Granular Treasury Demand with Arbitrageurs

Kristy A. E. Jansen Wenhao Li Lukas Schmid

June 2026

Abstract

We show that understanding the Treasury market requires both estimating granular investor demand and structurally modeling arbitrageurs. Using a new dataset of sector-level U.S. Treasury holdings, we estimate demand functions that exhibit strong cross-maturity substitution. Embedding these estimates in an equilibrium model with risk-averse arbitrageurs yields two main findings. First, Treasury market elasticity is steeply downward sloping in maturity, with very high elasticity in the T-bill market; without structurally modeling arbitrageurs, a pure demand system implies implausibly low T-bill elasticity. Second, cross-maturity substitution implies that monetary tightening raises term premia; without it, as in baseline preferred habitat models, the prediction reverses.

Keywords: Treasury demand; arbitrage; term premium; demand elasticity; monetary policy.

Jansen: USC Marshall School of Business, CEPR and DNB, kjansen@marshall.usc.edu. Li: USC Marshall School of Business and NBER, liwenhao@marshall.usc.edu. Schmid: USC Marshall School of Business and CEPR, lukas@marshall.usc.edu. We thank Viral Acharya (discussant), Yakov Amihud, Daniel Andrei (discussant), Adrien d'Avernas (discussant), Zefeng Chen, William Diamond (discussant), Greg Duffee (discussant), Darrell Duffie, Thomas Eisenbach (discussant), Chuck Fang (discussant), Paul Fontanier, Francois Gourio, Daniel Graves (discussant), Valentin Haddad, Samuel Hanson, Zhiguo He, Ben Hebert, Paul Huebner (discussant), Erica Jiang, Jay Kahn (discussant), Ralph Koijen, Arvind Krishnamurthy, Moritz Lenel (discussant), Martin Lettau, Karen Lewis, Jane Li, Tyler Muir, Anna Pavlova, Walker Ray (discussant), Eric Richert (discussant), Robert Richmond, Thomas Sargent, Alexi Savov, Alp Simsek, Adi Sunderam (discussant), Selale Tuzel, Quentin Vandeweyer (discussant), Dimitri Vayanos, Annette Vissing-Jorgensen, Olivier Wang, Motohiro Yogo, Xingtan Zhang, Geoffrey Zheng, and seminar participants at ANU, Bank for International Settlements, Bank of Italy, Berkeley Haas, Boston College, Chicago Fed, CKGSB, CUHK Shenzhen, De Nederlandsche Bank, ECB, Georgetown, HKUST, Lancaster, Lugano, Manchester, NY Fed, NYU Stern, Oxford, PBCSF Tsinghua, Peking University Guanghua, Purdue, SAIF, SF Fed, UCLA Macro-Finance, UIUC, University of Sydney, University of Technology Sydney, University of Toronto, UNSW, USC, Warwick, WashU Olin, Wharton, Yale SOM, as well as conference participants at 5th David Backus Memorial Conference, 11th International Conference on Sovereign Bond Markets, BI-SHoF Conference on Asset Pricing and Financial Econometrics, CEPR Asset Pricing Symposium, Chicago Treasury Market Conference, CMU Tepper-LAEF Conference, Fed Conference on Fixed Income Markets, Hong Kong International Finance Conference, JHU Carey Finance Conference, Junior Valuation Workshop at Wharton, Lake Dishui Finance Conference, LBS Summer Finance Symposium, MFA, Miami Conference on Fiscal and Monetary Interactions, NBER Asset Pricing, NBER Financial Market Frictions and Systemic Risks, Princeton Conference on Asset Demand Systems, Princeton Macro-Finance Conference, QRFE Workshop on Quantitative Finance, Rochester Financial Policy and Regulation Conference, Stanford Junior Macro-Finance Conference, Stanford SITE, STFM Conference, St. Louis Fed-WashU Olin Macrofinance Workshop, UBC Summer Finance Conference, UIC Finance Conference, and Zurich Quantitative Macroeconomics Workshop. We gratefully acknowledge financial support from NBER Financial Market Frictions and Systemic Risks initiative. We thank Winston Chen for excellent research assistance. Views expressed are those of the authors and do not necessarily reflect official positions of DNB or the Eurosystem.

1. Introduction

The U.S. Treasury market rests on two cornerstones. Banks, insurers, pension funds, money market funds, and foreign central banks hold Treasuries for safety, liquidity, liability matching, and regulatory compliance, creating maturity-specific demand that shapes the term structure. On the other hand, a smaller group of fixed-income arbitrageurs intermediates across these segments, providing elasticity where demand falls short. Despite the importance of both aspects, we lack quantitative estimates of how their interactions shape the Treasury market. In this paper, we integrate granular estimation of sector-level investor demand with a structural model of fixed-income arbitrage, and use the resulting framework to examine Treasury market elasticity and monetary policy transmission. We show that both components are necessary: granular demand estimation uncovers the cross-maturity substitution driving term premium dynamics, while structurally modeling arbitrageurs is essential for explaining the high elasticity of the T-bill market.

In the spirit of Koijen and Yogo (2019) and Vayanos and Vila (2021), we construct investor-specific Treasury demand functions at the maturity-bucket level using a novel dataset on Treasury holdings that covers about 80% of the market over 2011Q4–2022Q4. We classify players in the Treasury market into three groups: granular-demand investors, the Fed, and arbitrageurs. For granular-demand investors, including commercial banks, insurance companies and pension funds, money market funds, mutual funds, foreign officials, foreign private investors, and other U.S. investors, we estimate reduced-form demand functions that capture how holdings respond to yields and macroeconomic conditions. For the Fed, we estimate a balance-sheet policy function separately, given its distinct policy mandate. Arbitrageurs, empirically associated with hedge funds and broker-dealers, are structurally modeled as risk-averse agents who intermediate across investor sectors and respond endogenously to demand imbalances. We classify broker-dealers and hedge funds as arbitrageurs for three reasons: they exhibit behavior opposite to yield-seeking investors (Hanson and Stein 2015; Du et al. 2023b); they have broad access to trading instruments enabling sophisticated arbitrage; and a reduced-form regression reveals they are the only sectors with holdings that decrease in own yield but increase in other-maturity yields, consistent with absorbing demand imbalances rather than responding to yields directly. We embed these demand estimates into a dynamic equilibrium model of the Treasury market.

Our analysis reveals two main findings. First, the Treasury market exhibits a steeply downward-sloping term structure of market elasticity, with much stronger arbitrage forces at work at the short end because of lower risks involved in arbitrage trades. In particular, arbitrageurs are essential for explaining the highly elastic T-bill market: a pure demand system absent a structural representation of arbitrageurs yields a T-bill price impact per unit demand shock more than 70 times larger, counterfactually implying that the Fed cannot tightly control the short rate. Second, term premia

rise in response to monetary policy tightening, consistent with the empirical evidence (Hanson and Stein 2015; Bauer et al. 2023). The key driver is strong cross-maturity substitution: when the short rate rises, granular-demand investors rebalance toward higher-yielding short-term Treasuries, forcing arbitrageurs to absorb more long-term supply and raising the risk premium. Absent such substitution, as in a baseline Vayanos and Vila (2021) model, the prediction reverses.

Building on insights in Koijen and Yogo (2026) and the instrument in Fang et al. (2025), we identify own and cross-maturity yield sensitivities at the investor level using our panel dataset. The instrument exploits a pseudo market-clearing condition to isolate the yield component attributable to supply-demand imbalances driven by bond characteristics and macro variables, and uses it as an instrument for the actual yield. The key identifying assumption is that latent demand shocks (which in turn affect yields) do not contemporaneously affect macro variables, a plausible restriction at quarterly frequency given that interest rate changes take one to two years to transmit to the real economy. The validity of the instrument rests on economic structure: the power-function mapping from exogenous variables to yields follows directly from the market-clearing condition, and replacing it with ad hoc nonlinear transformations such as square or log transformations leads to a severe weak instrument problem. We show that demand elasticity estimates are stable using an alternative instrument construction based on Treasury supply shocks. While we consider the Fed separately, its demand aligns with that of granular-demand investors, increasing long-term holdings when yields are high and reducing them during monetary tightenings.

We embed the estimated demand functions into an equilibrium model of the Treasury market with granular-demand investors, the Fed, and risk-averse arbitrageurs, extending Vayanos and Vila (2021) in three key respects. First, granular-demand investors feature cross-substitution, which generates a positive reaction of term premia to monetary policy tightening, in contrast to the negative reaction in Vayanos and Vila (2021). Kekre et al. (2024) also generates this positive reaction by introducing an arbitrageur wealth effect into Vayanos and Vila (2021); we do not incorporate such a channel as the resulting nonlinearity poses numerical challenges beyond the scope of this paper. Second, we include a monetary policy rule that depends on macroeconomic dynamics rather than treating the short-term interest rate as exogenous, allowing us to identify the magnitude of monetary policy shocks. Third, we incorporate latent outside assets held by arbitrageurs, adding realism by recognizing that prices of risk are not entirely driven by their Treasury portfolios, and we let the data inform us how outside-asset risk exposure interacts with Treasury pricing. Because the term-premium response to monetary policy depends on arbitrageurs' total priced rate exposure rather than on their Treasury holdings alone, identifying their outside-asset exposure is essential, and this identification in turn requires estimating the demand functions of non-arbitrageur investors.

To build intuition, we first analyze a simplified, analytically tractable version of our model.

Two key insights emerge. First, the one-period bond market is perfectly elastic because arbitrageurs can absorb supply shocks of this segment against the monetary policy rate at zero duration cost, while this intermediation becomes increasingly costly at longer maturities as duration risk grows, generating a downward-sloping term structure of elasticity. Second, whether the term premium rises or falls with monetary tightening depends on the strength of cross-maturity substitution in investor demand. Without cross-maturity substitution, a rate hike increases long-term yields through the expectations hypothesis and attracts granular-demand investors, reducing arbitrageur exposure and compressing the term premium. With sufficient cross-substitution, investors instead move toward the relatively more attractive short end, forcing arbitrageurs to absorb larger long-term supply and raising the risk premium. Whether or not the cross-substitution force dominates is a quantitative question answered by our demand estimation.

These insights guide our identification and estimation. We estimate the full model by jointly minimizing fitting errors in yield curve dynamics and arbitrageur Treasury holdings. The key insight that allows for sharp parameter identification is that arbitrageur risk aversion and outside-portfolio parameters are disciplined by distinct sources of variation: macro-yield covariation identifies the composite price of risk and recovers the outside-portfolio loadings; Treasury-specific latent demand shocks then move both yields and arbitrageur positions, identifying risk aversion separately from outside portfolio exposure. This separation ensures that outside portfolio risk exposure does not absorb the variation needed to discipline arbitrageur risk aversion.

Quantitatively, we reconcile seemingly contradictory findings in the literature: Greenwood et al. (2015b) find that T-bill supply shocks have very small yield effects, while Krishnamurthy and Vissing-Jorgensen (2011) and D'Amico and King (2013) find that QE purchases of long-term bonds substantially compress term premia. Our model explains both through the asymmetry in arbitrageur intermediation costs across maturities. In particular, the highly elastic T-bill market hinges on arbitrageurs. Without arbitrageurs, the T-bill price impact per unit demand shock rises by more than 70 times, a counterfactual inconsistent with the Fed's well-documented ability to tightly control the short rate. Beyond reconciling these findings, averaging across all maturities, the Treasury market is more elastic (i.e., smaller price impact of demand shocks) than the equity market (Gabaix and Koijen 2021) and the corporate bond market (Chaudhary et al. 2025), with a price multiplier of 0.31 compared to 5 and 3.5, respectively.[1]

Our second quantitative finding resolves the tension between the empirical finding that monetary tightening raises term premia (Hanson and Stein 2015; Bauer et al. 2023) and the prediction of standard preferred-habitat models. We trace this contrast to the absence of cross-maturity substitution. Starting from a pure habitat model without cross-maturity substitution, neither adding

[1]The price multiplier is the inverse of market elasticity. This implies that the Treasury market elasticity is 3.23.

an arbitrageur to it nor further enriching the arbitrageur model by giving it access to outside-asset exposure changes the sign of the term premium response. However, introducing the estimated cross-maturity substitution flips it from negative to positive. The intuition is that investors rebalance toward the short end and reduce long-term Treasury holdings when the short rate rises, forcing arbitrageurs to absorb more long-term supply and raising the term premium.

Related Literature

Our paper contributes to a growing literature that analyzes granular asset demand in fixed-income markets, building on the seminal work by Koijen and Yogo (2019). Specifically, Bretscher et al. (2025), Chaudhary et al. (2025), Siani (2025), and Darmouni et al. (2025) apply demand systems to corporate bond markets, Fang et al. (2025) to global government bond markets, Koijen et al. (2021) to the euro area government bond market, Jansen (2025) to the Dutch government bond market, and Jiang et al. (2024c) to international bond and currency markets. Allen et al. (2020) analyze the demand at T-bill auctions and find that auction format matters for portfolio allocations. Doerr et al. (2023) and Stein and Wallen (2025) provide a granular analysis of the demand by money-market funds for near-money assets. Closest to ours, Eren et al. (2026) and Chaudhary et al. (2024) apply a demand system to the overall U.S. Treasury market using Flows of Funds data. Consistent with their studies, we find that investment funds and banks are more price elastic than ICPFs and foreign officials within the U.S. Treasury market. We contribute to this literature by using more granular data on U.S. Treasury holdings by different institutions *across* maturity buckets, allowing us to estimate cross-elasticities. Our methodology can also serve as a building block for quantitative models of the market macrostructure of financial markets (Haddad and Muir 2025) that account for heterogeneity in investors' objectives, mandates, and constraints in shaping asset prices.

Furthermore, our paper is related to the preferred habitat view of the term structure of interest rates, e.g., Culbertson (1957), Modigliani and Sutch (1966), Guibaud et al. (2013), Greenwood and Vayanos (2014), and Vayanos and Vila (2021). Recent papers have started to build a tighter connection between data and theory. Droste et al. (2024) identify demand shocks from Treasury auctions and calibrate the model in a New Keynesian framework to study the impact of QE. Kaminska and Zinna (2020) estimate a structural term-structure model where arbitrageurs accommodate official-sector demand pressures, using data on both yields and official holdings. Hanson et al. (2024) quantify the demand and supply shocks in the interest-rate swap market. Khetan et al. (2023) leverage more detailed data on interest-rate swaps and find a high level of segmentation. Bahaj et al. (2023) utilize transaction-level data on UK inflation swaps to quantify a model of inflation risks. Greenwood et al. (2023) and Gourinchas et al. (2025) connect preferred-habitat demand with exchange rate dynamics. Our contribution is to build and estimate a quantitative

version of Vayanos and Vila (2021) that accounts for empirically estimated demand functions and actual arbitrageurs' Treasury holdings.

Our estimates of investor demand are consistent with the hypothesis of "yield-oriented investors" in Hanson and Stein (2015). We confirm, both theoretically and quantitatively, the rationale in Hanson and Stein (2015) that cross-substitution drives the positive term premium response to monetary policy tightening. This also addresses a broad literature that shows that risk premia overall rise with monetary policy tightening (Bernanke and Kuttner 2005; Gertler and Karadi 2015; Bekaert et al. 2013; Kekre et al. 2024).

Our paper is also related to the recent literature on the specialty of U.S. government debt. Krishnamurthy and Vissing-Jorgensen (2012) show that there is a downward-sloping aggregate demand curve for the convenience provided by Treasuries. The literature shows that Treasury convenience yield is closely connected to financial crises (Del Negro et al. 2017; Li 2024), monetary policy (Nagel 2016; Drechsler et al. 2018; Diamond and Van Tassel 2023), exchange rates (Jiang et al. 2021), inflation (Cieslak et al. 2024), pricing of stocks (Di Tella et al. 2025), hedging properties of Treasuries (Brunnermeier et al. 2024; Acharya and Laarits 2023), banking (Diamond 2020; Li et al. 2023; Krishnamurthy and Li 2023), financial regulation (Payne et al. 2025; Payne and Szőke 2024), and government debt valuation (Jiang et al. 2024a,b). We contribute to the above literature by unpacking the demand for Treasuries and sources of demand variation across investors.

Finally, arbitrageurs are critical to our analysis, in the same spirit as in a growing literature on financial intermediaries (He and Krishnamurthy 2013; Adrian et al. 2014; He et al. 2017; Du et al. 2018; Wallen 2020; Jermann 2020; Haddad and Muir 2021; Fang and Liu 2021; Kargar 2021; Favara et al. 2022; Siriwardane et al. 2025; Du et al. 2023a; Diamond et al. 2024; An and Huber 2024). Haddad and Sraer (2020) show that banks' interest income gap significantly predicts Treasury returns. d'Avernas and Vandeweyer (2024) and d'Avernas et al. (2023) provide theories of how different types of intermediaries together with the central bank affect Treasury market dynamics. Duffie et al. (2023) use dealer-level data on Treasury holdings to show that dealer balance sheet utilization is important for Treasury pricing. Du et al. (2023b) quantitatively show that balance sheet frictions of intermediaries are important in pricing Treasuries. A key contribution relative to this literature is that we cover the majority of the Treasury market beyond intermediaries and explicitly link the pricing kernel with intermediation activities.

2. Data

One of our contributions is the construction of a novel, granular dataset of U.S. Treasury holdings at the sector level, capturing the majority of the market. Indeed, our dataset covers all major institutional holders of U.S. Treasuries, including banks, the Federal Reserve, primary dealers and hedge funds, money market and mutual funds, ETFs, and both foreign official and private investors. We next describe these data sources, the construction of our dataset, and stylized facts about U.S. Treasury holders.

2.1. Treasury Holdings Data Sources

The Flow of Funds (FoFs) is the standard data source for extant research on investors in U.S. Treasuries (e.g., Krishnamurthy and Vissing-Jorgensen (2012), Eren et al. (2026), Chaudhary et al. (2024)). While the FoFs provides information about Treasury holdings at the investor sector level, the holdings are aggregated across all maturities, which limits the ability to conduct a more granular analysis of, for example, the term structure of interest rates or the cross-substitution across maturities. To address these limitations, we compile a richer and more detailed dataset by leveraging multiple data sources to obtain U.S. Treasury holdings with the highest level of granularity available. Table 1 summarizes our primary data sources, with further details provided in Appendix A.1.

Table 1. **Data sources**

This table provides a summary of the different data sources that we use in this paper.

Investor Type	**Data Source**	**Frequency**	**Period**	**Detail**
Banks	CALL Reports	Quarterly	1976Q1-2022Q4	Maturity bucket
Fed	Federal Reserve	Weekly	2003W1-2022W52	Security
Primary Dealers	Federal Reserve	Weekly	1998W5-2022W52	Maturity bucket
Hedge Funds	Form PF SEC	Quarterly	2011Q4-2022Q4	Aggregate
Insurers and Pension Funds	eMAXX	Quarterly	2010Q1-2022Q4	Security
Money Market Funds	IMoneyNet	Monthly	2011M8-2022M12	Security
	Flow of Funds	Quarterly	1993Q1-2022Q4	Aggregate
Mutual Funds	Morningstar	Monthly/Quarterly	2000M1-2022M12	Security
ETFs	ETF Global	Daily/Monthly	2012M1-2022M12	Security
Foreign Official and Private	Public TIC	Quarterly	2011Q4-2022Q4	T-bill/non T-bill

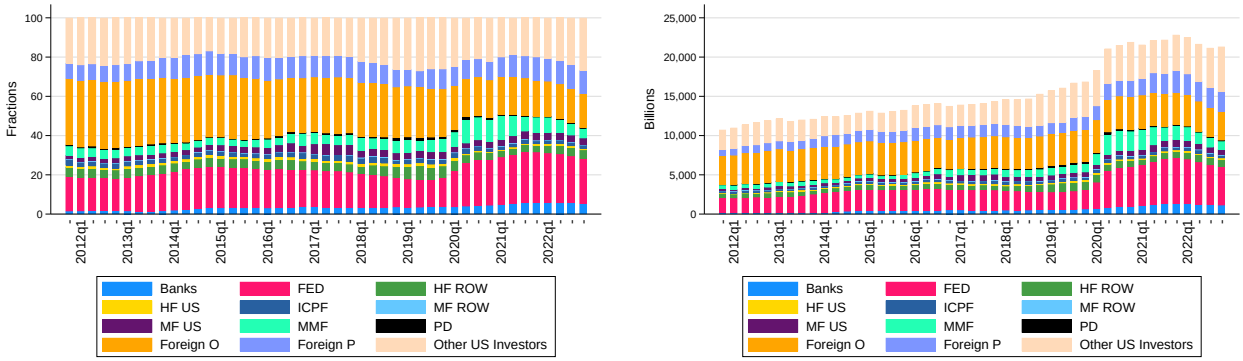
2.2. Data Aggregation

The reporting frequency and granularity differ across data sources. In constructing our final dataset, we therefore make aggregation choices to ensure consistency. Specifically, we analyze data at a quarterly frequency from 2011Q4 (the earliest available period for foreign investors and hedge funds) to 2022Q4. We then group Treasuries into three maturity buckets. Denoting remaining time to maturity as τ, the buckets are $\tau < 1Y$, $1Y \leq \tau < 5Y$, and $\tau \geq 5Y$, indexed by $m \in \{1,2,3\}$. The choice of these three maturity buckets is motivated by two considerations: First, this division reflects commonality across portfolio holdings data availability for different sectors. Second, as we show later, we need sufficient cross-sectional variation across maturity buckets to apply our instrument, and using more than three buckets complicates identification by reducing variation across buckets. To ensure stationarity, we scale all quantities by the ratio of potential GDP at the end of our sample period to potential GDP at that particular quarter. We provide details on the data aggregation process in Appendix A.2. In our analysis, we also incorporate macroeconomic dynamics, and we provide details on the macro variables in Appendix A.3.

2.3. Stylized Facts about Treasury Holdings

Figure 1. **Holdings of U.S. Treasuries by Investor Type**

Panel (a) plots the fraction of total U.S. Treasury outstanding (TAO) held by each investor type over time. Panel (b) plots the corresponding market values (billions). Sectors are U.S. banks (Banks), Federal Reserve (FED), hedge funds outside the U.S. (HF ROW), U.S. hedge funds (HF US), U.S. insurance companies and pension funds (ICPF), mutual funds outside the U.S. (MF ROW), U.S. mutual funds (MF US), U.S. money market funds (MMF US), U.S. and foreign primary dealers (PD), foreign official (Foreign O), foreign private (Foreign P), and other U.S. investors (Other U.S. Investors). Other U.S. Investors is defined as the total U.S. Treasuries outstanding minus the holdings of all the other sectors. We report market values, and the quarterly sample period is 2011Q4–2022Q4.



(a) % Total Holdings of TAO

(b) Total Holdings (billions)

Figure 1 shows the dollar values and the fraction of total outstanding of U.S. Treasuries held by each investor type from 2011Q4 to 2022Q4. On average, our dataset covers about 80% of the holdings of U.S. Treasuries. Based on FoF data, the remaining 20% consists of U.S. households (11%), pension funds (5%), local governments (4%), and non-financial corporations (2%).[2]

In Figure 2, we plot maturity-bucket level Treasury holdings of each investor type over the same period. The figure reveals several notable facts. First, MMFs are only active in maturity bucket 1 and hold between 10% and 35% of outstanding short-term Treasuries. Second, at the other end of the spectrum, ICPFs barely hold short-term Treasuries but hold around 3-5% of the Treasuries with maturities beyond 5 years. Third, the Fed holds substantially more of the intermediate and long-term bonds outstanding as opposed to short-term bonds. Fourth, mutual funds hold little in short-term bonds, but are equally spread between maturity buckets 2 and 3. Fifth, only primary dealers and hedge funds report negative holdings across maturity buckets (also see Appendix Table A2), which is a sign of fixed-income arbitrage. Finally, foreign official holdings have significantly declined, mainly in the short and medium-maturity buckets.

Appendix Table A1 further examines which investor types are marginal in trading U.S. Treasuries when supply changes. A key finding is that broker-dealers and hedge funds show disproportionately high trading activity relative to their average holdings across all maturity buckets, a characteristic indicative of arbitrage.

3. Empirical Results

Our data reveal substantial heterogeneity in Treasury holdings across sectors, reflecting two distinct investor types. We refer to the first type as *granular-demand investors*: banks, insurance companies, and pension funds (ICPFs), mutual funds, money market funds (MMFs), foreign official investors, and foreign private investors. These sectors' Treasury demand reflects significant non-pecuniary reasons such as liquidity regulation, liability matching, and reserve management. As we show in Appendix E.1, a mean-variance portfolio approach to Treasury demand, with such non-pecuniary attributes combined with expected returns that depend on yields, justifies a linear demand function of the form:

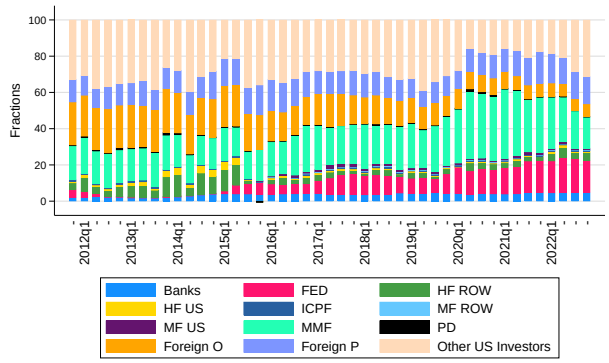
$$Z_t^l = \theta_0^l + B^l y_t - \theta^l \beta_t + u_t^l, \tag{1}$$

where y_t is the vector of Treasury yields and β_t is a vector of macroeconomic states. B^l is a full matrix, so demand for each maturity responds to yields of other maturities as well. This cross-maturity substitution follows when investors' expectations of returns depend on yield levels

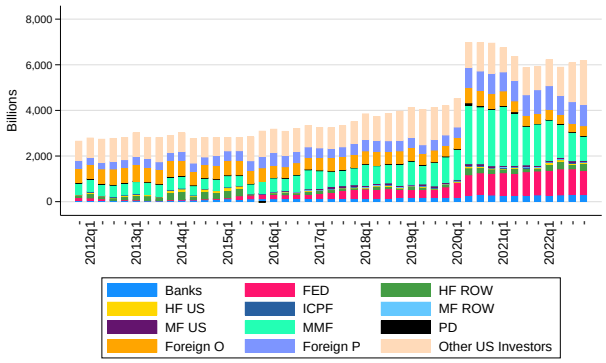
[2]These shares are computed from the Federal Reserve's Financial Accounts of the United States (Z.1 release), Table L.210 (Treasury Securities), averaged over our sample period 2011Q4–2022Q4.

Figure 2. **Holdings of U.S. Treasuries by Maturity Bucket**

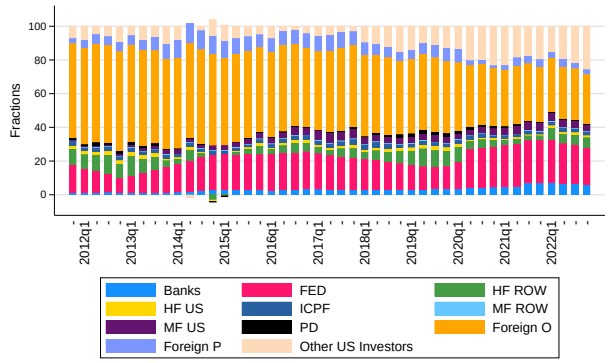
Left panels display the fraction of total U.S. Treasury outstanding (TAO) held by each investor type by maturity buckets. Right panels plot the corresponding market values (billions). Sectors are U.S. banks (Banks), Federal Reserve (FED), hedge funds outside the U.S. (HF ROW), U.S. hedge funds (HF US), U.S. insurance companies and pension funds (ICPF), mutual funds outside the U.S. (MF ROW), U.S. mutual funds (MF US), U.S. money market funds (MMF US), U.S. and foreign primary dealers (PD), foreign official (Foreign O), foreign private (Foreign P), and other U.S. investors (Other U.S. Investors). Other U.S. Investors is defined as the total U.S. Treasuries outstanding minus the holdings of all the other sectors. We report market values, and the quarterly sample period is 2011Q4–2022Q4.



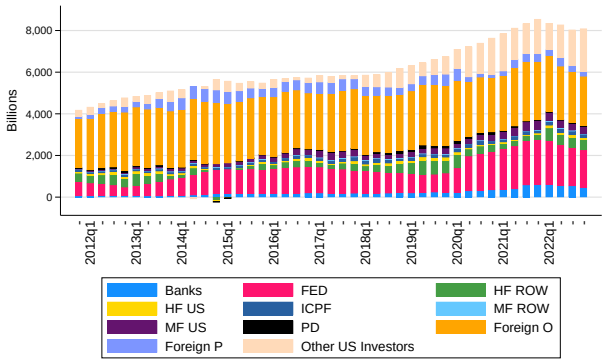
(a) $\tau < 1Y$: % Holdings as of TAO



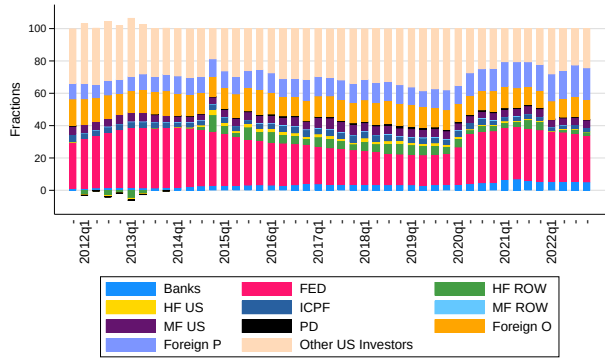
(b) $\tau < 1Y$: Total Holdings (billions)



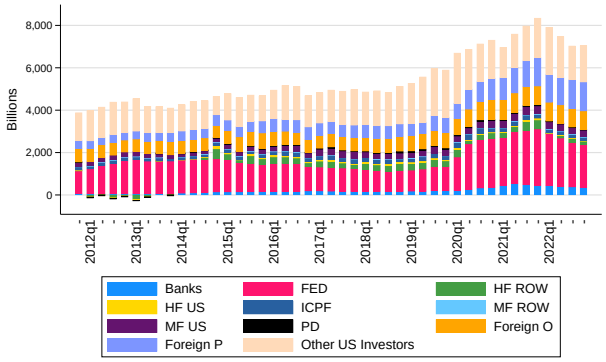
(c) $1 \leq \tau < 5Y$: % Holdings of TAO



(d) $1 \leq \tau < 5Y$: Total Holdings (billions)



(e) $\tau \geq 5Y$: % Holdings of TAO



(f) $\tau \geq 5Y$: Total Holdings (billions)

(e.g., reaching-for-yield or yield-curve extrapolation), and is validated in the data in Section 3. In Appendix E.1, we derive these demand functions formally and discuss their microfoundations. The Fed's Treasury demand is estimated separately using a similar approach, reflecting its distinct policy objectives; together with granular-demand investors, it constitutes the non-arbitrageur side of the market.

We refer to the second type as *arbitrageurs*, following Hanson and Stein (2015) and Du et al. (2023b): broker-dealers and hedge funds do not exhibit demand functions in the same sense. They condition their portfolio on the underlying state of the Treasury market, including macroeconomic dynamics and demand imbalances, rather than on yields directly. Applying a reduced-form demand regression to arbitrageurs therefore produces misleading results. Instead, we structurally model arbitrageurs and estimate their risk-bearing capacity in the equilibrium model of Section 4.

3.1. Demand System Specification

We estimate granular-demand investor ι's demand for U.S. Treasuries according to (1). In practice, we make two slight modifications. First, we group Treasuries into three maturity buckets, consistent with the empirical aggregation of Treasury holdings, and denote a maturity bucket as $m \in \{1,2,3\}$. Second, we add bond characteristics to the demand specification as control variables, although those bond characteristics will not be directly modeled in Section 4. In particular, we implement the following regression:

$$Z_t^\iota(m) = \theta_0^\iota + b_1^\iota y_t(m) + b_2^\iota y_t(-m) + (b_3^\iota)' \mathbf{x}_t(m) + (b_4^\iota)' \mathbf{Macro}_t + u_t^\iota(m), \tag{2}$$

where $y_t(m)$ is the yield for maturity bucket m and $y_t(-m)$ denotes the weighted-average yield of the other maturity buckets. The vector $\mathbf{x}_t(m)$ is a vector of value-weighted bond characteristics for maturity bucket m: coupon, maturity bucket fixed effects, and bid-ask spread. The vector $\mathbf{Macro}_t$ denotes a set of macro variables, including GDP gap, debt/GDP, core inflation, and credit spread. We residualize the coupon and the bid-ask spread with respect to the maturity fixed effects to address multicollinearity issues and ensure that maturity preferences are not confounded with either of these two characteristics. We provide summary statistics for this set of variables in Table A9 and the correlation table in Table A10.

