterly lags. The first column ($\zeta_{1,0}$) displays the average elasticity for investors with average active share. The second column ($\zeta_{1,\text{Active Share}}$) displays how much the average elasticity varies with active share. The third column ($\zeta_{2,0}$) displays how much elasticity changes with the magnitude of past accumulated price changes for the average investor. The fourth column ($\zeta_{2,\text{Active Share}}$) displays how much this sensitivity of elasticity to the magnitude of past accumulated price average elasticity varies with active share. The estimation controls for the same set of controls as in Table 2, as well as five latent stock-characteristics estimated using PCA. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

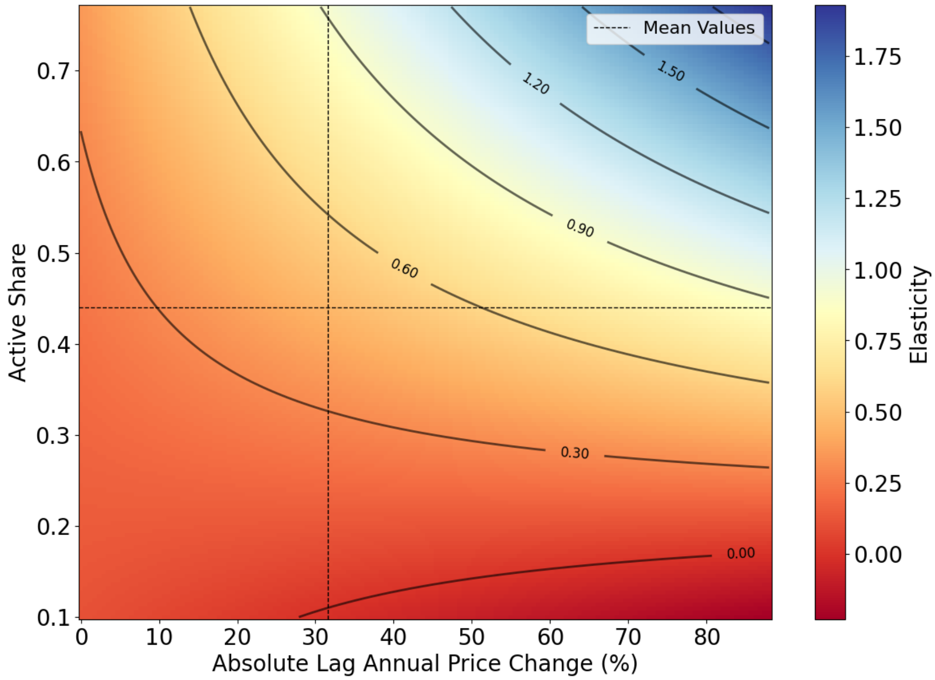
— the parameters that control the elasticity for the average stock (i.e. when the demeaned absolute lag accumulated price change $\tilde{P}_{n,t} = 0$). In the first column, we find $\zeta_{1,0} \approx 0.46$: on average investors reduce their quantity demanded by 0.46% in reponse to a 1% ceteris paribus price increase. This estimate is similar to the investor-level elasticity estimates from Koijen and Yogo (2019). In the second column we find a positive $\zeta_{1,\text{Active Share}}$: more active investors have higher elasticities. The magnitude of $\zeta_{1,\text{Active Share}} \approx 0.69$: impies that a one standard deviation increase in active share above the mean raises elasticity by about $1.39 \times 0.2 \approx 0.28$, or from $\zeta_{1,0} = 0.46$ for the average investor to 0.74, an economically significant increase.

The last two columns display the estimates of $\zeta_{2,0}$ and $\zeta_{2,\text{Active Share}}$, the parameters of interest that govern how demand elasticity varies with past cumulative price changes. In the third column, we find $\zeta_{2,0} > 0$: the average investor becomes more elastic as accumulated past price changes grow in magnitude. In the fourth column we find $\zeta_{2,\text{Active Share}} > 0$, which implies that more active investors display this behavior to a greater degree — they become even more elastic (compared to the average investor) as accumulated past price changes grow in magnitude.

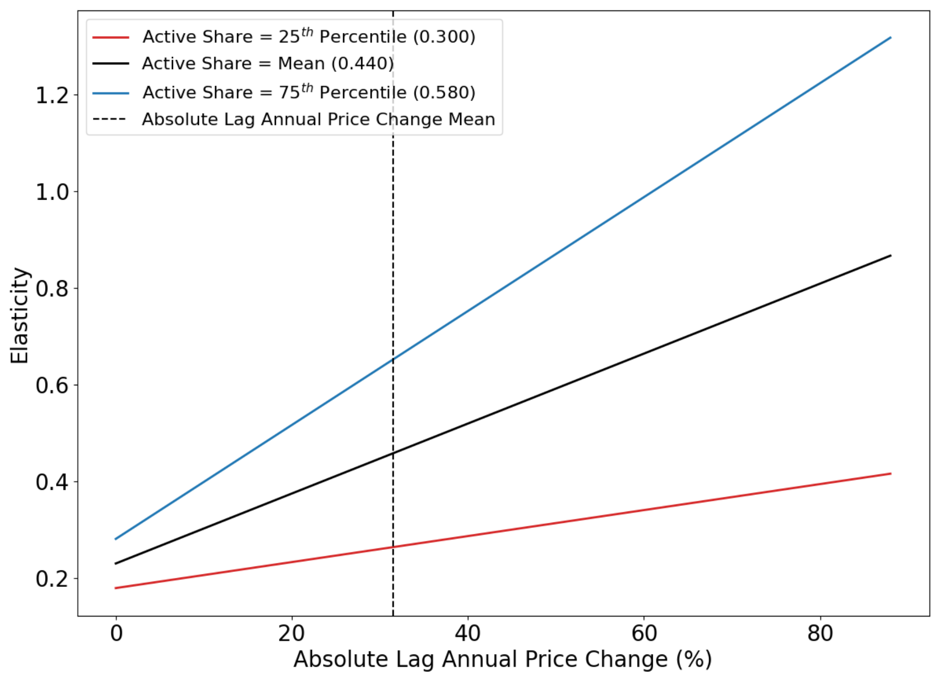
The magnitudes of $\zeta_{2,0}$ and $\zeta_{2,\text{Active Share}}$ are economically significant. Figure 3 Panel (a) displays total elasticity as a function of active share and absolute past price change. Figure 3

Panel (b) displays total elasticity as a function of absolute past price change for fixed active shares. For the average investor, moving from the 5th to the 95th percentile of absolute past annual price change (from approximately 2% to 88% from Table 5) raises the total elasticity from 0.23 to 0.87. For more active investors, such as those in the 75th percentile of active share, moving from the 5th to the 95th percentile of absolute past annual price change raises the total elasticity from 0.28 to 1.32. Even for relatively inactive investors — those in the 25th percentile of active share — moving from the 5th to the 95th percentile of absolute past annual price change raises the total elasticity from 0.18 to 0.42. Thus, we find that most investors indeed become more elastic in the face of larger past price dislocations and that this behavior is particularly pronounced among more active investors.

Crucially, since we restrict our sample to intensive-margin quantity changes, these results provide evidence that the endogenous multiplier results from Sections 3 and 4 are not driven solely by investors entering or exiting on the extensive margin. Investors become more elastic as past accumulated price changes grow in magnitude even for stocks they already hold and continue to hold. Moreover, since we estimate demand curve (12) in the cross-section of each investor's holdings, these results provide evidence that our endogenous multiplier results are not driven solely by reallocation of AUM to more elastic investors when past price changes grow in magnitude.



(a) Elasticity as a Function of Active Share and Past Price Changes



(b) Elasticity as a Function of Past Price Changes for Fixed Active Shares

Figure 3. Visualizing variation in demand elasticities

We visualize the variation in institutional demand elasticities estimated in Section 5.2. Panel (a) uses a heatmap to show the estimated demad elasticities as a function of the absolute value of lagged four-quarter price change ($\sum_{l=1}^{L=4} \Delta p_{n,t-l}$) on the x-axis and the active share of institutions on the y-axis. Panel (b) plots the model-implied demand elasticity as a function of past price changes for different fixed active shares.

6 Implications for the Microfoundations of Inelastic Demand

There are many mechanisms that can generate inelastic demand. In this section, we use our empirical findings to shed light on the applicability of various mechanisms. Many mechanisms fail to explain at least some aspect of our findings, which includes the static multiplier results in Section 3, the dynamic multiplier results in Section 4, and the within-investor, intensive-margin holdings-based results in Section 5. Our results therefore shed light on the microfoundations of inelastic demand.

In this section, we first present a stylized model that nests many different frictions that can give rise to inelastic demand. We then specialize the model to particular frictions, derive the resulting implications, and examine whether those implications are consistent with our empirical results.

6.1 Stylized Model

The model has three periods: $t = 0, 1, 2$. At $t = 0$, investors receive a fixed endowment. At $t = 1$, a supply shock occurs, the asset market clears, and the equilibrium price is determined. At $t = 2$, the asset's payoff is realized. We explore the static and dynamic properties of demand elasticities and price multipliers by characterizing how these objects at $t = 1$ relate to investor holdings at $t = 0$ and $t = 1$.

Assets. There is one risky asset that pays a dividend at $t = 2$:

$$\tilde{D} \sim N\left(\bar{D}, \sigma_D^2\right).$$

Unless otherwise noted, all investors are endowed with Θ_0 shares of the asset at $t = 0$. At $t = 1$, an exogenous supply shock changes total supply to Θ_1. This supply should be interpreted as the *residual* supply faced by optimizing investors: the total fixed supply less the demand from noise traders outside the model. The risk-free rate is normalized to zero.

Agents and Preferences: There are two types of investors. First, there is mass $1-\epsilon$ of type-I (inelastic) investors who choose portfolios at $t=1$ to maximize mean-variance utility over $t=2$ (i.e. terminal) wealth subject to a reduced-form cost $C(Q, \mathbb{E}\left[\tilde{D}-P\right], \Theta_0)$:

$$\max_{Q_I} Q_I \mathbb{E}\left[\tilde{D}-P\right] - \frac{\gamma}{2} Q_I^2 \mathbb{V}\left[\tilde{D}-P\right] - C\left(Q_I, \underbrace{\mathbb{E}[\tilde{D}-P]}_{\equiv \mu}, \Theta_0\right) \tag{18}$$

The cost can in general depend on the quantities held in both this and last period (Q_I and Θ_0), as well as the expected return μ. This cost renders these investors' demand inelastic and can nest many different frictions, as discussed in Section 6.2.[10]

Taking the first-order condition, the type-I investors' optimal quantity satisfies

$$P = \bar{D} - Q \cdot \gamma \sigma_D^2 - \underbrace{\frac{\partial}{\partial Q} C\left(Q, \mu, \Theta_0\right)}_{\equiv MC(Q, \mu, \Theta_0)}. \tag{19}$$

$MC\left(Q, \mu, \Theta_0\right)$ is the marginal cost of holding one more share of the asset.

Second, there is a fringe of type-E (elastic) investors with mass $0<\epsilon<1$ who do not face cost $C(\cdot)$:

$$\max_{Q_E} Q_E \mu - \frac{\gamma}{2} Q_E^2 \sigma_D^2 \rightarrow Q_E = \frac{\mu}{\gamma \sigma_D^2}$$

These investors ensure the market clears even if the type-I investors are constrained.

[10]Assume $\frac{\partial^2 C}{\partial Q^2} > 0$ so the investor's objective function is strictly concave.

Equilibrium: Market-clearing implies

$$(1-\epsilon)Q_I + \epsilon Q_E = \Theta_1. \tag{20}$$

Substituting this equation into the type-I first-order condition (19) yields equilibrium price

$$P = \bar{D} - \Theta_1 \cdot \gamma\sigma_D^2 - \frac{1}{1-\epsilon} MC\left(\Theta_1 - \epsilon Q_E, \mu, \Theta_0\right). \tag{21}$$

Price depends on marginal cost. If the marginal cost of holding an additional share is high, then price must be low to incentivize the type-I investors to do so.

Differentiating both sides of (21) with respect to Θ_1 yields the following price multiplier as $\epsilon \to 0$ (the per-unit change in price due to a change in Θ_1, expressed as a positive number):

$$M \equiv -\frac{dP}{d\Theta_1} = \frac{\gamma\sigma_D^2 + \frac{\partial}{\partial\Theta_1} MC\left(\Theta_1, \mu, \Theta_0\right)}{1 - \frac{\partial}{\partial\mu} MC\left(\Theta_1, \mu, \Theta_0\right)}. \tag{22}$$

The multiplier M can be large for two reasons. First, M is large if marginal cost increases with quantity held ($\frac{\partial MC}{\partial\Theta_1} > 0$). If holding more shares requires paying a higher marginal cost, the price must adjust (i.e. drop) to incentivize investors to do so. Second, M is large if marginal cost increases as expected return rises ($\frac{\partial MC}{\partial\mu} > 0$). In this case, when supply rises and expected return rises because price falls, investors' marginal cost rises. Thus, price must fall even more to compensate investors.

The endogenous behavior of the multiplier depends on the nonlinear relationships between marginal cost, current and past supply shocks, and expected returns. These relationships vary across different microfoundations for the cost function, as discussed in the next section.

6.2 Endogenous Multiplier Behavior with Different Frictions

Our empirical results discipline the set of underlying frictions for the cost function $C(\cdot)$ that can be used to microfound inelastic demand. We examine which frictions can and cannot explain each of our empirical results, as summarized in Table 7.

	Price Multiplier Results		**Demand Elasticity Results**	
	Multiplier Decreases with Contemporaneous Shock Size	Multiplier Decreases with Accumulated Shock Size	Elasticity Increases Within Investor	Elasticity Increases on Intensive Margin
Position Constraints	**No**	**No**	**No**	**No**
Benchmarking & Investment Mandates	**No**	**No**	**No**	**No**
Convex Adjustment Costs	**Yes**	**No**	**No**	**No**
Reallocation of Capital Across Investors	**Yes**	**Yes**	**No**	**No**
Participation Costs	**Yes**	**Yes**	**Yes**	**No**
Fixed Adjustment Costs	**Yes**	**Yes**	**Yes**	**Yes**
Costly Information Acquisition	**Yes**	**Yes**	**Yes**	**Yes**

Table 7. The Empirical Facts that Can or Cannot be Explained by Mechanisms
This table summarizes the implications of different mechanisms for explaining variation in price multipliers and demand elasticities. Each row corresponds to one mechanism and each column corresponds to an empirical finding in this paper.

6.2.1 Mechanisms that Do Not Explain Static Multiplier Results

In Section 3, we find that price multipliers decrease with contemporaneous shock size. Mechanisms emphasizing position constraints, as well as those emphasizing benchmarking and investment mandates, cannot explain this result (rows 1 and 2 in Table 7).