We focus on the dollar value of holdings rather than portfolio weights, because dynamics in total portfolio demand are crucial for the term structure of interest rates; modeling only portfolio weights is not sufficient. For example, inflows into money-market mutual funds will cause extra demand for short-maturity Treasuries, yet their below-one-year Treasury portfolio weight remains at 100%, failing to capture such fluctuations. Moreover, we use market values rather than face

values because our model in Section 4 indicates that market values are the relevant signals for investors, so our specification in (2) has a direct mapping to our dynamic quantitative model.

Unlike Koijen and Yogo (2019), but following our model in Section 4.2, we include "other yield" to capture cross substitution across the maturity structure. We find that excluding other yield from the estimation leads to a downward bias in the coefficient on own yield. The reason is that own yield and other yield are correlated, while demand increases when own yield rises but decreases when other yield rises. Hence, when not accounting for other yield, b_1^l absorbs both the positive and negative effects, leading to a coefficient that is biased toward zero.[3]

In our specification, we assume that the macro variables are contemporaneously exogenous to investors, as, for instance, in Fang et al. (2025). Precisely, our identifying assumption is not that all demand innovations are idiosyncratic, but that the *residual* demand variation u_t^l unexplained by the included macro controls does not contemporaneously affect those same macro variables. The macro controls already absorb common co-movement between demand and the macroeconomy at a quarterly frequency. What remains in u_t^l reflects sector-specific shocks, such as changes in regulation, institutional mandates, or fund flows, that are plausibly orthogonal to contemporaneous macro outcomes. Moreover, even if Treasury demand were to feed back into macro aggregates, the transmission works through yields and then through the real economy, a process that takes many quarters. Romer and Romer (2004) find that a monetary policy shock takes roughly 22 to 27 months for output to reach its peak response and nearly two years for inflation to decline significantly; Christiano et al. (1999) document similarly slow output and price dynamics across VAR specifications. These lags far exceed one quarter, making contemporaneous feedback from sectoral Treasury demand to GDP or inflation implausible at our data frequency. We also assume that bond characteristics such as the pre-determined coupon rate are exogenous to latent demand. We maintain these assumptions throughout our analysis.

Under these assumptions, bond characteristics and macro variables are valid controls, and we can estimate the demand system specified in equation (2) by GMM if in addition it satisfies the moment condition:

$$\mathbb{E}[u_t^l(m)|y_t(m), y_t(-m), \mathbf{x}_t(m), \mathbf{Macro}_t] = 0. \tag{3}$$

The remaining concern is that the error term may not be orthogonal to *yields*. If certain sectors have a large latent demand for Treasuries, this demand is likely to also suppress the yield. As such, we need an instrument for bond yields.

[3]Table A17 shows that the coefficient on own yield is attenuated closer to zero when not accounting for other yield.

3.2. Instrument

We propose an instrument designed to isolate the yield component attributable to supply-demand imbalances driven by bond characteristics and macro variables, rather than relying on an instrument based on investment mandates in the spirit of Koijen and Yogo (2019). The reason is that, within the Treasury market, most institutional investors are not subject to investment mandates, except for MMFs, which have a clear mandate to operate only in the short-maturity bucket. Even seemingly long-term investors, such as pension funds, invest across the entire yield curve. Consistent with this, Jansen (2025) shows that for safe European government bond markets, only 53% of the government bonds held by euro area investors were also in their portfolios in the previous quarter, compared to over 90% for equity markets (Koijen and Yogo 2019).

Construction. We build on insights in Koijen and Yogo (2026) and the instrument used in Fang et al. (2025) and use the following three step procedure.[4]

First, we estimate demand for each granular-demand investor type ι as in equation (2), but excluding the yield. We exclude arbitrageurs (broker-dealers and hedge funds) from this step, because, as discussed above, their holdings are not well described by a reduced-form demand function of the type in equation (2). To preserve market clearing in the pseudo equilibrium, we scale Treasury supply proportionally by the share of the market accounted for by non-arbitrageurs (granular demand investors and the Fed), so that the excluded arbitrageur holdings do not create a mechanical imbalance.

We then extract, in a second step, the predicted values $\hat{Z}_t^{\iota}(m)$. We also follow steps (1) and (2) for the nominal value of Treasury supply at each maturity bucket, whereby we regress it on the bond characteristics, the interest on reserves (IOR), and macro variables, consistent with the specification of our US Treasury model introduced in Section 4. We use IOR rather than the federal funds rate because IOR is set administratively by the Federal Reserve, alleviating the endogeneity concern that the federal funds rate responds endogenously to Treasury market conditions.

In a third step, we impose market clearing and extract the imposed yield that sets the implied demand equal to the implied market value of supply:

$$\sum_{\iota} \hat{Z}_t^{\iota}(m) = \frac{\hat{S}_t(m)}{(1+\tilde{y}_t(m))^{\tau(m)}}, \tag{4}$$

where $\hat{S}_t(m)$ is the predicted nominal value of supply for maturity bucket m, and $\tau(m)$ the cor-

[4]Fang et al. (2025) impose market clearing at the country level by regressing investor demand and supply in a given country on the macroeconomic variables of that country. Rather than taking a cross-country approach, we adapt their methodology across the maturity spectrum by regressing holdings in each maturity bucket m on bond characteristics and macroeconomic variables in maturity bucket m.

responding maturity. We take $\tau(m)$ as the average bond duration for maturity bucket m. We then extract the pseudo yield $\tilde{y}_t(m)$ that clears the market at each point in time t and use it as an instrument for the actual yield $y_t(m)$:

$$\tilde{y}_t(m) = \left(\frac{\hat{S}_t(m)}{\sum_\iota \hat{Z}_t^\iota(m)} \right)^{\frac{1}{\tau(m)}} - 1. \tag{5}$$

We apply the same logic to the value-weighted yield of the other buckets, for which the instrument is given by the value-weighted pseudo yield for the other maturity buckets: $\tilde{y}_t(-m)$.

Economic Interpretation. The pseudo yield in (5) generates nonlinear variation between yields and bond and macro factors, providing identifying power beyond the linear demand functions we estimate. The nonlinear IV literature (Kelejian 1971) addresses this type of problem using a Taylor-series approximation to the unknown conditional expectation. Newey (1990) shows that the efficiency-achieving instrument is exactly this conditional expectation, and that it must be estimated nonparametrically when economic structure does not pin down the functional form. Our setting is the opposite: the market-clearing condition uniquely determines the power-function transformation $\tilde{y}_t = (\hat{Z}_t/S)^{-1/\tau} - 1$, so there is no researcher choice over functional form. This uniqueness is testable: if instrument power came from generic functional-form flexibility, alternative nonlinear transformations such as X^2 or $\log(|X|)$ should work equally well. They do not, yielding Kleibergen–Paap F-statistics of 0.6 and 0.45, respectively, confirming that identifying power comes from the economic structure rather than from nonlinearity per se.

To visually confirm the strength of our identification strategy, we examine the correlation between the endogenous yield and the non-linear instrument after partialing out the linear control variables. Figure 3 shows the resulting scatter plot. The robust positive relationship between the instrument and the residualized yields confirms that the non-linear transformation captures significant yield variation orthogonal to the controls.

Instrument Relevance. Empirically, this instrument is relevant. The first stage estimates of the demand system are summarized in Table A4. The corresponding Kleibergen–Paap (KP) F-statistic[5] is 10.48. Because our setting features two endogenous variables, HAC errors, and generated instruments, the standard Stock and Yogo (2005) critical value of 10 does not formally apply. In Appendix C.2, we derive simulation-based critical values calibrated to our exact data-generating process: the size-controlled threshold (90th percentile of the KP distribution at the 10% Nagar bias level) is 7.28, which our baseline KP of 10.48 exceeds.

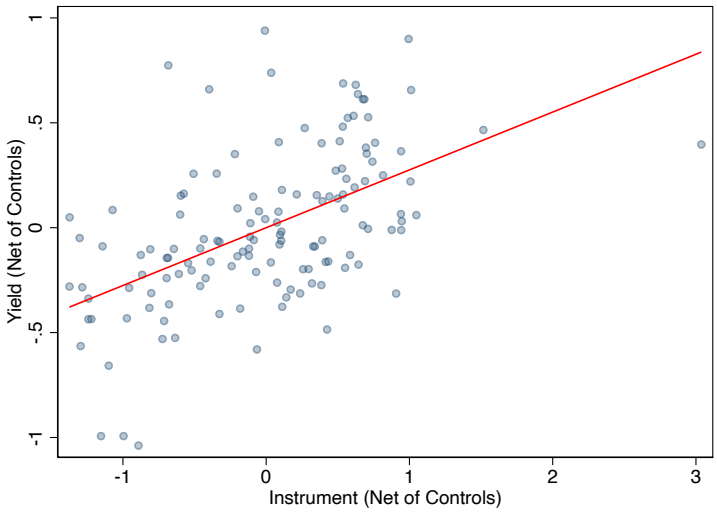
Instrument Validity. A potential concern for identification is if investor demand itself is highly nonlinear in bond characteristics and macro variables.[6] Appendix C tests this concern directly by

[5]The first stage is the same for all sectors, except MMFs, for which the statistic equals 25.24.

[6]In this instance, the coefficient on the pseudo yield would suffer from the classical omitted variable bias, because

Figure 3. **First-Stage Correlation: Residualized Yields and Pseudo Yields (Instrument)**

This scatter plot shows the relationship between the endogenous yield and the non-linear instrument, where both variables have been orthogonalized against the full set of exogenous controls, including maturity bucket indicators, coupon rate, bid-ask spread, IOR, inflation, debt-to-GDP, GDP gap, and credit spread. The robust positive slope confirms that the instrument provides significant explanatory power beyond the linear controls.



re-estimating the IV demand system with squared macroeconomic terms and squared bond characteristics added as second-stage controls, while keeping the pseudo-yield instruments unchanged. The own-yield and cross-yield coefficients are essentially unchanged across all eight sectors, and the KP F-statistic is also stable (10.48 in the baseline versus 10.23 with the squared controls added). Any quadratic effects in demand are therefore too small to materially affect the baseline IV; the instrument's identifying power comes from the nonlinear transformation in its construction, not from any nonlinearity in the demand functions. As a further validity check, Section 3.3 shows that an alternative instrument based on supply shocks yields broadly similar demand estimates.

Appendix C also presents a stylized example that illustrates the source of identifying variation. The key intuition is that demand and supply respond asymmetrically to macro variables, so pseudo market clearing generates yield variation that is driven by fundamentals rather than by latent demand shocks. For instance, foreign investors reduce their demand for short- and medium-term Treasuries when inflation is high, while supply does not materially respond to inflation (see Table A11). High-inflation periods therefore coincide with high pseudo yields, providing instrument variation that is distinct from idiosyncratic demand fluctuations.

In summary, the idea behind the instrument is that the pseudo yield isolates the component of the yield that is driven by supply and demand imbalances in bond characteristics and macro

the pseudo yields and bond characteristics and macro variables are correlated following equation (4).

variables, which are exogenous to latent demand shocks. This instrument satisfies the exclusion restriction under the identifying assumption that bond characteristics and macro variables are exogenous to investor latent demand, and that linear demand functions capture investor demand well. With the instrument, we can weaken moment condition (3) to:

$$\mathbb{E}[u_t^l(m)|\tilde{y}_t(m),\tilde{y}_t(-m),\mathbf{x}_t(m),\mathbf{Macro}_t]=0. \tag{6}$$

3.3. Demand Functions of Granular-Demand Investors

Table 2. **Demand System Results - IV**

This table reports IV estimates of the demand system in equation (2). The dependent variable is the market value ($bn) of U.S. Treasuries held by sector ι in maturity bucket m, scaled by the ratio of end-of-sample to current-quarter GDP potential; for sector abbreviations, refer to Figure 1. Own yield $y_t(m)$ and other yield $y_t(-m)$ are instrumented with pseudo yields from Section 3.1. Bond Controls include coupon rate and bid-ask spread (both orthogonalized with respect to maturity bucket indicators) and maturity bucket FE; excluded for MMFs, which hold only short-term Treasuries. Macro Controls include GDP gap, debt/GDP, core inflation, and credit spread. The full table showing all control coefficients is in Appendix Table A12. Sample: 2011Q4–2022Q4. KP: Kleibergen–Paap first-stage F-statistic. HAC standard errors with optimal lags are in brackets; $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	Banks	**ICPF**	**MF ROW**	**MF U.S.**	**MMF**	**Other U.S.**	**Foreign O**	**Foreign P**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$y_t(m)$	55.815**	-6.939	6.330**	122.951***	275.646**	181.738	-122.170	56.168
	[25.022]	[11.387]	[3.008]	[40.845]	[116.979]	[195.228]	[110.004]	[109.121]
$y_t(-m)$	-50.822*	7.000	-2.309	-123.275***	-302.708**	-117.855	-42.737	-69.378
	[28.109]	[13.286]	[3.103]	[41.929]	[149.146]	[238.685]	[141.245]	[136.417]
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Maturity Bucket FE	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	135	135	135	135	45	135	135	135
KP F-Statistic (first stage)	10.484	10.484	10.484	10.484	25.241	10.484	10.484	10.484

Table 2 shows the results using the IV methodology outlined in the previous section.[7] We find that all investors have downward-sloping demand curves, except for the foreign official and ICPF sectors, whose coefficients on both own and other yield are insignificant, which we interpret as indicating that both sectors are insensitive to yields.[8] For the remaining sectors, granular-demand investors demand more U.S. Treasuries of maturity bucket m when the yield (price) is high (low). In addition, investors load negatively on the yield of other maturity buckets, meaning that their

[7] Appendix Table A12 reports the full table showing all coefficients and the results of the OLS estimates are in Appendix Table A14.

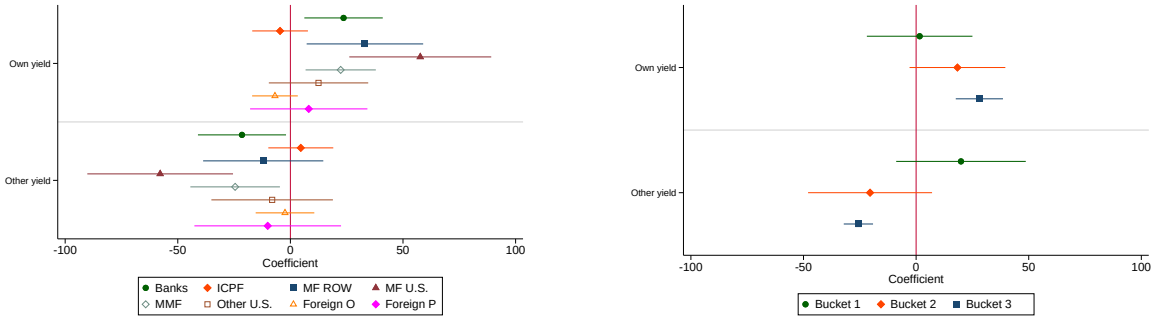
[8] For ICPFs, a negative coefficient might be explained by the "hunt-for-duration" behavior (Domanski et al. 2017): when interest rates decline, ICPFs often increase their demand for long-dated assets to hedge the increasing value of their long-term liabilities.

demand for maturity bucket m decreases when the yields of other buckets rise. Generally, we find that other elasticity is similar in magnitude to own elasticity. This is consistent with the findings in Chaudhary et al. (2025). They find a ratio between cross-elasticity and own-elasticity of close to 1 at the CUSIP level and for portfolios at the rating $\times$ quarter-to-maturity level for corporate bonds, the latter aggregation closely resembling ours. This ratio implies that own and cross-elasticities have the same magnitude but opposite signs.

As we show in Section 6.3, this cross-maturity substitution is the key ingredient that generates a positive term premium response to monetary tightening: when the short rate rises, granular investors rebalance toward the short end, forcing arbitrageurs to absorb more long-term supply and raising the risk premium.

Figure 4. **Yield Elasticities by Investor Type**

Panel (a) plots the coefficients on own and other yields for different granular-demand investors, scaling holdings for each sector by the average holding across time and maturity buckets for that sector to allow for comparison of coefficients across investor types. A coefficient of 50 implies that for a one percentage point increase in yield, the demand goes up by 50%. For explanations of sector abbreviations, refer to the notes of Appendix Table A1. Panel (b) shows the yield sensitivities for the Federal Reserve by maturity bucket, whereby we scale the holdings in each bucket by the time-series average holding in that bucket. We use market values scaled by GDP potential and the quarterly sample period is 2011Q4–2022Q4.



(a) Granular-Demand Investors

(b) Federal Reserve

Since Table 2 reports holdings in market values, it does not allow for direct comparison of price elasticities across investor types. Therefore, we scale the holdings for each sector by the average holding of that sector, across buckets and time. Figure 4a plots the coefficients on own and other yield for each investor type. Interestingly, mutual funds and banks appear to be the most price elastic, followed by MMFs.[9] ICPFs and foreign official investors are the least price elastic. We will present a demand analysis for the Fed and discuss panel (b) in the next subsection.

The heterogeneous cross elasticities across sectors revealed by Table 2 and Figure 4 are consistent with various mechanisms in the literature (see Appendix D.2 for sector-by-sector details).

[9]Eren et al. (2026) also find that banks and investment funds are more price elastic.

Banks show strong negative cross elasticity, reallocating toward higher-yield maturities in line with reaching-for-yield behavior (Hanson and Stein 2015), and a strong substitution across Treasuries due to liquidity regulation rules (e.g., all Treasuries are high quality liquid assets). In contrast, ICPFs display price-insensitive demand, reflecting preferred-habitat behavior (Vayanos and Vila 2021). Mutual funds, particularly U.S. funds, respond elastically to relative yields. This behavior reflects both active management and potentially behavioral factors, such as confusion between short and long rates (Shue et al. 2024), and return-chasing retail flows in bond mutual funds (Hanson et al. 2021).

MMFs, while restricted to short maturities by law, exhibit negative cross elasticity via investor flows: as longer-term yields rise, money flows out of MMFs, reducing their T-bill demand. These flow dynamics potentially reflect extrapolative beliefs (Barberis et al. 2015). On the other hand, "Other U.S. investors," primarily households that directly hold Treasuries and corporations, show muted cross-substitution. Unlike investors holding Treasuries through mutual funds or MMFs, households directly holding U.S. Treasuries benefit from tax advantages that accrue over time, and they do not continuously observe market-value fluctuations, reducing their propensity for frequent portfolio adjustments to prices.

Finally, foreign official investors exhibit strongly segmented demand, favoring shorter maturities due to safety, liquidity needs, and reserve management guidelines. Their behavior is price-inelastic and consistent with the global savings glut narrative (Bernanke 2005). In contrast, foreign private investors are more yield-sensitive and may substitute across maturities or borders to reach for yield, akin to their domestic counterparts.

Overall, despite this heterogeneity in mechanisms, negative cross-elasticities are pervasive across nearly all sectors. This broad pattern matters in the aggregate: regardless of whether cross-substitution stems from reaching-for-yield, regulatory liquidity rules, or investor flows, the aggregate effect is that investors rebalance toward higher-yielding maturities, a feature that is crucial for our structural model to generate a positive term premium response to monetary tightening.

We conduct several stability checks for the instrument and the resulting demand estimates, with results in Appendix D.3. These checks complement the specification test for linear demand in macro variables and bond characteristics in Appendix C (Table A3) by showing that the elasticity estimates do not hinge on which macro variables enter the pseudo-yield construction. First, Table A15 re-estimates the demand system using a pseudo yield built from a restricted set of variables, excluding bid-ask spread, credit spread, and core inflation. Second, Table A16 augments the macro set with the MBS spread (the 10-year MBS rate minus the 10-year Treasury yield) and the swap spread (the 10-year swap rate minus the 10-year Treasury yield); the estimates remain qualitatively and quantitatively robust, ruling out confounding effects from substitution toward MBS or swaps.

Supply-Shock Instrument. As an additional robustness check, Appendix C.4 reports results using an instrument based on legislated emergency spending (for the short bucket) and high-frequency Treasury auction supply shocks (for the medium and long buckets). This instrument is constructed entirely without reference to bond characteristics or macro variables. Although the first stage is weaker and we therefore retain the pseudo-yield as our primary instrument, the estimated demand patterns remain broadly similar across sectors, with own-yield coefficients positive and cross-yield coefficients negative, consistent with the baseline.

3.4. Demand Functions of the Fed

For the Fed, we estimate its demand curves separately for each maturity bucket. The reason is that the Fed implements unconventional monetary policies mainly via long-term Treasuries. We should, therefore, expect the Fed to respond to yields for its long-term Treasury holdings, but not for its short- and medium-term Treasury holdings. In contrast, we do not have a strong prior that granular-demand investors have significantly different responses to yields across maturities.

Table 3. **Demand System Results - Fed**

This table reports IV estimates of the demand system in equation (2), estimated separately for each maturity bucket. The dependent variable is the market value ($bn) of U.S. Treasuries held by the Fed in maturity bucket m, scaled by the ratio of end-of-sample to current-quarter GDP potential. Own yield $y_t(m)$ and other yield $y_t(-m)$ are instrumented with pseudo yields from Section 3.1. Bond Controls include coupon rate and bid-ask spread. Macro Controls include GDP gap, debt/GDP, core inflation, and credit spread. The full table showing all control coefficients is in Appendix Table A13. Sample: 2011Q4–2022Q4. KP: Kleibergen–Paap first-stage F-statistic. HAC standard errors in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	$\tau < 1Y$	$1Y \leq \tau < 5Y$	$\tau \geq 5Y$
	(1)	(2)	(3)
$y_t(m)$	7.308	280.506	564.069***
	[66.386]	[197.647]	[127.987]
$y_t(-m)$	92.827	-312.573	-514.711***
	[81.503]	[255.812]	[79.403]
Bond Controls	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes
Observations	45	45	45
KP F-Statistic (first stage)	25.241	5.214	14.360

Table 3 presents the results. Notably, in the long-term maturity bucket, the Fed behaves similarly to granular-demand investors by increasing its long-term Treasury holdings when long-term yields are elevated. Controlling for macroeconomic conditions and bond-specific features, this pattern suggests that the Fed expands its holdings in response to yield movements that are not

driven by macro or bond fundamentals. This behavior aligns with qualitative evidence from Fed communications and QE episodes. For instance, in a March 2013 speech, Fed Chair Ben Bernanke (Bernanke 2013) referenced a "growing body of research" showing that large-scale asset purchases reduce term premia and thus lower long-term interest rates. These premia are largely shaped by financial market conditions rather than macroeconomic fundamentals. The Fed has also deployed its balance sheet to offset increases in long-term yields perceived as technical, such as during the March 2020 Treasury market turmoil. Supporting this view, Haddad et al. (2024) incorporate a "QE rule" explicitly responsive to yield levels. Theoretically, Caballero et al. (2024) show that such yield-based interventions (termed "financial conditions targeting") can be optimal, even if financial conditions are not direct policy objectives.

Moreover, Table 3 also indicates significant "cross elasticity" for the Fed in the long-maturity bucket. This is consistent with the practice of pairing rate hikes with reductions in long-term asset holdings for policy consistency. The Fed's own documentation explicitly frames a "dual tightening" process. The 2014 and 2017 FOMC statements (Federal Reserve 2018) made clear that short-term rate increases would come first, and balance sheet runoff would follow. In October 2018, Simon Potter, Executive Vice President of the Federal Reserve Bank of New York, observed that "the FOMC had increased the federal funds target rate from 0% to 2–2.25%" and "the FOMC has reduced the size of the portfolio from nearly $4.3 trillion to about $4.0 trillion" (Potter 2018). Financial markets well understood these policies. For example, Bloomberg News (2022) highlighted that the Fed was signaling a balance sheet reduction to begin shortly after the first rate increase, quoting Fed communications that this combination would strengthen the impact on financial conditions.

Despite the significant price elasticity in long-term Treasury holdings, the Fed's medium- and short-term Treasury holdings are not responsive to Treasury yields, consistent with the focus of QE/QT on long-term securities (Appendix Table A13 reports the full table showing all control coefficients).

To assess whether our main results are affected by the zero lower bound (ZLB) or inflation expectations, both prominent drivers of QE policy in the macroeconomic literature, we re-estimate the Fed regressions with additional controls. Specifically, we include the 5-year inflation swap rate as a proxy for inflation expectations and the spread between the federal funds rate (FFR) and the shadow rate from Wu and Xia (2016) to account for the ZLB. In untabulated regressions, the coefficient on own yield decreases from 564.1 to 522.9, and the coefficient on other yield shifts from -514.7 to -483.3, with both remaining statistically significant. These results suggest that the Fed's estimated demand elasticities are robust to concerns about confounding effects from the ZLB or inflation expectations.

Figure 4b shows the relative yield sensitivities of the Fed across maturity buckets. Clearly,

the Fed's short-term Treasury holdings are price inelastic, while its long-term Treasury holdings exhibit significant price elasticity, comparable in magnitude to that of banks.

3.5. Distinct Demand of Arbitrageurs

Arbitrageurs do not have a "demand function" in the same way that granular-demand investors do. Granular-demand investors respond directly to yields, which reflects institutional frictions (such as mandates or capital regulation) or heuristic beliefs (such as reaching for yield). In contrast, arbitrageurs are different: they are rational, forward-looking agents who price Treasuries based on expectations of macroeconomic conditions, interest rate dynamics, and future demand-supply imbalances. More fundamentally, demand depends on yields if and only if conditional expected returns depend on yields. For non-arbitrageur sectors, yields are the visible signal absent more sophisticated information, so expected returns load on yields directly and the regression in equation (1) is the mean-variance solution. Arbitrageurs condition on the underlying state, including macroeconomic dynamics and demand imbalances, so their expected returns do not load on yields directly and equation (1) cannot represent their portfolio rule.

The data confirm this directly. Appendix Table A8 runs the demand regression of equation (2) for arbitrageurs. Their yield loadings have the opposite sign from those of granular-demand investors: holdings fall when own yields increase but rise when other-maturity yields increase. This sign reversal is in line with the forces dictated by market-clearing, rather than a direct response to yields in the sense of a demand function. When a maturity bucket offers higher yields, granular investors demand more of it, reducing the residual supply that arbitrageurs must absorb. Du et al. (2023b) document the same pattern: higher term spreads are associated with lower arbitrageur holdings, the opposite of yield-seeking behavior. Appendix Table A2 shows that arbitrageurs are by far the most frequent short sellers, with shorting concentrated at longer maturities, consistent with absorbing supply imbalances at the long end.

We therefore model arbitrageurs structurally as risk-averse investors who solve a portfolio problem, and estimate their risk aversion and outside-portfolio exposure from the joint price and quantity responses of the Treasury market to demand shocks (Section 4).

4. An Equilibrium Model of the Treasury Market

The previous section highlighted three key findings. First, granular-demand investors and the Fed exhibit downward-sloping demand curves. Second, their demand displays significant cross-maturity substitution. Third, arbitrageurs absorb supply-demand imbalances and systematically

take short positions in Treasuries, reflecting their active arbitrage activities distinct from the broader market.

Building on these empirical results, we now develop a model where arbitrageurs interact with granular-demand investors and the Fed in the Treasury market, in the spirit of Vayanos and Vila (2021). We explicitly model the optimization problem of arbitrageurs but use demand functions to capture granular-demand investors and the Fed. After we set up the model, we provide a simplified version that allows us to derive analytical results to obtain intuition regarding the fundamental mechanisms. Finally, we estimate the full model using our dataset.

To capture the rich economics in the Treasury market, we deviate from Vayanos and Vila (2021) mainly in three respects. First, we incorporate cross-substitution in investor demand, a critical feature that generates realistic term premium responses to monetary policy shocks. Second, we include a monetary-policy rule that depends on macroeconomic dynamics, rather than treating the short-term interest rate as exogenous, allowing us to identify the magnitude of monetary policy shocks. Third, we account for latent outside assets held by arbitrageurs, adding realism by recognizing that arbitrageurs also hold outside assets, so the price of risk is not entirely driven by their Treasury portfolios.

4.1. Model Setup

The model is discrete-time and infinite-horizon. There are four types of agents in the economy: a competitive arbitrageur sector, the Fed, a set of granular-demand investors, and the government. We explicitly model only the optimal decisions of arbitrageurs while we capture the behavior of other agents through policy rules that directly correspond to our estimated demand functions. Model dynamics are driven by macroeconomic shocks, monetary policy shocks, and demand shocks.

Consider zero-coupon bonds of maturities $\tau \in \{1,2,\cdots,N\}$ that all pay a face value of 1 at maturity. Denote by $P_t^{(\tau)}$ and $y_t^{(\tau)}$, respectively, the time-t price and yield of the bond with maturity τ. We use "prime" to denote the transpose of vectors and matrices, and all vectors are column vectors. Define the log price vector as

$$p_t = \left(\log(P_t^{(1)}), \log(P_t^{(2)}), \cdots, \log(P_t^{(N)})\right)'. \tag{7}$$

For simplicity, we denote the yield of a one-period bond as r_t, defined as $r_t = -\log(P_t^{(1)})$.

We consider r_t as directly controlled by monetary policy. All other bond yields and prices are endogenously determined in equilibrium. Denote the total return from holding a Treasury of

maturity τ as

$$R_{t+1}^{(\tau)} = \frac{P_{t+1}^{(\tau-1)} - P_t^{(\tau)}}{P_t^{(\tau)}}. \tag{8}$$

Accordingly, the total return of a one-period Treasury is $R_{t+1} \equiv R_{t+1}^{(1)} = (1 - P_t^{(1)})/P_t^{(1)} = \exp(r_t) - 1 \approx r_t$.

The dynamics of the economy are driven by a K-dimensional vector of macro factors,

$$\beta_t = (\beta_{1,t}, \beta_{2,t}, \cdots, \beta_{K,t})', \tag{9}$$

which follows a VAR(1) process,

$$\beta_{t+1} = \bar{\beta} + \Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}. \tag{10}$$

In the above expression, ε_{t+1} is a K-dimensional vector that follows an i.i.d. standard normal distribution, $\bar{\beta}$ represents the steady-state, and Φ is a matrix determining the persistence of the process.

We interpret the vector β_t as macro states of the economy that drive the monetary policy stance in equilibrium and also expectations regarding future economic states. Monetary policy depends on contemporaneous economic variables,

$$r_{t+1} = \bar{r} + \phi_r'(\beta_{t+1} - \bar{\beta}) + \rho_r r_t + \sigma_r \varepsilon_{t+1}^r, \tag{11}$$

where ρ_r captures monetary policy inertia, as discussed, for example, in Clarida et al. (2000) and Stein and Sunderam (2018), and ε_{t+1}^r reflects monetary policy shocks. We assume that monetary policy shocks ε_{t+1}^r are independent from ε_{t+1}, i.e., monetary policy shocks are not subsumed by public information on macro dynamics.

Denote the set of institutions and investors (including granular-demand investors and the Fed) excluding arbitrageurs as $\mathscr{I}$. Sector-ι's ($\iota \in \mathscr{I}$) demand for bonds with maturity $\tau \in \{1, \cdots, N\}$ takes the form

$$Z_t^\iota(\tau) = \theta_0^\iota(\tau) - \alpha^\iota(\tau)' p_t - \theta^\iota(\tau)'\beta_t + u_t^\iota(\tau), \tag{12}$$

where we use log prices instead of yields for consistency with Vayanos and Vila (2021); the two are related by $p^{(\tau)} = -\tau y^{(\tau)}$, so the demand system has the same functional form in either parameterization up to a maturity scaling of the coefficients. The parameter vector $\alpha^\iota(\tau)$ loads on the whole log-price vector p_t and reflects not only the demand sensitivity to the price of maturity τ itself but also sensitivities to prices of other maturities $\tau' \neq \tau$, capturing cross elasticities. We lump the demand for bonds from granular-demand investors and the Fed together, and refer to it as

the "non-arbitrageur demand", defined as

$$Z_t(\tau) = \sum_{\iota \in \mathscr{I}} Z_t^{\iota}(\tau). \tag{13}$$

Accordingly, we define $\theta_0(\tau)$, $\alpha(\tau)$, $\theta(\tau)$, and $u_t(\tau)$ as the sums of corresponding values from each sector $\iota \in \mathscr{I}$. We use column vector forms to express our setup in a more convenient and compact notation. In vector form, we can write (13) as

$$Z_t = \theta_0 - \alpha p_t - \theta \beta_t + u_t, \tag{14}$$

where $\theta_0 = (\theta_0(1), \theta_0(2), \cdots, \theta_0(N))'$ is an N-dimensional vector, $\alpha = (\alpha(1), \alpha(2), \cdots, \alpha(N))'$ is an $N \times N$ matrix, and $\theta = (\theta(1), \theta(2), \cdots, \theta(N))'$ is an $N \times K$ matrix. The unobservable, maturity-specific latent demand shock, $u_t = (u_t(1), u_t(2), \cdots, u_t(N))'$, reflects the non-systematic component of demand shocks. We assume that u_t is i.i.d., with mean zero and covariance matrix Σ^u.

On the supply side, we assume that the government issues Treasuries depending on macroeconomic conditions and the monetary policy rate. Accordingly, we specify the aggregate value of government bond supply, or, more precisely, the supply to the public market, i.e., marketable Treasury securities, as

$$S_t(\tau) = \bar{S}(\tau) + \zeta(\tau)' \beta_t + \zeta_r(\tau) r_t, \tag{15}$$

or, in vector form, as

$$S_t = \bar{S} + \zeta \beta_t + \zeta_r r_t, \tag{16}$$

where $\zeta = (\zeta(1), \zeta(2), \cdots, \zeta(N))'$ is an $N \times K$ matrix. We can interpret total government debt supply, which is $\sum_\tau S_t(\tau)$, as coming from a budget equation of the government, where Treasury supply adjusts to meet the need for government financing driven by macroeconomic conditions and the interest rate. Moreover, (15) captures the maturity-specific issuance, which as discussed in Greenwood et al. (2015a) is determined by fiscal needs (large immediate deficits often financed with short-term bills) and market conditions including the prevailing rate and the term premium (which is largely captured by the macro state β_t and Fed policy rate r_t). Together, (15) captures fiscal dynamics: both how total government financing responds to macroeconomic conditions and the interest rate, and how the government's maturity composition decisions vary across buckets through the freely estimated loadings $\zeta(\tau)$ and $\zeta_r(\tau)$.

We model a representative arbitrageur who maximizes mean-variance utility over Treasury positions and an outside asset, with rational expectations and no non-pecuniary motive. We denote arbitrageur positions in Treasuries of maturity τ as $X_t(\tau)$, and the outside asset position as $\tilde{X}_t$.

We view modeling outside assets as adding an important element of realism to models in the spirit of Vayanos and Vila (2021), since arbitrageurs' risk-bearing capacity in the Treasury market plausibly depends on their positions in other markets. We will estimate arbitrageurs' outside-asset risk exposure with a revealed preference approach.

Accordingly, arbitrageurs' wealth dynamics evolve as

$$W_{t+1} = W_t(1+R_t) + \sum_{\tau=2}^{N} X_t(\tau)(R_{t+1}^{(\tau)} - R_t) + \tilde{X}_t(\tilde{R}_{t+1} - R_t). \tag{17}$$

We assume that the return of the outside asset is normally distributed and depends on the state of the economy, in that

$$\tilde{R}_{t+1} = \tilde{\phi}'\beta_t + \tilde{\phi}_r r_t + \tilde{\sigma}'\varepsilon_{t+1} + \tilde{\sigma}_r \varepsilon_{t+1}^r, \tag{18}$$

where $\tilde{\phi}$ is a $K \times 1$ vector, $\tilde{\phi}_r$ is a scalar, $\tilde{\sigma}$ is a $K \times 1$ vector, and $\tilde{\sigma}_r$ is a scalar.

The objective of arbitrageurs is to maximize a mean-variance utility,

$$\max_{\{X_t(\tau)\}_\tau, \tilde{X}_t} \mathbb{E}_t[W_{t+1}] - \frac{\gamma}{2}\mathbb{V}_t(W_{t+1}), \tag{19}$$

subject to the wealth dynamics specified in (17).

Finally, for each maturity τ, there is a market-clearing condition,

$$Z_t(\tau) + X_t(\tau) = S_t(\tau). \tag{20}$$

We conjecture that there is an affine equilibrium in the form of

$$p_t = A\beta_t + A_r r_t + A_u u_t + C, \tag{21}$$

where $A = (A(1), A(2), \cdots, A(N))'$ is an $N \times K$ matrix, $A_r = (A_r(1), A_r(2), \cdots, A_r(N))'$ is an $N \times 1$ vector, $A_u = (A_u(1), A_u(2), \cdots, A_u(N))'$ is an $N \times N$ matrix, and $C = (C(1), C(2), \cdots, C(N))'$ is an $N \times 1$ vector.

4.2. A Simplified Version with Analytical Solutions

To gain intuition, we analyze a simplified version of the model with $N = 2$ maturities representing "short" and "long". We parameterize the demand-response matrix α in (14) as

$$\alpha = \begin{pmatrix} a & -b/2 \\ -b & a/2 \end{pmatrix}. \tag{22}$$

Since $p_t^{(\tau)} = -\tau y_t^{(\tau)}$, the demand responses to yields are symmetric with own-yield coefficient a and cross-yield coefficient $-b$. We assume $a, b > 0$, so that Treasury demand increases in its own yield but decreases in the other-maturity yield, consistent with the aggregate granular-demand investor demand we estimate in Section 3.

We set $K = 1$ so that the macro factor β_t is one-dimensional, and we interpret it as a "supply" factor that drives total debt supply. We also set $\phi_r = 0$ so that the monetary policy process does not depend on the macro factor, and $\bar{r} = 0$ for simplicity. We further set $\zeta_r = 0$ so that debt supply is

$$S_t(\tau) = \bar{S}(\tau) + \zeta(\tau)'\beta_t, \tag{23}$$

for $\tau = \{1, 2\}$. We impose a regularity condition that $\zeta(2) > -\theta(2)$ so that any supply expansion does not automatically get overshadowed by the expansion of demand in response to such supply expansion. Finally, for simplicity, we shut off all outside portfolio exposure by setting $\tilde{X}_t = 0$.

Using the first-order conditions and the market-clearing condition, we find the following unique equilibrium solution for log prices,

$$\begin{aligned} p_t^{(1)} &= -r_t, \\ p_t^{(2)} &= -\frac{1 + \rho_r + \gamma\sigma_r^2 b}{1 + \frac{a}{2}\gamma\sigma_r^2} r_t - \frac{\gamma\sigma_r^2(\zeta(2) + \theta(2))}{1 + \frac{a}{2}\gamma\sigma_r^2}\beta_t + \frac{\gamma\sigma_r^2}{1 + \frac{a}{2}\gamma\sigma_r^2} u_t(2) + \frac{\frac{1}{2} - \gamma\bar{S}(2) + \gamma\theta_0(2)}{\frac{1}{\sigma_r^2} + \frac{a}{2}\gamma}, \end{aligned} \tag{24}$$

where the first equation reflects that one-period bonds are priced at par against the policy rate, and the second equation comes from arbitrageurs accommodating the imbalance between Treasury supply and non-arbitrageur demand subject to risk aversion. Detailed derivations are provided in Appendix E.3, which also contains proofs of all the following propositions in this section.

The two equations in (24) already reveal a sharp asymmetry between the short and long end of the yield curve: the one-period price is pinned entirely by monetary policy, while the two-period price depends on macro shocks, latent demand, and the arbitrageur's risk aversion. To see the role of arbitrageurs more clearly, consider the two extreme cases.

When $\gamma \to \infty$, arbitrageurs drop out of the market. Taking the limit of $p_t^{(2)}$ in (24) gives the long-term price

$$p_t^{(2)} = -\frac{2b}{a} r_t - \frac{2}{a}(\zeta(2) + \theta(2))\beta_t + \frac{2}{a} u_t(2) + \frac{2}{a}(\theta_0(2) - \bar{S}(2)), \tag{25}$$

where market clearing at the long end requires granular-demand investors to absorb all long-term supply independently. On the other hand, short-end supply imbalances must instead be absorbed by granular-demand investors, who require a price concession of $1/a$ per unit. The short end loses

its special elasticity, and the sharp maturity asymmetry disappears.

When $\gamma \to 0$, arbitrageurs are risk-neutral and arbitrage to the full extent, so the long-term price converges to

$$p_t^{(2)} = -(1+\rho_r)r_t + \frac{1}{2}\sigma_r^2, \tag{26}$$

the log Treasury price under the expectations hypothesis, where the second term is a Jensen's convexity adjustment.

These two limits clarify the role of arbitrageurs and motivate the following proposition.

Proposition 1 (Arbitrageurs and the Term Structure of Market Elasticity)**.** *The Treasury market exhibits a downward-sloping term structure of elasticity: the one-period market is perfectly elastic, while the two-period market has finite elasticity decreasing in γ. Arbitrageurs are the source of this asymmetry. Without arbitrageurs ($\gamma \to \infty$), the one-period market loses this special status and the term structure of elasticity flattens.*

In practice, T-bills are not identical to the overnight policy rate, so the short end is highly elastic but not perfectly so. The full quantitative model captures this: the short end retains much higher elasticity than the long end, consistent with evidence that T-bill supply shocks have very small price impact (Greenwood et al. 2015b).

The equilibrium holdings follow directly from market clearing: $X_t(2) = S_t(2) - Z_t(2)$, where $Z_t(2)$ is obtained by substituting the price solution (24) into the demand equation (14). As $\gamma \to \infty$, arbitrageur holdings vanish and granular-demand investors absorb all supply. As $\gamma \to 0$, arbitrageurs fully absorb demand shocks from other investors. Closed-form expressions are in Appendix E.3.

We next turn to the term premium response to monetary policy, which is where the role of cross-maturity substitution is key.

Proposition 2 (Cross-Substitution and the Term Premium Response to Monetary Policy)**.** *In a plain-vanilla preferred-habitat model without cross-maturity substitution ($b = 0$), a positive monetary policy shock reduces the term premium on long-term Treasuries. With cross-maturity substitution ($b > 0$), the term premium rises if and only if cross elasticity is strong enough and satisfies $2b/a > 1 + \rho_r$.*

The plain-vanilla case $b = 0$ delivers the standard result in Vayanos and Vila (2021): when the monetary policy rate rises, long-term Treasuries become cheaper, which induces granular-demand investors to absorb more of them. This reduces the quantity arbitrageurs must hold, lowers the risk premium, and dampens the long-term yield increase. The term premium always falls.

With cross-substitution, a competing force emerges. When the short rate rises, granular-demand investors facing attractive short-term yields reduce their long-term Treasury holdings and reallocate toward the short end. This forces arbitrageurs to absorb a larger long-term position, raising the risk premium. The proposition characterizes exactly when this rebalancing force dominates: the condition $2b/a > 1+\rho_r$ requires cross-substitution to be large relative to own-elasticity and to the persistence of monetary policy shocks.

In Section 3, we find that for most investor sectors, cross-elasticity is of comparable magnitude to own-elasticity. Aggregating across sectors gives $2b/a > 1+\rho_r$, so the strong cross-elasticity condition is satisfied in the data. Proposition 2 therefore indicates overreaction of long-term yields to monetary tightening, consistent with the empirical evidence in Hanson and Stein (2015) and Bauer et al. (2023). Critically, this result is not obtained in a plain-vanilla VV model regardless of parameterization: cross-maturity substitution is necessary to flip the sign.

Propositions 1 and 2 provide sharp characterizations, but they are derived under a simplification of the full model. In the richer full model, we consider more than two maturities, the demand elasticity matrix α is more general than equation (22), and arbitrageurs hold an outside portfolio that interacts with Treasury pricing through r_t and β_t. Nevertheless, we believe this simplified model provides the key intuition that guides our quantitative analysis: arbitrageurs are what make the short end of the Treasury market highly elastic, and cross-maturity substitution is what generates the positive term premium response to monetary tightening.

4.3. Full Model Solution

We conjecture an affine solution of the form (21). Given this conjecture, we solve the mean-variance problem in (19) and derive arbitrageurs' first-order conditions. For tractability, we assume idiosyncratic latent demand shocks carry no systematic risk exposure and hence do not contribute to the prices of risk. However, they still affect Treasury prices through demand pressure.

Define $\mu_t^{(\tau)} \equiv \mathbb{E}_t[R_{t+1}^{(\tau)}]$ as the expected return on maturity-τ Treasuries, where $R_{t+1}^{(\tau)} \approx r_{t+1}^{(\tau)} + \frac{1}{2}\mathbb{V}_t[r_{t+1}^{(\tau)}]$ and $r_{t+1}^{(\tau)} = p_{t+1}^{(\tau-1)} - p_t^{(\tau)}$.[10] Solving the optimization problem (19) gives the first-order conditions

$$\mu_t^{(\tau)} - r_t = \hat{A}(\tau-1)'\underbrace{\gamma\left(\sum_{\hat{\tau}=2}^{N}\Sigma\hat{A}(\hat{\tau}-1)X_t(\hat{\tau})+\Sigma^{1/2}\tilde{\sigma}\tilde{X}_t\right)}_{\lambda_{\beta,t}} + A_r(\tau-1)\underbrace{\gamma\left(\sum_{\hat{\tau}=2}^{N}\sigma_r^2 A_r(\hat{\tau}-1)X_t(\hat{\tau})+\sigma_r\tilde{\sigma}_r\tilde{X}_t\right)}_{\lambda_{r,t}}, \tag{27}$$

where $\lambda_{\beta,t}$ and $\lambda_{r,t}$ are the time-varying prices of macro and interest-rate risk, and $\hat{A}(\tau-1) \equiv A(\tau-1)+\phi_r A_r(\tau-1)$ is the total risk exposure of the τ-period bond (see Appendix E.2).

[10] The approximation becomes exact in continuous time; see Greenwood et al. (2024).

The outside-asset position $\tilde{X}_t$ enters the price of risk through $\Sigma^{1/2}\tilde{\sigma}\tilde{X}_t$ and $\sigma_r\tilde{\sigma}_r\tilde{X}_t$. Since we cannot separately identify all parameters governing $\tilde{X}_t$'s dynamics, we assume these composites are affine in the state vector:

$$\begin{aligned}\Sigma^{1/2}\tilde{\sigma}\tilde{X}_t &= \Psi\beta_t + \Lambda r_t + \psi,\\ \sigma_r\tilde{\sigma}_r\tilde{X}_t &= \Psi_r\beta_t + \Lambda_r r_t + \psi_r,\end{aligned} \tag{28}$$

where $\{\Psi, \Psi_r, \Lambda, \Lambda_r, \psi, \psi_r\}$ are constant matrices and vectors. Intuitively, outside-asset exposure captures non-Treasury risk borne by arbitrageurs, including positions in interest-rate swaps (Du et al. 2023b), Treasury futures (Barth and Kahn 2025), and other fixed-income instruments. As we explain in Section 5.1, $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$ are recovered in closed form once γ is identified.

Equilibrium arbitrageur holdings follow directly from market clearing:[11]

$$X_t(\tau) = \underbrace{\bar{S}(\tau) + \zeta(\tau)'\beta_t + \zeta_r(\tau)r_t}_{\text{supply}} - \underbrace{\big(\theta_0(\tau) - \alpha(\tau)'p_t - \theta(\tau)'\beta_t + u_t(\tau)\big)}_{\text{non-arbitrageur demand}}. \tag{29}$$

Substituting the affine price conjecture into (27) and matching coefficients on $(\beta_t, r_t, u_t, 1)$ yields iterative equations for (A, A_r, A_u, C) given in Appendix E.2.

The structure of equations (27)–(29) reveals a clean separation between two channels through which γ affects Treasury pricing.

The first channel operates through the *pricing of common risk factors*. Macro shocks (β_t, r_t) affect both Treasury prices and the outside portfolio, so their impact on term premia is a composite of Treasury-specific and non-Treasury risk exposures. In the FOC (27), this is reflected in the price-of-risk loadings λ, which encapsulate γ, the outside-portfolio loadings $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$, and the covariance structure of arbitrageur holdings. Outside-portfolio exposure and Treasury risk aversion are non-separable from macro-yield data alone.

The second channel operates through the *absorption of Treasury-specific demand shocks*. Latent demand shocks $u_t(\tau)$ are idiosyncratic to the Treasury market: they shift supply-demand imbalances of Treasuries without directly affecting the outside asset risk exposure. When such a shock arrives, arbitrageurs must absorb the residual imbalance, and the required yield adjustment is determined entirely by how costly it is for them to hold more Treasuries, i.e., by γ alone. The loading A_u therefore depends on γ directly through the market-clearing condition, with no outside-portfolio degree of freedom. An arbitrageur with higher γ demands a larger yield to absorb a given quantity shock. Because the outside portfolio is unaffected by these idiosyncratic shocks, the yield movement per unit of Treasury absorbed is directly driven by γ. We exploit this to identify γ in

[11] This is an equilibrium outcome, not a demand function. As discussed in Section E.1, rational arbitrageurs do not respond directly to yields.

Section 5.1.

5. Identification and Model Estimation

This section describes how we identify and estimate the model. Separating arbitrageur risk aversion γ from outside-portfolio exposure is the central identification challenge: both affect Treasury pricing, and without a clean separation, γ would not be recoverable from yield data alone. Section 5.1 shows that the two channels are identified from distinct sources of variation, with macro-driven yield dynamics pinning down the prices of risk λ and idiosyncratic demand shocks pinning down γ. Section 5.2 implements this separation as a three-step estimation procedure. Sections 5.3 and 5.4 report the estimates and confirm that the model fits both the yield curve and the time series of arbitrageur Treasury holdings.

5.1. Identification: Arbitrageur Risk Aversion vs. Outside Portfolio Exposure

The granular demand estimation in Section 3 is essential not only for documenting investor behavior but also for identifying the arbitrageur risk aversion in the model. The Treasury-specific latent demand shocks u_t recovered from sector-level estimation are the moments that disentangle arbitrageur risk aversion from the outside-portfolio loadings; observed arbitrageur Treasury holdings alone do not deliver these moments.

Section 4.3 shows that arbitrageurs affect Treasury pricing through two distinct channels. The first is their *Treasury holdings*. By absorbing supply-demand imbalances, arbitrageurs bear duration risk and demand compensation, with the required premium scaling with risk aversion γ through the market-clearing loadings $A_u(\gamma)$. The second is their *outside portfolio exposure*. Arbitrageurs also hold non-Treasury assets, including interest-rate swaps (Du et al. 2023b) and Treasury futures (Barth and Kahn 2025), whose covariance with Treasury returns shapes the common-factor price of risk λ. The outside-portfolio loadings $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$ in equation (28) capture this channel. Both are economically important. Omitting either would misattribute term premium variation to the wrong source.

A natural concern is that these two channels are not separately identified. The outside portfolio, by providing additional flexibility in fitting yield dynamics, could absorb all the pricing variation and render γ unidentifiable. We show that this is not the case. The two channels are identified from distinct sources of variation in the data, as established in Section 4.3.

Macro-driven yield variation identifies the prices of risk λ. As equation (27) shows, λ is a

composite of γ and the outside-portfolio loadings $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$. Any value of γ can be offset by a different outside-portfolio exposure to produce the same λ. Consequently, γ is *not* identified from macro-yield covariation alone. Instead, macro-driven variation identifies λ directly. Given λ, the systematic price loadings (A, A_r) are immediately determined by their recursion equations. Intuitively, λ is a sufficient statistic for (A, A_r). Whatever combination of γ and outside-portfolio loadings produces a given λ, it implies the same systematic yield loadings. Once γ is subsequently pinned down from Treasury-specific shocks, the outside-portfolio loadings $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$ follow in closed form.

Demand-shock-driven variation identifies γ. Given the price loadings (A, A_r), the price impact matrix $A_u(\gamma)$ is uniquely pinned down by γ through the market-clearing condition, with no outside-asset degree of freedom. When a latent demand shock $u_t(\tau)$ hits a specific maturity, it triggers both a yield response and a quantity adjustment by arbitrageurs. The ratio of these two responses is driven by γ. A more risk-averse arbitrageur requires a larger yield movement to absorb the same quantity shock. These intuitions imply that γ is sharply identified by matching the idiosyncratic components of Treasury yields and arbitrageur holdings, that is, the parts unexplained by macro factors and systematic supply and demand.

We further validate this separation in Section 6.2, where we introduce a Treasury-holding penalty κ and hold outside-asset parameters fixed. We show that term premia and market elasticity respond strongly as Treasury absorption changes. If the outside portfolio were absorbing all the pricing work, these responses would be negligible. The data show the opposite, confirming that γ is genuinely disciplined by Treasury quantities.

5.2. Estimation Procedure

We denote observed Treasury log prices as $p_t^o(\tau)$ and observed arbitrageur holdings as $X_t^o(\tau)$, and estimate the model in three steps that mirror the identification logic above. Step 1 exploits macro-driven yield variation to recover the prices of risk λ, following directly from the first identification channel. Step 2 then exploits Treasury-specific demand shocks to identify γ, following from the second channel. Step 3 jointly refines all parameters, with the yield-curve intercept $C(\tau)$ determined by the equilibrium recursion rather than treated as a free parameter. Full details are in Appendix E.6.

Step 1 recovers the prices of macro and interest-rate risk. Under assumption (28), the prices of risk become affine in (β_t, r_t):

$$\lambda_{\beta,t} = \lambda_{\beta\beta}\beta_t + \lambda_{\beta r} r_t + c_\beta, \qquad \lambda_{r,t} = \lambda_{r\beta}\beta_t + \lambda_{rr} r_t + c_r, \tag{30}$$

where $\lambda \equiv \{\lambda_{\beta\beta}, \lambda_{\beta r}, \lambda_{r\beta}, \lambda_{rr}\}$ are constant slope matrices. We estimate λ and a maturity-specific free intercept $C(\tau)$ by minimizing pricing errors:

$$\min_{\lambda, C(\cdot)} \sum_t \sum_\tau \left(A(\lambda)\beta_t + A_r(\lambda) r_t + C(\tau) - p_t^o(\tau)\right)^2. \tag{31}$$

As discussed in Section 5.1, this step does not identify γ.

Step 2 identifies γ from the joint price and quantity response to latent demand shocks. Given $\{A, A_r\}$ from Step 1, A_u is uniquely determined by γ. We estimate γ by minimizing

$$\min_\gamma \left\{ \underbrace{\sum_t \sum_\tau \left(\frac{\eta_t^o(\tau) - e_\tau' A_u(\gamma)\hat{u}_t}{\bar{\eta}^o(\tau)} \right)^2}_{\mathscr{L}^{\text{price}}(\gamma)} + \underbrace{\sum_t \sum_m \left(\frac{X_t^{o,resid}(m) - X_t^{resid}(m)}{\bar{X}_{abs}^{o,resid}(m)} \right)^2}_{\mathscr{L}^{\text{qty}}(\gamma)} \right\}, \tag{32}$$

where $\eta_t^o(\tau) \equiv p_t^o(\tau) - A\beta_t - A_r r_t - C(\tau)$ is the demand-driven yield residual, $\hat{u}_t$ is the estimated latent demand shock from Section 3, e_τ selects the τ-th element of $\hat{u}_t$, and $X_t^{o,resid}$, X_t^{resid} are observed and model-implied arbitrageur position residuals after removing systematic macro-driven components. The normalization terms $\bar{\eta}^o(\tau)$ and $\bar{X}_{abs}^{o,resid}(m)$ are average absolute values that equalize the weight of each maturity bucket. Since both $\mathscr{L}^{\text{price}}$ and $\mathscr{L}^{\text{qty}}$ are driven by the same $A_u(\gamma)$, they jointly pin down γ.

Step 3 completes the estimation by jointly minimizing

$$\min_{\lambda, \gamma, \psi, \psi_r} \underbrace{\sum_t \sum_\tau \left(\frac{A\beta_t + A_r r_t + A_u u_t + C(\tau) - p_t^o(\tau)}{\bar{p}^o(\tau)} \right)^2}_{\mathscr{L}^{\text{price,full}}} + \underbrace{\sum_t \sum_m \left(\frac{X_t(m) - X_t^o(m)}{\bar{X}^o(m)} \right)^2}_{\mathscr{L}^{\text{qty,full}}}, \tag{33}$$

where $\mathscr{L}^{\text{price,full}}$ fits the full model-implied price, including the demand-shock component $A_u u_t$, against observed yields, and $\mathscr{L}^{\text{qty,full}}$ fits model-implied arbitrageur holdings against observed positions across maturity buckets. Having both objectives is important. $\mathscr{L}^{\text{price,full}}$ disciplines the price loadings and the intercept $C(\tau)$, while $\mathscr{L}^{\text{qty,full}}$ ensures the implied portfolio positions remain consistent with data, jointly anchoring γ and the outside-portfolio parameters. $C(\tau)$ is solved from the model's market-clearing recursion at every candidate parameter vector, ensuring full equilibrium consistency. The outside-portfolio loadings $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$ are recovered in closed form from the Step 3 estimates. All reported estimates are from this final step.

5.3. Estimation Results

In line with our empirical analysis and motivated in Appendix A.3, we set the macro state vector $\beta_t =$ (credit spread, GDP gap, core inflation, debt/GDP). We estimate a VAR of the form (10) over the main sample. Core inflation and debt/GDP are highly persistent, but the maximum eigenvalue of the VAR is 0.87, so macro variables converge to their long-run averages.

The monetary policy rule is estimated over the post-Volcker period (1990–2024) excluding the ZLB episode (2008–2015); over this sample, monetary policy exhibits sufficient variation for identification. Estimated coefficients on GDP gap and inflation have the same signs as in the classical Taylor rule (Taylor 1993), and the inertia coefficient on the lagged policy rate is 0.78 (see Appendix equation (A52)). This high inertia generates a strong expectations channel through which monetary policy affects long-term yields, playing a critical role in how the yield curve responds to monetary policy shocks.

The full model is estimated over 2011Q4–2022Q4. The resulting risk-aversion parameter is $\gamma = 0.03$. As shown in Section 6, this implies a relatively elastic Treasury market. Figure 5 plots the Step 2 objective $\mathscr{L}^{\text{price}}(\gamma) + \mathscr{L}^{\text{qty}}(\gamma)$ as a function of γ: the profile is convex with a unique minimum near the baseline estimate, confirming that γ is sharply identified. Appendix E.7 maps γ into a coefficient of relative risk aversion and shows that the estimate implies economically plausible values.

To assess robustness, we use a pairs bootstrap that jointly resamples the time series of prices, latent demand shocks, macro factors, and arbitrageur holdings, then re-runs the structural estimation. The standard deviation of the bootstrap distribution for γ is 0.016, indicating a tight distribution around the baseline (see Appendix E.8).

Moreover, we implement an additional robustness check in Appendix F.3. The subsample analysis re-estimates the full pipeline on a pre-COVID period (2011Q4–2019Q4) and a post-ZLB period (2016Q1–2022Q4); the market elasticity ($\mathscr{E} = 2.72$–4.28) is relatively stable across all three samples despite variation in reduced-form IV loadings.

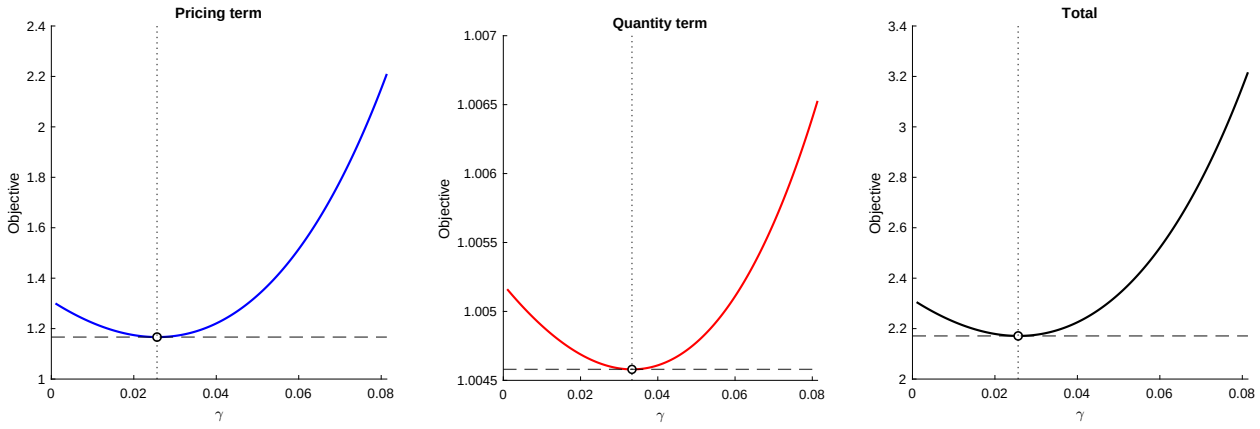
5.4. Model Fit: Yields and Arbitrageur Holdings

The Step 3 objective jointly targets yield dynamics and arbitrageur positions, so model fit along both dimensions is informative about the quality of the estimates.

Figure 6 reports model-implied yields against the data at four maturities. Each panel plots the data (black), the model fit using macro fundamentals and monetary policy alone ($u_t = 0$, blue dashed), and the full fit including latent demand shocks (u_t, red dash-dot). The macro-only fit

Figure 5. **Step 2 Objective Profiles for Identifying γ.**

Left panel: $\mathscr{L}^{\text{price}}(\gamma)$. Middle panel: $\mathscr{L}^{\text{qty}}(\gamma)$. Right panel: $\mathscr{L}^{\text{price}}(\gamma)+\mathscr{L}^{\text{qty}}(\gamma)$. The convex shape and unique minimizer in each panel indicate clear identification and a well-defined minimization problem.



captures the broad level and trend at all maturities. Including u_t substantially tightens the fit, particularly at long maturities where demand-driven fluctuations are most important.

Figure 7 compares model-implied arbitrageur Treasury holdings with observed data. The close alignment across all three maturity buckets validates both the estimated γ and the identification strategy. This tight fit is not mechanical: it requires γ to be set at exactly the level where the implied arbitrageur positions match what hedge funds and broker-dealers actually hold.

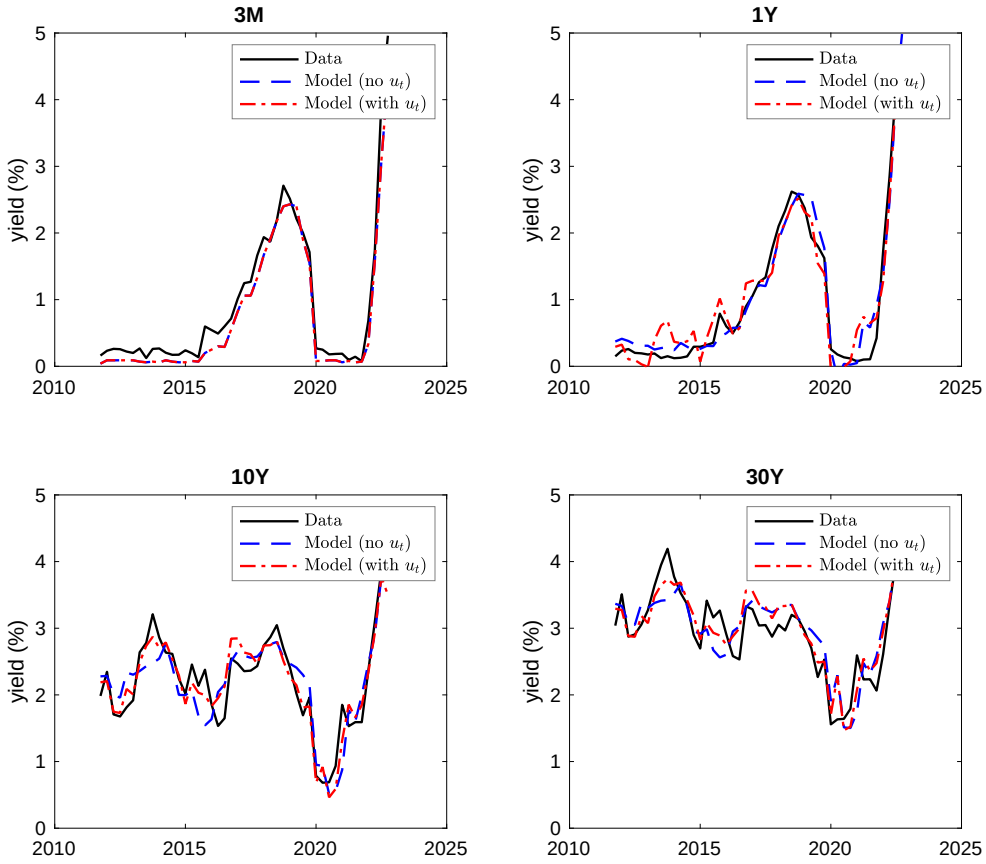
6. How Arbitrageurs and Granular Demand Shape the Treasury Market

In this section, we put our model to work and quantify how arbitrageurs and granular investor demand jointly shape Treasury pricing and monetary transmission. Moreover, we contrast the quantitative implications of our model with arbitrageurs and cross-maturity substitution from those of extant preferred-habitat models (Vayanos and Vila 2021) and pure demand-system approaches to the term structure.

The analysis delivers three results. First, the term structure of market elasticity is steeply downward-sloping, and arbitrageur intermediation is its primary driver: removing arbitrageurs collapses market elasticity by about 10 times in the aggregate market multiplier and amplifies the T-bill price impact of demand shocks by more than 70 times (Section 6.1). Second, augmenting the analysis with a Treasury holdings cost confirms that this arbitrageur role reflects genuine pricing discipline from the quantities they hold, not a parameterization artifact (Section 6.2). Third, a model comparison across three nested specifications shows that cross-maturity substitution is the

Figure 6. **Model Fit on the Dynamics of Treasury Yields.**

Each panel plots observed yields (black), model-implied yields without latent demand shocks ($u_t = 0$, blue dashed), and model-implied yields including u_t (red dash-dot), constructed via equation (21).



single ingredient that flips the sign of the term premium response to monetary tightening from negative to positive (Section 6.3). The plain-vanilla Vayanos-Vila model, with or without outside-asset exposure, produces the empirically wrong sign. Only the full model with estimated cross-maturity elasticities reverses it.