Position constraints. Contrary to our findings, models with position constraints, such as leverage constraints (e.g., Shleifer and Vishny (1997); He and Krishnamurthy (2013)) and short sale constraints (e.g., Diamond and Verrecchia (1987); Duffie et al. (2002)), often imply price multipliers *increase* — not decrease — with supply shock magnitude. To show this, Appendix D.1 considers a model in which type-I investors face a quantity limit: $|Q| \leq \alpha$. In this case, the cost function can be written as

$$C(Q, \mu) = -\lambda(\mu)(\alpha - |Q|).$$

where $\lambda(\mu)$ is the shadow cost of the constraint. For small shocks ($|\Theta_1| \leq \alpha$), the constraint does not bind and $\lambda(\mu) = 0$. For large shocks ($|\Theta_1| > \alpha$), the constraint binds, and $\lambda(\mu)$ increases with the magnitude of the expected return μ. Intuitively, when supply shocks are large, expected returns are also large, so these investors want to take large positions but cannot due to the constraint. Thus, per (22), the multiplier increases with the supply shock magnitude $|\Theta_1|$:

$$M = \begin{cases} \gamma\sigma_D^2 & , |\Theta_1| \leq \alpha \\ \gamma\sigma_D^2/\epsilon & , |\Theta_1| > \alpha \end{cases}$$

Large shocks tighten investors' constraints, lower their elasticities, and thus raise multipliers. Hence, position constraints do not explain the static results in Section 3.

Benchmarking and Investment Mandates Contrary to our findings, models with benchmarking or investment mandates often imply constant price multipliers (e.g., Petajisto (2009); Gabaix and Koijen (2022); Pavlova and Sikorskaya (2023)). Appendix D.2 considers a model in which type-I investors maximize mean-variance utility over compensation, where compensation is a linear combination of both absolute return and return relative to a benchmark (as in Pavlova and Sikorskaya (2023)):

$$c(Q_I) = \underbrace{aQ_I\left(\tilde{D}-P\right)}_{\text{Absolute Return}} + \underbrace{b\left[Q_I\left(\tilde{D}-P\right)-\bar{Q}\left(\tilde{D}-P\right)\right]}_{\text{Return Relative to Benchmark}},$$

where $\bar{Q}$ represents the benchmark holding. In this case, the cost function can be written as

$$C(Q_I,\mu) = \mu(\lambda(Q_I)+Q_I) + \frac{\gamma}{2}\sigma^2_{D_\epsilon}\left[\lambda(Q_I)^2 - Q_I^2\right],$$

where $\lambda(Q_I)$ is linear and decreasing in Q_I. This cost represents the welfare loss to an investor who would prefer to maximize mean-variance utility over absolute returns but is instead incentivized to tilt toward the benchmark. Since mean-variance utility (over both absolute returns and compensation) is quadratic in quantity, so is this cost function. As a result, the marginal cost is linear, and its derivatives with respect to quantity and expected return are constant. Thus, per (22), the multiplier is constant:

$$M = \frac{\gamma\sigma^2_D(a+b)}{1+(a+b-1)\epsilon},$$

Intuitively, an investor's unwillingness to deviate from a benchmark renders their demand inelastic, resulting in large price multipliers. However, because this unwillingness does not vary with quantity or expected return, the price multiplier is constant. Hence, benchmarking or investment mandates cannot explain the static multiplier results in Section 3.

6.2.2 Mechanisms That Do Not Explain Dynamic Multiplier Results

In Section 4, we find that price multipliers decrease with the size of past accumulated shocks, a result that standard convex adjustment cost models cannot explain (row 3 in Table 7).

(Contemporaneous) Convex Adjustment Costs Contrary to our findings, models in which investors face costs to adjust their portfolios imply multipliers do not depend on past accumulated supply shocks (e.g., Gârleanu and Pedersen (2013); Van der Beck (2025)). Appendix D.3 considers a model with a general adjustment cost function

$$C(Q_I, \Theta_0) = \lambda \left| Q_I - \Theta_0 \right|^{\alpha},$$

that penalizes type-I investors for adjusting from their initial holdings. The cost is convex for $\alpha > 1$. For $\alpha < 2$, the *marginal* adjustment cost is concave (i.e., $\frac{\partial MC}{\partial Q_I}$ decreases with Q_I). Thus, per (22), such adjustment costs can generate multipliers that decrease with the *contemporaneous* supply shock size ($|\Theta_1 - \Theta_0|$). Concretely, for $\epsilon \to 0$, the multiplier is

$$M = \gamma\sigma_D^2 + \alpha(\alpha - 1)\lambda|\Theta_1 - \Theta_0|^{\alpha-2}.$$

Intuitively, for $\alpha < 2$, large adjustments are relatively cheaper than small ones, so M declines with shock size. Thus, certain convex adjustment costs can explain the static results in Section 3. However, since the cost depends only on the quantity *change* ($Q_I - \Theta_0$) and not on the level of past holdings (Θ_0), multipliers do not depend on past shocks. Hence, convex adjustment costs cannot explain the dynamic multiplier results in Section 4.

6.2.3 Mechanisms That Do Not Explain Within-Investor Elasticity Results

In Section 5, we find that investor-specific demand elasticities increase with the magnitude of accumulated price dislocations, a finding that models of capital reallocation across investors cannot explain (row 4 in Table 7).

Reallocation of Capital Across Investors Contrary to our findings, models with endogenous reallocation of capital between investors (e.g., Duffie and Strulovici (2012)) do not imply that *investor-specific* elasticities increase as accumulated price dislocations grow.

Appendix D.4 considers a model in which type-I investors face a quadratic inventory cost

$$C(Q) = a \cdot Q^2$$

that raises their effective risk aversion and renders their demand inelastic. Crucially, the mass of type-E investors, ϵ, is endogenous:

$$\epsilon(\mu) = 1 - \left(1 + \mu^2\right)^{-\frac{1}{2}}.$$

That is, capital flows to the more elastic type-E investors as the magnitude of the expected return rises. Thus, the aggregate elasticity in the economy increases with the magnitude of the expected return, causing the multiplier to decrease:

$$M = \frac{\gamma(\gamma + a)\sigma_D^2}{\gamma + a\left(1 - (1 + \mu^2)^{-\frac{3}{2}}\right)}.$$

Since the expected return magnitude ($|\mu|$) increases with the accumulated supply shock magnitude ($|\Theta_1|$), this model can explain both the static and dynamic results in Sections 3 and 4. However, each investor's elasticity is fixed. Hence, reallocation of capital cannot explain the within-investor elasticity results in Section 5.

6.2.4 Mechanisms That Do Not Explain Intensive-Margin Elasticity Results

In Section 5, we find that demand elasticities increase on the intensive margin of investor holdings as accumulated price dislocations grow, a finding that models with participation cost cannot explain (row 5 in Table 7).

Participation (Entry/Exit) Costs Contrary to our findings, models with participation costs (e.g., Vissing-Jorgensen (2002); Gomes and Michaelides (2005); Alan (2006)) do not imply that investor-specific elasticities increase *on the intensive margin* due to larger accumulated price dislocations. Appendix D.5 considers a model in which type-I investors are endowed with zero holdings at $t = 0$ and face a fixed participation cost at $t = 1$:

$$C(Q_I) = \begin{cases} 0 & , Q_I = 0 \\ \lambda & , Q_I \neq 0 \end{cases}$$

For small cumulative supply shocks, the expected return is insufficient to overcome the fixed cost of entry. For large shocks, however, the expected return is high enough that type-I investors optimally choose to enter and take non-zero positions. That is, when expected returns grow sufficiently large, type-I investors become more elastic. As a result, the multiplier decreases with the cumulative shock size ($|\Theta_1|$):

$$M = -\frac{dP}{d\Theta} = \begin{cases} \frac{\gamma\sigma_D^2}{\epsilon} & , |\Theta_1| \leq \sqrt{\frac{2\lambda}{\gamma\sigma_D^2}} \\ \gamma\sigma_D^2 & , |\Theta_1| > \sqrt{\frac{2\lambda}{\gamma\sigma_D^2}} \end{cases}$$

Thus, participation costs can explain the static and dynamic multiplier results in Sections 3 and 4, as well as the within-investor results in Section 5. However, the only variation in investor-specific elasticities in this model is on the extensive margin — when investors decide whether to enter or exit an asset. On the intensive margin (i.e., conditional on holding the asset), each investor's elasticity is fixed. Hence, participation costs cannot explain the intensive-margin elasticity results in Section 5.

6.2.5 Mechanisms That Can Explain Our Findings

Our results can be explained by models in which investors face frictions that make their demand inelastic, but can overcome these frictions to become more elastic when expected returns are large. Two mechanisms that generate this behavior are fixed adjustment costs and costly information acquisition (rows 6 and 7 in Table 7).

Fixed Adjustment Costs Appendix D.6 considers a model in which type-I investors face a fixed adjustment cost λ at $t = 1$ for deviating from their $t = 0$ quantity Θ_0 (e.g., as in Constantinides (1986); Grossman and Laroque (1987)):

$$C(Q_I) = \begin{cases} 0 & , Q_I = \Theta_0 \\ \lambda & , Q_I \neq \Theta_0 \end{cases}$$

This cost gives rise to (S, s)-style adjustment behavior (Scarf (1960)). For small shocks, the initial position is close to the optimum, so the benefit of adjustment does not exceed the fixed cost. Large shocks, however, significantly alter the optimal quantity (since they alter expected returns), making adjustment worthwhile. Thus, for sufficiently large shocks, type-I investors become more elastic. As a result, the multiplier decreases with the shock size ($|\Theta_1 - \Theta_0|$):

$$M = \begin{cases} \frac{\gamma\sigma_D^2}{\epsilon} & , |\Theta_1 - \Theta_0| \leq \sqrt{2\lambda\gamma\sigma_D^2} \\ \gamma\sigma_D^2 & , |\Theta_1 - \Theta_0| > \sqrt{2\lambda\gamma\sigma_D^2} \end{cases}$$

In this three-period model, M does not depend on past shocks conditional on the current shock ($\Theta_1 - \Theta_0$). Yet in a fully dynamic model, the adjustment decision and the multiplier would depend on accumulated shocks since the last adjustment, which need not have occurred in the previous period. Hence, fixed adjustment costs can explain the static and dynamic multiplier results in Sections 3 and 4, and the within-investor, intensive-margin elasticity results in Section 5.

Costly Information Acquisition Appendix D.7 considers a model in which type-I investors have uncertainty about cash flows, which renders their demand inelastic. However, these investors can pay to acquire information to reduce this uncertainty.[11] The marginal

[11] E.g., as in Grossman and Stiglitz (1980); Kyle (1989); Van Nieuwerburgh and Veldkamp (2009, 2010); Kacperczyk et al. (2016); Han (2018).

benefit of reducing uncertainty is larger when investors believe expected returns are higher: lower uncertainty allows them to take larger positions to more aggressively exploit the high expected returns without incurring excessive perceived risk.

The portfolio choice cost takes the form

$$C(Q, \mu) = \text{Posterior Variance} \cdot Q^2$$

where Posterior Variance is a decreasing function of $|\mu|$. The multiplier is

$$M = \gamma \cdot \left(\sigma_D^2 + \text{Posterior Variance}\right),$$

which decreases with the accumulated supply shock size since $|\mu|$ increases with shock size. These multiplier reductions arise from increases in investor-specific, intensive-margin elasticities. Hence, costly information acquisition can explain the static and dynamic multiplier results in Sections 3 and 4, and the within-investor, intensive-margin elasticity results in Section 5.

7 Conclusion

In this paper, we document a new stylized fact about inelastic demand in stock markets: larger uninformed demand shocks have smaller price multipliers. This pattern is robust across three cash flow-unrelated demand shocks from previous work: index-reconstitution induced changes in benchmarking intensity, fund flow-induced trading, and order flow imbalance. We show that this finding is unlikely to be explained by mismeasurement or unobserved variation in liquidity. Dynamically, we find price multipliers decrease both with the magnitude of current shocks and with the accumulation of past shocks. Consistent with these results, when examining institutional investor holdings data, we find that investor-specific price elasticities of demand increase when past cumulative price changes are larger.

Overall, our findings illuminate an important dimension of endogenous variation in price elasticities: demand is more elastic in the face of larger price dislocations that create more opportunities for active investment. This is important for the purpose of quantifying the effect of large demand shocks such as government asset purchases, passive investing, and green asset demand. The existing literature estimates price multipliers using small demand shocks, and those price effects cannot be extrapolated linearly to larger-scale phenomenon.

Our results also shed light on the mircofoundations of inelastic demand in financial markets. We show that several prominent mechanisms — such as those based on arbitrageur financial constraints, or those emphasizing capital reallocation across investors — cannot explain our findings. Our results are best explained by models in which investors allocate more attention and capital to securities with greater profit opportunities.

References

Alan, S. 2006. Entry costs and stock market participation over the life cycle. *Review of Economic Dynamics* 9:588–611.

Alfonsi, A., A. Fruth, and A. Schied. 2010. Optimal execution strategies in limit order books with general shape functions. *Quantitative finance* 10:143–157.

Appel, I., T. A. Gormley, and D. B. Keim. 2021. Identification using Russell 1000/2000 index assignments: A discussion of methodologies. *Critical Finance Review, Forthcoming* .

Appel, I. R., T. A. Gormley, and D. B. Keim. 2019. Standing on the shoulders of giants: The effect of passive investors on activism. *The Review of Financial Studies* 32:2720–2774.

Ben-David, I., F. Franzoni, and R. Moussawi. 2019. An improved method to predict assignment of stocks into Russell indexes. Tech. rep., National Bureau of Economic Research.

Bouchaud, J.-P., J. Bonart, J. Donier, and M. Gould. 2018. *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press.

Chakrabarty, B., P. C. Moulton, and A. Shkilko. 2012. Short sales, long sales, and the Lee–Ready trade classification algorithm revisited. *Journal of Financial Markets* 15:467–491.

Chang, Y.-C., H. Hong, and I. Liskovich. 2015. Regression discontinuity and the price effects of stock market indexing. *The Review of Financial Studies* 28:212–246.

Chaudhary, M., Z. Fu, and H. Zhou. 2024. Anatomy of the Treasury Market: Who Moves Yields? *Available at SSRN* .

Chaudhry, A. 2023. The Impact of Prices on Analyst Cash Flow Expectations. *Working paper* .

Chen, A. Y., and T. Zimmermann. 2022. Open Source Cross-Sectional Asset Pricing. *Critical Finance Review* 27:207–264.

Chen, H. 2024. Diversification Driven Demand for Large Stock. *Available at SSRN 4288953* .

Constantinides, G. M. 1986. Capital market equilibrium with transaction costs. *Journal of political Economy* 94:842–862.

Coval, J., and E. Stafford. 2007. Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics* 86:479–512.

Crane, A. D., S. Michenaud, and J. P. Weston. 2016. The effect of institutional ownership on payout policy: Evidence from index thresholds. *The Review of Financial Studies* 29:1377–1408.

Davis, S. J., and J. Haltiwanger. 1992. Gross job creation, gross job destruction, and employment reallocation. *The Quarterly Journal of Economics* 107:819–863.

Diamond, D. W., and R. E. Verrecchia. 1987. Constraints on short-selling and asset price adjustment to private information. *Journal of financial economics* 18:277–311.

Donier, J., J. Bonart, I. Mastromatteo, and J.-P. Bouchaud. 2015. A fully consistent, minimal model for non-linear market impact. *Quantitative finance* 15:1109–1121.

Duffie, D., N. Garleanu, and L. H. Pedersen. 2002. Securities lending, shorting, and pricing. *Journal of Financial Economics* 66:307–339.

Duffie, D., and B. Strulovici. 2012. Capital mobility and asset pricing. *Econometrica* 80:2469–2509.

Fama, E. F., and J. D. MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of political economy* 81:607–636.

Frazzini, A., R. Israel, and T. J. Moskowitz. 2018. Trading Costs. *Working paper* .