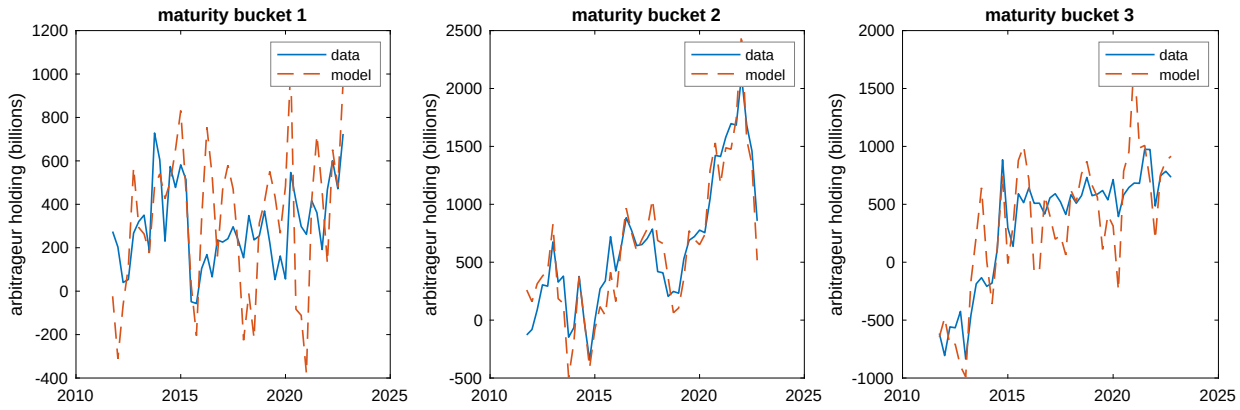
6.1. The Term Structure of Market Elasticity

In our model, the response of Treasury yields to demand shocks is shaped not only by granular investors' individual demand elasticities, but critically also by arbitrageur risk aversion, as shown by Proposition 1. To quantify this, we compare the baseline case with estimated γ to the limit $\gamma \to \infty$ where arbitrageurs exit the market entirely. In the latter case, market clearing implies

$$p_t = \alpha^{-1}\left((\theta_0 - \theta\beta_t + u_t) - \left(\bar{S} + \zeta\beta_t + \zeta_r r_t\right)\right), \tag{34}$$

Figure 7. **Arbitrageur Holdings: Model vs. Data.**

Model-implied arbitrageur holdings $X_t(\tau) = S_t(\tau) - Z_t(\tau)$ versus data aggregated from hedge fund and broker-dealer holdings.



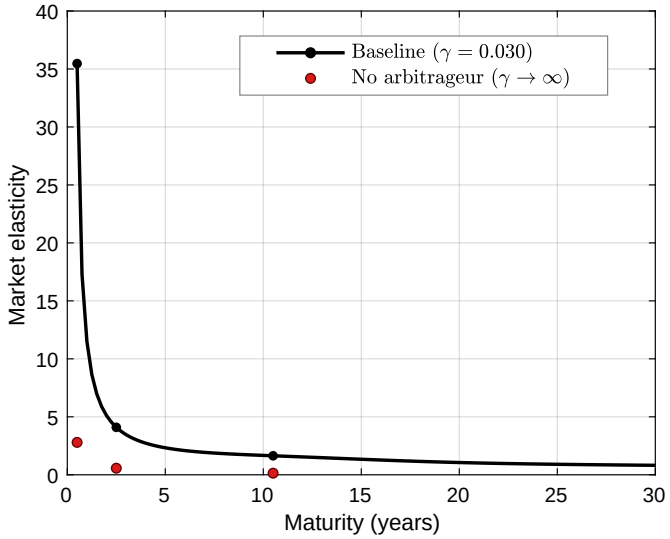
so the price response to a demand shock is simply α^{-1}, the inverse of the non-arbitrageur demand elasticity matrix. With arbitrageurs present, the equilibrium elasticity additionally depends on γ, macroeconomic shock volatility Σ, monetary policy uncertainty σ_r and inertia ρ_r, and the persistence of macro dynamics Φ.

Figure 8 plots the term structure of market elasticity $\mathscr{E}(\tau)$, defined as the inverse of the price multiplier: a market elasticity of $\mathscr{E}$ at maturity τ means that a 1% demand shock at τ moves *total Treasury valuation* by $1/\mathscr{E}\%$. Higher elasticity thus means smaller price impact. For example, an elasticity of 10 implies that a 1% demand shock moves total valuation by only 0.1%, consistent with a highly liquid market. The baseline curve (solid black) reveals a sharply downward-sloping term structure, directly confirming Proposition 1. Note that the elasticity here measures the total market valuation response, aggregating price changes across all maturities weighted by supply. The own-price impact of a T-bill shock alone would be far smaller, since arbitrageurs absorb T-bill shocks at near-zero duration cost and pass them almost entirely into longer-maturity prices rather than T-bill prices. The three red dots mark the no-arbitrageur elasticities at the three bucket maturities (0.5, 2.5, and 10.5 years): they collapse to 2.8, 0.6, and 0.1, compared to baseline values of 35.5, 4.1, and 1.6 at the same maturities. Thus, without arbitrageurs, elasticities are about 12 times smaller on average.

Table 4 quantifies this more precisely, reporting the equilibrium bond price impact (in %) at each maturity bucket to a latent demand shock equal to 1% of the outstanding amount in that bucket, in the spirit of the price multiplier in Gabaix and Koijen (2021). Panel (a) shows the full model with arbitrageurs. For a given demand shock, the price impact is substantially larger at longer maturities: a short-maturity demand shock produces a long-maturity price impact 15 times

Figure 8. **The Term Structure of Market Elasticity.**

The solid black line shows the baseline term structure of market elasticity ($\gamma = 0.03$). The red dots mark the no-arbitrageur ($\gamma \to \infty$) elasticities at the three maturity buckets. Market elasticity at maturity τ is defined as the percentage change in demand at maturity τ relative to total Treasury outstanding required to move total Treasury valuation by 1%.



larger. Panel (b) removes arbitrageurs ($\gamma = \infty$), and Panel (c) reports the ratio. Price impacts are one to two orders of magnitude larger without arbitrageurs, with the T-bill own-price impact rising 75 times. The T-bill price impact in Panel (b) is far too large to be consistent with a world in which the Fed effectively controls the short rate. With arbitrageurs, the Fed pins the one-period rate and arbitrageurs propagate dynamics through the term structure with weakening effects at longer maturities.

Using the average supply of Treasuries in each maturity bucket as weights, the full-model multipliers aggregate to a total market multiplier of 0.31, implying that a $100 billion demand shock raises total Treasury valuation by $31 billion.[12] For comparison, Chaudhary et al. (2025) report a corporate bond market multiplier of 3.5, and Gabaix and Koijen (2021) find a stock market multiplier of 5. Thus, the Treasury market has significantly lower price impact per dollar of demand, so it is a much more elastic market in aggregate. In Appendix E.8, a bootstrap procedure confirms that with near certainty the Treasury market is more elastic than both the corporate bond and equity markets. However, without arbitrageurs, the multiplier rises to 3.26.

The downward slope reconciles seemingly contradictory findings in the literature. Krishnamurthy and Vissing-Jorgensen (2011) and D'Amico and King (2013) show that QE purchases of

[12]This is for a latent demand shock. The multiplier for permanent demand shocks is 0.81. See Appendix F.2 for details on calculating market multipliers for both latent and permanent demand shocks.

Table 4. Price Impact of Latent Demand Shocks with and without Arbitrageurs

Each cell reports the equilibrium bond price impact (%) at the column maturity bucket to a latent demand shock equal to 1% of outstanding at the row maturity bucket. Panel (a) reports responses under the baseline estimated model with arbitrageurs. Panel (b) reports responses in the counterfactual without arbitrageurs ($\gamma \to \infty$), where market clearing is determined solely by granular-demand investors. Panel (c) reports the ratio of Panel (b) to Panel (a). The three maturity buckets are: short ($\tau < 1$ year), medium ($1 \leq \tau < 5$ years), and long ($\tau \geq 5$ years). Latent demand shocks u_t are the residual component of investor demand not explained by observable macro variables and yields, as defined in Section 3.1. Model parameters are estimated as described in Section 5.2.

	Price response (%)		
	short	medium	long
Panel (a): With Arbitrageur			
shock on short maturity	0.001	0.006	0.015
shock on medium maturity	0.009	0.077	0.188
shock on long maturity	0.016	0.142	0.432
Panel (b): Without Arbitrageur			
shock on short maturity	0.057	0.486	1.917
shock on medium maturity	0.191	0.720	3.827
shock on long maturity	0.109	0.555	1.509
Panel (c): Ratio (b)/ (a)			
shock on short maturity	75.239	76.243	128.615
shock on medium maturity	21.457	9.307	20.392
shock on long maturity	7.007	3.919	3.494

long-term bonds significantly compress term premia, while Greenwood et al. (2015b) find that large T-bill supply shifts move convenience yields by only a few basis points. Our model explains both: high elasticity at the short end (arbitrageurs absorb T-bill shocks at near-zero duration cost) and low elasticity at the long end (duration risk requires substantial compensation). Our model implies a supply-weighted average non-T-bill elasticity of 3, falling to 1.6 at the long end. These estimates align well with recent empirical studies that report roughly 2 in Eren et al. (2026) and around 1 in Chaudhary et al. (2024).

6.2. The Role of Arbitrageur Treasury Holdings

Next, we show that Treasury pricing is disciplined by the quantities arbitrageurs hold. To trace out this sensitivity, we introduce a quadratic penalty κ on arbitrageur Treasury holdings,

$$\max_{\{X_t(\tau)\},\tilde{X}_t} \mathbb{E}_t[W_{t+1}] - \frac{\gamma}{2}\mathbb{V}_t(W_{t+1}) - \frac{\kappa}{2}\sum_{\tau=2}^{N} X_t(\tau)^2, \tag{35}$$

and ask how yields change as κ rises and arbitrageurs are progressively forced toward smaller positions. We do not interpret κ as a structural feature of the economy; it is purely a diagnostic device for mapping arbitrageur quantities into prices. Setting $\kappa = 0$ recovers the baseline; $\kappa \to \infty$ forces $X_t(\tau) \to 0$ and collapses the model to the pure habitat-demand benchmark. Full derivations are in Appendix E.9.

Table 5. Effect of the Arbitrageur Penalty Parameter κ on Key Model Outcomes

This table traces how key model outcomes change as the penalty parameter κ forces arbitrageurs toward smaller Treasury positions. The term $\frac{\kappa}{2}\sum_\tau X_t(\tau)^2$ is added to the arbitrageur objective as a diagnostic device; κ is not a structural feature of the model. *Baseline* ($\kappa = 0$) is the estimated model. *Re-estimated* re-optimizes all model parameters under $\kappa = 10^{-6}$, giving the model full flexibility to adjust. Mean $|X|$ (in \$billions) is the time-series average of absolute arbitrageur Treasury holdings. γ is the arbitrageur risk aversion parameter. T-bill price impact per T-bill shock reports the equilibrium bond price response (%) to a 1% T-bill supply shock; the empirical benchmark from Greenwood et al. (2015b) is approximately 7.2×10^{-4}%.

	Baseline $\kappa = 0$	Re-estimated $\kappa = 10^{-6}$
Mean $\lvert X \rvert$ (arbitrageur holdings)	527.23	519.14
Estimated γ	0.0298	0.0296
T-bill price impact per T-bill shock ($\times 10^{-4}$%)	7.64	67.06

Table 5 compares the baseline to a re-estimated model under $\kappa = 10^{-6}$, where all parameters are re-optimized to give the model maximum flexibility to absorb the balance-sheet penalty. The key diagnostic is the T-bill price response to a T-bill supply shock. Greenwood et al. (2015b) estimate that a 1% increase in the T-bill quantity outstanding reduces T-bill price by 7.2×10^{-4}%;[13] consistent with this empirical target, our baseline reproduces a price impact of 7.64×10^{-4}% per unit shock (this is reported as 0.001 in Panel (a) of Table 4, rounded to three decimal places).

The results in Table 5 are striking: despite a decline in arbitrageur holdings of only 1.5% and essentially no change in the estimated risk aversion γ, the T-bill price impact rises from 7.64×10^{-4}% to 67×10^{-4}%, nearly nine times larger and well above the Greenwood et al. (2015b) benchmark. The no-arbitrageur limit $\kappa \to \infty$ pushes the T-bill price impact further to 570×10^{-4}% (the 0.057 entry in Panel (b) of Table 4), more than 70 times the baseline estimate. Treasury pricing at the short end is therefore genuinely disciplined by the quantities arbitrageurs hold, and this role cannot be absorbed through the adjustment of outside portfolio parameters.

[13] Their Table I, Panel A, column (15) reports an IV coefficient of 10.44 bps per percentage-point increase in BILLS/GDP. With mean BILLS/GDP of approximately 9% in their sample, a 1% increase in T-bill quantity outstanding corresponds to a 0.09 percentage-point change in BILLS/GDP, implying a yield response of $10.44 \times 0.09 \approx 0.94$ bps. Converting to a price impact using the 4-week T-bill duration of $4/52 \approx 0.077$ years gives $0.94 \times 0.077/100 \approx 7.2 \times 10^{-4}$%. Our baseline model implies a T-bill price impact of 7.64×10^{-4}%, computed directly as the short-bucket price response to a 1% short-bucket supply shock.

6.3. Cross-Substitution and the Term Premium Response to Monetary Tightening

We now turn to the second major implication of our framework: the response of term premia to monetary policy shocks. As Proposition 2 establishes, the sign of this response depends critically on whether cross-maturity substitution in non-arbitrageur demand is sufficiently strong.

To see why this matters, consider the two polar cases. In models where the expectations hypothesis holds, the term premium does not respond to monetary policy at all. In the plain-vanilla preferred-habitat model of Vayanos and Vila (2021) without cross-maturity substitution, a positive monetary policy shock lowers Treasury prices at the long end, which induces granular-demand investors to absorb more long-term bonds, reducing the supply held by arbitrageurs and thereby compressing the term premium. The result is an *underreaction* of long-term yields relative to the expectations hypothesis. This is the empirically wrong sign: a large literature documents that long-term yields *overreact* to monetary tightening (Hanson and Stein 2015; Bauer et al. 2023).

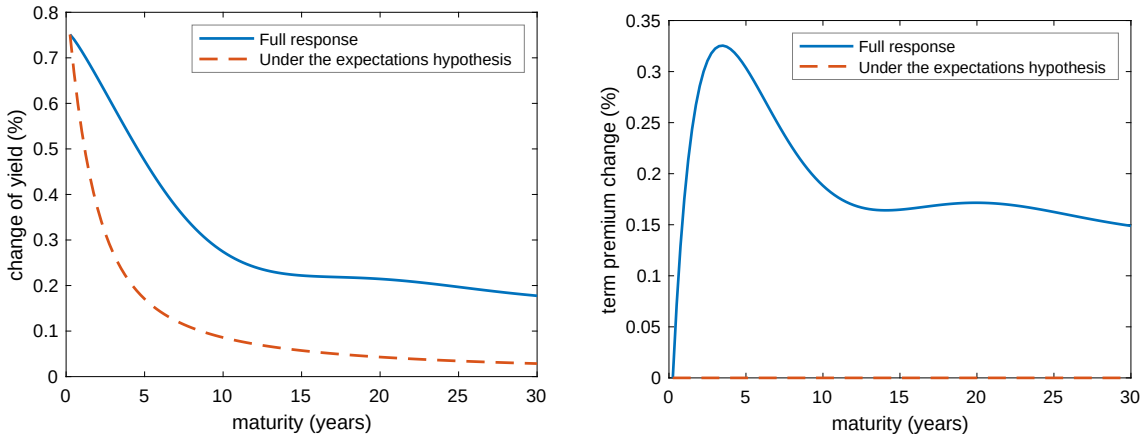
Our model reverses this result through cross-maturity substitution. When the short rate rises, granular-demand investors who substitute across maturities *reduce* their long-term demand rather than increase it, because the more attractive short-term yield attracts them toward the short end. This forces arbitrageurs to absorb more long-term supply, expanding their balance-sheet exposure and raising the long-end risk premium. Proposition 2 shows the term premium rises with monetary tightening whenever the cross-substitution elasticity is strong enough relative to monetary policy inertia. Whether this condition holds in the data is a quantitative question that requires estimating the demand system from data. Our estimated demand elasticities satisfy the condition, so the cross-substitution force dominates.

Figure 9 confirms the quantitative prediction. Panel (a) shows the yield curve response to a one-standard-deviation monetary tightening ($\varepsilon_r = 1$, equivalent to a 0.75% increase in the short rate (see Appendix equation (A52))). The full model response exceeds the expectations-hypothesis (EH) component at every maturity, with the gap widening at longer horizons as the EH component declines toward zero. Panel (b) shows the implied term premium response, defined as the difference between the full yield response and the EH component. It is uniformly positive, peaking near 30 basis points at the 3 to 4 year range and settling around 13 basis points at long maturities.

Next, we show that this positive term premium response is driven by cross-maturity substitution. Figure 10 compares three nested specifications that are estimated on the same data and use the same structural algorithm, but differ in one input: the estimated demand system. Cases (1) and (2) impose the Vayanos and Vila (2021) diagonal-demand restriction, ignoring all cross-maturity elasticities in the first-stage IV regressions (i.e., no "other yield" in the regression). Case (1) additionally excludes outside-asset exposure from the arbitrageur's portfolio; Case (2)

Figure 9. **Contemporaneous Yield Curve Response to a Monetary Policy Shock.**

A one standard deviation monetary policy shock ($\varepsilon_t^r = 1$) under two cases: the full model and the expectations hypothesis (risk-neutral arbitrageurs). Left panel: yield curve responses. Right panel: term premium responses.



(a) Yield curve response

(b) Term premium response

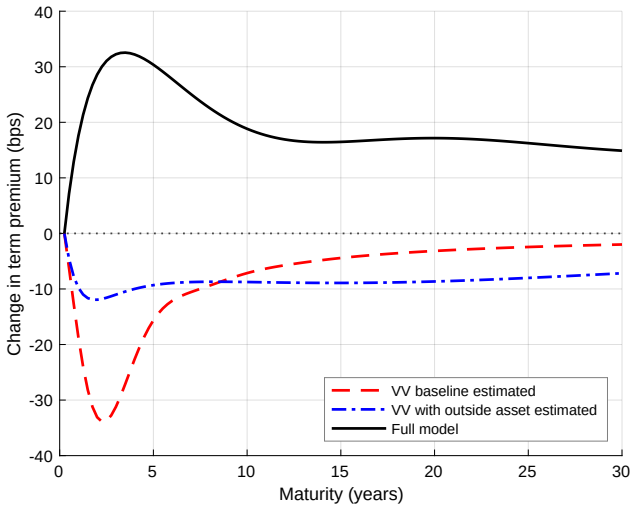
adds it back. Case (3) is our full model, which uses the unrestricted demand system in which cross-maturity elasticities are freely estimated. Because Cases (2) and (3) share the same structural estimation algorithm and the same outside-asset structure, any difference in their term-premium responses is attributable solely to whether the demand function inputs for the model allow for cross-maturity substitution.

Under both diagonal-demand variants in Cases (1) and (2), the term-premium response to monetary tightening is negative across the entire maturity spectrum. Adding the outside-asset channel in Case (2) moves the response only mildly and does not alter the sign. Only the full model in Case (3), which incorporates estimated cross-maturity elasticities, reverses the sign, generating a uniformly positive response that peaks near 30 basis points at the 3 to 4 year range and remains around 13 basis points at long maturities.

The mechanism is transparent. Under diagonal demand, a rate hike lowers long-term bond prices and induces granular-demand investors to absorb more of them, reducing arbitrageur exposure and compressing the term premium. Under full demand with strong cross-maturity substitution, the same rate hike draws investors toward the short end of the curve, forcing arbitrageurs to absorb more long-term supply and raising the risk premium. Cross-maturity substitution is the key force that reverses the sign of monetary transmission into term premia.

Figure 10. **Term Premium Response to Monetary Tightening: Model Comparison.**

Term-premium response (in bps) to a one-standard-deviation monetary policy shock across three estimation specifications: plain-vanilla VV (diagonal demand, no outside asset), VV with outside-asset exposure, and the full model with cross-maturity demand substitution. Cases (2) and (3) share the same estimation algorithm and outside-asset structure; their responses differ in sign solely because of the diagonal-versus-full demand function.



7. Conclusion

In this paper, we estimate an equilibrium model of the U.S. Treasury market, nesting granular-demand investors, whose Treasury demand we flexibly estimate from a novel dataset on granular Treasury holdings in the spirit of Koijen and Yogo (2019), and risk-averse arbitrageurs, who absorb demand imbalances as in Vayanos and Vila (2021). A key finding of our paper is that rationalizing the empirical evidence necessarily requires accounting for both granular-demand investors and arbitrageurs.

Our first main finding is that arbitrageurs are essential for the highly elastic T-bill market that a pure demand approach cannot explain. In the no-arbitrageur limit, the T-bill price impact per unit demand shock rises by more than 70 times, counterfactually implying that the Fed could not tightly control the short rate through conventional policy. Arbitrageurs absorbing T-bills at near-zero duration cost are the primary source of short-end elasticity.

Our second main finding is that granular demand estimation is critical to obtain a realistic term premium response to monetary tightening. Absent the cross-maturity demand substitution recovered from our granular demand estimation, term premia fall with monetary tightening, as in preferred-habitat settings such as Vayanos and Vila (2021). In contrast, in our setting, a one-standard-deviation monetary tightening raises term premia by up to 30 basis points at the 3 to

4 year range and 13 basis points at long maturities. The mechanism is that when the short rate rises, granular-demand investors rebalance toward higher-yielding short-term Treasuries, forcing arbitrageurs to absorb more long-term supply and raising the risk premium.

We view our paper as a stepping stone to building quantitative models of the macrostructure of financial markets (Haddad and Muir 2025) that recognize how heterogeneity in investors' objectives, mandates, and constraints shapes asset prices. Our approach allows us to combine novel micro data with equilibrium asset-pricing models to study government bond markets. The granular demand system and structural estimation approach developed here can be incorporated into macroeconomic models to study how investor heterogeneity shapes the transmission of fiscal and monetary policy through the yield curve, or extended to broader markets and asset classes.

References

Acharya, V. V. and Laarits, T. (2023). When do Treasuries earn the convenience yield?: A hedging perspective. *National Bureau of Economic Research Working Paper*.

Adrian, T., Etula, E., and Muir, T. (2014). Financial intermediaries and the cross-section of asset returns. *The Journal of Finance*, 69(6):2557–2596.

Allen, J., Kastl, J., and Wittwer, M. (2020). Estimating demand systems for Treasuries. *Working paper*.

An, Y. and Huber, A. (2024). Intermediary elasticity. *SSRN Electronic Journal*.

Bahaj, S., Czech, R., Ding, S., and Reis, R. (2023). The market for inflation risk. *Bank of England Working Paper*.

Banegas, A. and Monin, P. (2023). Hedge fund treasury exposures, repo, and margining. *FEDS Notes*. Board of Governors of the Federal Reserve System.

Barberis, N., Greenwood, R., Jin, L., and Shleifer, A. (2015). X-CAPM: An extrapolative capital asset pricing model. *Journal of Financial Economics*, 115(1):1–24.

Barth, D. and Kahn, R. J. (2025). Hedge funds and the Treasury cash-futures disconnect. *Journal of Monetary Economics*, 151:103744.

Bauer, M. D., Bernanke, B. S., and Milstein, E. (2023). Risk appetite and the risk-taking channel of monetary policy. *Journal of Economic Perspectives*, 37(1):77–100.

Becker, B. and Ivashina, V. (2015). Reaching for yield in the bond market. *The Journal of Finance*, 70(5):1863–1902.

Bekaert, G., Hoerova, M., and Duca, M. L. (2013). Risk, uncertainty and monetary policy. *Journal of Monetary Economics*, 60(7):771–788.

Bernanke, B. (2005). The global saving glut and the U.S. current account deficit. *Federal Reserve Board speech*.

Bernanke, B. S. (2013). Long-term interest rates. Speech, Annual Monetary/Macroeconomics Conference, Federal Reserve Bank of San Francisco, March 2013. `https://www.federalreserve.gov/newsevents/speech/bernanke20130301a.htm`.

Bernanke, B. S. and Kuttner, K. N. (2005). What explains the stock market's reaction to federal reserve policy? *The Journal of Finance*, 60(3):1221–1257.

Bi, H., Phillot, M., and Zubairy, S. (2026). Treasury supply shocks: Propagation through debt expansion and maturity adjustment. Technical report, National Bureau of Economic Research.

Bloomberg News (2022). Fed signals liftoff soon, sees asset reduction start afterward. `https://www.bloomberg.com/news/articles/2022-01-26/fed-signals-liftoff-soon-sees-asset-reduction-start-afterward`.

Bretscher, L., Schmid, L., Sen, I., and Sharma, V. (2025). Institutional corporate bond pricing. *Review of Financial Studies*.

Brunnermeier, M. K., Merkel, S., and Sannikov, Y. (2024). Safe assets. *Journal of Political Economy*.

Caballero, R. J., Caravello, T. E., and Simsek, A. (2024). Financial conditions targeting. *National Bureau of Economic Research Working Paper*.

Campbell, J. Y. (2006). Household finance. *The Journal of Finance*, 61(4):1553–1604.

Chaudhary, M., Fu, Z., and Li, J. (2025). Corporate bond multipliers: Substitutes matter. *SSRN Electronic Journal*.

Chaudhary, M., Fu, Z., and Zhou, H. (2024). Anatomy of the Treasury market: Who moves yields? *SSRN Electronic Journal*.

Choi, J. and Kronlund, M. (2018). Reaching for yield in corporate bond mutual funds. *The Review of Financial Studies*, 31(5):1930–1965.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? In Taylor, J. B. and Woodford, M., editors, *Handbook of Macroeconomics*, volume 1, pages 65–148. Elsevier.

Cieslak, A., Li, W., and Pflueger, C. (2024). Inflation and Treasury convenience. *National Bureau of Economic Research Working Paper*.

Clarida, R., Gali, J., and Gertler, M. (2000). Monetary policy rules and macroeconomic stability: evidence and some theory. *The Quarterly Journal of Economics*, 115(1):147–180.

Culbertson, J. (1957). The term structure of interest rates. *The Quarterly Journal of Economics*, 71(4):485–517.

Darmouni, O., Siani, K., and Xiao, K. (2025). Nonbank fragility in credit markets: Evidence from a two-layer asset demand system. *The Journal of Finance*.

d'Avernas, A., Petersen, D., and Vandeweyer, Q. (2023). The central bank's balance sheet and Treasury market disruptions. *SSRN Electronic Journal*.

d'Avernas, A. and Vandeweyer, Q. (2024). Treasury bill shortages and the pricing of short-term assets. *The Journal of Finance*, 79(6):4083–4141.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, 98(4):703–738.

Del Negro, M., Eggertsson, G., Ferrero, A., and Kiyotaki, N. (2017). The great escape? a quantitative evaluation of the Fed's liquidity facilities. *American Economic Review*, 107(3):824–857.

Di Tella, S., Hébert, B., Kurlat, P., and Wang, Q. (2025). The zero-beta interest rate. *Journal of Political Economy*.

Diamond, W. (2020). Safety transformation and the structure of the financial system. *The Journal of Finance*, 75(6):2973–3012.

Diamond, W., Jiang, Z., and Ma, Y. (2024). The reserve supply channel of unconventional monetary policy. *Journal of Financial Economics*, 159:103887.

Diamond, W. and Van Tassel, P. (2023). Risk-free rates and convenience yields around the world. *Federal Reserve Bank of New York Staff Reports*, 1032.

Doerr, S., Eren, E., and Malamud, S. (2023). Money market funds and the pricing of near-money assets. *Swiss Finance Institute Research Paper 23-04*.

Domanski, D., Shin, H., and Sushko, V. (2017). The hunt for duration: Not waving but drowning? *IMF Economic Review*, 65(1):113–153.

Drechsler, I., Savov, A., and Schnabl, P. (2018). A model of monetary policy and risk premia. *The Journal of Finance*, 73(1):317–373.

Droste, M., Gorodnichenko, Y., and Ray, W. (2024). Unbundling quantitative easing: Taking a cue from Treasury auctions. *Journal of Political Economy*, 132(9):3115–3172.

Du, W., Hébert, B., and Huber, A. W. (2023a). Are intermediary constraints priced? *The Review of Financial Studies*, 36(4):1464–1507.

Du, W., Hébert, B., and Li, W. (2023b). Intermediary balance sheets and the Treasury yield curve. *Journal of Financial Economics*, 150(3):103722.

Du, W., Tepper, A., and Verdelhan, A. (2018). Deviations from covered interest rate parity. *The Journal of Finance*, 73(3):915–957.

Duffie, D., Fleming, M. J., Keane, F., Nelson, C., Shachar, O., and Van Tassel, P. (2023). Dealer capacity and U.S. Treasury market functionality. *Federal Reserve Bank of New York Staff Reports*.

D'Amico, S. and King, T. B. (2013). Flow and stock effects of large-scale Treasury purchases: Evidence on the importance of local supply. *Journal of Financial Economics*, 108(2):425–448.

Eren, E., Schrimpf, A., and Xia, F. D. (2026). The demand for government debt. *Management Science*.

Fang, X., Hardy, B., and Lewis, K. K. (2025). Who holds sovereign debt and why it matters. *The Review of Financial Studies*, 38.

Fang, X. and Liu, Y. (2021). Volatility, intermediaries, and exchange rates. *Journal of Financial Economics*, 141(1):217–233.

Favara, G., Infante, S., and Rezende, M. (2022). Leverage regulations and treasury market participation: Evidence from credit line drawdowns. *Working Paper*.

Federal Reserve (2018). Policy normalization discussions and communications history. `https://www.federalreserve.gov/monetarypolicy/policy-normalization-discussions-communications-history.htm`.

Gabaix, X. and Koijen, R. S. (2021). In search of the origins of financial fluctuations: The inelastic markets hypothesis. *National Bureau of Economic Research Working Paper*.

Gertler, M. and Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.

Gilchrist, S. and Zakrajšek, E. (2012). Credit spreads and business cycle fluctuations. *American Economic Review*, 102(4):1692–1720.

Gourinchas, P.-O., Ray, W., and Vayanos, D. (2025). A preferred-habitat model of term premia, exchange rates, and monetary policy spillovers. *American Economic Review*, 115(11):3788–3824.

Greenwood, R., Hanson, S., Stein, J. C., and Sunderam, A. (2023). A quantity-driven theory of term premia and exchange rates. *The Quarterly Journal of Economics*, 138(4):2327–2389.

Greenwood, R., Hanson, S., and Vayanos, D. (2024). Supply and demand and the term structure of interest rates. *Annual Review of Financial Economics*, 16:115–151.

Greenwood, R., Hanson, S. G., Rudolph, J. S., and Summers, L. (2015a). The optimal maturity of government debt. *The $13 trillion question: How America manages its debt*, pages 1–41.

Greenwood, R., Hanson, S. G., and Stein, J. C. (2015b). A comparative-advantage approach to government debt maturity. *The Journal of Finance*, 70(4):1683–1722.

Greenwood, R. and Vayanos, D. (2014). Bond supply and excess bond returns. *Review of Financial Studies*, 27(3):663–713.

Guibaud, S., Nosbusch, Y., and Vayanos, D. (2013). Bond market clienteles, the yield curve, and the optimal maturity structure of government debt. *The Review of Financial Studies*, 26(8):1914–1961.

Haddad, V., Moreira, A., and Muir, T. (2024). Asset purchase rules: How QE transformed the bond market. *Working paper, UCLA and University of Rochester*.

Haddad, V. and Muir, T. (2021). Do intermediaries matter for aggregate asset prices? *The Journal of Finance*, 76(6):2719–2761.

Haddad, V. and Muir, T. (2025). Market macrostructure: Institutions and asset prices. *Annual Review of Financial Economics*, 17:133–150.

Haddad, V. and Sraer, D. (2020). The banking view of bond risk premia. *The Journal of Finance*, 75(5):2465–2502.

Hanson, S. G., Lucca, D. O., and Wright, J. H. (2021). Rate-amplifying demand and the excess sensitivity of long-term

rates. *The Quarterly Journal of Economics*, 136(3):1719–1781.

Hanson, S. G., Malkhozov, A., and Venter, G. (2024). Demand-and-supply imbalance risk and long-term swap spreads. *Journal of Financial Economics*, 154:103814.

Hanson, S. G. and Stein, J. C. (2015). Monetary policy and long-term real rates. *Journal of Financial Economics*, 115(3):429–448.

He, Z., Kelly, B., and Manela, A. (2017). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics*, 126(1):1–35.

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

Jansen, K. A. (2025). Long-term investors, demand shifts, and yields. *Review of Financial Studies*, 38(1):114–157.

Jermann, U. (2020). Negative swap spreads and limited arbitrage. *The Review of Financial Studies*, 33(1):212–238.

Jiang, W., Sargent, T. J., Wang, N., and Yang, J. (2024a). A p theory of government debt and taxes. *The Journal of Finance*.

Jiang, Z., Krishnamurthy, A., and Lustig, H. (2021). Foreign safe asset demand and the dollar exchange rate. *The Journal of Finance*, 76(3):1049–1089.

Jiang, Z., Lustig, H., Van Nieuwerburgh, S., and Xiaolan, M. Z. (2024b). The U.S. public debt valuation puzzle. *Econometrica*, 92(4):1309–1347.

Jiang, Z., Richmond, R. J., and Zhang, T. (2024c). A portfolio approach to global imbalances. *The Journal of Finance*, 79(3):2025–2076.

Kaminska, I. and Zinna, G. (2020). Official demand for U.S. debt: Implications for real rates. *Journal of Money, Credit and Banking*, 52(2-3):323–364.

Kargar, M. (2021). Heterogeneous intermediary asset pricing. *Journal of Financial Economics*, 141(2):505–532.

Kekre, R., Lenel, M., and Mainardi, F. (2024). Monetary policy, segmentation, and the term structure. *National Bureau of Economic Research Working Paper*.

Kelejian, H. H. (1971). Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association*, 66(334):373–374.

Khetan, U., Li, J., Neamtu, I., and Sen, I. (2023). The market for sharing interest rate risk: Quantities and asset prices. *SSRN Electronic Journal*.

Koijen, R. and Yogo, M. (2019). A demand system approach to asset pricing. *Journal of Political Economy*, 127(4):1475–1515.

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

Koijen, R. S. and Yogo, M. (2026). Exchange rates and asset prices in a global demand system. *Journal of Political Economy*.

Krishnamurthy, A. and Li, W. (2023). The demand for money, near-money, and treasury bonds. *The Review of Financial Studies*, 36(5):2091–2130.

Krishnamurthy, A. and Muir, T. (2025). How credit cycles across a financial crisis. *The Journal of Finance*, 80(3):1339–1378.

Krishnamurthy, A. and Vissing-Jorgensen, A. (2011). The effects of quantitative easing on interest rates: Channels and implications for policy. *Brookings Papers on Economic Activity*, 43(2):215–287.

Krishnamurthy, A. and Vissing-Jorgensen, A. (2012). The aggregate demand for Treasury debt. *Journal of Political Economy*, 120(2):233–267.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pages 1315–1335.

Lewis, D. J. and Mertens, K. (2025). A robust test for weak instruments for 2SLS with multiple endogenous regressors. *Review of Economic Studies*. Advance article, DOI: 10.1093/restud/rdaf103.

Li, W. (2024). Public liquidity and financial crises. *American Economic Journal: Macroeconomics*.

Li, W., Ma, Y., and Zhao, Y. (2023). The passthrough of Treasury supply to bank deposit funding. *Columbia Business School Research Paper, USC Marshall School of Business Research Paper*.

Maggiori, M., Neiman, B., and Schreger, J. (2020). International currencies and capital allocation. *Journal of Political Economy*, 128(6):2019–2066.

Mehra, R. and Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, 15(2):145–161.

Modigliani, F. and Sutch, R. (1966). Innovations in interest rate policy. *The American Economic Review*, 56(1/2):178–197.

Montiel Olea, J. L. and Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369.

Nagel, S. (2016). The liquidity premium of near-money assets. *The Quarterly Journal of Economics*, 131(4):1927–1971.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–837.

Payne, J. and Szőke, B. (2024). Convenience yields and financial repression. *Working paper*.

Payne, J., Szőke, B., Hall, G., and Sargent, T. J. (2025). Costs of financing U.S. federal debt under a gold standard: 1791–1933. *The Quarterly Journal of Economics*, 140(1):793–833.

Phillot, M. (2025). U.S. Treasury auctions: A high-frequency identification of supply shocks. *American Economic Journal: Macroeconomics*, 17(1):245–273.

Potter, S. (2018). Money markets at a crossroads: Policy implementation at a time of structural change. Speech, Federal Reserve Bank of New York, October 2018.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.

Shue, K., Townsend, R., and Wang, C. (2024). Categorical thinking about interest rates. *CESifo Working Paper Series*.

Siani, K. (2025). Raising bond capital in segmented markets. *The Review of Financial Studies*.

Siriwardane, E., Sunderam, A., and Wallen, J. L. (2025). Segmented arbitrage. *The Journal of Finance*, 80(5):2543–2590.

Stein, J. C. and Sunderam, A. (2018). The Fed, the bond market, and gradualism in monetary policy. *The Journal of Finance*, 73(3):1015–1060.

Stein, J. C. and Wallen, J. (2025). The imperfect intermediation of money-like assets. *The Journal of Finance*, 80(6):3185–3221.

Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, chapter 5. Cambridge: Cambridge Univ. Press.

Tabova, A. M. and Warnock, F. E. (2021). Foreign investors and U.S. Treasuries. *National Bureau of Economic Research Working Paper*.

Taylor, J. B. (1993). Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy*, volume 39, pages 195–214. Elsevier.

Vayanos, D. and Vila, J.-L. (2021). A preferred-habitat model of the term structure of interest rates. *Econometrica*, 89(1):77–112.

Wallen, J. (2020). Markups to financial intermediation in foreign exchange markets. Working paper, Harvard University.

Wu, J. C. and Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit and Banking*, 48(2-3):253–291.

Internet Appendix of "Granular Treasury Demand with Arbitrageurs"

Kristy A.E. Jansen Wenhao Li Lukas Schmid

A. Data Sources and Aggregation

This appendix details the various data sources used to construct our dataset of granular U.S. Treasury holdings and explains how these datasets are merged. Specifically, in Section A.1 we report the data sources of U.S. Treasury holders, in Section A.2 we discuss the process of merging datasets of Treasury holdings, and in Section A.3 we provide data sources for macro variables.

A.1. Treasury Holders

A. Banks - CALL Reports

Banks are major investors in the U.S. Treasury market. We obtain banks' holdings of U.S. Treasuries at the maturity bucket level from CALL reports. CALL reports are regulatory filings required for all U.S. banks and include detailed information on a bank's assets, liabilities, income, and expenses. The CALL reports are filed on a quarterly basis and cover the period from the first quarter of 1976 to the end of 2022. Banks report their aggregate U.S. Treasury holdings and their holdings in different maturity buckets of U.S. Treasuries and U.S. Agency bonds combined. The maturity buckets are: $\tau < 3M$, $3M \leq \tau < 1Y$, $1Y \leq \tau < 3Y$, $3Y \leq \tau < 5Y$, $5Y \leq \tau < 15Y$, $\tau \geq 15Y$. To obtain their allocation to U.S. Treasuries for different maturities, we assume that the fraction of Treasuries versus Agency bonds is fixed across maturities at a given point in time. Hence, at each point in time, we multiply the total maturity bucket holdings by the fraction of Treasuries relative to the sum of Treasuries and Agency bonds.

B. Fed - Federal Reserve

In the aftermath of the Great Financial Crisis, the Federal Reserve has become a major player in the U.S. Treasury market. The Federal Reserve System Open Market Account (SOMA) reports security holdings that are acquired through open market operations by the Fed. These data are obtained through the website of the Federal Reserve Bank of New York.[1] The holdings are at the security (CUSIP) level and reported on a weekly basis since the start of 2003.

[1] `https://www.newyorkfed.org/markets/soma-holdings`

C. Primary Dealers - Federal Reserve

To maintain transparency of U.S. and foreign primary dealers' trading activities, their total weekly positions are made available through the website of the Federal Reserve Bank of New York.[2] Primary dealers report their holdings for conventional maturity buckets since early 1998. However, the specific maturity buckets reported change over time. The time frames with the same reporting standards are: January 1998 to June 2001, July 2001 to March 2013, April 2013 to December 2014, January 2015 to December 2021, and from January 2022 onward. Generally, more recent data report finer maturity buckets. To be consistent across time, we treat July 2001 to March 2013 as the baseline and aggregate the maturity buckets of subsequent periods to match that of this time frame. The final maturity buckets are: T-bills, Treasuries with $1Y \leq \tau \leq 3Y$, $3Y < \tau \leq 6Y$, $6Y < \tau \leq 11Y$, and $\tau > 11Y$.

D. Hedge Funds - Form PF

We obtain aggregate U.S. and foreign hedge fund Treasury positions from Form PF that hedge funds file with the SEC.[3] As of 2011Q4, hedge funds must file Form PF if they are registered or are required to register with the SEC, manage private funds, and have at least $150 million in total assets. The Fed reports the totals separately for domestic and foreign hedge funds. We only observe the aggregate Treasury positions, so we rely on the maturity distribution obtained from primary dealers to infer the maturity bucket holdings. That is, we multiply the maturity bucket weights of primary dealers by the aggregate hedge fund Treasury positions at each point in time to obtain maturity bucket specific hedge fund holdings. The reason we rely on primary dealers to infer the maturity distribution is twofold. First, we define both as arbitrageurs, consistent with the literature (Du et al. 2023b; Vayanos and Vila 2021). Second, corroborating the idea that both hedge funds and primary dealers act as arbitrageurs, the aggregate Treasury holdings of primary dealers and hedge funds align closely in that higher aggregate Treasury holdings for primary dealers tend to come with higher holdings for hedge funds (see Figure A4 of the Appendix).

E. Insurers and Pension Funds - eMAXX

eMAXX provides broad coverage of fixed-income holdings of institutional investors at the security (CUSIP) level. The database predominantly covers the holdings of insurance companies, mutual funds, and pension funds (Becker and Ivashina 2015; Bretscher et al. 2025). We only use the data

[2]The data and the list of primary dealers that must report can be found here: `https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics`. Specifically, the Fed allows certain foreign-owned institutions to operate as primary dealers in the U.S. Treasury market if they meet specific criteria.

[3]We thank Moritz Lenel for directing us to this data source.

on insurance companies and pension funds, and rely on Morningstar for mutual funds. Due to the voluntary nature of reporting by pension funds, the coverage of pension funds in eMAXX is limited, unlike the mandatory reporting by insurance companies. Additionally, we focus on the U.S. eMAXX database, which covers the holdings of North American investors. The holdings data are quarterly and cover the period from the first quarter of 2010 to the end of 2022.

F. Money Market Funds - IMoneyNet and FoFs

IMoneyNet provides a wide coverage of asset holdings (predominately fixed income and cash) by U.S. money market funds (MMFs) at the security (CUSIP) level. We focus on both holdings reported by MMFs domiciled in the U.S. as well as on their offshore holdings. The holdings are reported on a monthly basis since August 2011.

To obtain a larger coverage of the total MMF population, we augment the data with FoFs from the Federal Reserve. Using our security-level database, we verify that on average 99.6% of MMF holdings are in either T-bills or U.S. Treasuries with remaining time to maturity less than 1 year. Hence, we can reasonably assume that MMF Treasury holdings reported in FoF have remaining maturities below 1 year.

G. Mutual Funds - Morningstar

We obtain holdings data on domestic and foreign mutual funds from Morningstar, Inc. The funds report all their positions including stocks, bonds, and cash at the security (CUSIP) level. We focus on both fixed-income and allocation funds. Funds either report monthly or quarterly, and to maintain consistency across the funds and other data sets we use data at quarter ends. Figure A5 reports the aggregate holdings in USD (trillions) over time. These aggregates align closely with the numbers reported in Maggiori et al. (2020).

H. ETFs - ETF Global

We obtain the holdings of U.S. Exchange-Traded Funds (ETFs) at the security (CUSIP) level from ETF Global. ETF Global contains extensive coverage of securities held by U.S. ETFs and in our analysis we focus on fixed-income funds. Funds either report daily or monthly, and to maintain consistency with the other datasets we use data at quarter ends. As U.S. ETFs only hold a small fraction of U.S. Treasuries outstanding, we merge them with the U.S. mutual fund sector.

I. Foreign Official and Private - Public TIC

We obtain quarterly U.S. Treasury holdings by foreign investors from the Treasury International Capital Reporting System (TIC). Specifically, we obtain the public TIC Form SLT that exists as of September 2011. As of this date, TIC also provides a breakdown of the total amount held in T-bills versus non T-bills. As of December 2011, TIC also distinguishes between foreign official and foreign private investors. Moreover, to avoid double counting, we subtract from the private foreign Treasury holdings the holdings of foreign mutual funds that we obtain through Morningstar and foreign hedge funds that we obtain through Form PF.

A.2. Data Aggregation

For the data sources in Table 1 that are at the security level, we observe the corresponding CUSIP identifiers that we use to match the holdings data with the CRSP U.S. Treasury Database. The CRSP U.S. Treasury Database contains detailed bond-level information on U.S. Treasuries, including bond yields, prices, bond type, coupon rate, maturity date, issue date, and issuance size. We use the bond prices to convert nominal holdings to market values. For the sectors that report at a more aggregate level (banks, foreign investors, hedge funds, and primary dealers), we use their reported market value holdings directly.

For investors that report at the CUSIP level, including insurers and pension funds, mutual funds, ETFs, money market funds, and the Fed, it is straightforward to divide their holdings in the respective maturity buckets: $\tau < 1Y, 1Y \leq \tau < 5Y, \tau \geq 5Y$. For banks, we aggregate maturity buckets $\tau < 3M$ and $3M \leq \tau < 1Y$ to obtain the first bucket, $1Y \leq \tau < 3Y$ and $3Y \leq \tau < 5Y$ to obtain the second bucket, and $5Y \leq \tau < 15Y$ and $\tau \geq 15Y$ to obtain the third bucket. We follow a similar approach for the primary dealers, whereby we assign T-bills to bucket 1, $1Y \leq \tau \leq 3Y$ and $3Y < \tau \leq 6Y$ to bucket 2, and $6Y < \tau \leq 11Y$ and $\tau > 11Y$ to bucket 3. As motivated earlier, we assume that hedge funds have the same maturity bucket distribution as primary dealers.

For foreign investors, we only observe the fraction that is held in T-bills versus non T-bills. To allocate the foreign holdings to different maturity buckets, we first multiply the T-bill holdings by the inverse of the fraction of the total amount outstanding in maturity bucket 1 of the CRSP universe that is in T-bills, at each point in time. The reason is that on average only 60% of the total amount outstanding in maturity bucket 1 consists of T-bills, while the remaining 40% are bonds and notes with remaining time to maturity below 1 year. This adjustment is meant to more accurately reflect the remaining maturity structure, but our estimations for foreign investors are similar when we assume that T-bills are the only securities held in bucket 1. We then subtract the additional fraction we attribute to maturity bucket 1 from the total non T-bill holdings to compute

the total holdings in the remaining maturity buckets. To further determine the fraction in maturity bucket 2 versus 3, we choose the fraction such that the average duration of both the foreign official and foreign private investors' Treasury portfolio is consistent with Tabova and Warnock (2021) at each point in time. To assign the fractions, we take the bond durations of a 6-month, 3-year, and 15-year bond, respectively, as representative bonds for each maturity bucket. However, our main results do not depend on this choice. For instance, the results are qualitatively and quantitatively similar if we choose instead a 10-year or a 20-year bond for the third bucket.

To obtain the residual sector, we subtract the holdings of all investors from the total amount outstanding in each bucket. Since we observe the total foreign investor position, the residual sector consists of U.S. based investors only and hence we will refer to this sector as "Other U.S. Investors".

Finally, in our growing economic environment, portfolio holdings in dollar values will not be stationary. For stationarity, we scale all quantities in our regressions and in the model by the ratio of potential GDP (ticker `NGDPPOT` in FRED, which is nominal potential gross domestic product) at the end of our sample period to the potential GDP at that particular quarter. For example, the ratio of potential GDP in 2022 Q4 to that in 2011 Q4 is 1.6. The dollar value of total debt supply in 2011 Q4 is 10.7 trillion, but we use a scaled value, namely $10.7 \times 1.6 = 17.1$ trillion. We use nominal values so that the scaling adjusts for the inflation effect. Moreover, using a GDP adjuster rather than just inflation ensures that we account for the growing scale of the economy. Finally, we use nominal potential GDP rather than nominal GDP to avoid cyclical fluctuations in nominal GDP that cause mechanical correlations among the variables due to the scaling. The underlying assumption is that after accounting for the scaling effect, all quantities are stationary in the fundamental state variables. An alternative scaling is to use a constant exponential growth rate that matches the overall economic growth during our sample period, and we find that this approach leads to similar results.

A.3. Macro Data

We complement our dataset with a number of macroeconomic variables that capture relevant drivers of monetary and fiscal policy stances, as well as aggregate economic conditions. Specifically, we obtain four macro variables from the Federal Reserve Economic Data (FRED).

First, we include the GDP gap and core inflation to capture aggregate economic conditions as well as the response of monetary policy to macroeconomic dynamics. They together reflect aggregate demand and supply fluctuations in the economy, and they are also the variables that drive monetary policy in the Taylor rule.

Second, we include the debt/GDP ratio to capture the overall supply and dynamics of govern-

ment debt. As an indicator of the government's fiscal policy stance, the debt/GDP ratio is plausibly connected to the GDP gap, as well as inflation.

Finally, as an indicator of financial market conditions relevant to the aggregate economy, we include credit spreads, which have been widely shown to predict macroeconomic movements (Gilchrist and Zakrajšek 2012; Krishnamurthy and Muir 2025).

B. Stylized Facts of Treasury Holdings

This appendix presents stylized facts on U.S. Treasury holdings. Specifically, in Section B.1 we document which investor types are marginal buyers of Treasuries across maturity buckets, and in Section B.2 we characterize the prevalence of short positions across sectors and maturities.

B.1. Marginal Buyers of U.S. Treasuries

Table A1 examines which investor types are marginal in trading U.S. Treasuries when supply changes. Similar to Fang et al. (2025), but with a focus on maturity buckets, we decompose the marginal holders of Treasuries. For each maturity bucket m and sector ι, we regress changes in holdings on changes in the total supply of debt:

$$\frac{Z_t^\iota(m) - Z_{t-1}^\iota(m)}{S_{t-1}(m)} = a^\iota(m) + b^\iota(m)\frac{S_t(m) - S_{t-1}(m)}{S_{t-1}(m)} + \varepsilon_t^\iota(m), \quad \forall \iota \tag{A1}$$

where $Z_t^\iota(m)$ denotes the total market value of Treasury holdings by sector ι in maturity bucket m at time t (in billions) and $S_t(m)$ the total market value supply of Treasuries in maturity bucket m at time t (in billions). The accounting identity in equation (A1) implies that the sum across sectors must add up to the total so that $\sum_\iota b^\iota(m) = 100\%$ for all m. We also aggregate the holdings of each sector across maturities and estimate the total marginal contribution of each sector.

Panel (a) shows the results on marginal holdings. Aggregate regressions highlight the Fed and MMFs as the largest absorbers of U.S. Treasuries, but a maturity breakdown reveals significant differences. In short maturities, MMFs dominate, with hedge funds (HF ROW + HF US) surpassing the Fed. In intermediate maturities, foreign officials are the primary absorbers, but their contribution is negligible in long maturities.

Panel (b) shows average holdings. HF ROW, HF US, and PD exhibit substantially higher prominence in marginal holdings than in their average holdings across all maturity buckets.

Panel (c) quantifies this gap by reporting the ratio of marginal to average holdings by sector and maturity. HF ROW, HF US, and PD show high trading activity relative to their average holdings, a

characteristic indicative of arbitrage. This distinction is obscured at the aggregate level, as shown by the much smaller aggregate ratio in the first row of Panel (c).

Table A1. **Marginal Holders of U.S. Treasuries**

Panel (a) reports the marginal holders of U.S. Treasuries that we obtain by regressing percentage changes in holdings as a fraction of total outstanding (TAO) on the contemporaneous percentage changes in TAO. Panel (b) reports the average fraction of TAO held by each sector over our sample period. Panel (c) reports the ratio between Panel (a) and Panel (b), with an additional column "Avg. Abs. Ratio" that calculates the average of absolute ratios across the three maturity buckets. We report results for both the aggregate and by maturity bucket. Sectors are U.S. banks (Banks), Federal Reserve (FED), hedge funds outside the U.S. (HF ROW), U.S. hedge funds (HF US), U.S. insurance companies and pension funds (ICPF), mutual funds outside the U.S. (MF ROW), U.S. mutual funds (MF US), U.S. money market funds (MMF US), U.S. and foreign primary dealers (PD), foreign official (Foreign O), foreign private (Foreign P), and other U.S. investors (Other U.S. Investors). The numbers are in percentage points and the quarterly sample period is from 2011Q4 to 2022Q4. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel (a): Marginal Holders (% of outstanding)

	Banks	Fed	HF ROW	HF US	ICPF	MF ROW	MF US	MMF	PD	Other US	Foreign O	Foreign P
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Aggregate	4.4***	39.5***	-1.9	-0.5	1.1	0.1	1.9	30.2***	2.0**	7.4	6.3***	9.5***
$\tau < 1Y$	2.5***	8.0***	12.7***	3.1***	0.3	0.0	0.7**	51.0***	3.4***	-6.8**	8.4***	16.7***
$1Y \leq \tau < 5Y$	7.7***	34.5***	-24.7	-8.6**	-4.5	0.2	9.5**		0.7	30.1	36.1***	19.0**
$\tau \geq 5Y$	1.7	38.2***	15.7**	4.0**	3.4**	0.5**	6.0***		2.8***	25.6**	-1.7	3.9

Panel (b): Average Holders (% of outstanding)

	Banks	Fed	HF ROW	HF US	ICPF	MF ROW	MF US	MMF	PD	Other US	Foreign O	Foreign P
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Aggregate	3.4	19.6	4.4	1.2	2.3	0.3	3.2	5.9	0.7	21.7	27.2	10.2
$\tau < 1Y$	3.7	8.0	4.3	1.2	0.8	0.1	0.7	22.6	0.6	28.8	16.1	13.2
$1Y \leq \tau < 5Y$	3.3	18.9	5.6	1.4	2.4	0.3	3.4		0.9	10.5	46.5	6.7
$\tau \geq 5Y$	3.4	28.6	3.2	0.9	3.3	0.4	4.7		0.5	29.9	13.1	12.0

Panel (c): Marginal/Average Ratio

	Banks	Fed	HF ROW	HF US	ICPF	MF ROW	MF US	MMF	PD	Other US	Foreign O	Foreign P
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Aggregate	1.30	2.01	-0.43	-0.45	0.49	0.28	0.60	5.14	2.87	0.34	0.23	0.93
$\tau < 1Y$	0.67	1.00	2.95	2.56	0.41	0.14	1.03	2.25	6.10	-0.24	0.52	1.27
$1Y \leq \tau < 5Y$	2.32	1.83	-4.37	-6.01	-1.87	0.56	2.82		0.72	2.87	0.78	2.83
$\tau \geq 5Y$	0.50	1.34	4.88	4.38	1.03	1.14	1.27		5.65	0.86	-0.13	0.33
Avg. Abs. Ratio	1.17	1.39	**4.07**	**4.32**	1.10	0.61	1.71	2.25	**4.15**	1.32	0.48	1.47

B.2. Short Positions in U.S. Treasuries

Table A2 reports the fraction of sample periods in which each sector held short positions in U.S. Treasuries, by maturity bucket. Arbitrageurs (broker-dealers and hedge funds) account for nearly all observed shorting activity, and the frequency rises with maturity. Granular-demand investors

essentially never go short. Other U.S. Investors register short positions only in the medium maturity bucket, but as this is a residual sector we cannot rule out that it contains arbitrageurs.

Table A2. **Short positions in U.S. Treasuries**

This table reports the fraction of periods in which investors held short positions in U.S. Treasuries, categorized by sector and maturity bucket. The fraction is calculated as the number of periods in which a given sector was short in a specific maturity bucket, divided by the total number of periods in the sample. Arbitrageurs report the fraction of periods in which the aggregate position of foreign hedge funds, domestic hedge funds, and primary dealers is negative. For explanations of other sector abbreviations, refer to the notes of Table A1. The numbers are in percentage points and the quarterly sample period is from 2011Q4 to 2022Q4.

	Banks	**Fed**	**ICPF**	**MF ROW**	**MF U.S.**	**MMF**	**Arbitrageurs**	**Other U.S.**	**Foreign O**	**Foreign P**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\mathbf{1}\{\tau < 1Y\}$	0	0	0	0	0	0	2	0	0	0
$\mathbf{1}\{1Y \leq \tau < 5Y\}$	0	0	0	0	0	0	4	2	0	0
$\mathbf{1}\{\tau \geq 5Y\}$	0	0	0	0	0	0	16	0	0	0

C. Identification of the Instrument

This appendix provides formal support for the instrument. Section C.1 tests demand linearity and documents first-stage relevance. Section C.2 derives simulation-based critical values for the Kleibergen–Paap F-statistic calibrated to our exact data-generating process. Section C.3 reuses the same simulation pipeline to verify the consistency of the IV estimator as the panel time dimension grows. Section C.4 provides robustness using an independent supply-shock instrument.

C.1. Instrument Validity and Relevance

To illustrate the identification of our instrument, we assume a simplified setting of one asset with maturity τ and price $P_t = \frac{1}{(1+y_t)^\tau}$. We also assume one investor and fixed supply S.

Assume that the data-generating process of demand is given by:

$$Z_t = \theta + b_1 y_t + (b_2)'\mathbf{x}_t + (b_3)'\mathbf{Macro}_t + u_t \tag{A2}$$

The predicted demand $\hat{Z}_t = \hat{\theta} + (\hat{b}_2)'\mathbf{x}_t + (\hat{b}_3)'\mathbf{Macro}_t$ is obtained by projecting demand on bond characteristics and macro variables, excluding yield. The pseudo yield $\tilde{y}_t$ is then defined as the yield that would clear the market given $\hat{Z}_t$ and fixed supply S:

$$\hat{Z}_t = \frac{S}{(1+\tilde{y}_t)^\tau} \tag{A3}$$

Solving for $\tilde{y}_t$, we obtain:

$$\tilde{y}_t = \left(\frac{\hat{Z}_t}{S}\right)^{-\frac{1}{\tau}} - 1 = \left(\frac{\hat{\theta} + (\hat{b}_2)'\mathbf{x}_t + (\hat{b}_3)'\mathbf{Macro}_t}{S}\right)^{-\frac{1}{\tau}} - 1 \tag{A4}$$

Plugging back into Equation (A2), we have:

$$Z_t = \theta + b_1 \left(\left(\frac{\hat{\theta} + (\hat{b}_2)'\mathbf{x}_t + (\hat{b}_3)'\mathbf{Macro}_t}{S}\right)^{-\frac{1}{\tau}} - 1 \right) + (b_2)'\mathbf{x}_t + (b_3)'\mathbf{Macro}_t + u_t \tag{A5}$$

Hence, the pseudo yield $\tilde{y}_t$ is a nonlinear (power-function) transformation of the linear combination $\hat{\theta} + (\hat{b}_2)'\mathbf{x}_t + (\hat{b}_3)'\mathbf{Macro}_t$, so it is not collinear with the linear macro controls already included in the demand regression. This nonlinearity is an equilibrium object: macro variables enter predicted demand $\hat{Z}_t$ linearly through the demand regression, but inverting the market-clearing condition converts this linear combination into a nonlinear mapping from macro variables to the pseudo yield. The structural demand equation itself remains linear in yields; the nonlinearity arises solely from the equilibrium price-quantity identity. In our main specification, we also obtain predicted supply based on the IOR and the macro and bond variables, so the denominator of equation (A5) also contains macro and bond variables, adding a further layer of nonlinearity.

Our model is linear throughout: demand is linear in yields, bond characteristics, and macro variables. The instrument's identifying power comes from the nonlinear transformation in the instrument construction via pseudo market clearing, which makes the pseudo yield nonlinear, in contrast with the linear macro and bond controls already included in the demand regression. A concern would arise if the true demand function contained substantial nonlinear terms in either bond or macro variables, because those would generate systematic nonlinear patterns in the residuals u_t, violating the exclusion restriction.

Table A3 addresses this concern directly. We re-estimate the full IV demand system after augmenting the second-stage controls with squared macroeconomic terms (Credit Spread^2, Debt/GDP^2, GDP Gap^2, Core Inflation^2) and squared bond characteristics (Coupon Rate^2, Bid-Ask Spread^2), keeping the pseudo yields as the instruments. If the demand residual u_t contained nonlinear terms in either set correlated with the pseudo yield, controlling for those terms in the second stage would absorb the contaminating variation and shift the IV estimates.

Panel A reports the baseline IV (reproducing Table 2), and Panel B reports the IV with both sets of squared controls added. The own-yield and cross-yield coefficients across Panels A and B are essentially unchanged across all eight sectors, with matched signs and significance levels. The Kleibergen–Paap F-statistic is also stable in the seven sectors with maturity bucket fixed

Table A3. **Demand Elasticities: Baseline IV vs. IV with Squared Macro and Bond Controls**

This table compares the IV estimates of the demand system from Equation (2) under two specifications. Panel A is the baseline IV (identical to Table 2). Panel B re-estimates the same demand system after augmenting the second-stage controls with squared macroeconomic terms (Credit Spread^2, Debt/GDP^2, GDP Gap^2, Core Inflation^2) and squared bond characteristics (Coupon Rate^2, Bid-Ask Spread^2); the instruments (own and other pseudo yields) are unchanged. The dependent variable is the market value of U.S. Treasuries held by sector ι in maturity bucket m at time t, adjusted for GDP potential. "Bond Controls" refers to Coupon Rate and Bid-Ask Spread; "Macro Controls" refers to Credit Spread, Debt/GDP, GDP Gap, and Core Inflation; "Squared Macro Controls" and "Squared Bond Controls" refer to the corresponding quadratic terms. The bottom row group reports three joint Wald χ^2 tests of the squared coefficients in Panel B: the four squared macro terms (4 d.f.), the two squared bond terms (2 d.f.), and all six squared terms jointly (6 d.f.). The quarterly sample period is from 2011Q4–2022Q4. HAC standard errors with optimal lags are reported in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Banks	**ICPF**	**MF ROW**	**MF U.S.**	**MMF**	**Other U.S.**	**Foreign O**	**Foreign P**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Baseline IV								
$y_t(m)$	55.815**	-6.939	6.330**	122.951***	275.646**	181.738	-122.170	56.168
	[25.022]	[11.387]	[3.008]	[40.845]	[116.979]	[195.228]	[110.004]	[109.121]
$y_t(-m)$	-50.822*	7.000	-2.309	-123.275***	-302.708**	-117.855	-42.737	-69.378
	[28.109]	[13.286]	[3.103]	[41.929]	[149.146]	[238.685]	[141.245]	[136.417]
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Maturity Bucket FE	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	135	135	135	135	45	135	135	135
KP F-Statistic (first stage)	10.484	10.484	10.484	10.484	25.241	10.484	10.484	10.484
Panel B: IV with Squared Macro and Bond Controls								
$y_t(m)$	53.137**	-7.532	5.988**	116.070***	342.122***	175.229	-137.235	40.015
	[23.367]	[11.452]	[2.683]	[37.020]	[119.186]	[198.316]	[110.145]	[106.907]
$y_t(-m)$	-46.282*	5.014	-4.171	-119.708***	-336.139*	-147.633	5.284	-43.932
	[27.905]	[14.180]	[2.980]	[39.900]	[173.336]	[240.307]	[137.441]	[135.351]
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Squared Macro Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Squared Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Maturity Bucket FE	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	135	135	135	135	45	135	135	135
KP F-Statistic (first stage)	10.227	10.227	10.227	10.227	15.725	10.227	10.227	10.227
Joint tests of squared terms (Panel B only)								
Wald χ^2 (sq macro, 4 d.f.)	3.64	2.63	8.71	5.41	26.70	2.26	0.55	2.91
p-value (sq macro)	0.457	0.622	0.069	0.248	0.000	0.688	0.968	0.573
Wald χ^2 (sq bond, 2 d.f.)	4.50	0.90	5.72	4.53	0.33	0.04	10.27	5.15
p-value (sq bond)	0.106	0.638	0.057	0.104	0.848	0.980	0.006	0.076
Wald χ^2 (all squared, 6 d.f.)	7.38	3.78	15.75	7.82	30.65	2.28	11.43	7.59
p-value (all squared)	0.287	0.706	0.015	0.251	0.000	0.893	0.076	0.270

Table A4. **First Stage IV**

This table shows the first-stage estimates of the IV methodology specified in Equation (2). The dependent variable in Column (1) is $y_t(m)$, the value-weighted yield of maturity bucket m, and in Column (2) is $y_t(-m)$, the value-weighted yield of the other maturity buckets $-m$. We instrument own and other yield using pseudo yields specified in Section 3.1. Additional variables include Coupon Rate, Bid-Ask Spread, an indicator variable if the holdings are in maturity bucket 2 ($\mathbf{1}\{1Y \leq \tau < 5Y\}$), an indicator variable if the holdings are in maturity bucket 3 ($\mathbf{1}\{\tau \geq 5Y\}$), Credit Spread, Debt/GDP, GDP Gap, and Core Inflation. We orthogonalize the coupon and the bid-ask spread with respect to the maturity fixed effects. The quarterly sample period is from 2011Q4–2022Q4. HAC standard errors with optimal lags are reported in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	$\tilde{y}_t(m)$	$\tilde{y}_t(-m)$
	(1)	(2)
$\tilde{y}_t(m)$	0.618***	0.363***
	[0.038]	[0.031]
$\tilde{y}_t(-m)$	0.804***	0.903***
	[0.048]	[0.055]
Coupon Rate	0.780***	-0.110
	[0.128]	[0.118]
Bid-Ask Spread	0.164***	0.019
	[0.045]	[0.050]
$1Y \leq \tau < 5Y$	0.420***	0.060
	[0.079]	[0.058]
$\tau \geq 5Y$	1.343***	-0.474***
	[0.082]	[0.074]
Credit Spread	-0.228	0.005
	[0.142]	[0.127]
Debt/GDP	4.119***	2.344***
	[0.476]	[0.542]
GDP Gap	-0.072***	-0.053***
	[0.022]	[0.020]
Core Inflation	0.243***	0.197***
	[0.042]	[0.035]
Constant	-2.821***	-0.758*
	[0.347]	[0.413]
Observations	135	135

effects (10.48 in Panel A versus 10.23 in Panel B); for MMF, where the regression has only 45 observations, the statistic falls from 25.24 to 15.73. In general, this stability indicates that demand is well approximated as linear in both macro variables and bond characteristics.