Frazzini, A., and O. A. Lamont. 2008. Dumb money: Mutual fund flows and the cross-section of stock returns. *Journal of financial economics* 88:299–322.

Gabaix, X. 2019. Behavioral inattention. In *Handbook of behavioral economics: Applications and foundations 1*, vol. 2, pp. 261–343. Elsevier.

Gabaix, X., and R. S. Koijen. 2022. In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis. *Working paper* .

Gabaix, X., and R. S. Koijen. 2024. Granular instrumental variables. *Journal of Political Economy* 132:2274–2303.

Gârleanu, N., and L. H. Pedersen. 2013. Dynamic trading with predictable returns and transaction costs. *The Journal of Finance* 68:2309–2340.

Gatheral, J. 2010. No-dynamic-arbitrage and market impact. *Quantitative finance* 10:749–759.

Glossner, S. 2019. Russell index reconstitutions, institutional investors, and corporate social responsibility. *Available at SSRN 3180776* .

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift. 2020. Bartik instruments: What, when, why, and how. *American Economic Review* 110:2586–2624.

Gomes, F., and A. Michaelides. 2005. Optimal life-cycle asset allocation: Understanding the empirical evidence. *The Journal of Finance* 60:869–904.

Grossman, S. J., and G. Laroque. 1987. Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods.

Grossman, S. J., and J. E. Stiglitz. 1980. On the impossibility of informationally efficient markets. *The American economic review* 70:393–408.

Haddad, V., P. Huebner, and E. Loualiche. 2024. How competitive is the stock market? theory, evidence from portfolios, and implications for the rise of passive investing. *Working paper* .

Han, B. 2018. Dynamic information acquisition and asset prices. *Unpublished Working Paper, University of Maryland* .

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society* pp. 1029–1054.

Harris, L., and E. Gurel. 1986. Price and volume effects associated with changes in the S&P 500 list: New evidence for the existence of price pressures. *the Journal of Finance* 41:815–829.

Hartzmark, S. M., and D. H. Solomon. 2022. Predictable price pressure. Tech. rep., National Bureau of Economic Research.

He, Z., and A. Krishnamurthy. 2013. Intermediary asset pricing. *American Economic Review* 103:732–770.

Huang, S., Y. Song, and H. Xiang. 2024. Noise trading and asset pricing factors. *Management Science* .

Huberman, G., and W. Stanzl. 2004. Price manipulation and quasi-arbitrage. *Econometrica* 72:1247–1275.

Huebner, P. 2023. The making of momentum: A demand-system perspective. In *Proceedings of the EUROFIDAI-ESSEC Paris December Finance Meeting*.

Kacperczyk, M., S. Van Nieuwerburgh, and L. Veldkamp. 2016. A rational theory of mutual funds' attention allocation. *Econometrica* 84:571–626.

Kelly, B. T., S. Pruitt, and Y. Su. 2019. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics* 134:501–524.

Koijen, R. S., F. Koulischer, B. Nguyen, and M. Yogo. 2021. Inspecting the mechanism of quantitative easing in the euro area. *Journal of Financial Economics* 140:1–20.

Koijen, R. S., R. J. Richmond, and M. Yogo. 2024. Which investors matter for equity valuations and expected returns? *Review of Economic Studies* 91:2387–2424.

Koijen, R. S., and M. Yogo. 2019. A demand system approach to asset pricing. *Journal of Political Economy* 127:1475–1515.

Koijen, R. S., and M. Yogo. 2020. Exchange rates and asset prices in a global demand system. Tech. rep., National Bureau of Economic Research.

Kyle, A. S. 1989. Informed speculation with imperfect competition. *The Review of Economic Studies* 56:317–355.

Lee, C. M., and M. J. Ready. 1991. Inferring trade direction from intraday data. *The Journal of Finance* 46:733–746.

Li, J. 2022. What drives the size and value factors? *Review of Asset Pricing Studies* 12:845–885.

Li, J., and Z. Lin. 2023. Prices are less elastic at more aggregate levels. *Working paper* .

Lou, D. 2012. A flow-based explanation for return predictability. *The Review of Financial Studies* 25:3457–3489.

Pástor, L., R. F. Stambaugh, and L. A. Taylor. 2022. Dissecting green returns. *Journal of financial economics* 146:403–424.

Pavlova, A., and T. Sikorskaya. 2023. Benchmarking intensity. *The Review of Financial Studies* 36:859–903.

Petajisto, A. 2009. Why do demand curves for stocks slope down? *Journal of Financial and Quantitative Analysis* 44:1013–1044.

Pollet, J. M., and M. Wilson. 2008. How does size affect mutual fund behavior? *The Journal of finance* 63:2941–2969.

Scarf, H. 1960. The optimality of (S, s) policies in the dynamic inventory problem .

Schmickler, S., and P. Tremacoldi-Rossi. 2022. Spillover effects of payouts on asset prices and real investment. *Available at SSRN 4287300* .

Shleifer, A. 1986. Do demand curves for stocks slope down? *The Journal of Finance* 41:579–590.

Shleifer, A., and R. W. Vishny. 1997. The limits of arbitrage. *The Journal of finance* 52:35–55.

Tóth, B., Y. Lemperiere, C. Deremble, J. De Lataillade, J. Kockelkoren, and J.-P. Bouchaud. 2011. Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X* 1:021006.

Van der Beck, P. 2021. Flow-driven ESG returns. *Swiss Finance Institute Research Paper* .

Van der Beck, P. 2025. Short versus Long-run Demand Elasticities in Asset Pricing. *Working paper* .

Van Nieuwerburgh, S., and L. Veldkamp. 2009. Information immobility and the home bias puzzle. *The Journal of Finance* 64:1187–1215.

Van Nieuwerburgh, S., and L. Veldkamp. 2010. Information acquisition and under-diversification. *The Review of Economic Studies* 77:779–805.

Vissing-Jorgensen, A. 2002. Towards an explanation of household portfolio choice heterogeneity: Nonfinancial income and participation cost structures.

Wei, W., and A. Young. 2021. Selection bias or treatment effect? A re-examination of Russell 1000/2000 Index reconstitution. *Critical Finance Review* 1000.

Zhou, Y., and Y. Chen. 2025. Deflated HeteroPCA: Overcoming the curse of ill-conditioning in heteroskedastic PCA. *The Annals of Statistics* 53:91–116.

APPENDIX

A Supplemental Results for Demand Measurement

This section provides supplemental results and discussions about the three demand measures used in Section 2.

A.1 Description of Russell Banding Methodology Starting in 2007

Prior to 2007, firms with market capitalizations on the May rank date that fell between ranks 1 and 1000 were assigned to the Russell 1000, and those with market caps ranked between 1001 and 3000 were assigned to the Russell 2000.

To reduce turnover, since 2007 FTSE Russell has used a "banding policy" under which there are two separate cutoffs for stocks in the Russell 1000 and 2000 in the previous year, both of which are based on a mechanical function of the firm size distribution in the year. Under this policy:

- Stocks in the Russell 2000 in the previous year are assigned to the Russell 1000 if they're rank date market cap ranks fall between 1 and $1000 - c_1$.
- Stocks in the Russell 1000 in the previous year are assigned to the Russell 2000 if they're rank date market cap ranks fall between $1000 + c_2$ and 3000.

To calculate c_1 and c_2 Russell first computes the cumulative market cap of the largest 1000 stocks (i.e. those with ranks 1 through 1000). Let $C(N)$ represent the cumulative market cap of the largest N stocks. c_1 is calculated such that $C(1000 - c_1) = 0.95 \cdot C(1000)$. c_2 is calculated such that $C(1000 + c_2) = 1.05 \cdot C(1000)$. That is, the band of stocks between ranks $1000 - c_1$ and $1000 + c_2$ constitutes a 5% band around the cumulative market cap of the largest 1000 stocks.

Thus, even after the introduction of the banding policy, assignment to the Russell 1000 or 2000 is still based on a mechanical rule. After the introduction of the banding policy, this mechanical rule changes each year with the distribution of firm sizes.

A.2 Robustness in FIT Measurement

This section describes how we compute cleaned versions of flow-induced trading (FIT) to help alleviate measurement error concerns. Specifically, as Section 3.3 discusses, we want to guard against our main result — that multipliers are smaller when the magnitude of FIT is large — arise from attenuation bias. Therefore, we want to alleviate the concern of overestimating large FIT values in the tails.

Where would mismeasurement come from? Under the simplest specification where we assume that trades respond one-to-one to flows, the FIT of fund n in stock n is given by

$$\text{SharesHeld}_{i,n,t-1} \times f_{i,t}.$$

From this perspective, there are two main mismeasurement concerns, and we tackle them in turn.

1. **Heterogeneous trading response to flows.** Trades respond less than one-to-one to flows, and this may be especially true when flows are large or when pre-existing positions are large. When a fund faces large inflows, it may use a larger fraction of the flows to buy new stocks, which reduces the need to purchase stocks in the existing holdings (e.g. Pollet and Wilson, 2008; Lou, 2012). Further, for diversification purposes, it may reduce its purchase if a stock already occupies a large part of the portfolio (Chen, 2024). Both of these considerations may lead us to over-estimate large FIT values.

   We address this by explicitly estimating heterogeneous trade-to-flow responses. We estimate a panel regression of trades on dummy variables:

$$\text{Trade}_{i,n,t} = \sum_b \sum_f \beta_{b,f} \cdot \text{I}_{w_{i,n,t-1} \in \text{ bin } b} \times \text{I}_{f_{i,t} \in \text{ bin } f} + \epsilon_{i,n,t} \tag{A.1}$$

   where $\text{Trade}_{i,n,t} = \frac{\text{SharesHeld}_{i,n,t}}{\text{SharesHeld}_{i,n,t-1}} - 1$ and $w_{i,n,t-1}$ is the lagged portfolio weight of stock n for fund i. To study heterogeneous responses, we sort the sample by fund flows into $f = -20, ..., 0, ..., 20$ bins, with the first (last) 20 bins covering the flows below -1% (above 1%) realizations, and bin 0 is defined by $f_{j,t} \in [-1\%, 1\%]$. Panel (a) of Figure A.1 plots the average flow by bins and show that they cover a large range from approximately -30% to +100%.

   Similarly, for each fund in each period, we sort its stock holdings by the existing portfolio weight into $b = 1, ..., 20$ bins. To reduce the impact of outliers due to "dividing by a small number" in the dependent variable, we estimate (A.1) via a weighted regression with weights equal to $w_{i,n,t-1}$.[1]

Panel (b) of Figure A.1 plots the point estimates of $\beta_{b,f}$ for bins $b = 1, 10$, and 20. The result does indicate that when facing in flows, funds tend to trade less than one-to-one in stocks that they already have large holdings in, but the heterogeneity is limited in magnitude. This effect is less pronounced for out flows, a finding that is consistent with Lou (2012).

To account for the heterogeneity of trade-to-flow responses, we compute FIT as:

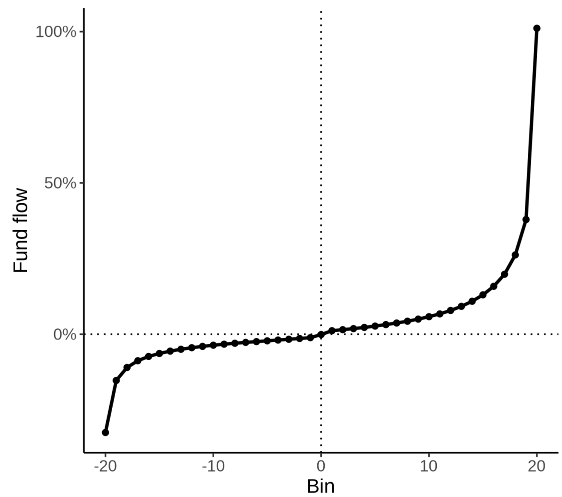
$$\text{FIT}_{n,t} = \sum_{\text{fund } i} \frac{\text{SharesHeld}_{i,n,t-1}}{\text{Share Outstanding}_{n,t-1}} \cdot \underbrace{\beta(w_{i,n,t-1}, f_{i,t})}_{\text{trade response}}$$

where the heterogeneous responses $\beta(w_{i,n,t-1}, f_{i,t})$ are based on the estimates in regression (A.1). Specifically, we first sort holdings into 20 bins by $w_{i,n,t-1}$ for each fund-quarter, and then apply a third-order polynomial-estimated curve based on the regression estimates.

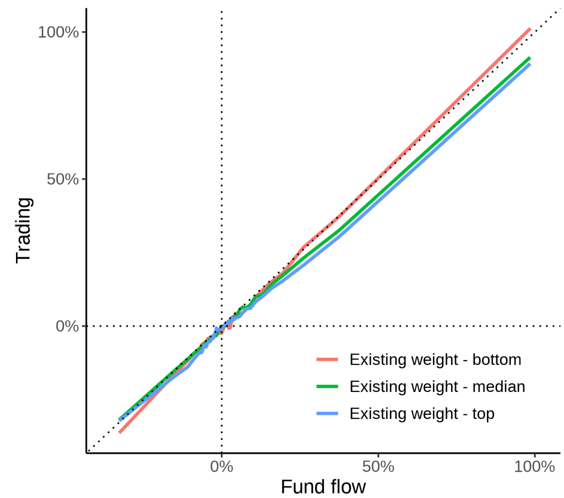
2. **Winsorize fund flows.** Another possible concern is that the extreme fund flows may be misestimated. To alleviate this concern, we recompute FIT after winsorizing 1%, 5%, or 10% of fund flows, with equal fraction of winsorization on each side. The winsorization thresholds are illustrated in Panel (c) of Figure A.1. The black line plots the density of fund flows, and the colored vertical dashed lines represent the cutoffs. By removing extreme values of fund flows, we ensure that our FIT meaures are not subject to mismeasured large flow values.

Panel (d) of Figure A.1 shows the effect of applying these cleaning and winsorization steps. We sort the sample into 100 bins by "raw FIT," which does not winsorize flows and assumes a one-to-one trade-to-flow response, and we plot it on the horizontal axis. On the vertical axis, the red line plots the average FIT after taking into account heterogeneous trade-to-flow responses. The remaining lines plot the results after further applying fund flow winsorization. The plot suggests that applying these cleaning steps serves to gradually dampen the large

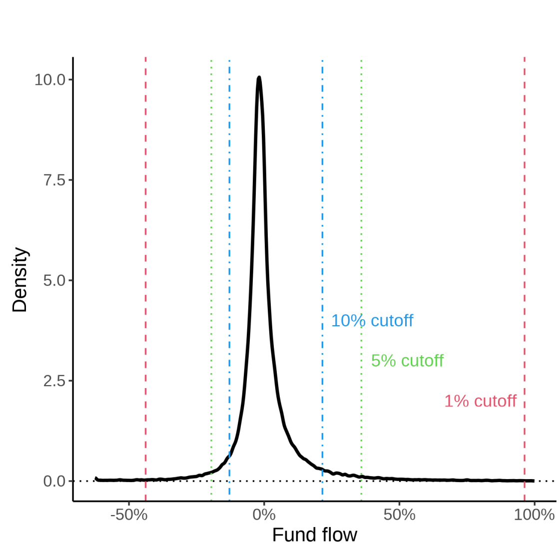
[1] To control for the fact that, even without flows, portfolio weights tend to mean-revert — that is, the largest (smallest) positions tend to be reduced (increased) subsequently — we estimate a first-stage regression where we regress the dependent variable on indicators of existing portfolio weight bin b. We use the resulting residuals as the dependent variable in regression (A.1).



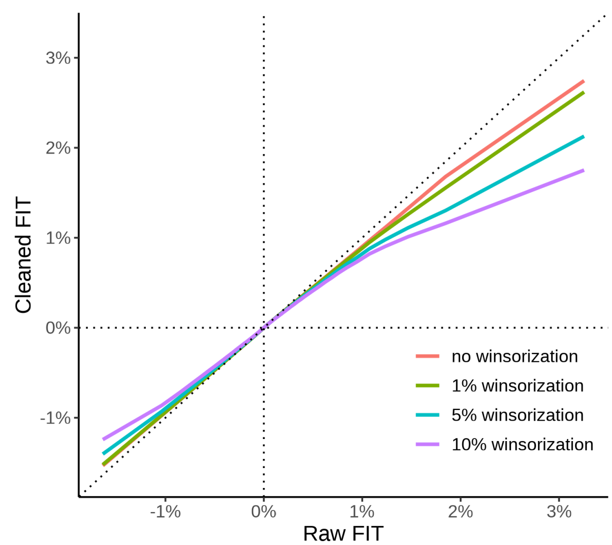
(a) Fund flow by bin



(b) Trade response to flow



(c) Fund flow winsorization



(d) Comparing FIT measures

Figure A.1. Cleaning the flow-induced trading (FIT) measure

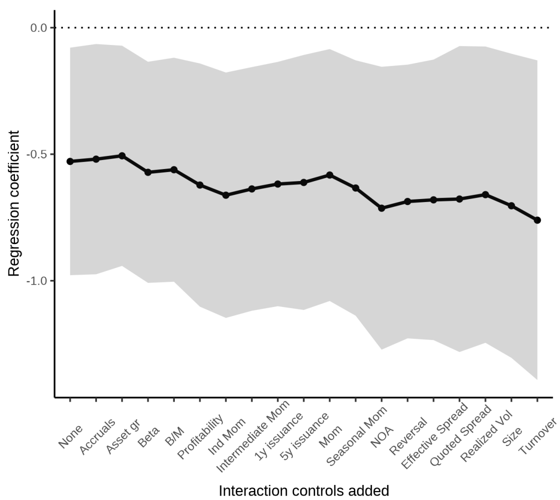
Panel (a) plots fund flows by bins. Panel (b) plots the average fund trade as a function of fund flows by the size of the pre-existing portfolio weights, and the dashed diagonal line is the 45 degree line. Panel (c) plots the kernel density of fund flows, and the vertical dashed lines represent the various winsorization cutoffs. Panel (d) plots the cleaned versions of FIT against the "raw FIT" which does not winsorize fund flows and assumes that trades respond one-to-one to flows.

values of FIT.

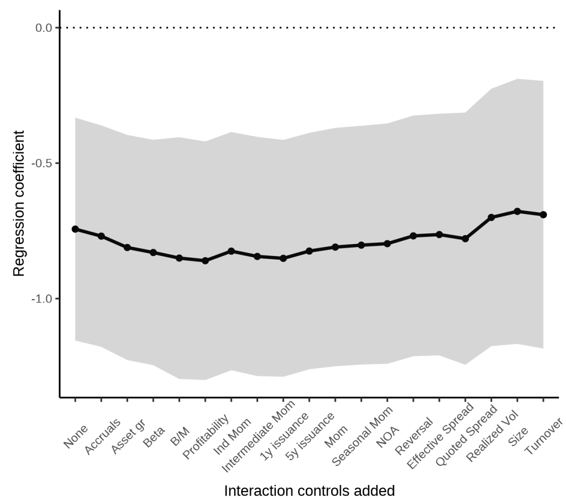
B Additional Empirical Results for Section 3

B.1 Interaction with Stock Characteristics

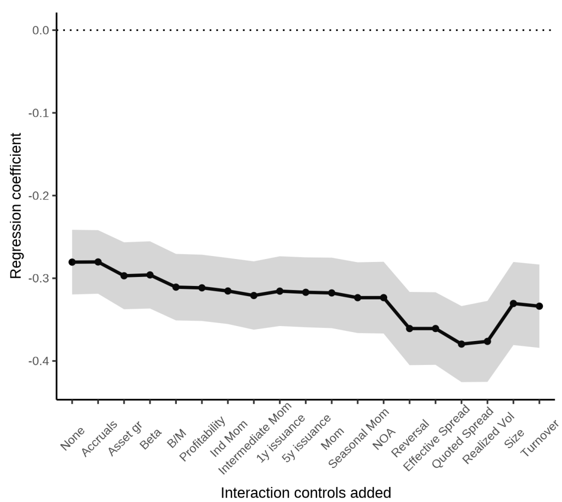
To further visualize the impact of progressively adding each specific characteristic, Figure B.2 displays the estimated b_1 and b_2 coefficients from these regressions. Each point represents the coefficient estimates from regression (4) when adding an additional characteristic (i.e. the right-most points represent coefficient estimates from the regression including all characteristics). We find evidence that price multipliers decrease with shock size across all specifications for all three demand shocks: $b_2 < 0$ for all specifications. Moreover, the coefficient estimates are quantitatively stable across specifications.



(a) ΔBMI



(b) FIT



(c) OFI

Figure B.2. Nonlinear Price Impact Coefficient: Adding One Control at a Time

Panels (a), (b), and (c) plot the estimated interaction coefficient (b_2) in regression (4) for the ΔBMI, FIT, and OFI demand shocks, respectively. The left-most points represent estimates when no interactions between the demand shock and any stock characteristics are included. Each subsequent point represents the estimates from regression (4) when adding the interaction with characteristic labelled on the x-axis (i.e. the right-most points represent coefficient esimtates from the regression including all liquidity and stock characteristics).

B.2 Measurement Error

We argue that measurement errors are unlikely to explain the findings in Section 3.

FIT. This concern is potentially applicable for FIT because it is an *imputed* demand measure rather than a realized one. Where could mismeasurement come from? Recall that the flow-induced-trading of fund n in stock n is given by

$$b_{i,n,t} \times \text{SharesHeld}_{i,n,t-1} \times f_{i,t}.$$

This formula indicates that not accounting for heterogeneity in flow-to-trade coefficient $b_{i,n,t}$ may pose a concern. Specifically, funds may exhibit lower response coefficient when experiencing large fund flow magnitudes. It is also possible that funds also exhibit lower response coefficients when handling stocks that is already a large part of their portfolios. Both of these concerns can lead to over-estimation of FIT.

We tackle this concern in Appendix A.2 and summarize our findings here. We explicitly estimate flow-to-trade responses that depend on both realized fund flow $f_{i,t}$ and $w_{i,n,t-1}$, the lagged portfolio weight of stock n in fund n's portfolio. We do find evidence of slightly weaker responses to large inflows, consistent with Lou (2012). We also find heterogeneous responses for stocks with higher $w_{i,n,t-1}$, but the degree of heterogeneity is limited. Therefore, when we apply these heterogeneous responses in computing FIT, all results are qualitatively unchanged.

As a further robustness step to guard against potential mismeasurement in fund flows $f_{i,t}$, we winsorize it at the 1%, 5%, or even the 10% levels (symmetrically on both sides). This also had minimal effects on our main findings. Overall, we do not find measurement errors to be a likely explanation of our findings.

ΔBMI and OFI. We argue that the concern is less relevant for these two two demand measures. OFI is directly computed from observed trades, rather than imputed. The only inference step is assigning trade directions via the Lee-Ready algorithm. Prior work find that the Lee-Ready algorithm contain errors at the individual trade level, but the errors wash out when aggregating at the daily frequency (Chakrabarty et al., 2012). Therefore, there is no clear reason why mismeasurement will become more pronounced at the quarterly frequency and for larger OFI realizations. In unreported robustness checks, we also estimate our OFI-based findings at monthly and weekly frequencies. We find evidence for concave price impacts at all those frequencies.

As described in Section 2, the main variation of ΔBMI is driven by differences in market capitalization between stocks. Because market capitalization is measured relatively accurately, we also do not consider measurement error to be a significant concern.

B.3 Removing News Days

It is widely accepted that stock returns are often driven by news releases, which leads to the concern that perhaps our demand shocks *per se* do not impact prices, but rather are just correlate with news releases. Further, if demand is positively correlated with news-driven returns, and if the magnitude of demand happens to scale sublinearly with the magnitude of news-driven returns, then this can explain our finding.[2] We argue this concern is less relevant for BMI and FIT which exploit variation in trading by specific investors in specific circumstances that are plausibly unrelated to cash flow news. However, this concern is potentially more relevant for OFI, which captures trading behavior by many investors over a whole quarter.

To assess this concern, we examine whether we observe concave price impact when focusing on days with less news releases. We use three measures of news releases at the daily frequency. The first is whether there is an earnings release. The second is the number of analyst updates from IBES, and the third is the number of media reports about that company in Ravenpack. For the latter two, we convert them into quintile indicators after sorting within each quarter-stock to adjust for the fact that larger stocks and later sample periods tend to have more updates. This conversion makes the news measures more comparable across stocks and over time.

We first verify that these measures do capture price-relevant news. In Table B.1, we regress the absolute value of daily stock returns on news indicators, and the results indicate that days with earnings releases, as well as days in the top quintile of either IBES or Ravenpack news releases, are associated with higher return volatility. To see the magnitudes, in the last column of Table B.1 which contains all news indicators, we see that earnings days are associated with 0.93% higher return absolute values. Days in the top quintile of IBES (and Ravenpack) news releases are associated with 0.35% (and 0.49%) higher absolute returns. There is also some evidence of higher return volatility in quintiles 4 for IBES and Ravenpack, but the economic magnitudes are much smaller. These lead us to conclude that, when it

[2] For instance, consider the possibility that when news-driven return is 1%, demand shock is 1%. When news-driven return is 2%, demand shock is 1.5% rather than 2%.

comes to using IBES and Ravenpack to look for news days, it is most important to filter out the top quintile of days. As a baseline, we note that the dependent variable's average value is 2.57% in the full sample.

	Dependent variable: $\lvert r_{n,t}\rvert$ (%)			
	(1)	(2)	(3)	(4)
Earnings	1.29***			0.93***
	(0.03)			(0.04)
IBES bin 2		0.00		−0.00
		(0.00)		(0.00)
IBES bin 3		0.00		0.00
		(0.00)		(0.00)
IBES bin 4		0.01***		0.01***
		(0.00)		(0.00)
IBES bin 5		0.36***		0.35***
		(0.01)		(0.01)
Ravenpack bin 2			0.00**	0.00*
			(0.00)	(0.00)
Ravenpack bin 3			0.01***	0.01**
			(0.00)	(0.00)
Ravenpack bin 4			0.03***	0.04***
			(0.01)	(0.01)
Ravenpack bin 5			0.56***	0.49***
			(0.02)	(0.02)
Quarter-stock FE	Y	Y	Y	Y
Obs	36,461,790	27,145,392	21,239,328	16,125,776
Within R^2	0.003	0.003	0.006	0.015

Table B.1. Information measures and return variability

We use daily panel regressions to estimate the relationship between the absolute value of stock returns and indicators of news releases. The baseline average $\lvert r_{n,t}\rvert$ is 2.57%. All regressions control for quarter-stock fixed effects, and standard errors are clustered by quarter and stock. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

We now re-estimate the results for OFI in Table 2 but remove days with news from our computation of quarterly returns and demand. The results are reported in Table B.2. Column (1) reproduces the result in column (9) of Table 2 where no days are filtered out. In the subsequent columns, we filter out earnings days. Starting from column (3), we impose increasingly strong filters based on IBES and Ravenpack, where the filters are indicated in the last three rows of the table. Throughout these specifications, we continue to see significant evidence for nonlinearity. For instance, in the last column, we filter out earnings days, as well as days in the top two quintiles of either IBES or Ravenpack indicators. The result continue to indicate a similar degree of nonlinearity in the price impact curve. To summarize, we do not find evidence that our discovery of nonlinear price impact is due to news.

Dependent variable: stock return $r_{n,t}$						
	(1)	(2)	(3)	(4)	(5)	(6)
$d_{n,t}$	4.65^{***}	3.83^{***}	3.33^{***}	3.01^{***}	3.06^{***}	3.13^{***}
	(0.18)	(0.16)	(0.15)	(0.15)	(0.15)	(0.16)
$d_{n,t} \times \|d_{n,t}\|$	-0.28^{***}	-0.18^{***}	-0.19^{***}	-0.17^{***}	-0.19^{***}	-0.25^{***}
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Controls	Y	Y	Y	Y	Y	Y
Obs	333,771	315,606	279,612	204,703	204,703	204,703
R^2	0.144	0.146	0.131	0.115	0.095	0.083
Marg $R^2(d_{n,t})$	0.054	0.055	0.048	0.038	0.038	0.039
Sample selection:						
Earnings days	Y	N	N	N	N	N
IBES filter	N	N	$< 80\%$	$< 80\%$	$< 60\%$	$< 60\%$
Ravenpack filter	N	N	N	$< 80\%$	$< 80\%$	$< 60\%$

Table B.2. OFI Price Impact Regressions in Periods with Less News
We re-estimate the Fama-MacBeth regressions in Table 2 for OFI, but only use days with less news. That is, from column (1) through (6), we progressively remove certain days with more news releases in computing quarterly OFI and return. Column (1) does not apply any news filter and is the same as column (9) in Table 2. Starting from column (2), we remove the earnings days. Column (3) further removes days on which the IBES news indicator is above 80% within each quarter-stock, and column (4) also removes days on which the Ravenpack news indicator is above 80% within each quarter-stock. Columns (5) and (6) make more stringent removals and the filters applied are indicated by the bottom three rows. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

B.4 Additional Specifications

Table B.3 estimates the same specification as in Table 2 using panel regressions.

	Dependent variable: stock return $r_{n,t}$								
$d_{n,t} =$	Δ BMI			FIT			OFI		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$d_{n,t}$	2.32***	2.18***	2.11***	4.84***	4.83***	4.83***	4.92***	4.93***	4.95***
	(0.78)	(0.72)	(0.73)	(0.64)	(0.62)	(0.62)	(0.24)	(0.24)	(0.25)
$d_{n,t} \times \lvert d_{n,t} \rvert$	−0.56**	−0.54**	−0.53**	−0.52***	−0.52***	−0.52***	−0.25***	−0.25***	−0.25***
	(0.25)	(0.23)	(0.24)	(0.13)	(0.12)	(0.12)	(0.02)	(0.02)	(0.02)
Time FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Predictor controls	N	Y	Y	N	Y	Y	N	Y	Y
Liquidity controls	N	N	Y	N	N	Y	N	N	Y
Obs	9,914	9,914	9,914	561,405	561,405	561,405	333,772	333,772	333,772
R^2	0.141	0.157	0.158	0.153	0.156	0.157	0.220	0.223	0.224
Marginal $R^2(d_{n,t})$	0.007	0.006	0.005	0.004	0.004	0.004	0.043	0.043	0.043

Table B.3. Interacted price impact: panel regressions
This Table is simliar to 2 except that we estimate panel regressions instead of Fama-MacBeth regressions. All regressions control for time fixed effect and cluster standard errors by time and stock. Column (1) reports results using the BMI measure of Pavlova and Sikorskaya (2023) and monthly returns. Columns (2) reports results using quarterly returns, whereas the demand is based on the FIT measure of Lou (2012) and the OFI measure in Li and Lin (2023), respectively. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

As discussed in Sections 3.3 and 4.1, for both the static and dynamic price multiplier regressions, we also estimate piecewise linear specifications. This section reports the results. Table B.4 report the static regression results and provide more details to Figure 1 in Section 3.3. Table B.5 report results that support Figure 2.