The bottom row group reports three joint Wald χ^2 tests of the squared coefficients in Panel B. The squared-macro test (4 d.f.) is significant at the 5% level only for MMF ($p < 0.001$). The squared-bond test (2 d.f.) is significant only for Foreign O ($p = 0.006$). The full six-term test (6 d.f.) is significant for MMF ($p < 0.001$) and MF ROW ($p = 0.015$), and marginal for Foreign O ($p = 0.076$). The corresponding elasticity remains stable in sign and order of magnitude across the two panels: the MMF own-yield coefficient strengthens rather than weakens; the MF ROW own-yield coefficient stays positive; and the Foreign O elasticities are statistically insignificant in both panels, so the rejection in the squared test carries no implication for the reported elasticities. The qualitative conclusions on demand elasticities are unaffected.

Further support comes from the robustness checks in Appendix D.3, where demand elasticity estimates are stable across alternative pseudo-yield constructions that vary which macro variables enter the instrument (Table A15), and from the subsample stability in Appendix F.3.

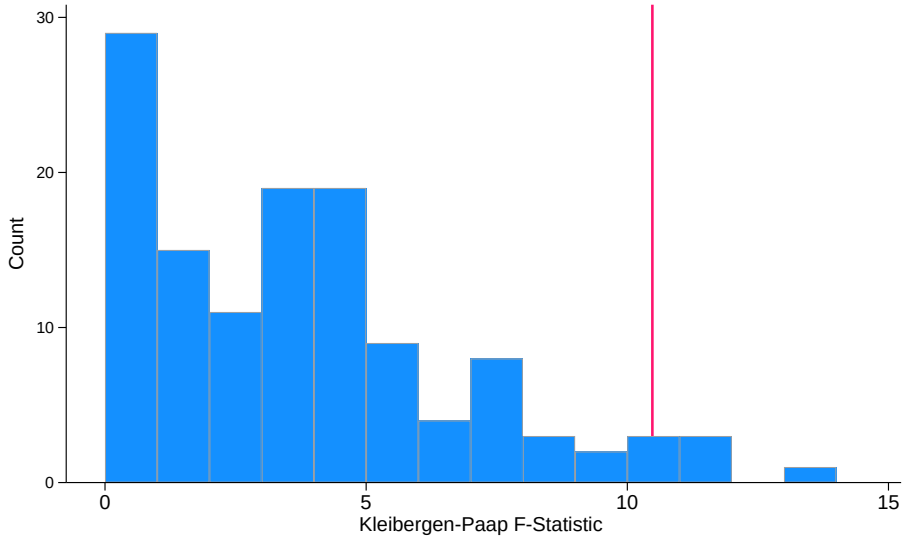
We check the relevance of our IV and show the first-stage results in Table A4. The first stage is strong, with a high R^2. Figure 3 in the main text shows the scatter plot of the residualized own yield against the residualized pseudo yield, confirming that the instrument provides significant explanatory power beyond the linear macroeconomic baseline.

We further validate the instrument by examining what happens when sectors are omitted from the pseudo-yield construction. The instrument is built from pseudo market clearing: both demand (summed across all granular-demand investors and the Fed) and supply are regressed on bond characteristics and macroeconomic variables, and the fitted values are inverted to obtain the pseudo yield. This procedure requires capturing the full market. When a portion of the investor base is omitted, the remaining demand no longer reflects the full set of bond characteristics and macro demand forces driving yields, so the pseudo yield loses its identifying power and the first stage weakens.

To document this formally, we randomly draw all possible combinations of four sectors to exclude from the pseudo-yield construction and re-estimate the first stage for each combination. When dropping sectors, we proportionally scale back Treasury supply by the market share of the excluded sectors. For instance, dropping sectors that together account for 40% of market holdings also reduces the supply measure by 40%. This ensures that the supply-demand balance is preserved mechanically, so any deterioration in first-stage strength reflects a failure to capture the true demand forces, not a spurious imbalance. Figure A1 shows the distribution of Kleibergen–Paap F-statistics across all combinations. The first stage fails in most cases: most combinations

yield a weak instrument, with F-statistics far below conventional thresholds. This confirms that instrument relevance depends on accurately capturing the full set of demand forces, and is not a mechanical artifact of the construction.

Figure A1. **First-Stage Strength When Dropping Four Sectors.** This figure shows the distribution of Kleibergen–Paap (KP) F-statistics from re-estimating the first stage across all combinations of four sectors excluded from the pseudo-yield construction. The red vertical line indicates the KP F-statistic of our main instrument, as in Table 2. Both demand and supply are regressed on bond characteristics and macro variables; when sectors are dropped, supply is proportionally scaled back by their market share to preserve the supply-demand balance mechanically. The instrument nonetheless loses relevance in most combinations, confirming that first-stage strength requires capturing all demand sources.



C.2. Simulation-Based Instrument Strength Testing

How large does the Kleibergen–Paap (KP) F-statistic need to be before we can trust our 2SLS estimates? The standard benchmarks do not answer this question for our setting.

- The Stock and Yogo (2005) critical values assume i.i.d. errors and a single endogenous regressor. We have serially correlated quarterly data, two endogenous variables ($y_t(m)$ and $y_t(-m)$), and their threshold of 10 has no formal justification here.
- The Montiel Olea and Pflueger (2013) effective F thresholds extend to HAC errors but are derived for a single endogenous regressor ($p = 1$). Our exactly identified system has $p = 2$ endogenous variables and $k = 2$ instruments. Lewis and Mertens (2025) generalize this to

$p \geq 2$, but their worst-case critical values remain conservative relative to our setting (see below).

- All three frameworks assume observed instruments. Our pseudo yields are generated instruments, constructed from first-step OLS regressions of aggregate demand and supply.

We therefore follow the same logic as these papers and compute simulation-based critical values calibrated to our exact data-generating process. The idea is to vary instrument strength from very weak to very strong within a calibrated data-generating process, map each strength level to the resulting KP F-statistic and 2SLS bias, and read off the KP value at which the bias falls to 10% of the OLS endogeneity bias. The resulting thresholds are specific to our setting, with $p = 2$ endogenous variables, HAC errors, and the estimated covariance structure.

Simulation Design

We want to find the KP F-value at which the 2SLS Nagar bias equals a given fraction c of the OLS bias. Define:

$$\text{Nagar ratio} = \frac{|\text{median}[\hat{b}_{2SLS} - b^*]|}{|\text{median}[\hat{b}_{OLS} - b^*]|} = \frac{|\text{2SLS median bias}|}{|\text{OLS median bias}|}, \tag{A6}$$

where b^* is the true structural parameter. A Nagar ratio of $c = 0.10$ means the instruments remove 90% of the endogeneity bias.

The pseudo yield is constructed from aggregate demand (equation (5)), not from any single sector's demand. The first stage, which determines instrument strength, reflects how well the aggregate demand-supply equilibrium affects actual yields. To calibrate a data-generating process that is internally consistent with the instrument construction, we therefore use the *aggregate demand equation*,

$$\sum_{j \in \mathscr{J}} Z_t^j(m) = \alpha_0 + \alpha_1 y_t(m) + \alpha_2 y_t(-m) + \delta' W_t(m) + u_t(m), \tag{A7}$$

as the structural equation for the simulation. The first stage relates the two endogenous variables to the two pseudo yields and the controls:

$$\begin{pmatrix} y_t(m) \\ y_t(-m) \end{pmatrix} = \Pi' \begin{pmatrix} \tilde{y}_t(m) \\ \tilde{y}_t(-m) \end{pmatrix} + \Delta' W_t(m) + V_t(m),$$

where Π is the 2×2 matrix of coefficients on the excluded instruments and $V_t(m) = (V_{1t}, V_{2t})'$ collects the first-stage residuals (Table A4). Together, these give us $\hat{\Pi}$, the structural coefficients

(α_1, α_2), and the joint distribution of (u_t, V_{1t}, V_{2t}), where u_t is the structural residual. The correlation $\text{corr}(u, V)$ determines the endogeneity.

Following the approach used to derive critical values in the weak-instrument literature (Stock and Yogo 2005; Montiel Olea and Pflueger 2013; Lewis and Mertens 2025), we trace out the relationship between instrument strength and bias by varying the first-stage coefficient matrix over a range of strengths:

$$\Pi(\lambda) = \lambda \cdot \hat{\Pi}, \qquad \lambda \in (0, 1].$$

At $\lambda = 1$ the data-generating process matches the data; at $\lambda < 1$ the instruments are progressively weakened while everything else (the endogeneity structure, error distributions, exogenous variation) is held fixed. This generates a mapping from instrument strength to bias, from which we read off the minimum KP F-statistic needed to ensure that the 2SLS bias is acceptably small.

For each λ on a grid from 0.05 to 1.00 (steps of 0.025 up to 0.50, then 0.05 up to 1.00), we draw $S = 10{,}000$ simulated datasets:

1. **Block-resample residuals.** Draw the joint residuals (u_t, V_{1t}, V_{2t}) by moving-block bootstrap with block length $\ell = 3$, within each panel group (maturity bucket). The block length matches the HAC bandwidth used in our main estimation (bw = 2.5, i.e., 3 lags). This preserves serial correlation, the u–V correlation that generates endogeneity, and the panel structure.

2. **Construct simulated endogenous variables.** The pseudo yields $\tilde{y}_t$ are held fixed at their original-data values; there is no need to re-run the pseudo market-clearing inversion because instrument strength is controlled directly through λ. Using the first-stage equation with scaled coefficients,

$$Y_{jt}^{(s)} = \tilde{y}_t'[\lambda \hat{\Pi}_{\cdot j}] + W_t' \hat{\Delta}_{\cdot j} + V_{jt}^{(s)}, \qquad j = 1, 2,$$

   where $\tilde{y}_t = (\tilde{y}_t(m), \tilde{y}_t(-m))'$ is the vector of pseudo yields, $\hat{\Pi}_{\cdot j}$ is the j-th column of $\hat{\Pi}$ (the instrument coefficients for the j-th endogenous variable), $\hat{\Delta}_{\cdot j}$ is the estimated control coefficients from the j-th first-stage regression, and W_t collects the controls (also fixed at their original-data values).

3. **Construct simulated dependent variable.** Using the structural equation (A7),

$$\text{dep}_t^{(s)} = \hat{\alpha}_1 Y_{1t}^{(s)} + \hat{\alpha}_2 Y_{2t}^{(s)} + W_t' \hat{\delta} + u_t^{(s)}.$$

4. **Compute statistics.** On each simulated dataset, compute the KP F-statistic, the 2SLS

estimate of α_1, and the OLS estimate of α_1.

Computing the Threshold

Let α_1^* denote the true parameter value used in the data-generating process (the 2SLS estimate of α_1 from equation (A7)). At each λ:

- 2SLS bias $=$ median$_s$ $(\hat{\alpha}_1^{(s),2SLS} - \alpha_1^*)$. We use the median rather than the mean to measure central tendency of the bias, following Lewis and Mertens (2025), who recommend median bias for exactly-identified models ($k = p$). The median is robust to extreme 2SLS draws in the weak-instrument region without requiring an arbitrary winsorization cutoff.
- OLS bias $=$ median$_s$ $(\hat{\alpha}_1^{(s),OLS} - \alpha_1^*)$ evaluated at full instrument strength ($\lambda = 1$). Because OLS does not use the instruments, its bias is unchanged by λ, so we fix the denominator at $\lambda = 1$.
- Nagar ratio $= |\text{2SLS bias}|/|\text{OLS bias}|$.

Threshold concept. At each λ, the KP F-statistic varies across simulation draws. Following the size-control logic of Stock and Yogo (2005), we use the 90th percentile of the KP distribution to define the critical value. Specifically, for a given bias target c (e.g., 10%), we find the λ at which the Nagar ratio crosses c and read off the 90th percentile of the KP distribution at that λ. This KP F-value is the critical value KP^*: it is the minimum KP F-statistic such that, for any data-generating process with Nagar ratio at most c, a KP realization below KP^* would occur no more than 10% of the time. This mirrors the logic of Stock and Yogo (2005), who set critical values so that under the boundary data-generating process the false-rejection rate is controlled at a given α.

Results: Simulation-Based Critical Values

Table A5 reports the simulation results. As instruments strengthen, the 90th percentile of the KP distribution rises and the Nagar ratio falls sharply: by $\lambda = 0.35$ the Nagar ratio is already below 5.1%, and for $\lambda \geq 0.40$ it is negligible ($< 2\%$). The 90th percentile of KP is roughly 2–3 times the median at each λ.

Figure A2 plots the relationship. As instrument strength increases (higher λ), the KP F-statistic rises and the Nagar ratio falls.

The simulation-based critical value at the 10% Nagar bias level is:

$$\text{Size-controlled } (\alpha = 10\%): \quad KP^* = 7.28. \tag{A8}$$

Table A5. **Simulation-Based KP-to-Bias Mapping**

$S = 10{,}000$ simulations per λ. Data-generating process is calibrated from the aggregate demand equation ($N = 135$). "Median KP" and "90th pctile KP" are the 50th and 90th percentiles of the KP distribution across simulations at each λ. Nagar ratio = $|\text{median}_s[\hat{\alpha}_1^{(s),2SLS} - \alpha_1^*]|/|\text{median}_s[\hat{\alpha}_1^{(s),OLS} - \alpha_1^*]|$, where median_s denotes the median across simulation draws. Smaller λ implies weaker instruments in the simulation. Block length $\ell = 3$ (matching the HAC bandwidth in our main estimation).

λ	Median KP	90th pctile KP	Nagar ratio
0.20	0.9	4.3	0.467
0.225	1.1	4.9	0.350
0.25	1.4	5.6	0.262
0.275	1.7	6.2	0.185
0.30	2.0	6.9	0.120
0.325	2.4	7.6	0.084
0.35	2.8	8.4	0.051
0.375	3.2	9.2	0.027
0.40	3.7	10.0	0.012
0.425	4.2	10.9	0.001
0.45	4.7	11.9	0.006
0.50	5.8	13.8	0.012
0.60	8.4	18.2	0.013
0.70	11.5	23.2	0.012
0.80	15.1	28.9	0.011
0.90	19.2	35.1	0.010
1.00	23.8	42.2	0.009

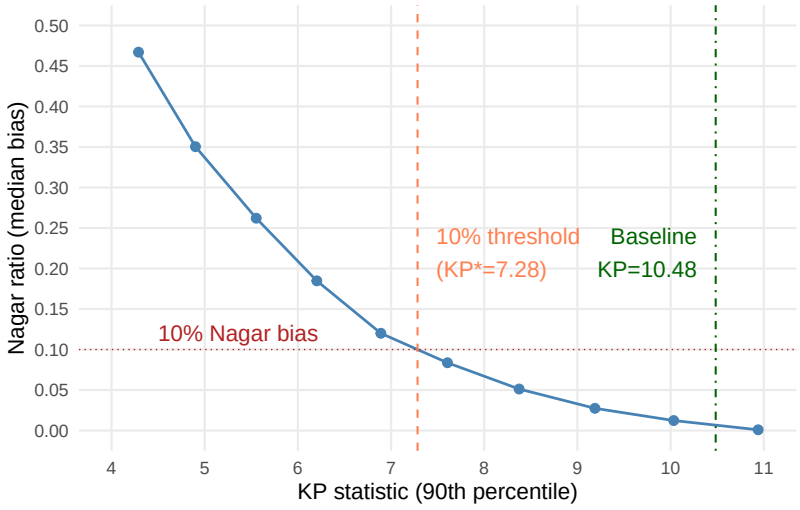


Figure A2. **Simulation-Based KP-to-Nagar Ratio Mapping**

Notes: Each point corresponds to one value of λ ($\lambda \geq 0.20$). The curve plots the *90th percentile* of the KP distribution across $S = 10{,}000$ simulations at each λ against the Nagar ratio (median 2SLS bias divided by median OLS bias at full instrument strength). The horizontal dotted line marks the 10% bias target. The dashed vertical line marks the size-controlled 10% threshold ($KP^* = 7.28$); the dot-dashed vertical line marks the baseline KP from the original sample (10.48). The x-axis is truncated at 11 because the Nagar ratio is negligible beyond that point.

This is a conservative threshold: the median-based critical value (the KP at which the typical data-generating process has a 10% Nagar ratio) is only 2.21. We use the 90th percentile to ensure that our conclusion is robust to sampling variability in the KP F-statistic itself.

Our baseline KP of 10.48 (Table 2) exceeds the size-controlled threshold (7.28). We can therefore formally reject that the median Nagar bias exceeds 10% at the $\alpha = 10\%$ significance level.

For comparison, the Lewis and Mertens (2025) generalization of Montiel Olea and Pflueger (2013) to $p \geq 2$ endogenous variables gives a critical value of 22.15 (10% worst-case median bias, $\alpha = 10\%$) for our specification.[4] Our baseline KP of 10.48 falls short of that threshold, which would lead to a failure to reject the null of weak instruments under the Lewis–Mertens framework. However, their critical value is a worst-case bound over *all* possible data-generating processes, whereas ours is calibrated to the specific covariance structure of the estimated model. The Stock and Yogo (2005) critical values for $p = 2$, $k = 2$ are available only for Wald-test size distortion, which is 7.03 for $\alpha = 10\%$. Our simulation-based size-controlled threshold (7.28) is close to that Stock–Yogo benchmark, which provides reassurance that our procedure yields sensible magnitudes.

[4]Computed using `weakivtest2` after `ivreg2` on the aggregate demand equation with Bartlett kernel, bw = 2.5; the test statistic is $g_{\min} = 10.42$. Because the model is exactly identified ($k = p = 2$), the Lewis–Mertens test is for median bias rather than mean bias.

C.3. Consistency of the IV Estimator

A separate question from the weak-instrument question above is whether the IV estimator is *consistent* as the sample size grows. We address this directly by reusing the simulation pipeline of Appendix C.2 but holding instrument strength fixed at its calibrated value ($\lambda = 1$, so the true Π equals the OLS first-stage $\hat{\Pi}$ from the data) and varying the simulated panel time dimension T.

To extend T beyond the observed sample, we take the joint moving-block bootstrap of Appendix C.2 and apply it to the *full* time-varying input vector: pseudo yields, controls, the structural residual, the first-stage residuals, and the per-row control contributions in each equation. All variables are drawn at the same row indices within each maturity bucket (block length 3), preserving the empirical joint distribution as the sample is enlarged. The structural parameters are calibrated from the data IV estimate. For each $T \in \{45, 90, 180, 360, 720\}$ quarters per maturity bucket, we run 2000 replications; within each replicate we simulate the panel, construct the implied yields and dependent variable through the structural equations, and compute the IV estimator in closed form.

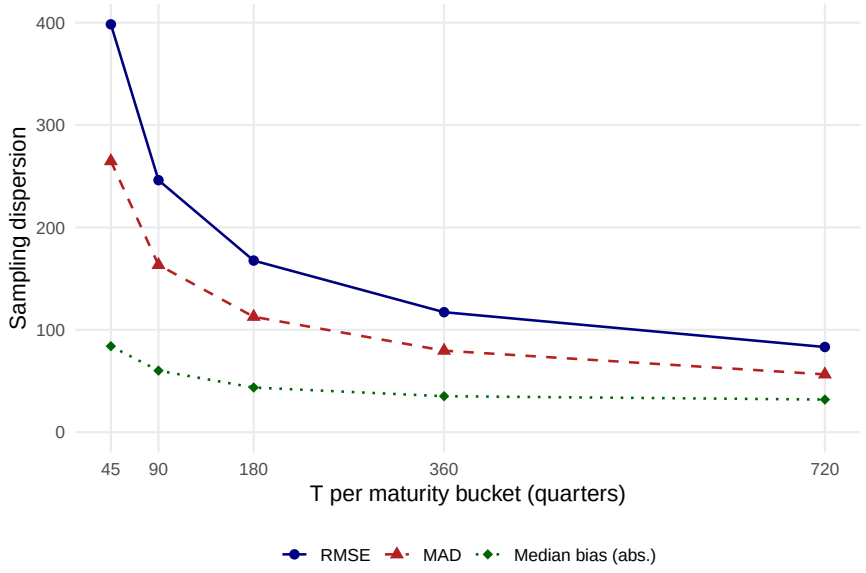


Figure A3. **Sampling distribution of the IV estimator as the panel time dimension T grows.**

Notes: For each T in $\{45, 90, 180, 360, 720\}$ quarters per maturity bucket, $S = 2000$ simulated panels are drawn by joint moving-block bootstrap (block length 3) of the time-varying input vector (pseudo yields, controls, structural residual, first-stage residuals, per-row control contributions), with all series drawn at the same indices within each bucket. On each replicate the IV estimate of the own-yield coefficient is computed in closed form. The plot reports the root-mean-square error (RMSE), median absolute deviation (MAD), and absolute median bias of the estimator against the true value $b_1^* = 1046.7$.

Figure A3 reports the result for the own-yield coefficient. All three statistics decline monotonically as the sample lengthens: the root-mean-square error falls from 398.4 at $T = 45$ to 83.2

at $T = 720$ (log-log slope -0.56), the median absolute deviation declines at slope -0.55, and the absolute median bias of the IV estimator falls from 8.0% of b_1^* at $T = 45$ to 3.0% at $T = 720$. The simulation confirms that sampling variability and finite-sample bias both vanish as T grows, implying consistency of the IV estimator.

C.4. Alternative Instrument

This section lays out an alternative instrument based on supply shocks coming from emergency spending and Treasury auctions. Although supply shocks may be the preferred instrument for Treasury yields in our setting, we use them only as a robustness test because we are unable to find a strong enough first stage. Nevertheless, the findings indicate that our results are robust to alternative instruments.

Instrument design and intuition. Our goal is to generate variation in yields that is plausibly exogenous to sector-specific Treasury demand shocks, while preserving maturity segmentation. The construction uses (i) emergency-spending-driven debt-supply pressure for the short bucket, and (ii) auction-supply shocks for the medium and long buckets.

Step 1: Build the emergency-spending component (bucket 1). We use emergency spending (or appropriations) as an instrument for the short-term bucket ($T \leq 1$). The reason is that we believe most emergency spending is immediate and will (in large part) be financed by issuing T-bills (Greenwood et al. 2015a).[5] A sudden increase in short-term debt is likely to increase short-term interest rates (Krishnamurthy and Vissing-Jorgensen 2011). The exclusion restriction requires that emergency appropriations affect sector-level Treasury demand only through the short-term yield. This is plausible because the timing and magnitude of appropriations are determined by the severity of crisis events (natural disasters, pandemics, wars) and are therefore orthogonal to investors' latent demand. Any broad macroeconomic stress coinciding with such emergencies is absorbed by the macro controls (GDP gap, Debt/GDP) included in the demand regression.

More specifically, we obtain (annual) emergency spending directly from the CBO: `https://www.cbo.gov/system/files/2024-06/2024-04-24b-Supplemental-Appropriations.pdf`. We then take total emergency spending outlays over the current and previous three quarters:

$$EmergencySpending_t = \frac{1}{4}\sum_{j=0}^{3} EmergencySpendingQ_{t-j}.$$

[5]For instance, the Treasury quickly raised trillions of dollars to fund the federal response to COVID-19 and dramatically increased its issuance of bills (`https://www.gao.gov/products/gao-21-606`).

The 4-quarter average captures gradual pass-through into financing needs.

We scale by lagged nominal GDP potential to express the shock in relative GDP units:

$$EmergencyInstrument_t = 100 \times \frac{EmergencySpending_t}{GDPPotential_{t-4}}.$$

Step 2: Treasury auction shock instrument (buckets 2 and 3). We identify high-frequency Treasury supply shocks using the 30-minute price reactions of Treasury futures to auction announcements, following the methodology of Phillot (2025) and Bi et al. (2026).[6] To align these intra-day surprises with our quarterly data frequency, we aggregate the shocks at the quarterly level using the Gertler and Karadi (2015) weighting scheme; this approach allocates each shock between the current and subsequent quarters based on the fraction of the period remaining at the time of the announcement. Phillot (2025) and Bi et al. (2026) use these shocks to study the term structure of interest rates, finding that unanticipated increases in the total volume of issuances shift the yield curve upward across all tenors (2-year to 30-year). The exclusion restriction requires that these auction surprises affect sector-level Treasury demand only through the yield. This is plausible because the shocks are constructed from 30-minute price reactions around auction announcements, capturing the unanticipated component of issuance volume rather than any persistent macroeconomic signal. Sectors' portfolio rebalancing decisions are driven by the resulting yield change, not by the auction size per se, and any correlated macro conditions are again absorbed by the GDP gap and Debt/GDP controls in the demand regression.

Step 3: Map instruments to maturity buckets (own-yield instrument). Let $b \in \{1,2,3\}$ denote maturity buckets (short, medium, long). We define

$$z_{b,t}^{\text{own,raw}} = \begin{cases} \text{EmergencySpending}_t, & b = 1, \\ \text{Supply Shock Maturity T=5}_t, & b = 2, \\ \text{Supply Shock Maturity T=10}_t, & b = 3. \end{cases}$$

Hence, bucket 1 is moved by emergency-spending financing pressure, while buckets 2 and 3 are moved by auction-supply shocks at corresponding maturities.

Because instrument magnitudes differ across buckets, we standardize within each bucket:

$$z_{b,t}^{\text{own}} = \frac{z_{b,t}^{\text{own,raw}} - \mu_b}{\sigma_b},$$

where μ_b, σ_b are the mean and standard deviation of $z_{b,t}^{\text{own,raw}}$ in bucket b.

[6] We thank Huixin, Maxime, and Sarah for sharing the updated data with us.

Step 4: Construct the instrument for other-bucket yield. For each bucket b, define the "other-yield" instrument as the TAO-weighted average of the *other buckets'* own-yield instruments:

$$z_{b,t}^{\text{other}} = \frac{\sum_{b' \neq b} \text{TAO}_{b',t}\, z_{b',t}^{\text{own}}}{\sum_{b' \neq b} \text{TAO}_{b',t}}.$$

We estimate the demand system over the period from 2010Q1 to 2022Q4. This choice increases power because we extend the sample period by two years of data, while still being able to estimate the demand system for four sectors: Banks, MF US, MF ROW, and ICPFs. We do not estimate the demand system for the Fed and MMFs, as the weak-instrument problem becomes too severe when estimating demand functions on a bucket-by-bucket basis.

The first-stage results are reported in Table A6, and the second-stage results are reported in Table A7. Relative to the baseline specification (Table 2), this specification has a much weaker first stage (Kleibergen–Paap $F = 1.69$), so precision is lower. Because the first stage is weak, we also report Anderson–Rubin (AR) p-values for each sector. The AR test remains valid under weak instruments because it does not rely on precise first-stage estimation. Instead, for a given null hypothesis $\beta = \beta_0$, it tests whether the instruments have explanatory power for the transformed outcome $y - \beta_0 x$. Under the null, $y - \beta_0 x$ equals the error term, so the test amounts to checking whether the instruments are uncorrelated with the error.

For Banks (Column 1), the AR p-value is 0.014, allowing us to reject the null of no yield effects at the 5% level. For MF U.S. (Column 4), the AR p-value is 0.074, providing some evidence against the null (significant at the 10% level). For the remaining sectors, the AR test does not reject the null at conventional levels.

Even so, the estimated demand patterns look similar to the baseline: for most sectors, own yield affects holdings positively and cross yield affects holdings negatively.

- **Banks:** We estimate an own-elasticity of 228.2 and a cross-elasticity of -140.4, compared to 55.8 and -50.8 in the baseline specification.
- **MF ROW:** For mutual funds in the rest of the world, we find estimates of 8.9 and -4.1 for own versus cross-elasticity, respectively, mirroring the 6.3 and -2.3 found in the baseline.
- **MF US:** For domestic mutual funds, we find estimates of 162.6 versus -143.0, which align closely with the 123.0 and -123.3 reported in the main specification.
- **ICPF:** For insurance companies and pension funds, we find estimates of -45.2 and 42.4, compared to the earlier estimates of -6.9 and 7.0.

Despite the reduced power of the supply-shock instrument and the longer sample period, the sign patterns and relative magnitudes are consistent across specifications, reinforcing the baseline demand system identification.

Table A6. **First Stage IV: Supply-shock pseudo yields**

First-stage estimates for the specification that instruments own and other yields with pseudo yields constructed from emergency spending and Treasury auction supply shocks (Section C.4). The dependent variables are $y_t(m)$ in column (1) and $y_t(-m)$ in column (2). Controls match the second-stage demand system (bond characteristics, maturity-bucket indicators, and macro variables). Sample: 2010Q1–2022Q4. HAC standard errors with bandwidth 4 in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	$\tilde{y}_t(m)$	$\tilde{y}_t(-m)$
	(1)	(2)
$\tilde{y}_t(m)$	0.076	0.204***
	[0.069]	[0.053]
$\tilde{y}_t(-m)$	0.221***	0.168***
	[0.075]	[0.065]
Coupon Rate	0.700***	-0.556**
	[0.256]	[0.235]
Bid-Ask Spread	0.073	0.103
	[0.090]	[0.093]
$1Y \leq \tau < 5Y$	1.342***	-0.452
	[0.345]	[0.342]
$\tau \geq 5Y$	2.614***	-1.480***
	[0.403]	[0.352]
Credit Spread	0.561	0.586*
	[0.423]	[0.319]
Debt/GDP	-5.882***	-8.828***
	[1.093]	[0.825]
GDP Gap	0.185***	0.133***
	[0.059]	[0.043]
Core Inflation	0.305***	0.274***
	[0.103]	[0.070]
Constant	6.397***	7.715***
	[0.889]	[0.716]
Observations	156	156

Table A7. **Second Stage IV: Supply-shock instrument**

IV estimates of the demand system for Banks, ICPFs, MF ROW, and MF U.S., using emergency spending and auction-based pseudo yields as described in the first stage above. The dependent variable is Treasury holdings at market value (GDP-adjusted). Bond and macro controls are included in all regressions. Sample: 2010Q1–2022Q4. HAC standard errors in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Banks	**ICPF**	**MF ROW**	**MF U.S.**
	(1)	(2)	(3)	(4)
$y_t(m)$	228.162**	-45.147	8.891	162.550*
	[101.232]	[45.611]	[7.603]	[90.290]
$y_t(-m)$	-140.362*	42.384	-4.068	-142.995**
	[79.681]	[33.036]	[5.901]	[71.678]
Constant	-790.125***	-64.294	-47.837***	-209.807
	[218.952]	[90.651]	[14.384]	[195.476]
Bond Controls	Yes	Yes	Yes	Yes
Macro Controls	Yes	Yes	Yes	Yes
Observations	156	156	156	156
KP F-Statistic (first stage)	1.687	1.687	1.687	1.687
Anderson-Rubin p-value	0.014	0.219	0.384	0.074

D. Additional Empirical Analysis

D.1. Demand Regressions for Arbitrageurs

Table A8 reports the demand regression of equation (2) estimated for arbitrageurs (hedge funds and primary dealers). As discussed in Section 3.3, their yield loadings have the opposite sign from those of granular-demand investors, reflecting market-clearing behavior rather than direct response to yields.