Table B.6 report static regression results where we use $d_{n,t}$ standardized by its own stock-specific rolling standard deviations. Those results support the discussion in Section 3.3.

Panel A: price impact regressions									
	Dependent variable: stock return $r_{n,t}$								
	BMI			FIT			OFI		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$M_{\|d_{n,t}\|<\sigma}$	2.41***	1.75***	1.74***	4.78***	3.96***	3.92***	5.13***	4.71***	4.67***
	(0.30)	(0.28)	(0.28)	(0.48)	(0.36)	(0.30)	(0.12)	(0.09)	(0.10)
$M_{\|d_{n,t}\|\in[\sigma,2\sigma]}$	1.41***	1.06***	1.05***	3.69***	3.17***	3.33***	4.14***	4.02***	4.00***
	(0.41)	(0.34)	(0.33)	(0.27)	(0.21)	(0.18)	(0.10)	(0.09)	(0.09)
$M_{\|d_{n,t}\|>2\sigma}$	0.88***	0.67***	0.69***	3.13***	2.67***	2.69***	2.86***	2.77***	2.78***
	(0.17)	(0.17)	(0.17)	(0.19)	(0.16)	(0.15)	(0.09)	(0.09)	(0.09)
Predictor controls	N	Y	Y	N	Y	Y	N	Y	Y
Liquidity controls	N	N	Y	N	N	Y	N	N	Y
Obs	9,914	9,914	9,914	561,405	561,405	561,405	333,772	333,772	333,772
R^2	0.051	0.142	0.172	0.009	0.064	0.080	0.063	0.127	0.144
Panel B: Coefficient differences									
$M_{\|d_{n,t}\|\in[\sigma,2\sigma]} - M_{\|d_{n,t}\|<\sigma}$	−1.00***	−0.69*	−0.69**	−1.10***	−0.78***	−0.59***	−0.99***	−0.69***	−0.66***
	(0.38)	(0.35)	(0.34)	(0.28)	(0.22)	(0.20)	(0.08)	(0.06)	(0.06)
$M_{\|d_{n,t}\|>2\sigma} - M_{\|d_{n,t}\|\in[\sigma,2\sigma]}$	−0.53	−0.39	−0.36	−0.56***	−0.50***	−0.64***	−1.29***	−1.25***	−1.22***
	(0.40)	(0.35)	(0.31)	(0.17)	(0.14)	(0.14)	(0.07)	(0.07)	(0.07)
$M_{\|d_{n,t}\|>2\sigma} - M_{\|d_{n,t}\|<\sigma}$	−1.53***	−1.07***	−1.05***	−1.65***	−1.29***	−1.23***	−2.27***	−1.94***	−1.89***
	(0.29)	(0.27)	(0.26)	(0.38)	(0.29)	(0.26)	(0.10)	(0.08)	(0.08)

Table B.4. Price multiplier by demand shock sizes

In panel A, we report results for Fama-MacBeth regressions:

$$r_{n,t} = \sum_{\text{bin } b} M_{b,t} \cdot I_{|d_{n,t}|\in b} \cdot d_{n,t} + \boldsymbol{c}_t^{'} \boldsymbol{x}_{n,t-1} + \tau_t + \epsilon_{n,t} \tag{B.2}$$

where $r_{n,t}$ is stock return and $d_{n,t}$ represent demand shocks. In each time period, we split the sample into three bins b by comparing $|d_{n,t}|$ against its cross-sectional standard deviation. Columns (1) through (3) report results using the ΔBMI demand shock and monthly returns. Columns (4) through (6) and (7) through (9) report results using quarterly returns and the FIT and OFI demand measures, respectively. For each demand shock measure, same as in Table 2, we vary the set of regression controls in $\boldsymbol{x}_{n,t-1}$. Panel B reports pairwise differences in price multipliers and the standard errors are computed using the Delta method. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01. Figure 1 plots coefficient estimates in columns (3), (6), and (9) in Panel A.

Panel A: price impact regressions								
	Dependent variable: stock return $r_{n,t}$							
$d_{n,t} =$	FIT				OFI			
	$h = 1$	2	3	4	$h = 1$	2	3	4
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$M_{\|d_{n,t}\|<\sigma}$	4.10***	3.93***	3.92***	3.72***	3.77***	3.68***	3.66***	3.66***
	(0.38)	(0.33)	(0.32)	(0.32)	(0.26)	(0.29)	(0.33)	(0.35)
$M_{\|d_{n,t}\|\in[\sigma,2\sigma]}$	2.98***	2.92***	2.84***	2.80***	3.29***	3.33***	3.30***	3.21***
	(0.29)	(0.33)	(0.33)	(0.34)	(0.26)	(0.30)	(0.28)	(0.31)
$M_{\|d_{n,t}\|>2\sigma}$	1.78***	1.90***	1.81***	2.01***	2.63***	2.74***	2.72***	2.63***
	(0.34)	(0.34)	(0.30)	(0.32)	(0.24)	(0.28)	(0.37)	(0.35)
Predictor controls	Y	Y	Y	Y	Y	Y	Y	Y
Liquidity controls	Y	Y	Y	Y	Y	Y	Y	Y
Obs	538,398	517,033	497,185	478,324	314,110	296,241	279,914	264,872
R^2	0.081	0.082	0.084	0.085	0.145	0.146	0.148	0.148
Panel B: Coefficient differences								
$M_{\|d_{n,t}\|\in[\sigma,2\sigma]} - M_{\|d_{n,t}\|<\sigma}$	−1.12***	−1.01***	−1.08***	−0.93***	−0.49***	−0.34**	−0.36**	−0.45***
	(0.24)	(0.26)	(0.27)	(0.29)	(0.13)	(0.16)	(0.16)	(0.13)
$M_{\|d_{n,t}\|>2\sigma} - M_{\|d_{n,t}\|\in[\sigma,2\sigma]}$	−1.20***	−1.01***	−1.02***	−0.78***	−0.66***	−0.59**	−0.58**	−0.59**
	(0.28)	(0.28)	(0.29)	(0.25)	(0.17)	(0.23)	(0.23)	(0.23)
$M_{\|d_{n,t}\|>2\sigma} - M_{\|d_{n,t}\|<\sigma}$	−2.32***	−2.03***	−2.11***	−1.71***	−1.14***	−0.93***	−0.95***	−1.03***
	(0.38)	(0.38)	(0.32)	(0.36)	(0.17)	(0.25)	(0.32)	(0.30)

Table B.5. Price Impact Regressions Interacted with Past Demand

For each demand shocks FIT and OFI, and for each lookback horizon $h = 1$ to 4 quarters, we estimate cross-sectional regressions:

$$r_{n,t} = \sum_{\text{bin } b} M_{b,t} \cdot I_{\left|\sum_{l=1}^{h} d_{n,t-l}\right| \in b} \cdot d_{n,t} + \boldsymbol{c}_t' \boldsymbol{x}_{n,t-1} + \tau_t + \epsilon_{n,t} \tag{B.4}$$

where $r_{n,t}$ is stock return and $d_{n,t}$ is the demand shock. In each period, we split the sample into three bins b by comparing $\left|\sum_{l=1}^{h} d_{n,t-l}\right|$ with its cross-sectional standard deviation. The controls $\boldsymbol{x}_{n,t-1}$ are the same as in Table 2. The first four columns use FIT as the demand shock and the last four columns use OFI. Panel A reports regression estimates, and the results are also plotted in Figure 2. Panel B reports pairwise coefficient differences with standard errors computed using the Delta method. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

	Dependent variable: stock return $r_{n,t}$							
$d^{\text{std}}_{n,t}$ =	FIT				OFI			
lookback window	N/A	4	8	12	N/A	4	8	12
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$d^{\text{std}}_{n,t}$	3.97***	3.47***	3.76***	3.56***	4.65***	5.01***	4.80***	4.56***
	(0.37)	(0.28)	(0.32)	(0.32)	(0.18)	(0.25)	(0.26)	(0.25)
$d^{\text{std}}_{n,t} \times \lvert d^{\text{std}}_{n,t}\rvert$	−0.71***	−0.68***	−0.96***	−0.71***	−0.28***	−0.49***	−0.52***	−0.52***
	(0.21)	(0.09)	(0.15)	(0.17)	(0.02)	(0.03)	(0.03)	(0.04)
Predictor controls	Y	Y	Y	Y	Y	Y	Y	Y
Liquidity controls	Y	Y	Y	Y	Y	Y	Y	Y
Obs	561,405	491,638	436,750	392,262	333,772	264,872	214,766	175,835
R^2	0.080	0.082	0.085	0.086	0.144	0.138	0.144	0.145
Marginal $R^2(d^{\text{std}}_{n,t})$	0.005	0.003	0.004	0.004	0.054	0.040	0.041	0.037

Table B.6. Contemporaneous Price Impact Regressions: Standardized Demand

This Table reports regression results for FIT and OFI and is similar to Table 2, except that it uses standardized demand ($d^{\text{std}}_{n,t}$) instead of raw demand ($d_{n,t}$). Standardized demand is estimated as $d^{\text{std}}_{n,t} = \frac{d_{n,t}}{\hat{\sigma}_{(h)}(d_{n,t})} \times \sigma(d_{n,t})$, where $\hat{\sigma}_{(h)} = \sqrt{\frac{1}{h-1}\sum_{l=1}^{h}(d_{n,t-l} - \bar{d}_{n,t-1})^2}$, and where $h = 4, 8,$ or 12 is the lookback window. We multiply the standardized demand with the full-sample volatility ($\sigma(d_{n,t})$) so that the resulting variable has the same amount of variation as the original one. For comparison purposes, columns (1) and (5) reproduces columns (6) and (9) of Table 2 where the demand is not standardized. In the other columns, we report results using standardized demand with different lookback windows. Levels of significance are presented as follows: *p<0.1; **p<0.05; ***p<0.01.

C Supplements to Section 5

C.1 Estimation Details

C.1.1 Setup

We seek to estimate the following nonlinear demand curve

$$\Delta q_{i,n,t} = -\underbrace{\left(\zeta_{1,i,t} + \zeta_{2,i,t}\tilde{P}_{n,t}\right)}_{\text{Price Elasticity of Demand}} \Delta p_{n,t} + \underbrace{\boldsymbol{\lambda}'_{i,t}\boldsymbol{\eta}_{n,t} + u_{i,n,t}}_{\text{Demand Shock}}$$

$$\tilde{P}_{n,t} = \left|\sum_{l=1}^{L} \Delta p_{n,t-l}\right| - \mathbb{E}_t\left[\left|\sum_{l=1}^{L} \Delta p_{n,t-l}\right|\right],$$

$\Delta q_{i,n,t}$ is the percentage change in quantity of shares demanded by investor i for stock n in quarter t. $\Delta p_{n,t}$ is change of log price for stock n in quarter t. L is the number of lagged quarterly price changes to consider ($L = 4$ in the baseline analysis).

The first step is to residualize $\Delta q_{i,n,t}$, $\Delta p_{n,t}$, and $\left|\sum_{l=1}^{L} \Delta p_{n,t-l}\right|$ with respect to stock characteristics $\boldsymbol{\eta}_{n,t}$. Note that by the Frisch-Waugh-Lovell theorem, this residualization is equivalent to controlling for $\boldsymbol{\eta}_{n,t}$. $\boldsymbol{\eta}_{n,t}$ contains both observed characteristics (as well as investor-quarter fixed effects) and latent characteristics estimated using principal component analysis on the within-quarter investor $\times$ stock panel of $\Delta q_{i,n,t}$. Since this panel is unbalanced and the residual demand shocks are assumed to be heteroskedastic, we estimate the latent characteristics using the deflated heteroskedastic PCA methodology of Zhou and Chen (2025). Let $\check{q}_{i,n,t}$, $\check{p}_{n,t}$, and $\check{x}_{n,t}$ be the residualized counterparts of $\Delta q_{i,n,t}, \Delta p_{n,t}$, and $\Delta p_{n,t} \cdot \tilde{P}_{n,t}$. So we have the following residualized demand curve

$$\check{q}_{i,n,t} = -\zeta_{1,i,t}\check{p}_{n,t} - \zeta_{2,i,t}\check{x}_{n,t} + u_{i,n,t}. \tag{C.1}$$

To increase statistical power, we parameterize $\zeta_{1,i,t}$ and $\zeta_{2,i,t}$ as linear functions of each investor's active share:

$$\zeta_{k,i,t} = \zeta_{k,0,t} + \zeta_{k,\text{Active Share},t}\left(\text{Active Share}_{i,t-1-L} - \mathbb{E}_t\left[\text{Active Share}_{i,t-1-L}\right]\right)$$

$$\zeta_{k,i,t} = \boldsymbol{\zeta}'_{k,t}\boldsymbol{X}_{i,t}$$

where $\boldsymbol{X}_{i,t} = [1, \text{Active Share}_{i,t-1-L}]$.

The identifying assumption of GIV in our setting is that the residual demand shocks $u_{i,n,t}$ are conditionally independent across investors both within and across quarters:

$$\forall i \neq j, \forall t, \forall l = 1, \ldots, L : \mathbb{E}[u_{i,n,t} \cdot u_{j,n,t-l}] = 0.$$

From this conditional independence assumption, if we let

$$\begin{aligned} u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) &= \Delta q_{i,n,t} + \left(\zeta_{1,i,t} + \zeta_{2,i,t}\tilde{P}_{n,t}\right)\Delta p_{n,t} - \boldsymbol{\lambda}_{i,t}^{'}\boldsymbol{\eta}_{n,t} \\ &= \Delta q_{i,n,t} + \left(\boldsymbol{\zeta}_{1,t}^{'}\boldsymbol{X}_{i,t} + \boldsymbol{\zeta}_{2,t}^{'}\boldsymbol{X}_{i,t}\cdot\tilde{P}_{n,t}\right)\Delta p_{n,t} - \boldsymbol{\lambda}_{i,t}^{'}\boldsymbol{\eta}_{n,t} \end{aligned} \tag{C.2}$$

then we obtain the following two sets of moment conditions that allow us to identify $\boldsymbol{\zeta}_{1,t}$ and $\boldsymbol{\zeta}_{2,t}$:

$$\begin{aligned} \forall i,t : \mathbf{0} &= \mathbb{E}\left[\boldsymbol{X}_{i,t}\cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\cdot\underbrace{\sum_{j\neq i} S_{j,n,t-1}u_{j,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)}_{\equiv z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)}\right] \\ &= \mathbb{E}\left[\boldsymbol{X}_{i,t}\cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\right] \end{aligned} \tag{C.3}$$

and

$$\begin{aligned} \forall i,t : \mathbf{0} &= \mathbb{E}\left[\boldsymbol{X}_{i,t}\cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\cdot\left(\left|\sum_{l=1}^{L}\underbrace{\sum_{j\neq i} S_{j,n,t-1-l}u_{j,n,t-l}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)}_{\equiv z_{i,n,t-l}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)}\right| - \mathbb{E}\left[\left|\sum_{l=1}^{L}\underbrace{\sum_{j\neq i} S_{j,n,t-1-l}u_{j,n,t-l}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)}_{\equiv z_{i,n,t-l}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)}\right|\right]\right)\cdot\underbrace{\sum_{j\neq i} S_{j,n,t-1}u_{j,n,t}}_{\equiv z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)}\right] \\ &= \mathbb{E}\left[\boldsymbol{X}_{i,t}\cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\cdot\underbrace{\left(\left|\sum_{l=1}^{L} z_{i,n,t-l}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\right| - \mathbb{E}\left[\sum_{l=1}^{L} z_{i,n,t-l}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\right]\right)}_{\equiv \tilde{Z}_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)}\cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\right] \\ &= \mathbb{E}\left[\boldsymbol{X}_{i,t}\cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\cdot\tilde{Z}_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)\right] \end{aligned} \tag{C.4}$$

All expectations are taken in the cross section of stocks n. Note that the constructed instruments $z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)$ and $\tilde{Z}_{i,n,t}\left(\boldsymbol{\zeta}_{1,t},\boldsymbol{\zeta}_{2,t}\right)$ depend on the elasticity parameters $\boldsymbol{\zeta}_{1,t}$ and $\boldsymbol{\zeta}_{2,t}$ because knowledge of these parameters is required to recover the residual demand shocks from the (residualized) equilibrium quantity changes (as in (C.2)).