D.2. Sector-by-Sector Explanations of Demand Elasticities

Below, we provide more detailed explanations for our empirical findings in Table 2 at the sector level.

Banks

Banks show strong own-yield and negative cross elasticities, indicating that they shift toward maturities offering higher yields. This behavior is consistent with reaching-for-yield as theorized

Table A8. **Demand System Results - Hedge Funds and Primary Dealers**

This table shows the IV estimates of our demand system specified in equation (2). The dependent variable is the market value ($ billion) of U.S. Treasuries held by foreign hedge funds, U.S. hedge funds, or primary dealers in maturity bucket m at time t, adjusted by the ratio of GDP potential at the end of our sample period over the value at current quarter. The endogenous variables are: $y_t(m)$, which is the value-weighted yield of maturity bucket m, $y_t(-m)$, which is the value-weighted yield of the other maturity buckets excluding maturity bucket m. We instrument own and other yield using pseudo yields specified in Section 3.1. Additional variables include Coupon Rate, Bid-Ask Spread, maturity bucket indicators, Credit Spread, Debt/GDP, GDP Gap, and Core Inflation. We orthogonalize the coupon and the bid-ask spread with respect to maturity fixed effects. The quarterly sample period is from 2011Q4–2022Q4. HAC standard errors with optimal lags are reported in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	HF ROW	**HF U.S.**	**PD**
	(1)	(2)	(3)
$y_t(m)$	-200.196**	-48.919**	-21.067
	[90.398]	[22.594]	[17.449]
$y_t(-m)$	219.281**	55.702**	30.285
	[104.035]	[25.967]	[19.729]
Coupon Rate	13.399	-8.041	3.659
	[144.796]	[35.798]	[25.496]
Bid-Ask Spread	52.264	8.078	16.490**
	[48.819]	[13.754]	[6.559]
$1Y \leq \tau < 5Y$	301.888***	71.078***	51.957***
	[63.638]	[14.452]	[12.465]
$\tau \geq 5Y$	389.774**	97.607**	49.054
	[168.472]	[41.625]	[33.644]
Credit Spread	83.362	2.942	24.636
	[96.298]	[25.409]	[16.969]
Debt/GDP	162.261	-36.320	162.411***
	[291.075]	[77.578]	[48.236]
GDP Gap	7.494	1.920	-0.565
	[14.897]	[3.804]	[3.000]
Core Inflation	-11.594	-8.555	-11.479***
	[26.047]	[6.938]	[3.485]
Constant	-141.338	57.126	-121.492**
	[292.662]	[73.868]	[51.509]
Observations	135	135	135
KP F-Statistic (first stage)	10.484	10.484	10.484

in Hanson and Stein (2015): when longer maturities yield more, banks expand duration to enhance returns. Banks face minimal regulatory barriers to this behavior since all Treasuries qualify as high-quality liquid assets (HQLA) under liquidity coverage rules. While they are subject to interest rate risk, many hold Treasuries in held-to-maturity accounts or manage risk with hedging. Their behavior deviates from pure preferred habitat, instead reflecting yield-seeking optimization.

Insurance Companies and Pension Funds (ICPFs)

ICPFs display limited substitution across maturities. Both own- and cross-elasticities are imprecisely estimated and economically small, pointing to relatively weak price sensitivity in their demand. One plausible interpretation is that liability considerations, particularly the need to manage long-duration obligations, play an important role in line with the preferred-habitat framework (Vayanos and Vila 2021). While reaching-for-yield has been documented in insurers' corporate bond portfolios (Becker and Ivashina 2015), their Treasury holdings appear comparatively stable across maturities, consistent with a role for liability management alongside other considerations. This suggests that ICPFs' Treasury demand may partly reflect structural features of their balance sheets, rather than purely valuation-driven motives. Moreover, our findings are broadly consistent with the "hunt-for-duration" channel (Domanski et al. 2017), which predicts that ICPFs may increase demand for Treasuries when yields are low if asset duration falls short of liability duration. In line with this mechanism, the estimated own-yield coefficient for ICPFs is negative, indicating higher demand when yields decline, although the estimate is far from statistically significant.

Mutual Funds (U.S. and ROW)

Mutual funds, particularly U.S. funds, show strong cross-substitution. Table 2 indicates they significantly reallocate in response to relative yield changes. This is in line with active portfolio management and reaching for yield, where fund managers move across maturities to maximize returns. Another possibility is that investors confuse short rates with long rates (Shue et al. 2024): when the short-term rate increases, they expect the long-term rate to increase as well and therefore reduce long-term bond holdings.

While some funds face benchmark-based duration mandates (e.g., short-term or intermediate-term bond funds), sector-wide substitution likely reflects shifts across funds and return-chasing investor flows. Mutual funds behave as yield-sensitive investors, and their activity enforces relative pricing across the maturity spectrum.

Money Market Funds (MMFs)

MMFs are constrained by SEC Rule 2a-7 to invest in very short maturities, and they thus cannot directly substitute across the curve. However, MMFs show negative cross-elasticity: their Treasury holdings fall when long-term yields rise. This reflects investor-level substitution. When longer yields become more attractive, investors withdraw funds from MMFs, shrinking MMF demand for T-bills. Hence, MMFs' demand is indirectly yield-sensitive, reflecting their AUM response to relative yields across maturities.

The AUM response to yield could reflect various behavioral reasons. For example, with extrapolative beliefs (Barberis et al. 2015), investors may interpret a rise in short-term interest rates as a signal of a continuing upward trend in rates, leading them to expect further increases. Anticipating future rate increases and associated price declines for long-term bonds, they reduce long-term bond holdings and shift toward T-bills or other cash-like instruments.

Foreign Official Sector

Foreign official institutions (e.g., central banks, reserve managers) exhibit strong preferred-habitat behavior. These holders buy Treasuries for safety/liquidity or exchange-rate management, not to maximize returns, so we expect little substitution across maturities based on yield, as shown in Table 2. They favor shorter maturities (up to 5 or 10 years) to ensure liquidity.

Even if 30-year yields rise substantially, a central bank like the Bank of Japan is unlikely to start buying 30-year bonds, because that introduces too much volatility into their reserves. Foreign official demand is often described as "price inelastic" or "quantity-driven." Bernanke (2005) argued that the large foreign official purchases in the 2000s ("global savings glut") significantly depressed U.S. long-term yields, evidence that the official sector kept buying Treasuries even as yields fell and held them for reasons beyond yield chasing.

Additionally, institutional guidelines for reserve managers often include duration limits and liquidity requirements[7] that favor shorter-maturity Treasuries and limit the substitution with longer-maturity Treasuries. Tabova and Warnock (2021) assemble confidential security-level holdings data and show that official foreign reserve investors are largely price-insensitive and duration-constrained compared to private investors. In their sample (2003–2019), the average duration of foreign official Treasury portfolios was approximately 4 years.

[7] Refer to this IMF document, "Guidelines for Foreign Exchange Reserve Management: Accompanying Document and Case Studies" for more details.

Foreign Private Sector

Foreign private investors consist of a heterogeneous group of sectors, including foreign money market funds, banks, households, etc., but excluding foreign hedge funds and mutual funds. The insignificant elasticities we estimate likely reflect the aggregation of both inelastic investors, such as pension funds and wealthy foreign households who directly invest in Treasuries, and more yield-sensitive sectors, such as foreign banks.

Other U.S. Investors

As shown in Table 2, although money market funds (MMFs) and mutual funds exhibit strong behavioral cross-maturity substitution, the "Other U.S. Investors" sector shows notably muted responsiveness to yield differentials. This residual category, comprising U.S. households, non-profit organizations, and non-financial corporations, likely reflects a distinct investor base with different behavioral dynamics. Unlike MMFs and mutual funds, whose demand aggregates the actions of retail investors subject to return-chasing and extrapolation, direct holders of Treasuries tend to be wealthier households or corporations with more stable, longer-horizon objectives. First, Campbell (2006) shows that wealthier households are less subject to behavioral biases. Second, corporations' Treasury demand often serves purposes such as liquidity reserves, tax optimization, or balance sheet management. Moreover, in direct Treasury holdings, the absence of performance benchmarking reduces salience and reallocation incentives. As a result, the "Other U.S." sector contributes less to yield curve rebalancing compared to MMFs and the mutual fund sector.

D.3. Additional Empirical Tables and Figures

Table A9. **Summary Statistics**

This table provides summary statistics of the main variables of interest: $y_t(m)$, which is the value-weighted yield of maturity bucket m, $y_t(-m)$, which is the value-weighted yield of the other maturity buckets excluding maturity bucket m, Coupon Rate, Bid-Ask Spread, Credit Spread, Debt/GDP, GDP Gap, and Core Inflation.

	mean	sd	min	max
$y_t(m)$	1.400	1.081	0.041	4.291
$y_t(-m)$	1.469	0.902	0.132	4.289
Coupon Rate	2.039	0.883	0.750	4.158
Bid-Ask Spread	0.046	0.028	0.010	0.096
Credit Spread	0.949	0.233	0.550	1.490
Debt/GDP	0.762	0.095	0.654	0.974
GDP Gap	-1.329	1.910	-9.106	1.846
Core Inflation	2.461	1.322	1.173	6.429
Observations	135			

Table A10. **Correlation Table**

This table provides the correlation table of the main variables of interest: $y_t(m)$, which is the value-weighted yield of maturity bucket m, $y_t(-m)$, which is the value-weighted yield of the other maturity buckets excluding maturity bucket m, Coupon Rate, Bid-Ask Spread, Debt/GDP, Credit Spread, GDP Gap, and Core Inflation. We orthogonalize the coupon and the bid-ask spread with respect to the maturity fixed effects.

	$y_t(m)$	$y_t(-m)$	Coupon	Bid-Ask Spread	Credit Spread	Supply	GDP Gap	Core Inflation
$y_t(m)$	1.000	0.583	-0.092	0.019	-0.014	-0.102	0.465	0.404
$y_t(-m)$	0.583	1.000	-0.283	-0.035	-0.026	-0.150	0.552	0.486
Coupon	-0.092	-0.283	1.000	-0.313	0.289	-0.568	-0.397	-0.495
Bid-Ask Spread	0.019	-0.035	-0.313	1.000	-0.075	0.480	0.241	0.010
Credit Spread	-0.014	-0.026	0.289	-0.075	1.000	-0.138	-0.264	-0.013
Supply	-0.102	-0.150	-0.568	0.480	-0.138	1.000	0.171	0.426
GDP Gap	0.465	0.552	-0.397	0.241	-0.264	0.171	1.000	0.575
Core Inflation	0.404	0.486	-0.495	0.010	-0.013	0.426	0.575	1.000

Table A11. **Treasury supply and face value by maturity bucket**

Reduced-form relationships used in the identification discussion in Section 3.2.

	$\tau < 1Y$	$1Y \leq \tau < 5Y$	$\tau \geq 5Y$
	(1)	(2)	(3)
Coupon Rate	146.936	-1161.120***	-803.369***
	[585.558]	[274.071]	[132.100]
Bid-Ask Spread	-6.792	-55.386	-220.573***
	[169.102]	[56.797]	[42.041]
IOR	331.097***	149.211***	-14.906
	[73.855]	[37.113]	[32.393]
Credit Spread	-31.850	-149.419	-271.816**
	[239.135]	[125.011]	[105.918]
Debt/GDP	13231.183***	4388.389***	3613.496***
	[579.710]	[523.506]	[584.991]
GDP Gap	-162.631***	67.922***	4.081
	[36.333]	[20.684]	[18.327]
Core Inflation	-12.647	-31.245	270.380***
	[58.463]	[34.481]	[25.370]
Constant	-5378.321***	4680.724***	3340.696***
	[481.489]	[384.602]	[410.187]
R-squared	0.959	0.948	0.980
Observations	45	45	45

Table A12. **Demand System Results - IV (all controls)**

Full IV estimates corresponding to Table 2, including all bond and macro controls for each sector.

	Banks	**ICPF**	**MF ROW**	**MF U.S.**	**MMF**	**Other U.S.**	**Foreign O**	**Foreign P**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$y_t(m)$	55.815**	-6.939	6.330**	122.951***	275.646**	181.738	-122.170	56.168
	[25.022]	[11.387]	[3.008]	[40.845]	[116.979]	[195.228]	[110.004]	[109.121]
$y_t(-m)$	-50.822*	7.000	-2.309	-123.275***	-302.708**	-117.855	-42.737	-69.378
	[28.109]	[13.286]	[3.103]	[41.929]	[149.146]	[238.685]	[141.245]	[136.417]
Coupon Rate	-129.728***	12.195	-3.595	-111.236**	327.035	87.803	-418.792**	-342.891**
	[34.055]	[16.040]	[3.750]	[50.656]	[435.266]	[296.222]	[169.848]	[165.645]
Bid-Ask Spread	7.288	20.352***	3.053***	12.900	62.247	90.687	-86.743**	-43.246
	[7.677]	[5.059]	[1.145]	[14.701]	[104.071]	[68.674]	[43.115]	[43.545]
$1Y \leq \tau < 5Y$	57.783***	152.821***	13.098***	193.310***		-632.790***	2958.015***	-169.317**
	[15.011]	[4.788]	[1.999]	[24.183]		[113.334]	[88.088]	[77.147]
$\tau \geq 5Y$	-47.423	201.407***	11.061*	68.999		349.469	292.397	69.168
	[44.666]	[22.160]	[5.709]	[74.777]		[403.926]	[218.233]	[215.763]
Credit Spread	10.069	-13.921	0.464	-44.568	-397.419***	323.921*	84.930	-33.416
	[18.987]	[11.798]	[2.510]	[36.180]	[122.998]	[185.154]	[89.453]	[100.881]
Debt/GDP	729.280***	-8.345	46.501***	78.361	6490.224***	1751.992*	-1897.341***	493.919
	[69.776]	[48.986]	[10.009]	[113.740]	[607.060]	[896.989]	[540.332]	[549.832]
GDP Gap	9.257***	-4.007**	1.328***	10.252**	-75.819***	6.898	-4.224	4.589
	[3.358]	[1.726]	[0.444]	[4.761]	[20.920]	[31.812]	[17.128]	[18.499]
Core Inflation	13.497**	1.018	-2.431***	-6.555	-4.572	3.767	-57.698	9.249
	[6.867]	[3.292]	[0.800]	[9.874]	[52.991]	[52.844]	[37.179]	[36.455]
Observations	135	135	135	135	45	135	135	135
KP F-Statistic (first stage)	10.484	10.484	10.484	10.484	25.241	10.484	10.484	10.484

Table A12 extends Table 2 in the main text by reporting all control coefficients. Moving to the bond characteristics, ICPFs and foreign MFs have a higher demand for Treasuries when the bid-ask spreads are high; that is, when Treasuries are less liquid,[8] while foreign official investors reduce their demand at that time. Furthermore, ICPFs have a large demand for long-term Treasuries, while foreign officials have a strong preference for medium-term bonds, highlighting the importance of heterogeneity in maturity preferences across investors. Owing to the investment mandates of MMFs, they operate only in the shortest maturity bucket. Moving to the macro variables, we find that banks, MFs U.S., and MFs ROW increase their demand for Treasuries when the GDP gap is high, while MMFs and ICPFs reduce their demand. Foreign investors reduce their demand for Treasuries when core inflation is high, while banks increase their demand. Finally, we find that Banks, MF ROW, MMFs, and Other U.S. Investors increase demand for Treasuries when debt/GDP is high, while foreign officials heavily reduce their demand in response to a rise in the U.S. debt

[8]Bretscher et al. (2025) find that ICPFs' corporate bond demand has a positive loading on the bid-ask spread. For instance, ICPFs may prefer illiquid assets to keep their solvency positions appearing more stable. However, they find that MFs prefer liquid bonds in the *cross*-section. This finding does not necessarily contradict our result that picks up a preference for liquidity in the *time*-series by removing maturity fixed effects. Our finding should thus be interpreted as foreign MFs having a higher demand for Treasuries when market liquidity declines.

burden, consistent with the trends described in Appendix Table A1.

Table A13. **Demand System Results - Fed (all controls)**

Full IV estimates corresponding to Table 3, including all bond and macro controls for each Fed maturity bucket.

	$\tau < 1Y$	$1Y \leq \tau < 5Y$	$\tau \geq 5Y$
	(1)	(2)	(3)
$y_t(m)$	7.308	280.506	564.069***
	[66.386]	[197.647]	[127.987]
$y_t(-m)$	92.827	-312.573	-514.711***
	[81.503]	[255.812]	[79.403]
Coupon Rate	-22.803	-2422.808***	414.128
	[209.844]	[263.250]	[260.359]
Bid-Ask Spread	198.338***	57.629	-121.403**
	[65.630]	[73.002]	[55.465]
Credit Spread	5.579	59.967	-119.932
	[66.472]	[148.847]	[125.912]
Debt/GDP	3583.463***	41.906	5644.038***
	[209.750]	[785.398]	[914.015]
GDP Gap	-8.355	-27.390*	-40.446*
	[7.339]	[16.046]	[21.051]
Core Inflation	52.808**	-66.182	128.003***
	[26.081]	[49.981]	[35.021]
Observations	45	45	45
KP statistic	25.241	5.214	14.360

Table A13 extends Table 3 in the main text by reporting all control coefficients for the Fed. We find that the Fed reduces demand for long-term bonds when the GDP gap is high, indicating less need to support the economy via QE when the economy is doing well. The Fed significantly expands its Treasury holdings in all maturity buckets when Debt/GDP is higher, suggesting prominent fiscal accommodations by the Fed.

Table A14 shows the OLS counterpart to Table 2 in Section 3.3. The own-yield and cross-yield coefficients retain the same signs as in the IV estimates, but are attenuated toward zero or become more negative, consistent with downward simultaneity bias when yields are not instrumented.

Table A15 replicates Table 2 using pseudo yields constructed from a restricted set of variables (coupon, maturity, GDP gap, and Debt/GDP only, excluding bid-ask spread, credit spread, and core inflation). The own-yield and cross-yield patterns are qualitatively similar to the baseline,

Table A14. **Demand System Results OLS - Granular Demand Investors**

This table shows the OLS estimates of our demand system specified in Equation (2) for granular demand investors. The dependent variable is the market value of US Treasuries held by sector ι in maturity bucket m at time t. The independent variables are: $y_t(m)$, which is the value-weighted yield of maturity bucket m, $y_t(-m)$, which is the value-weighted yield of the other maturity buckets excluding maturity bucket m, Coupon Rate, Bid-Ask Spread, an indicator variable equal to 1 if the holdings are in maturity bucket 2 ($\mathbf{1}\{1Y \leq \tau < 5Y\}$), an indicator variable equal to 1 if the holdings are in maturity bucket 3 ($\mathbf{1}\{\tau \geq 5Y\}$), Credit Spread, Debt/GDP, GDP Gap, and Core Inflation. We orthogonalize the coupon and the bid-ask spread with respect to the maturity fixed effects. The quarterly sample period is from 2011Q4–2022Q4. HAC standard errors with optimal lags are reported in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Banks	**ICPF**	**MF ROW**	**MF U.S.**	**MMF**	**Other U.S.**	**Foreign O**	**Foreign P**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$y_t(m)$	56.676***	-4.351	3.845**	34.136**	26.371	92.795	-203.031***	-91.691*
	[14.325]	[6.016]	[1.574]	[15.409]	[82.959]	[102.895]	[61.419]	[49.132]
$y_t(-m)$	-57.315***	5.909	-0.143	-31.597**	105.763	-8.771	77.017	101.992
	[14.624]	[7.585]	[1.881]	[15.447]	[109.295]	[126.853]	[67.085]	[65.717]
Coupon Rate	-135.101***	10.707	-1.277	-17.017	512.046	195.903	-304.136***	-171.120*
	[23.444]	[12.861]	[2.925]	[33.696]	[390.244]	[228.667]	[114.897]	[100.147]
Bid-Ask Spread	7.806	19.847***	3.413***	24.162*	19.679	100.032	-80.549**	-26.595
	[7.638]	[4.834]	[1.143]	[13.148]	[96.878]	[64.175]	[38.206]	[35.499]
$1Y \leq \tau < 5Y$	58.001***	151.759***	14.013***	224.702***		-602.962***	2983.216***	-118.803*
	[12.694]	[5.180]	[1.626]	[20.429]		[111.183]	[68.496]	[65.110]
$\tau \geq 5Y$	-51.544**	197.389***	15.426***	231.527***		520.076**	456.849***	348.255***
	[23.479]	[12.992]	[2.969]	[26.707]		[219.118]	[114.795]	[103.546]
Credit Spread	11.640	-13.700	-0.022	-65.416**	-205.221	298.822	57.080	-72.705
	[18.426]	[12.313]	[2.458]	[32.228]	[151.589]	[183.543]	[84.304]	[87.447]
Debt/GDP	697.984***	-3.425	47.785***	200.882*	7587.435***	1967.493**	-1590.849***	798.598
	[63.868]	[43.153]	[9.359]	[106.771]	[537.329]	[922.567]	[511.343]	[548.851]
GDP Gap	10.028***	-4.249**	1.406***	11.035***	-69.571***	5.282	-8.555	3.288
	[3.480]	[1.753]	[0.435]	[4.187]	[24.891]	[30.980]	[16.147]	[17.528]
Core Inflation	15.072**	0.410	-2.171***	-1.397	-90.427*	3.924	-63.522*	12.401
	[6.703]	[3.189]	[0.766]	[8.157]	[52.450]	[53.290]	[35.251]	[31.670]
Constant	-327.753***	41.466	-25.053***	-13.218	-4414.299***	-420.218	2112.250***	25.316
	[52.157]	[39.458]	[7.506]	[80.676]	[426.581]	[727.048]	[404.866]	[422.772]
R-squared	0.903	0.914	0.843	0.855	0.946	0.707	0.979	0.537
Observations	135	135	135	135	45	135	135	135

Table A15. **Demand System Results - IV alternative pseudo yield**

This table shows the IV estimates of our demand system specified in Equation (2). The dependent variable is the market value of U.S. Treasuries held by sector ι in maturity bucket m at time t, adjusted for GDP potential. The endogenous variables are: $y_t(m)$, which is the value-weighted yield of maturity bucket m, $y_t(-m)$, which is the value-weighted yield of the other maturity buckets excluding maturity bucket m. We instrument own and other yield using pseudo yields specified in Section 3.1, but we leave out the bid-ask spread, credit spread, and core inflation in determining the pseudo yields. Additional control variables include Coupon Rate, an indicator variable equal to 1 if the holdings are in maturity bucket 2 ($\mathbf{1}\{1Y \leq \tau < 5Y\}$), an indicator variable equal to 1 if the holdings are in maturity bucket 3 ($\mathbf{1}\{\tau \geq 5Y\}$), Debt/GDP, and GDP Gap. We orthogonalize the coupon and the bid-ask spread with respect to the maturity fixed effects. The quarterly sample period is from 2011Q4–2022Q4. HAC standard errors with optimal lags are reported in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Banks	**ICPF**	**MF ROW**	**MF U.S.**	**MMF**	**Other U.S.**	**Foreign O**	**Foreign P**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$y_t(m)$	39.606	7.980	1.996	49.374*	157.087	234.910	39.524	7.323
	[27.052]	[11.206]	[3.607]	[26.894]	[108.499]	[185.324]	[122.990]	[125.415]
$y_t(-m)$	-36.439	-6.809	1.221	-52.425*	-133.407	-54.487	-231.405	-48.012
	[32.467]	[16.654]	[4.190]	[27.970]	[151.585]	[242.372]	[188.371]	[154.924]
Coupon Rate	-123.777***	-3.488	1.684	-46.721	467.716	206.656	-544.344***	-336.763**
	[38.157]	[14.826]	[4.353]	[36.540]	[324.792]	[273.336]	[195.295]	[161.313]
$1Y \leq \tau < 5Y$	63.731***	147.399***	14.717***	219.791***		-662.598***	2902.892***	-149.347*
	[14.847]	[6.020]	[2.496]	[22.532]		[118.162]	[94.033]	[78.450]
$\tau \geq 5Y$	-18.828	174.829***	18.563**	201.327***		305.872	-13.378	145.357
	[48.953]	[25.387]	[7.379]	[51.228]		[375.095]	[265.012]	[255.196]
Debt/GDP	823.030***	35.874	52.458***	218.293***	6857.024***	2438.882***	-2775.537***	367.740
	[92.570]	[50.676]	[11.741]	[81.696]	[410.663]	[857.300]	[676.954]	[535.727]
GDP Gap	15.005***	-3.171	1.105**	13.969***	-68.628***	-23.357	-31.012*	14.525
	[4.200]	[2.130]	[0.505]	[3.925]	[14.758]	[31.404]	[18.074]	[19.945]
Constant	-387.907***	11.337	-35.073***	-67.143	-3997.551***	-564.967	3180.269***	489.729
	[91.686]	[55.311]	[12.694]	[76.871]	[396.292]	[808.369]	[687.926]	[513.279]
Observations	135	135	135	135	45	135	135	135
KP F-Statistic (first stage)	10.552	10.552	10.552	10.552	9.044	10.552	10.552	10.552

Table A16. **Demand System Results - IV with MBS and swap spreads**

IV estimates augmenting macro controls with MBS and swap spread measures.

	Banks	**ICPF**	**MF ROW**	**MF U.S.**	**MMF**	**Other U.S.**	**Foreign O**	**Foreign P**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$y_t(m)$	56.488**	-6.809	6.104**	121.333***	245.520**	187.985	-119.946	54.150
	[24.541]	[11.427]	[2.705]	[37.841]	[115.330]	[190.819]	[112.302]	[107.053]
$y_t(-m)$	-47.585*	8.586	-1.994	-119.339***	-255.962*	-122.014	-49.203	-64.885
	[27.387]	[13.441]	[2.843]	[39.808]	[147.726]	[231.506]	[141.743]	[132.169]
Coupon Rate	-130.582***	12.181	-3.088	-107.334**	539.525	74.476	-424.323**	-338.090**
	[33.176]	[16.259]	[3.400]	[47.019]	[467.493]	[287.182]	[173.253]	[164.703]
Bid-Ask Spread	1.041	16.547***	1.362	-3.772	103.649	125.212	-60.968	-62.914
	[7.634]	[5.385]	[1.084]	[14.337]	[108.747]	[78.370]	[44.886]	[47.188]
$1Y \leq \tau < 5Y$	57.179***	152.614***	13.170***	193.671***		-635.211***	2957.618***	-168.829**
	[14.432]	[5.074]	[1.868]	[22.379]		[110.502]	[87.823]	[75.372]
$\tau \geq 5Y$	-46.869	201.950***	11.511**	72.989		339.077	286.434	73.954
	[43.349]	[22.308]	[5.286]	[69.333]		[391.293]	[221.310]	[208.752]
Credit Spread	-12.038	-26.419**	-4.112**	-91.762**	-355.932***	411.640**	158.848**	-88.712
	[18.360]	[12.664]	[1.954]	[36.234]	[119.887]	[186.676]	[76.186]	[99.357]
Debt/GDP	828.187***	46.018	64.710***	270.536**	6221.131***	1414.920	-2200.269***	718.329
	[78.294]	[58.826]	[10.744]	[114.666]	[626.418]	[1009.319]	[647.172]	[623.882]
GDP Gap	10.604***	-3.358*	1.442***	11.745***	-85.344***	5.589	-6.704	6.284
	[3.443]	[1.720]	[0.411]	[4.481]	[21.476]	[32.898]	[17.485]	[19.456]
Core Inflation	10.824	-0.207	-2.568***	-8.770	5.250	4.184	-53.872	6.792
	[7.091]	[3.313]	[0.737]	[9.366]	[49.989]	[58.455]	[39.548]	[39.021]
MBS Spread (10y)	82.579**	43.184*	11.989**	133.524	-267.438	-202.822	-213.481	154.741
	[41.817]	[25.399]	[5.123]	[85.233]	[341.234]	[412.638]	[204.972]	[214.866]
Swap Spread (10y)	-20.657	-21.035	-17.920***	-158.391***	386.404	415.626	236.475	-190.142
	[37.249]	[28.638]	[3.828]	[49.097]	[309.150]	[391.827]	[207.562]	[181.607]
Constant	-454.611***	-10.445	-42.449***	-58.979	-2854.197***	205.329	2759.655***	143.105
	[78.094]	[58.883]	[10.519]	[127.235]	[643.189]	[903.213]	[608.384]	[547.028]
Observations	135	135	135	135	45	135	135	135
KP F-Statistic (first stage)	10.209	10.209	10.209	10.209	31.651	10.209	10.209	10.209

confirming that the elasticity estimates do not hinge on which macro variables enter the pseudo-yield construction.

Table A16 replicates Table 2 augmenting the macro controls with the MBS spread and the swap spread. The own-yield and cross-yield estimates remain both qualitatively and quantitatively robust, confirming that substitution toward MBS or swaps does not confound the baseline demand patterns.

Table A17 replicates Table 2 omitting the other-yield regressor. As discussed in Section 3.1, excluding other yield attenuates the own-yield coefficient toward zero because own and other yields are positively correlated yet have opposing effects on demand.

Table A17. **Demand System Results - IV no other yield**

This table shows the IV estimates of our demand system specified in Equation (2), excluding other yield. The dependent variable is the market value of US Treasuries held by sector ι in maturity bucket m at time t. The endogenous variable is $y_t(m)$, which is the value-weighted yield of maturity bucket m. We instrument own and other yield using pseudo yields specified in Section 3.1. Additional variables include Coupon Rate, Bid-Ask Spread, an indicator variable equal to 1 if the holdings are in maturity bucket 2 ($\mathbf{1}\{1Y \leq \tau < 5Y\}$), an indicator variable equal to 1 if the holdings are in maturity bucket 3 ($\mathbf{1}\{\tau \geq 5Y\}$), Credit Spread, Debt/GDP, GDP Gap, and Core Inflation. We orthogonalize the coupon and the bid-ask spread with respect to the maturity fixed effects. The quarterly sample period is from 2011Q4–2022Q4. HAC standard errors with optimal lags are reported in brackets; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Banks	**ICPF**	**MF ROW**	**MF U.S.**	**MMF**	**Other U.S.**	**Foreign O**	**Foreign P**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$y_t(m)$	29.256**	-3.281	5.123***	58.530***	106.107**	120.149	-144.503***	19.912
	[12.683]	[5.351]	[1.580]	[19.774]	[44.769]	[90.466]	[50.778]	[47.272]
Coupon Rate	-82.927***	5.749	-1.469	2.285	536.993	196.333	-379.436***	-279.003***
	[21.071]	[12.245]	[2.138]	[37.187]	[424.290]	[177.478]	[87.870]	[87.142]
Bid-Ask Spread	8.037	20.249***	3.087***	14.715	7.274	92.422	-86.113**	-42.225
	[9.504]	[4.983]	[1.142]	[14.319]	[100.676]	[70.673]	[41.294]	[40.005]
$1Y \leq \tau < 5Y$	64.990***	151.828***	13.425***	210.792***		-616.077***	2964.076***	-159.479**
	[16.407]	[4.749]	[1.843]	[24.336]		[109.749]	[76.316]	[68.073]
$\tau \geq 5Y$	11.809	193.249***	13.752***	212.672***		486.826***	342.205***	150.026*
	[20.603]	[10.069]	[2.801]	[33.337]		[169.892]	[92.447]	[81.124]
Credit Spread	-1.886	-12.275	-0.079	-73.566*	-271.817**	296.198	74.877	-49.736
	[20.539]	[12.677]	[2.593]	[37.713]	[131.452]	[184.879]	[85.082]	[92.896]
Debt/GDP	891.673***	-30.710	53.878***	472.264***	7341.533***	2128.578**	-1760.784***	715.603
	[84.893]	[36.241]	[10.275]	[137.595]	[410.340]	[973.281]	[535.301]	[572.242]
GDP Gap	6.237*	-3.591**	1.191***	2.927	-73.917***	-0.104	-6.763	0.467
	[3.531]	[1.679]	[0.404]	[4.932]	[24.656]	[33.720]	[14.369]	[16.990]
Core Inflation	8.253	1.741	-2.669***	-19.274*	-67.512*	-8.393	-62.107	2.091
	[8.562]	[3.521]	[0.858]	[10.335]	[39.645]	[63.044]	[41.505]	[39.899]
Constant	-520.067***	67.051**	-29.948***	-248.786***	-4121.261***	-553.100	2299.574***	161.592
	[65.368]	[27.078]	[7.352]	[95.058]	[277.256]	[700.356]	[399.624]	[415.163]
Observations	135	135	135	135	45	135	135	135
KP F-Statistic (first stage)	89.96	89.96	89.96	89.96	369.21	89.96	89.96	89.96

Figure A4. **U.S. Treasury Holdings of Hedge Funds versus Primary Dealers**. This graph shows the aggregate holdings of U.S. Treasuries (in billions) by hedge funds (domestic and foreign, left y-axis) and primary dealers (right y-axis) over time.