Note that (C.3) and (C.4) give us an overidentified system. If I is the total number of investors, then there are 4 parameters to identify (since each $\boldsymbol{\zeta}_{k,t}$ has two elements) and $4I$ moment conditions. Chaudhary et al. (2024) prove that power is maximized by using a GMM weighting matrix that places more weight on the moment conditions of investors with less volatile residual demand shocks. Thus, following Chaudhary et al. (2024), we take the precision weighted averages of moment conditins (C.3) and (C.4) to obtain an exactly identified system:

$$\forall t : \mathbf{0} = \sum_i w_{i,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot \mathbb{E}\left[\boldsymbol{X}_{i,t} \cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right] \tag{C.5}$$

$$\mathbf{0} = \sum_i w_{i,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot \mathbb{E}\left[\boldsymbol{X}_{i,t} \cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot \tilde{Z}_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right] \tag{C.6}$$

$$w_{i,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) = \frac{\mathbb{V}^{-2}[u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)]}{\sum_j \mathbb{V}^{-2}[u_{j,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)]} \tag{C.7}$$

Empirically, we find the minimizers of the following objective function for each period t

$$\min_{\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}} \sum_k \left(\sum_i \hat{w}_{i,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot \hat{\mathbb{E}}\left[X_{i,t,k} \cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right]\right)^2 \tag{C.8}$$

$$+ \sum_k \left(\sum_i \hat{w}_{i,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot \hat{\mathbb{E}}\left[X_{i,t,k} \cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot \tilde{Z}_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right]\right)^2, \tag{C.9}$$

where $\hat{\mathbb{E}}$ indicates the empirical moments as opposed to the population moments and $X_{i,t,k}$ is the k-th element of $\boldsymbol{X}_{i,t}$.

C.1.2 Implementation Details

Time-Variation To improve numerical stability, we estimate the nonlinear demand curve (12) in rolling five-year windows as opposed to in each quarter. Specifically, the residualization process described in (C.1.1) is conducted in each quarter, but then the residual demand curve (C.1) is estimated in the five-year window.

Leave-One-Type Out We compute moment conditions (C.3) and (C.4) for investor i and all other investors j of a different 13 institution type (banks, investment advisors, insurance companies, mutual funds, pension funds, and other, as well as for the residual "household" sector that includes direct holdings by households as well as by non-13F institutions). That is, we construct the instruments $z_{i,n,t}$ and $\tilde{Z}_{i,n,t}$ in a leave-one-type out fashion as opposed to a leave-one-investor out fashion.

Removing Latent Characteristics The baseline analysis removes five latent stock characteristics using PCA from each within-quarter panel of stocks × pseudo investors. Appendix C.2 presents similar results when removing alternative numbers of latent characteristics.

Unbalanced Panel Empirically, the within-quarter investor × stock panel is highly unbalanced: most investors don't hold most stocks. Thus, to increase power we replace the precision weights in (C.7) with

$$w_{i,t}\left(\boldsymbol{a}_t, \boldsymbol{b}_t\right) = \frac{\mathbb{V}^{-2}[u_{i,n,t}]}{\sum_k \mathbb{V}^{-2}[u_{k,n,t}]} \cdot N_{i,t}$$

where $N_{i,t}$ is the number of stocks held by investor i in quarter t. That is, we put more weight on the moment conditions of investors with more holdings.

Root Selection In some quarters, the GMM minimization problem (C.9) has multiple roots — multiple solutions $\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)$ that set the minimized value of the objective function exactly equal to zero. This situation can arise since we are solving a nonlinear, exactly-identified system of equations, so there is no guarantee of a unique solution. In these situations, we use the folowing approach to select among these multiple roots:

1. Let $\mathcal{S}$ be the set of candidate roots. Let $\mathcal{S}_+$ and $\mathcal{S}_-$ be the subsets of candidate roots for which $\zeta_{1,0} > 0$ and $\zeta_{1,0} \leq 0$, respectively.

2. If only one candidate roots has $\zeta_{1,0} > 0$ (i.e. $|S_+| = 1$) use that root. Selecting roots with $\zeta_{1,0} > 0$ imposes the economic prior that, on average, demand curves should slope down.

3. Otherwise if there is more than one candidate root with $\zeta_{1,0} > 0$ (i.e. $|S_+| > 1$), then we use overidentifying restrictions to select amond these. While our baseline estimation solves an exactly-identified GMM problem, we can expand the problem to be overidentified using the following moment conditions:

$$\forall t: \mathbf{0} = \boldsymbol{g}_1\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \equiv \sum_i w_{i,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot \mathbb{E}\left[\boldsymbol{X}_{i,t} \cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right] \tag{C.10}$$

$$\forall t, l: \mathbf{0} = \boldsymbol{g}_{2,l}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \equiv \sum_i w_{i,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \tag{C.11}$$

$$\cdot \mathbb{E}\left[\boldsymbol{X}_{i,t} \cdot u_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) \cdot \left(\operatorname{sign}\left(\sum_{l'=1}^{L} z_{i,n,t-l'}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right) z_{i,n,t-l}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) - \frac{1}{L} \cdot \mathbb{E}\left[\sum_{l=1}^{L} z_{i,n,t-l}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right]\right) \cdot z_{i,n,t}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right] \tag{C.12}$$

That is, we keep moment condition (C.5) as is and split up moment condition (C.6) into L individual moment conditions (C.12) (that all sum to (C.6)). Letting

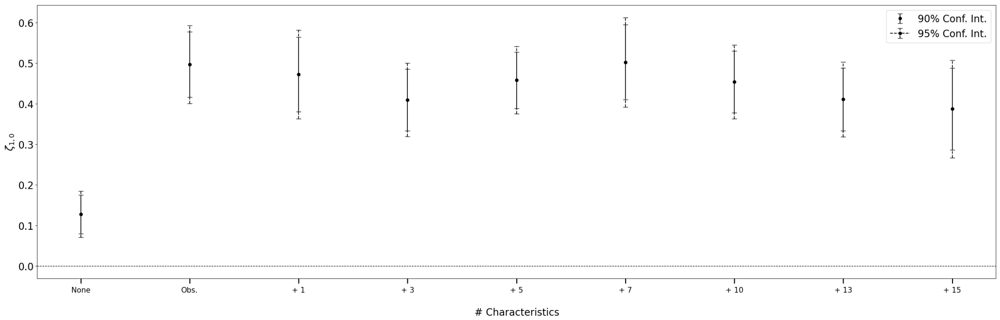
$$\boldsymbol{g}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) = \left[\boldsymbol{g}_1\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right), \boldsymbol{g}_{2,1}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right), \ldots, \boldsymbol{g}_{2,L}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)\right]$$

we choose the candidate root that minimizes this J statistic

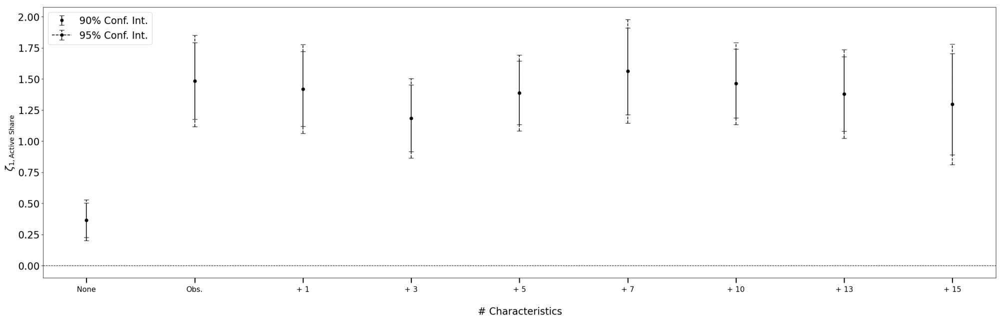
$$J\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right) = \boldsymbol{g}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right)' \boldsymbol{g}\left(\boldsymbol{\zeta}_{1,t}, \boldsymbol{\zeta}_{2,t}\right). \tag{C.13}$$

That is, we chose the candidate root that best fits the expanded set of moment conditions (C.10) and (C.12) using the identity weighting matrix. This approach is analogous to the Sargan-Hansen J-test (Hansen (1982)). The J-test is a test of overidentifying restrictions: given a solution that fits one set of moment conditions, how well does it fit a different ("held-out") set of moment conditions (where fit is evaluated based on the J-statistic (C.13), possibly with a more general weighting matrix). While the J-test is usually used to test if a model is correctly specified, here we use the J-statistic as a "model" selection criterion. Doing so ensures the selected solution is the most consistent with all available moment conditions (C.10) and (C.12), not only the particular linear combinations (C.5) and (C.6) used in the exactly-identified GMM problem (C.9).

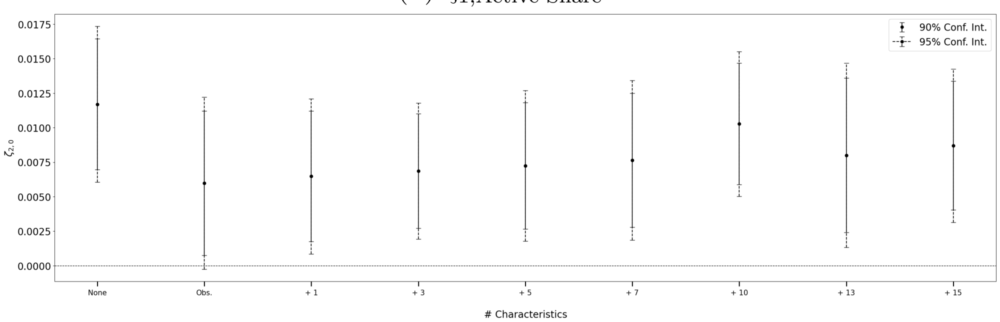
C.2 Supplemental Empirical Results



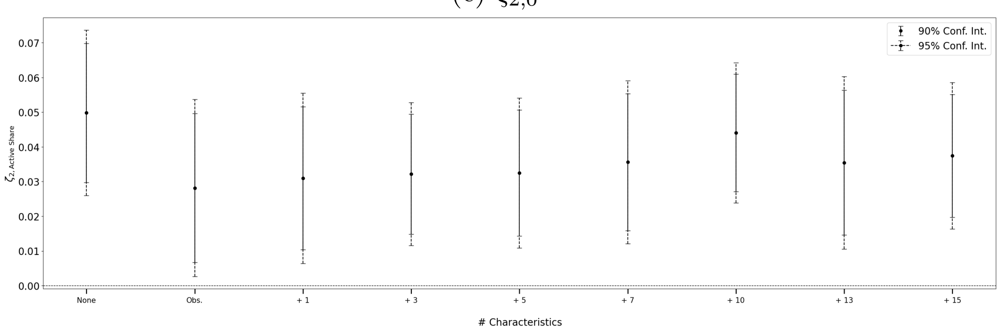
(a) $\zeta_{1,0}$



(b) $\zeta_{1,\text{Active Share}}$



(c) $\zeta_{2,0}$



(d) $\zeta_{2,\text{Active Share}}$

Figure C.3. Panels (a) through (d) display the estimation results for alternate specifications of the nonlinear demand curve (8). The "None" spefification displays the estimation results when controlling for no observed or latent characteristics. The "Obs." specification controls only for the observed stock characteristics and industry indicators described in Section 2. All subsequent specifications control for both the observed stock characteristics and industry indicators, as well the number of latent stock-quarter characteristics (estimated with PCA) indicated on the x-axis.

D Alternative Mechanisms

D.1 Position Constraints

Assume type-I investors maximize mean-variance utility over terminal wealth but is subject to a "leverage constraint" — a limit on position size:

$$\max_{Q_I} \mathbb{E}\left[Q_I\left(\tilde{D}-P\right)\right]-\frac{\gamma}{2}\mathbb{V}\left[Q_I\left(\tilde{D}-P\right)\right]$$
$$s.t. Q_I \leq \alpha$$

If the constraint does not bind ($QP < \alpha$), the investor demands

$$Q_I = \frac{\mathbb{E}[\tilde{D}]-P}{\gamma\mathbb{V}\left[\tilde{D}\right]},$$

, which is the same as the type-E investors. Thus,

$$P = \mathbb{E}\left[\tilde{D}\right]-\Theta\gamma\mathbb{V}\left[\tilde{D}\right].$$

If the constraint does bind, the type-I investors demand

$$Q = \alpha.$$

Market clearing then implies

$$\Theta_1 = (1-\epsilon)Q_I + \epsilon Q_E$$
$$\leftrightarrow P = \frac{\mathbb{E}\left[\tilde{D}\right]-\gamma\mathbb{V}\left[\tilde{D}\right]-(\Theta_1-(1-\epsilon)\alpha)}{\epsilon}$$

Note that the constraint binds if

$$Q_I \geq \alpha$$
$$\leftrightarrow \mathbb{E}\left[\tilde{D}\right]-P \geq \alpha\gamma\mathbb{V}\left[\tilde{D}\right].$$

That is, the constraint binds when the expected return is sufficiently large. We can plug in the unconstrained equilibrium price to reexpress this condition in terms of Θ_1:

$$\begin{aligned}\mathbb{E}\left[\tilde{D}\right]-P &\geq \alpha\gamma\mathbb{V}\left[\tilde{D}\right].\\ \leftrightarrow \mathbb{E}\left[\tilde{D}\right]-\left(\mathbb{E}\left[\tilde{D}\right]-\Theta\gamma\mathbb{V}\left[\tilde{D}\right]\right) &\geq \alpha\gamma\mathbb{V}\left[\tilde{D}\right]\\ \Theta &\geq \alpha.\end{aligned}$$

So the price multiplier is

$$M=-\frac{d}{d\Theta_1}P=\begin{cases}\gamma\mathbb{V}\left[\tilde{D}\right] & ,|\Theta_1|\leq\alpha\\ \gamma\mathbb{V}\left[\tilde{D}\right]/\epsilon & ,|\Theta_1|>\alpha\end{cases}$$

Thus, M increases at the point where the constraint binds ($|\Theta_1|=\alpha$).

We can reexpress this setup in the cost function notation of Section 6 as

$$\max_{Q_I}\mathbb{E}\left[Q_I\left(\tilde{D}-P\right)\right]-\frac{\gamma}{2}\mathbb{V}\left[Q_I\left(\tilde{D}-P\right)\right]-C(Q_I,\mu),$$

where

$$\begin{aligned}C(Q_I,\mu)&=-\lambda(\mu)\left(\alpha-|Q_I|\right)\\ \lambda(\mu)&=\begin{cases}0, & |\Theta_1|\leq\alpha\\ |\mu|-\gamma\alpha\sigma_D^2, & |\Theta_1|>\alpha\end{cases}\end{aligned}$$

$\lambda(Q_I,P)$ represents the shadow cost of relaxing the leverage constraint.

Hence, marginal cost is

$$MC(Q_I,\mu)\equiv\frac{\partial}{\partial Q_I}C(Q,\mu)=\begin{cases}0, & |\Theta_1|\leq\alpha\\ \text{sign}(Q_I)\lambda(\mu), & |\Theta_1|>\alpha\end{cases}$$

and so we have

$$\frac{\partial}{\partial\mu}MC(Q_I,\mu)=\begin{cases}0, & |\Theta_1|\leq\alpha\\ 1, & |\Theta_1|>\alpha.\end{cases}$$

Thus, as $\Theta_1 > 0$ increases from below to α and the constraint binds, $\frac{\partial}{\partial P} MC(Q_I, P)$ rises from 0 to 1, and so the multiplier rises.

D.2 Benchmarking

Let the type-I investors' compensation be linear combination of her absolute return and her return relative to a benchmark (as in Pavlova and Sikorskaya (2023))

$$c(Q_I) = \underbrace{aQ_I\left(\tilde{D} - P\right)}_{\text{Absolute Return}} + \underbrace{b\left[Q_I\left(\tilde{D} - P\right) - \bar{Q}\left(\tilde{D} - P\right)\right]}_{\text{Return Relative to Benchmark}},$$

where $\bar{Q}$ represents the benchmark holding.

Assume the investor maximizes mean-variance utility over compensation

$$\max_{Q_I} \mathbb{E}\left[c(Q_I)\right] - \frac{\gamma}{2}\mathbb{V}\left[c(Q_I)\right].$$

One can show the optimal quantity demanded is

$$Q_I = \frac{\mathbb{E}\left[\tilde{D}\right] - P}{\gamma \mathbb{V}\left[\tilde{D}\right] \cdot (a+b)} + \frac{b}{a+b}\bar{Q},$$

and so the equilibrium price is

$$\Theta_1 = (1-\epsilon)Q_I + \epsilon Q_E$$
$$P = \mathbb{E}\left[\tilde{D}\right] - \left(\Theta_1 - (1-\epsilon)\frac{b}{a+b}\bar{Q}\right)\frac{\gamma \mathbb{V}\left[\tilde{D}\right]}{\frac{1-\epsilon}{a+b} + \epsilon}.$$

The price multiplier is thus:

$$M = -\frac{d}{d\Theta_1}P = \frac{\gamma \mathbb{V}\left[\tilde{D}\right]}{\frac{1-\epsilon}{a+b} + \epsilon} = \frac{\gamma \mathbb{V}\left[\tilde{D}\right](a+b)}{1 + (a+b-1)\epsilon},$$

which is constant.

Reformulating this setup in the notation of Section 6, we can write the cost function as

$$\begin{aligned}
C &= \mathbb{E}\left[Q\left(\tilde{D}-P\right)\right] - \mathbb{E}\left[c(Q)\right] + \frac{\gamma}{2}\left[\mathbb{V}\left[c(Q)\right] - \mathbb{V}\left[Q\left(\tilde{D}-P\right)\right]\right] \\
&= (1-a-b)\mathbb{E}\left[Q\left(\tilde{D}-P\right)\right] - b\mathbb{E}\left[\bar{Q}\left(\tilde{D}-P\right)\right] \\
&\quad + \frac{\gamma}{2}\mathbb{V}\left[\tilde{D}\right]\left[\left((a+b)^2-1\right)Q^2 - 2(a+b)bQ\bar{Q}\right] \\
&= \mu\left(Q-(a+b)Q+b\bar{Q}\right) + \frac{\gamma}{2}\sigma_D^2\left[\left((a+b)Q-b\bar{Q}\right)^2 - Q^2\right] \\
\leftrightarrow C(Q,\mu) &= \mu(\lambda(Q)+Q) + \frac{\gamma}{2}\sigma_D^2\left[\lambda(Q)^2 - Q^2\right],
\end{aligned}$$

where $\lambda(Q) = -(a+b)Q + b\bar{Q}$ is linear and decreasing in Q.

Marginal cost is

$$MC(Q,\mu) = \mu(1+\lambda^{'}(Q)) + \gamma\sigma_D^2\left(\lambda(Q)\lambda^{'}(Q) - Q\right)$$

Thus, we have

$$\frac{\partial}{\partial\Theta_1}MC(\Theta_1,\mu) = \gamma\sigma_D^2\left(\left(\lambda^{'}(Q)\right)^2 - 1\right) = \gamma\sigma_D^2\left((a+b)^2-1\right)$$

which is constant. Similarly, we have

$$\frac{\partial}{\partial\mu}MC(\Theta_1,\mu) = 1+\lambda^{'}(Q) = -(1-a-b)$$

which is constant.

D.3 Convex Adjustment Costs

A representative investor maximizes mean-variance utility subject to an adjustment cost

$$\max_Q \mathbb{E}\left[Q\left(\tilde{D}-P\right)\right] - \frac{\gamma}{2}\mathbb{V}\left[Q\left(\tilde{D}-P\right)\right] - \lambda\left|Q-\Theta_0\right|^{\alpha}.$$

Note that an economy with a continuum of investors with different fixed costs can be represented as an economy with a representative investor economy who solves this problem for aggregate demand. The adjustment cost penalizes deviations from the number of shares the investor is originally endowed with Θ_0.

From (21), one can show that for $\epsilon \to 0$ the equilibrium price is:

$$P = \bar{D} - \Theta_1 \cdot \gamma\sigma_D^2 - \alpha\lambda(\Theta_1 - \Theta_0)^{\alpha-1} \operatorname{sign}|\Theta_1 - \Theta_0|$$

The price multiplier is:

$$M = -\frac{d}{d\Theta_1}P = \gamma\sigma_D^2 + \alpha(\alpha-1)\lambda|\Theta_1 - \Theta_0|^{\alpha-2}.$$

For $1 < \alpha < 2$, M decreases in the size of the $t = 2$ supply shock ($|\Theta_1 - \Theta_0|$). However, regardless of the value of α, M does not depend on Θ_0.

Reformulating this setup in the notation of Section 6, we can write the cost function as

$$C(Q, \Theta_0) = \lambda \left|Q - \Theta_0\right|^{\alpha}.$$

Marginal cost is given by

$$MC(Q, \Theta_0) = \frac{\partial}{\partial Q}C(Q, \Theta_0) = \alpha\lambda(Q - \Theta_0)^{\alpha-1} \operatorname{sign}\left|Q - \Theta_0\right|.$$

The derivative of marginal cost with respect to Q is

$$\frac{\partial}{\partial Q}MC(Q, \Theta_0) = \alpha(\alpha-1)\lambda|\Theta_1 - \Theta_0|^{\alpha-2},$$

which does not depend on Θ_0 conditional on $\Theta_1 - \Theta_0$.

D.4 Capital Reallocation Between Investors

Let tyep-I investors have a simple quadratic inventory cost that redners their demand inelastic:

$$C(Q_I) = aQ_I^2.$$

But now assume the mass ϵ of type-E investors in $t = 1$ is endogneous

$$\epsilon = 1 - (1 + \mu^2)^{-\frac{1}{2}}$$

where $\mu = \mathbb{E}[D - P]$ is the expected return. Thus, the mass of type-E investors increases with expected return.

Market clearing in each period $t = 1$ is

$$\Theta_1 = (1-\epsilon)\underbrace{Q_1}_{=\frac{\mu}{(\gamma+a)\sigma_D^2}} + (1-\epsilon)\underbrace{Q_2}_{=\frac{\mu}{\gamma\sigma_D^2}}. \tag{D.1}$$

So the equilibrium price is

$$P = \bar{D} - \underbrace{\frac{1}{\frac{\epsilon}{\gamma} + \frac{1-\epsilon}{\gamma+a}}}_{\equiv f(\mu)} \sigma_D^2 \Theta_1$$

Let

$$f(\mu) \equiv \frac{1}{\frac{\epsilon}{\gamma} + \frac{1-\epsilon}{\gamma+a}} = \frac{\gamma(\gamma+a)}{\gamma + a\epsilon}$$

be the effective aggregate risk aversion in the economy. Note that $f(\mu)$ is decreasing in μ since ϵ is increasing in μ.

The price multiplier is given by

$$\begin{aligned} M \equiv -\frac{dP}{d\Theta_1} &= \frac{f(\mu)\sigma_D^2}{1 - \Theta_1 \sigma_D^2 f'(\mu)} \\ &= \frac{f(\mu)\sigma_D^2}{1 - \mu \frac{f'(\mu)}{f(\mu)}} \end{aligned}$$

where the second line follows since the equilibrum price expression implies

$$\mu = f(\mu)\sigma_D^2 \Theta_1.$$

Note that

$$\begin{aligned} f^{'}(\mu) &= -\frac{\gamma(\gamma+a)}{(\gamma+a\epsilon)^2} a\epsilon^{'}(\mu). \\ &= -\frac{\gamma(\gamma+a)}{(\gamma+a\epsilon)^2} a(1+\mu^2)^{-\frac{3}{2}}\mu, \end{aligned}$$

so

$$
\begin{aligned}
\frac{f'(\mu)}{f(\mu)} &= \frac{-\frac{\gamma(\gamma+a)}{(\gamma+a\epsilon)^2}a(1+\mu^2)^{-\frac{3}{2}}\mu}{\frac{\gamma(\gamma+a)}{\gamma+a\epsilon}} \\
&= -\frac{1}{(\gamma+a\epsilon)}a(1+\mu^2)^{-\frac{3}{2}}\mu,
\end{aligned}
$$

Thus,

$$
\begin{aligned}
M &= \frac{\frac{\gamma(\gamma+a)}{\gamma+a\epsilon}\sigma_D^2}{1+\mu^2\frac{1}{(\gamma+a\epsilon)}a(1+\mu^2)^{-\frac{3}{2}}} \\
&= \frac{\gamma(\gamma+a)\sigma_D^2}{\gamma+a\epsilon+\mu^2 a(1+\mu^2)^{-\frac{3}{2}}} \\
&= \frac{\gamma(\gamma+a)\sigma_D^2}{\gamma+a-a(1+\mu^2)^{-\frac{1}{2}}+\mu^2 a(1+\mu^2)^{-\frac{3}{2}}} \\
&= \frac{\gamma(\gamma+a)\sigma_D^2}{\gamma+a\left(1-(1+\mu^2)^{-\frac{1}{2}}+\mu^2(1+\mu^2)^{-\frac{3}{2}}\right)} \\
&= \frac{\gamma(\gamma+a)\sigma_D^2}{\gamma+a\left(1-(1+\mu^2)^{-\frac{1}{2}}\left(1-\mu^2(1+\mu^2)^{-1}\right)\right)} \\
&= \frac{\gamma(\gamma+a)\sigma_D^2}{\gamma+a\left(1-(1+\mu^2)^{-\frac{3}{2}}\right)}
\end{aligned}
$$

which is decreasing in $|\mu|$.

Lastly, μ is increasing in Θ_1 since

$$
\begin{aligned}
& \mu = f(\mu)\sigma_D^2\Theta_1 \\
\rightarrow\ & \frac{d\mu}{d\Theta} = f(\mu)\sigma_D^2 + f'(\mu)\sigma_D^2\Theta_1\frac{d\mu}{d\Theta} \\
\leftrightarrow\ & \frac{d\mu}{d\Theta} = \frac{f(\mu)\sigma_D^2}{1+\frac{\gamma(\gamma+a)}{(\gamma+a\epsilon)^2}a(1+\mu^2)^{-\frac{3}{2}}\mu\sigma_D^2\Theta_1} \\
& \qquad > 0
\end{aligned}
$$

since μ and Θ_1 have the same sign (since $f(\mu) > 0$). A symmetric argument can be made to show that $-\mu$ is decreasing in Θ_1 for $\Theta_1 < 0$. Thus, $|\mu|$ is increasing in $|\Theta|$.

Thus, the $t = 2$ multiplier is decreasing in $|\Theta_1|$. However, each investor's elasticity is fixed, as can be seen from (D.1)

D.5 Participation Costs

Type-I investors have zero initial holdings in $t = 0$ and face a fixed participation cost λ in $t = 1$ if they choose a non-zero quantity:

$$C(Q_I) = \begin{cases} 0 & , Q_I = 0 \\ \lambda & , Q_I \neq 0 \end{cases}$$

Conditional on holding a non-zero quantitiy, all investors (both types I and E) choose

$$Q = \frac{\mathbb{E}\left[\tilde{D} - P\right]}{\gamma \mathbb{V}\left[\tilde{D}\right]} \tag{D.2}$$

Thus, type-I investors decides to enter the asset if

$$\frac{\mathbb{E}\left[\tilde{D} - P\right]^2}{2\gamma \mathbb{V}\left[\tilde{D}\right]} > \lambda.$$

So aggregate demand $(1 - \epsilon)Q_I + \epsilon Q_E$ is

$$Q_{\text{agg}} = \begin{cases} \frac{\epsilon \mathbb{E}\left[\tilde{D} - P\right]}{\gamma \mathbb{V}\left[\tilde{D}\right]} & , \mathbb{E}\left[\tilde{D} - P\right]^2 \leq 2\gamma \mathbb{V}\left[\tilde{D}\right] \lambda \\ \frac{\mathbb{E}\left[\tilde{D} - P\right]}{\gamma \mathbb{V}\left[\tilde{D}\right]} & , \mathbb{E}\left[\tilde{D} - P\right]^2 > 2\gamma \mathbb{V}\left[\tilde{D}\right] \lambda \end{cases}$$

By market clearing $Q_{\text{agg}} = \Theta_1$ we have the following equilibrium price:

$$P = \begin{cases} \bar{D} - \Theta_1 \frac{\gamma \sigma_D^2}{\epsilon} & , |\Theta_1| \leq \sqrt{\frac{2\lambda}{\gamma \sigma_D^2}} \\ \bar{D} - \Theta_1 \gamma \sigma_D^2 & , |\Theta_1| > \sqrt{\frac{2\lambda}{\gamma \sigma_D^2}} \end{cases}$$

So the price multiplier is

$$M \equiv -\frac{dP}{d\Theta} = \begin{cases} \frac{\gamma\sigma_D^2}{\epsilon} & , |\Theta_1| \geq \sqrt{\frac{2\lambda}{\gamma\sigma_D^2}} \\ \gamma\sigma_D^2 & , |\Theta_1| < \sqrt{\frac{2\lambda}{\gamma\sigma_D^2}} \end{cases}$$

which is decreasing in the supply shock magnitude $|\Theta_1|$.

However, each investor's elasticity on the intensive margin is fixed, as can be seen from (D.3).

D.6 Fixed Adjustment Costs

Type-I investors face a fixed adjustment cost λ in $t = 1$ if they adjust from their $t = 0$ quantity Θ_0:

$$C(Q_I) = \begin{cases} 0 & , Q_I = \Theta_0 \\ \lambda & , Q_I \neq \Theta_0 \end{cases}$$

Conditional on adjusting, all investors (both types I and E) choose

$$Q = \frac{\mathbb{E}\left[\tilde{D} - P\right]}{\gamma\mathbb{V}\left[\tilde{D}\right]} \tag{D.3}$$

Thus, type-I investors decides to adjust if

$$\begin{aligned} \lambda &< \frac{\mathbb{E}\left[\tilde{D} - P\right]^2}{2\gamma\mathbb{V}\left[\tilde{D}\right]} - \left[\Theta_0\mathbb{E}\left[\tilde{D} - P\right] - \frac{\gamma}{2}\Theta_0^2\mathbb{V}\left[\tilde{D}\right]\right] \\ \leftrightarrow 0 &< \frac{1}{2\gamma\sigma_D^2}\mu^2 - \Theta_0\mu + \frac{\gamma}{2}\Theta_0^2\sigma_D^2 - \lambda \\ \leftrightarrow 0 &< (\mu - \mu_+)(\mu - \mu_-) \end{aligned}$$

where

$$\begin{aligned} \mu_+ &= \Theta_0\gamma\sigma_D^2 + \sqrt{2\lambda\gamma\sigma_D^2} \\ \mu_- &= \Theta_0\gamma\sigma_D^2 - \sqrt{2\lambda\gamma\sigma_D^2} \end{aligned}$$

So aggregate demand $(1-\epsilon)Q_I + \epsilon Q_E$ is

$$Q_{\text{agg}} = \begin{cases} \frac{\epsilon\mu}{\gamma\mathbb{V}[\tilde{D}]} + (1-\epsilon)\Theta_0 & , \mu \in [\mu_-, \mu_+] \\ \frac{\mu}{\gamma\mathbb{V}[\tilde{D}]} & , \mu \notin [\mu_-, \mu_+] \end{cases}$$

By market clearing $Q_{\text{agg}} = \Theta_1$ we have the following equilibrium price:

$$P = \begin{cases} \frac{\bar{D} - \gamma\sigma_D^2(\Theta_1 - (1-\epsilon)\Theta_0)}{\epsilon} & , \mu \in [\mu_-, \mu_+] \\ \bar{D} - \gamma\sigma_D^2\Theta_1 & , \mu \notin [\mu_-, \mu_+] \end{cases}$$

Note that

$$\begin{aligned} \mu > \mu_+ &= \Theta_0\gamma\sigma_D^2 + \sqrt{2\lambda\gamma\sigma_D^2} \\ \leftrightarrow \Theta_1 &> \Theta_0 + \frac{\sqrt{2\lambda}}{\sqrt{\gamma\sigma_D^2}} \end{aligned}$$

and

$$\begin{aligned} \mu < \mu_- &= \Theta_0\gamma\sigma_D^2 - \sqrt{2\lambda\gamma\sigma_D^2} \\ \leftrightarrow \Theta_1 &< \Theta_0 - \frac{\sqrt{2\lambda}}{\sqrt{\gamma\sigma_D^2}}. \end{aligned}$$

Thus,

$$\mu \in [\mu_-, \mu_+] \leftrightarrow \Theta_1 \in \left[\underbrace{\Theta_0 - \sqrt{2\lambda\gamma\sigma_D^2}}_{\equiv\Theta_-}, \underbrace{\Theta_0 + \sqrt{2\lambda\gamma\sigma_D^2}}_{\equiv\Theta_+},\right]$$

Hence, we can write the equilibrium price as

$$P = \begin{cases} \frac{\bar{D} - \gamma\sigma_D^2(\Theta_1 - (1-\epsilon)\Theta_0)}{\epsilon} & , |\Theta_1 - \Theta_0| \leq \sqrt{2\lambda\gamma\sigma_D^2} \\ \bar{D} - \gamma\sigma_D^2\Theta_1 & , |\Theta_1 - \Theta_0| > \sqrt{2\lambda\gamma\sigma_D^2} \end{cases}$$

So the price multiplier is

$$M \equiv -\frac{dP}{d\Theta} = \begin{cases} \frac{\gamma\sigma_D^2}{\epsilon} & , |\Theta_1 - \Theta_0| \leq \sqrt{2\lambda\gamma\sigma_D^2} \\ \gamma\sigma_D^2 & , |\Theta_1 - \Theta_0| > \sqrt{2\lambda\gamma\sigma_D^2} \end{cases}$$

which is decreasing in the supply shock magnitude $|\Theta_1|$.

D.7 Costly Information Acquisition

D.7.1 Setup

Let the mass of type-E investors $\epsilon \to 0$ for simplicity.

There are three periods: $t = 1, 2, 3$. Information choice occurs in $t = 1$. The asset market clears in both $t = 1$ and $t = 2$. Asset payoffs are realized in $t = 3$.

Assets: There is one asset that pays a risky dividend in period $t = 3$:

$$\tilde{D} = \bar{D} + \eta + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma_\varepsilon^2\right), \quad \eta \sim N\left(0, \sigma_\eta^2\right).$$

where the shocks ε and η are independent, and the latter is learnable by paying a cost (to be explained later). The asset has exogenous, stochastic supply of Θ_t in period t:

$$\begin{aligned} \Theta_1 &= z_1 \\ \Theta_2 &= z_1 + z_2 \\ z_t &\sim N(0, \sigma_z^2), \end{aligned}$$

where z_1 is independent of z_2. The supply should be interpreted as the residual supply faced by the representative investor: the total fixed, positive net supply minus the exogenous demand shocks of some noise traders. The exogenous risk-free rate is normalized to zero.

Agents and Preferences: There is a unit mass of atomistic investors who choose their portfolios at $t = 1, 2$ to maximize mean-variance utility over terminal wealth

$$\max_{Q_t} \mathbb{E}_t\left[Q_t\left(\tilde{D} - P\right)\right] - \frac{\gamma}{2}\mathbb{V}_t\left[Q_t\left(\tilde{D} - P\right)\right] \tag{D.4}$$

Information: At $t = 1$, the investor knows the asset pricing parameters $\bar{D}, \sigma_\varepsilon^2$, and σ_Θ^2.

However, she does not know η. She has the objectively correct prior and believes

$$\eta \sim N\left(0, \sigma_\eta^2\right). \tag{D.5}$$

After observing Θ_1, the investor can pay a cost $\tilde{C}(G)$ (not the same as the portfolio choice cost function from Section 6) to acquire signal a noisy signal s for η at $t = 2$:[3]

$$s = \eta + u, u \sim N\left(0, \sigma_u^2\right) \tag{D.6}$$
$$G = \frac{\sigma_u^{-2}}{\sigma_u^{-2} + \sigma_\eta^{-2}}.$$

G denotes the Bayesian gain of signal s. With signal s, the investor's posterior distribution at $t = 2$ is

$$\eta \sim N\left(G \cdot s, (1 - G) \cdot \sigma_\eta^2\right). \tag{D.7}$$

We assume the cost function $\tilde{C}$ is such that the marginal cost of reducing uncertainty is positive ($\tilde{C}'(G) > 0$) and increasing ($\tilde{C}''(G) > 0$).

Since the investor does not know at $t = 1$ what signal s she will observe at $t = 2$, the investor chooses G to maximize *expected* utility at $t = 1$ (which integrates over all possible realizations of s):

$$\max_G \mathbb{E}_1 \left[\mathbb{E}_2 \left[Q_2 \left(\tilde{D} - P\right)\right] - \frac{\gamma}{2} \mathbb{V}_1 \left[Q_2 \left(\tilde{D} - P\right)\right]\right] - C(G) \tag{D.8}$$

The investor's beliefs are rational: her $t = 1$ beliefs about future prices and quantities are consistent with the true equilibrium distributions.

Equilibrium Definition: An equilibrium is defined as a set of portfolio choices (Q), information choices (G), and asset prices (P) such that:

1. The information choice G at $t = 1$ maximizes (D.8) given the investor's prior beliefs about η and her knowledge of the supply at $t = 1$ (Θ_1).

2. The portfolio choice Q_t at $t = 1$ and $t = 2$ maximizes (D.4) given the investor's information.

[3] As in e.g. Grossman and Stiglitz (1980); Kyle (1989); Van Nieuwerburgh and Veldkamp (2009, 2010); Kacperczyk et al. (2016); Han (2018).

3. The asset market clears in $t = 1, 2$:

$$Q_t = \Theta_t. \tag{D.9}$$

D.7.2 Solving the Model

We solve the model backwards. First, we solve the investor's portfolio choice problem (D.4) in $t = 1, 2$ holding the information set fixed in both periods. Second, we then impose market clearing (D.9) to solve for the equilibrium price in $t = 1, 2$ as a function of the investor's information choice. Third, we solve for the investor's ex-ante optimal gain G at $t = 1$ that maximizes (D.8), which pins down the equilibrium price and quantity in both periods.

Portfolio Choice: At $t = 1$, the investor solves (D.4) and obtains

$$Q_1 = \frac{\bar{D} - P_1}{\gamma \left(\sigma_\varepsilon^2 + \sigma_\eta^2\right)}.$$

Similarly, at $t = 2$ the investor chooses

$$Q_2 \left(\mathcal{I}_2\right) = \frac{\bar{D} + \bar{G} \cdot s - P_2}{\gamma \left(\sigma_\varepsilon^2 + (1 - \bar{G}) \cdot \sigma_\eta^2\right)}. \tag{D.10}$$

where Q_2 is a function of the equilibrium $t = 2$ information set $\mathcal{I}_2 = \left(s, \bar{G}\right)$: s is the random realization of (D.6) and $\bar{G}$ denote the equilibrium gain chosen in $t = 1$.

Market Clearing: From market clearing (D.9) we have

$$\begin{aligned} P_1 &= \bar{D} - \Theta_1 \gamma \left(\sigma_\varepsilon^2 + \sigma_\eta^2\right) \\ P_2 &= \bar{D} + \bar{G} \cdot s - \Theta_2 \gamma \left(\sigma_\varepsilon^2 + \sigma_\eta^2 \left(1 - \bar{G}\right)\right). \end{aligned} \tag{D.11}$$

Note that, if supply equals its mean ($\Theta_t = 0$), then price equals expected terminal dividend. Exogenous supply shocks distort price away from this fundamental value.

Information Choice: At $t = 1$ the investor chooses the optimal gain for $t = 2$ after observing Θ_1. She also knows the equilibrium distributions of quantity $Q\left(\mathcal{I}_2\right)$ and price P, which she takes as given.

Appendix D.7.4 shows the equilibrium first-order condition for the ex-ante information choice problem (D.8) is given by equalizing the marginal benefit and cost of information acquisition:

$$\frac{\gamma}{2}\sigma_\eta^2\left(\Theta_1^2+\sigma_z^2\right)=\tilde{C}'(G). \tag{D.12}$$

Since the cost C is convex in the chosen gain G, (D.12) implies that the optimal gain G^* is increasing in the magnitude of $t=1$ supply Θ_1.

D.7.3 Model Implications

The equilibrium quantities (D.10) and prices (D.11) are consistent with our empirical findings.

Price multipliers are smaller for larger (accumulated) supply shocks From (D.11), the price multiplier in $t=2$ is[4]

$$M(\Theta_1)\equiv\frac{\partial P_2}{\partial z_2}=\gamma\left(\sigma_\varepsilon^2+\sigma_\eta^2\left(1-\bar{G}\right)\right).$$

Since the equilibrium gain $\bar{G}$ is increasing in the magnitude of Θ_1 by (D.12), $M(\Theta_1)$ is decreasing in the magnitude of Θ_1. Thus, the price multiplier is smaller for larger accumulated supply shocks:

$$\frac{\partial M(\Theta_1)}{\partial\Theta_1}<0. \tag{D.13}$$

Since we interpret the (residual) supply Θ_t as the demand from noise traders outside the model, this theoretical result (D.13) as consistent with our dynamic empirical findings in Section 4. Moreover, we interpret (D.13) as consistent with our static empirical findings in Section 3 since one month or quarter is long enough for information acquisition and trading to have occured (i.e. one can think of $t=1,2$ as two sub-periods within a month or quarter).

Demand is more elastic within-investor and on the intensive margin for larger (accumulated) supply shocks From (D.10), the "price elasticity of demand" for the representative investor in $t=2$ is

$$\zeta(\Theta_1)\equiv\frac{\partial Q_2}{\partial P_2}=\frac{1}{\gamma\left(\sigma_\varepsilon^2+(1-\bar{G})\cdot\sigma_\eta^2\right)}$$

[4]This is technically not the price multiplier which is defined in elasticity units. However, in this model where portfolio choice is in the number of shares (instead of portfolio weights) and dollar expected returns (instead of percentage expected returns), our definition is arguably a more natural object for measuring price impact. The same comment applies to our later discussion of "demand elasticity".

Since the equilibrium gain $\bar{G}$ is increasing in the magnitude of Θ_1 by (D.12), $\zeta(\Theta_1)$ is increasing in the magnitude of Θ_1. Thus, the price elasticity is larger for larger accumulated supply shocks:

$$\frac{\partial \zeta(\Theta_1)}{\partial \Theta_1} > 0, \tag{D.14}$$

which is consistent with our empirical findings from institutional investor holdings data in Section 5. Note that (D.14) is a statement about how a particular investor's elasticity varies on the intensive margin of demand (i.e. for an asset that the investor already holds and continues to hold).

D.7.4 Proof of Equation (D.12)

Proof. At $t = 1$ the investor chooses the optimal gain for $t = 2$ after observing Θ_1. She also knows the equilibrium distributions of quantity $Q(\mathcal{I}_2)$ and price P, which she takes as given.[5]

Rewrite the information choice problem (D.8) in terms of the optimal $t = 2$ quantity, which is a function of the chosen gain G:

$$\max_G \mathbb{E}_1 \left[Q_2(G) \cdot \mathbb{E}\left[\bar{D} - P\right] - \frac{\gamma}{2} Q_2(G)^2 \cdot \mathbb{V}\left[\bar{D} - P\right] \right] - \tilde{C}(G)$$

Note that by the envelope theorem, $\partial Q_2(G)/\partial G = 0$. Thus, taking the first-order condition with respect to G yields

$$\begin{aligned}
0 &= \mathbb{E}_1 \left[-\frac{\gamma}{2} Q_2(G)^2 \cdot \frac{\partial}{\partial G} \mathbb{V}\left[\bar{D} - P\right] - \tilde{C}'(G) \right] \\
&= \mathbb{E}_1 \left[-\frac{\gamma}{2} Q_2(G)^2 \cdot \frac{\partial}{\partial G} \left[\sigma_\varepsilon^2 + \sigma_\eta^2 (1 - G)\right] - \tilde{C}'(G) \right] \\
&= \mathbb{E}_1 \left[\frac{\gamma}{2} Q_2(G)^2 \cdot \sigma_\eta^2 - \tilde{C}'(G) \right] \\
&= \mathbb{E}_1 \left[\frac{\gamma}{2} \Theta_2^2 \cdot \sigma_\eta^2 - \tilde{C}'(G) \right] \\
\leftrightarrow \frac{\gamma}{2} \sigma_\eta^2 \left(\Theta_1^2 + \sigma_z^2\right) &= \tilde{C}'(G)
\end{aligned}$$

where the fourth equation follows by market clearing (D.9), and the fifth equation follows since information choice is done after observing Θ_1.

[5] That is, the representative investor does not internalize the impact of her information choice on price, just as she does not internalize the impact of her portfolio choice on price.

□