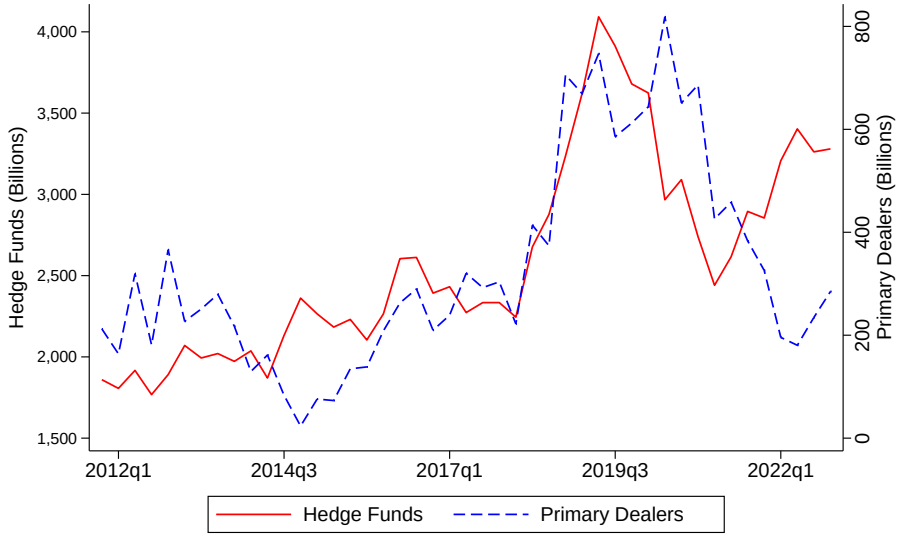
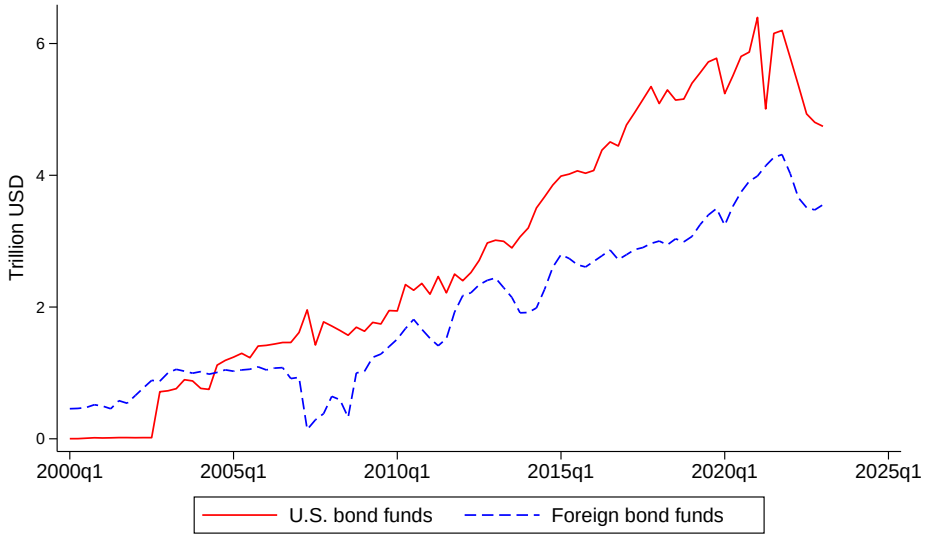


Figure A5. **Morningstar Aggregate Holdings by Domestic and Foreign Bond Funds**. This graph shows the aggregate holdings of U.S. and foreign bond funds in USD (trillions) over time.



E. Model Derivations and Estimation

E.1. A Motivating Model of Treasury Demand

To guide our empirical analysis, we start with a simple model of investor demand for U.S. Treasuries. We index investor groups by ι and denote their portfolio holdings of maturity $\tau \in \{1, \cdots, N\}$ as $Z_t^\iota(\tau)$ and stack all maturities into a vector Z_t^ι. We denote the return on a Treasury with maturity τ as $R_{t+1}^{(\tau)}$, and the risk-free rate as r_t. We allow for flexible beliefs and denote the beliefs of sector ι as $\mathbb{E}^\iota$ in expectations, and $\mathbb{V}^\iota$ in covariances. For the sake of realism, we accommodate that investors' portfolios extend beyond Treasuries and denote the non-Treasury holdings as $\tilde{Z}_t^\iota$ and the associated returns as $\tilde{R}_{t+1}^\iota$.

We model the optimization problem of investor ι with wealth W_t^ι as

$$\begin{aligned} &\max_{Z_t^\iota, \tilde{Z}_t^\iota} \mathbb{E}_t^\iota[W_{t+1}^\iota] - \frac{\gamma^\iota}{2}\mathbb{V}_t^\iota(W_{t+1}^\iota) + \underbrace{V^\iota(Z_t^\iota)}_{\text{non-pecuniary}} \\ &W_{t+1}^\iota = W_t^\iota(1+r_t) + \underbrace{\sum_{\tau=1}^{N} Z_t^\iota(\tau)(R_{t+1}^{(\tau)} - r_t)}_{\text{Treasury returns}} + \underbrace{\tilde{Z}_t^\iota(\tilde{R}_{t+1}^\iota - r_t)}_{\text{outside portfolio return}}, \end{aligned} \tag{A9}$$

where the objective function includes a non-pecuniary component that captures the special attributes of U.S. Treasuries, such as liquidity or safety, as in Krishnamurthy and Vissing-Jorgensen (2012). The non-pecuniary term can also reflect balance-sheet costs of holding cash securities, such as the supplementary leverage regulation on banks. Similarly, the term can represent an inconvenience for certain Treasuries, such as that of short-term Treasuries for pension funds or insurance companies. For tractability, we assume that the derivative of V^ι w.r.t. Z_t^ι is affine in the portfolio choice Z_t^ι,

$$\frac{\partial V^\iota(Z_t^\iota)}{\partial Z_t^\iota} = \bar{V}_0^\iota - \bar{V}^\iota Z_t^\iota. \tag{A10}$$

In the budget equation, the "outside portfolio" can capture institutional features such as the long-duration liabilities of pension funds and insurance companies. Moreover, we allow for heterogeneous risk-aversion γ^ι. Institutions such as money-market funds can be approximated as agents with extremely high γ^ι and thus unable to bear the risk of holding long-term bonds.

Denote the aggregate states of the economy as the vector β_t, and the vector of Treasury yields as $y_t = \left(y_t^{(1)}, y_t^{(2)}, \cdots, y_t^{(N)}\right)'$. We allow for flexible beliefs about asset returns,

$$\mathbb{E}^\iota[R_{t+1}^{(\tau)} - r_t] = \psi^\iota(\tau) \cdot \beta_t + \phi^\iota(\tau) \cdot y_t, \tag{A11}$$

where each sector ι may have different beliefs about how aggregate states affect expected returns. The direct dependence on yields may reflect heuristic expectations, such as yield curve extrapolation, reaching for yield (Hanson and Stein 2015), or confusion between short-term and long-term interest-rate expectations (Shue et al. 2024).

Solving for (A9), we obtain the first-order condition for $Z_t^\iota(\tau)$,

$$\begin{aligned}\psi^\iota(\tau)\cdot\beta_t+\phi^\iota(\tau)\cdot y_t+V_\tau'(Z_t^\iota)=\gamma^\iota\big(&\mathbb{V}^\iota(R_{t+1}^{(\tau)},R_{t+1})Z_t^\iota\\&+\mathbb{V}^\iota(R_{t+1}^{(\tau)},\tilde{R}_{t+1}^\iota)\tilde{Z}_t^\iota\big),\end{aligned}\tag{A12}$$

where we denote the vector of returns as $R_{t+1}=(R_{t+1}^{(1)},R_{t+1}^{(2)},\cdots,R_{t+1}^{(N)})'$. Stacking all the values of $\tau\in\{1,\cdots,N\}$ in (A12) and using the assumption in (A10), we obtain:

$$\begin{aligned}Z_t^\iota=&\left(\mathbb{V}^\iota(R_{t+1},R_{t+1})+\frac{1}{\gamma^\iota}\bar{V}^\iota\right)^{-1}\\&\times\left(\frac{1}{\gamma^\iota}(\psi^\iota\beta_t+\phi^\iota y_t+\bar{V}_0^\iota)-\mathbb{V}^\iota(R_{t+1},\tilde{R}_{t+1}^\iota)\tilde{Z}_t^\iota\right),\end{aligned}\tag{A13}$$

where we define the coefficient matrices $\psi^\iota=(\psi^\iota(1),\cdots,\psi^\iota(N))'$ and $\phi^\iota=(\phi^\iota(1),\cdots,\phi^\iota(N))'$. The outside portfolio covariance term $\mathbb{V}^\iota(R_{t+1},\tilde{R}_{t+1}^\iota)\tilde{Z}_t^\iota$ reflects the risk interaction between Treasuries and other holdings, such as corporate bonds or foreign securities, which may vary by institution. We assume it can be decomposed into a linear function of aggregate states β_t plus "noise", in the same spirit as market microstructure models (Kyle 1985; De Long et al. 1990). The noise term can reflect sector-level idiosyncratic risks, such as pension-specific regulation changes, or erroneous stochastic beliefs as in De Long et al. (1990).

Equation (A13) shows that investor demand can, in principle, be derived from an optimization problem with beliefs, risk preferences, and non-pecuniary portfolio considerations. However, the precise structure of these inputs likely varies widely across investor types and reflects a combination of institutional constraints, regulation, and behavioral forces. In practice, factors such as capital requirements, liquidity mandates, or yield-seeking behavior differ across sectors and are difficult to observe or model directly. Attempting to specify and estimate a fully structural model for each investor type would require strong assumptions and face severe data limitations. Instead, we follow the philosophy of Koijen and Yogo (2019) and estimate sector-level demand functions flexibly from data. More formally, we lump the noise term with the inverse of the matrix in (A13) as a normally distributed vector u_t^ι and express the solution in (A14) as a demand function:

$$Z_t^\iota=\theta_0^\iota+B^\iota y_t-\theta^\iota\beta_t+u_t^\iota.\tag{A14}$$

We note that the model allows for cross-elasticities in that $Z_t^\iota(\tau)$ may depend on $y_t(\tau')$ for $\tau' \neq \tau$, i.e., the τ-th row of B^ι may have a non-zero element at position τ'. According to (A13), this reflects a combination of asset return expectations depending on yields, and covariance across the term structure. Intuitively, the presence of cross-maturity elasticities allows granular investors to rebalance their portfolios toward higher-yielding maturities. Indeed, institutions such as insurance companies, mutual funds, and banks are known to exhibit "yield-seeking" behavior, actively reallocating across fixed-income instruments in pursuit of higher returns (Becker and Ivashina 2015; Hanson and Stein 2015; Choi and Kronlund 2018). However, we emphasize that alternative micro-foundations for cross-elasticities are possible, and our main results do not depend on the specific interpretation of these demand functions.

Arbitrageurs do not have a demand function in the same sense. Their holdings are not driven directly by yields but by the fundamental factors that determine yields: macroeconomic state variables, rational expectations of future returns, and equilibrium supply and demand imbalances. Formally, arbitrageur expected excess returns are rational expectations $\mathbb{E}[R_{t+1}^{(\tau)} - r_t]$, which in equilibrium are functions of β_t, r_t, and latent demand shocks rather than yields per se. This is why a reduced-form demand regression is not a structural representation of arbitrageur behavior, and why we instead solve their portfolio problem explicitly within the equilibrium framework in Section 4.

That said, if one were to run a reduced-form regression of arbitrageur holdings on yields, the estimated coefficients would have the opposite sign to those of granular-demand investors. Arbitrageurs absorb the imbalances that granular investors create: when yields rise and granular investors buy, arbitrageurs reduce their positions, and vice versa. This sign reversal is precisely what we use, along with the reduced-form regression evidence in Section 2, to classify broker-dealers and hedge funds as arbitrageurs rather than granular-demand investors.

The expression for Treasury holdings in (A13) also clarifies how the model accommodates the dependence of the optimal Treasury portfolio on other assets through the outside portfolio. Indeed, the portfolio depends on other assets if other assets' risk exposure comoves with Treasuries. To the extent that the state vector captures risks priced in other assets, innovations to these variables may transmit to Treasury demand fluctuations. For example, including the credit spread in the state vector allows credit market shocks to influence Treasury demand. In such a way, the model also captures substitution between corporate bonds and Treasuries.

Finally, we discuss the Federal Reserve's Treasury demand. Clearly, the Fed is not a profit-maximizing institution. The Fed's demand is driven by its policy decisions, for example, reducing long-term interest rates through its QE program. We find it useful to describe the Fed's Treasury demand also in the form of (A14), as this provides a flexible way to capture the Fed's policy-driven behavior.

E.2. Derivations for the Full Model

As noted, we conjecture an affine solution of the model of the form (21). In order to solve the model, we need to pin down the matrices A, A_r, and A_u, as well as the vector C. We next outline the critical steps in the model solution.

We start with the holding return of bonds with maturity τ from t to $t+1$, using (10) and (21),

$$\begin{aligned}
r_{t+1}^{(\tau)} &= p_{t+1}^{(\tau-1)} - p_t^{(\tau)} \\
&= A(\tau-1)'\beta_{t+1} + A_r(\tau-1)r_{t+1} - A(\tau)'\cdot\beta_t - A_r(\tau)r_t + A_u(\tau-1)'u_{t+1} - A_u(\tau)'u_t \\
&= A(\tau-1)'(\bar{\beta} + \Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}) \\
&\quad + A_r(\tau-1)(\bar{r} + \phi_r'(\Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}) + \rho_r r_t + \sigma_r \varepsilon_{t+1}^r) \\
&\quad - A(\tau)'\cdot\beta_t - A_r(\tau)r_t + A_u(\tau-1)'u_{t+1} - A_u(\tau)'u_t + C(\tau-1) - C(\tau).
\end{aligned} \tag{A15}$$

We can approximate the total holding return as

$$R_{t+1}^{(\tau)} = \exp(r_{t+1}^{(\tau)}) - 1 \approx r_{t+1}^{(\tau)} + \frac{1}{2}\mathbb{V}_t[r_{t+1}^{(\tau)}], \tag{A16}$$

which becomes exact when we take a continuous-time approach. Refer to Greenwood et al. (2024) for a more detailed discussion. Since there is no uncertainty regarding the current short rate, this approximation also leads to $R_{t+1} = R_{t+1}^{(1)} = \exp(r_t) - 1 \approx r_t$.

With (A15) and (A16), we can express the total return as

$$\begin{aligned}
R_{t+1}^{(\tau)} &= A(\tau-1)'(\bar{\beta} + \Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}) - A(\tau)'\cdot\beta_t + C(\tau-1) - C(\tau) \\
&\quad + \frac{1}{2}\left(A(\tau-1)' + A_r(\tau-1)\phi_r'\right)\Sigma\left(A(\tau-1) + \phi_r A_r(\tau-1)\right) \\
&\quad + A_r(\tau-1)(\bar{r} + \phi_r'(\Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}) + \rho_r r_t + \sigma_r\varepsilon_{t+1}^r) - A_r(\tau)r_t \\
&\quad + \frac{1}{2}(A_r(\tau-1)\sigma_r)^2 + A_u(\tau-1)'u_{t+1} - A_u(\tau)'u_t + \frac{1}{2}A_u(\tau-1)'\Sigma^u A_u(\tau-1).
\end{aligned} \tag{A17}$$

We note that the return $R_{t+1}^{(\tau)}$ in (A17) contains four important components. The first one reflects innovations to the macroeconomic factors β_t. The second one reflects innovations to latent demand u_t. The third one is the innovation to the monetary policy rate r_t. The final components are the Jensen terms for each type of risk, including the macroeconomic shocks, monetary policy shocks, and latent demand shocks.

To simplify expressions, we denote

$$\hat{A}(\tau-1) = A(\tau-1) + \phi_r A_r(\tau-1), \tag{A18}$$

so that $\hat{A}(\tau-1)' = A(\tau-1)' + A_r(\tau-1)\phi_r'$. Therefore, Equation (A17) can be simplified as

$$\begin{aligned} R_{t+1}^{(\tau)} &= A(\tau-1)'(\bar{\beta} + \Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}) - A(\tau)' \cdot \beta_t + C(\tau-1) - C(\tau) \\ &\quad + \frac{1}{2}\hat{A}(\tau-1)'\Sigma\hat{A}(\tau-1) + A_u(\tau-1)'u_{t+1} - A_u(\tau)'u_t + \frac{1}{2}A_u(\tau-1)'\Sigma^u A_u(\tau-1) \\ &\quad + A_r(\tau-1)(\bar{r} + \phi_r'(\Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}) + \rho_r r_t + \sigma_r \varepsilon_{t+1}^r) \\ &\quad\quad - A_r(\tau) r_t + \frac{1}{2}(A_r(\tau-1)\sigma_r)^2. \end{aligned} \tag{A19}$$

Wealth thus evolves as (using $R_t \approx r_t$ from (A16))

$$\begin{aligned} W_{t+1} &= W_t(1+r_t) + \sum_{\tau=2}^{N} X_t(\tau)(R_{t+1}^{(\tau)} - r_t) + \tilde{X}_t(\tilde{R}_{t+1} - r_t) \\ &= W_t(1+r_t) + \tilde{X}_t(\tilde{R}_{t+1} - r_t) \\ &\quad + \sum_{\tau=2}^{N} X_t(\tau) \begin{pmatrix} A(\tau-1)'(\bar{\beta} + \Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}) - A(\tau)' \cdot \beta_t + \frac{1}{2}\hat{A}(\tau-1)'\Sigma\hat{A}(\tau-1) \\ + A_u(\tau-1)'u_{t+1} - A_u(\tau)'u_t + \frac{1}{2}A_u(\tau-1)'\Sigma^u A_u(\tau-1) \\ + A_r(\tau-1)(\bar{r} + \phi_r'(\Phi(\beta_t - \bar{\beta}) + \Sigma^{1/2}\varepsilon_{t+1}) + \rho_r r_t + \sigma_r \varepsilon_{t+1}^r) - A_r(\tau) r_t \\ + C(\tau-1) - C(\tau) + \frac{1}{2}(A_r(\tau-1)\sigma_r)^2 - r_t \end{pmatrix} \\ &= W_t(1+r_t) + \sum_{\tau=2}^{N} X_t(\tau) \begin{pmatrix} A(\tau-1)'\left(\bar{\beta} + \Phi(\beta_t - \bar{\beta})\right) - A(\tau)'\beta_t + \frac{1}{2}\hat{A}(\tau-1)'\Sigma\hat{A}(\tau-1) \\ - A_u(\tau)'u_t + \frac{1}{2}A_u(\tau-1)'\Sigma^u A_u(\tau-1) + C(\tau-1) - C(\tau) \\ + A_r(\tau-1)(\bar{r} + \phi_r'\Phi(\beta_t - \bar{\beta}) + \rho_r r_t) - A_r(\tau) r_t + \frac{1}{2}(A_r(\tau-1)\sigma_r)^2 - r_t \end{pmatrix} \\ &\quad + \left(\sum_{\tau=2}^{N} X_t(\tau)\left(A(\tau-1)'\Sigma^{1/2} + A_r(\tau-1)\phi_r'\Sigma^{1/2}\right) + \tilde{X}_t\tilde{\sigma}'\right)\varepsilon_{t+1} + \left(\sum_{\tau=2}^{N} X_t(\tau)A_r(\tau-1)\sigma_r + \tilde{X}_t\tilde{\sigma}_r\right)\varepsilon_{t+1}^r \\ &\quad + \left(\sum_{\tau=2}^{N} X_t(\tau)A_u(\tau-1)'\right)u_{t+1} + \tilde{X}_t(\tilde{\phi}'\beta_t + \tilde{\phi}_r r_t - r_t). \end{aligned} \tag{A20}$$

To simplify notation, it is convenient to define the expected return on Treasuries of maturity τ as

$$\begin{aligned} \mu_t^{(\tau)} &= A(\tau-1)'\left(\bar{\beta} + \Phi(\beta_t - \bar{\beta})\right) - A(\tau)'\beta_t + \frac{1}{2}\hat{A}(\tau-1)'\Sigma\hat{A}(\tau-1) - A_u(\tau)'u_t + C(\tau-1) - C(\tau) \\ &\quad + \frac{1}{2}A_u(\tau-1)'\Sigma^u A_u(\tau-1) + A_r(\tau-1)(\bar{r} + \phi_r'\Phi(\beta_t - \bar{\beta}) + \rho_r r_t) \\ &\quad\quad - A_r(\tau) r_t + \frac{1}{2}(A_r(\tau-1)\sigma_r)^2. \end{aligned} \tag{A21}$$

In that case, we obtain expected next-period wealth

$$\mathbb{E}_t[W_{t+1}] = W_t(1+r_t) + \sum_{\tau=2}^{N} X_t(\tau)\left(\mu_t^{(\tau)} - r_t\right) + \tilde{X}_t(\tilde{\phi}'\beta_t + \tilde{\phi}_r r_t - r_t),$$

and variance of next-period wealth

$$\begin{aligned}
\mathbb{V}_t(W_{t+1}) &= \left(\sum_{\tau=2}^{N} X_t(\tau)\hat{A}(\tau-1)'\Sigma^{1/2} + \tilde{X}_t\tilde{\sigma}'\right)\left(\sum_{\tau=2}^{N} X_t(\tau)\Sigma^{1/2}\hat{A}(\tau-1) + \tilde{X}_t\tilde{\sigma}\right) \\
&+ \left(\sum_{\tau=2}^{N} X_t(\tau)A_r(\tau-1)\sigma_r + \tilde{X}_t\tilde{\sigma}_r\right)^2 \\
&+ \left(\sum_{\tau=2}^{N} X_t(\tau)A_u(\tau-1)'(\Sigma^u)^{1/2}\right)\left((\Sigma^u)^{1/2}\sum_{\tau=2}^{N} X_t(\tau)A_u(\tau-1)\right) \\
&= \sum_{\tau=2}^{N}\hat{A}(\tau-1)'\Sigma\hat{A}(\tau-1)(X_t(\tau))^2 + 2\sum_{\hat{\tau}\neq\tau}\hat{A}(\tau-1)'\Sigma\hat{A}(\hat{\tau}-1)X_t(\tau)X_t(\hat{\tau}) \\
&+ 2\sum_{\tau=2}^{N}\hat{A}(\tau-1)'\Sigma^{1/2}\tilde{\sigma}\cdot(X_t(\tau)\tilde{X}_t) + \tilde{\sigma}'\tilde{\sigma}(\tilde{X}_t)^2 + \left(\sum_{\tau=2}^{N} X_t(\tau)A_r(\tau-1)\sigma_r + \tilde{X}_t\tilde{\sigma}_r\right)^2 \\
&+ \sum_{\tau=2}^{N} A_u(\tau-1)'\Sigma^u A_u(\tau-1)(X_t(\tau))^2 + 2\sum_{\hat{\tau}\neq\tau} A_u(\tau-1)'\Sigma^u A_u(\hat{\tau}-1)X_t(\tau)X_t(\hat{\tau}).
\end{aligned}$$

We assume that the latent demand shocks themselves do not carry systematic risk compensation. Then the optimization problem of the arbitrageur implies the following FOC

$$\begin{aligned}
\mu_t^{(\tau)} - r_t &= \gamma\left(\sum_{\hat{\tau}=2}^{N}\hat{A}(\tau-1)'\Sigma\hat{A}(\hat{\tau}-1)X_t(\hat{\tau}) + \hat{A}(\tau-1)'\Sigma^{1/2}\tilde{\sigma}\tilde{X}_t\right) \\
&+ \gamma\left(\sum_{\hat{\tau}=2}^{N} A_r(\tau-1)\sigma_r^2 A_r(\hat{\tau}-1)X_t(\hat{\tau}) + A_r(\tau-1)\sigma_r\tilde{\sigma}_r\tilde{X}_t\right) \\
&= \hat{A}(\tau-1)'\gamma\left(\sum_{\hat{\tau}=2}^{N}\left(\Sigma\hat{A}(\hat{\tau}-1)X_t(\hat{\tau})\right) + \Sigma^{1/2}\tilde{\sigma}\tilde{X}_t\right) \\
&+ A_r(\tau-1)\gamma\left(\sum_{\hat{\tau}=2}^{N}\left(\sigma_r^2 A_r(\hat{\tau}-1)X_t(\hat{\tau})\right) + \sigma_r\tilde{\sigma}_r\tilde{X}_t\right).
\end{aligned} \tag{A22}$$

Define the prices of risk as

$$\lambda_{\beta,t} = \gamma\left(\sum_{\hat{\tau}=2}^{N}\left(\Sigma\hat{A}(\hat{\tau}-1)X_t(\hat{\tau})\right) + \Sigma^{1/2}\tilde{\sigma}\tilde{X}_t\right), \tag{A23}$$

$$\lambda_{r,t} = \gamma\left(\sum_{\hat{\tau}=2}^{N}\left(\sigma_r^2 A_r(\hat{\tau}-1)X_t(\hat{\tau})\right) + \sigma_r\tilde{\sigma}_r\tilde{X}_t\right), \tag{A24}$$

Using definitions (A23) and (A24), the first-order condition (A22) takes the compact form:

$$\mu_t^{(\tau)} - r_t = \hat{A}(\tau-1)'\lambda_{\beta,t} + A_r(\tau-1)\lambda_{r,t}. \tag{A25}$$

Ultimately, these coefficients are pinned down in equilibrium when markets clear. The market clearing condition is

$$Z_t(\tau) + X_t(\tau) = S_t(\tau). \tag{A26}$$

for maturity $\tau \in \{1,2,\cdots,N\}$. As a next step, using expressions for $Z_t(\tau)$ in (13) and $S_t(\tau)$ in (15), we express the equilibrium arbitrageur holdings solved from (A26) as

$$\begin{aligned} X_t(\tau) = &\left(\bar{S}(\tau) + \zeta(\tau)'\beta_t + \zeta_r(\tau)r_t\right) \\ &- \left(\theta_0(\tau) - \alpha(\tau)'p_t - \theta(\tau)'\beta_t + u_t(\tau)\right). \end{aligned} \tag{A27}$$

As a result, the prices of risk $\lambda_{\beta,t}$ and $\lambda_{r,t}$ vary over time and depend on Treasury supply $S_t(\tau)$, non-arbitrageur demand $Z_t(\tau)$, and outside portfolio exposure $\tilde{X}_t$.

Next, on the *right-hand side* of (A25), we expand the X_t term by substituting (A27) into the price-of-risk definitions (A23) and (A24), and we expand the $\tilde{X}_t$ term by applying assumptions (28) and the affine price in (21). This gives a price of risk decomposition

$$\begin{aligned} \lambda_{\beta,t} &= \lambda_{\beta\beta}\beta_t + \lambda_{\beta r}r_t + \lambda_{u\beta}u_t + c_\beta, \\ \lambda_{r,t} &= \lambda_{r\beta}\beta_t + \lambda_{rr}r_t + \lambda_{ur}u_t + c_r, \end{aligned} \tag{A28}$$

where

$$\lambda_{\beta\beta} \equiv \gamma\left(\sum_{\hat{\tau}=2}^{N} \Sigma\hat{A}(\hat{\tau}-1)\left(\zeta(\hat{\tau})' + \alpha(\hat{\tau})'A + \theta(\hat{\tau})'\right) + \Psi\right), \tag{A29}$$

$$\lambda_{\beta r} \equiv \gamma\left(\sum_{\hat{\tau}=2}^{N} \sigma_r^2 A_r(\hat{\tau}-1)\left(\zeta(\hat{\tau})' + \alpha(\hat{\tau})'A + \theta(\hat{\tau})'\right) + \Psi_r\right), \tag{A30}$$

$$\lambda_{r\beta} \equiv \gamma\left(\sum_{\hat{\tau}=2}^{N} \Sigma\hat{A}(\hat{\tau}-1)\left(\zeta_r(\hat{\tau}) + \alpha(\hat{\tau})'A_r\right) + \Lambda\right), \tag{A31}$$

$$\lambda_{rr} \equiv \gamma\left(\sum_{\hat{\tau}=2}^{N} \sigma_r^2 A_r(\hat{\tau}-1)\left(\zeta_r(\hat{\tau}) + \alpha(\hat{\tau})'A_r\right) + \Lambda_r\right), \tag{A32}$$

$$\lambda_{u\beta} \equiv \gamma\Sigma\left(\left(\sum_{\hat{\tau}=2}^{N} \hat{A}(\hat{\tau}-1)\alpha(\hat{\tau})'A_u\right) - \left(0,\hat{A}(1),...,\hat{A}(N-1)\right)\right) \tag{A33}$$

$$\lambda_{ur} \equiv \gamma\sigma_r^2\left(\left(\sum_{\hat{\tau}=2}^{N} A_r(\hat{\tau}-1)\alpha(\hat{\tau})'A_u\right) - (0,A_r(1),...,A_r(N-1))\right) \tag{A34}$$

and c_β, c_r are constant terms:

$$c_\beta = \gamma\left(\sum_{\hat{\tau}=2}^{N} \Sigma\hat{A}(\hat{\tau}-1)\left(\bar{S}(\hat{\tau}) - \theta_0(\hat{\tau}) + \alpha(\hat{\tau})'C\right) + \psi\right), \tag{A35}$$

$$c_r = \gamma\left(\sum_{\hat{\tau}=2}^{N} \sigma_r^2 A_r(\hat{\tau}-1)\left(\bar{S}(\hat{\tau}) - \theta_0(\hat{\tau}) + \alpha(\hat{\tau})'C\right) + \psi_r\right). \tag{A36}$$

The constant terms c_β and c_r depend on $(\gamma, \psi, \psi_r, C)$ and affect only the unconditional yield level, not the dynamic responses to β_t and r_t.

Similarly, we expand the *left-hand side* of (A25) as

$$\begin{aligned}\mu_t^{(\tau)} - r_t =& A(\tau-1)'\left(\bar{\beta} + \Phi(\beta_t - \bar{\beta})\right) - A(\tau)'\beta_t + \frac{1}{2}\hat{A}(\tau-1)'\Sigma\hat{A}(\tau-1) + C(\tau-1) - C(\tau) \\ &+ A_r(\tau-1)(\bar{r} + \phi_r'\Phi(\beta_t - \bar{\beta}) + \rho_r r_t) - A_r(\tau)r_t + \frac{1}{2}(A_r(\tau-1)\sigma_r)^2 - A_u(\tau)'u_t \\ &+ \frac{1}{2}A_u(\tau-1)'\Sigma^u A_u(\tau-1) - r_t.\end{aligned} \tag{A37}$$

Finally, we expand both the left side and right hand side of equation (A25) using the expected excess return in (A37) and the price of risk decomposition in (A28). Matching coefficients on β_t, r_t, u_t and the constant term, we obtain the iteration equations:

$$A(\tau-1)'\Phi - A(\tau)' + A_r(\tau-1)\phi_r'\Phi = \hat{A}(\tau-1)'\lambda_{\beta\beta} + A_r(\tau-1)\lambda_{\beta r}, \tag{A38}$$

$$A_r(\tau-1)\rho_r - A_r(\tau) - 1 = \hat{A}(\tau-1)'\lambda_{r\beta} + A_r(\tau-1)\lambda_{rr}, \tag{A39}$$

$$-A_u(\tau)' = \hat{A}(\tau-1)'\lambda_{u\beta} + A_r(\tau-1)\lambda_{ur}. \tag{A40}$$

$$\begin{aligned}&A(\tau-1)'(I-\Phi)\bar{\beta} + A_r(\tau-1)(\bar{r} - \phi_r'\Phi\bar{\beta}) + \frac{1}{2}\hat{A}(\tau-1)'\Sigma\hat{A}(\tau-1) \\ &+ \frac{1}{2}(A_r(\tau-1)\sigma_r)^2 + \frac{1}{2}A_u(\tau-1)'\Sigma^u A_u(\tau-1) + C(\tau-1) - C(\tau) \\ &= \hat{A}(\tau-1)'c_\beta + A_r(\tau-1)c_r.\end{aligned} \tag{A41}$$

E.3. Proofs of Results in the Simple Model

Since the simple model is a special case of the main model, we can use the derivations for the main model to help with proofs in the simple model. In particular, we will rely on the iteration equations in (A38), (A39), (A40), and (A41) in Section E.2 to help derive the simple model.

Derivations of Equilibrium Treasury Prices in Equation (24)

First, we note that due to perfect arbitrage, we must have $p_t^{(1)} = -r_t$, so that $A(1) = 0, A_r(1) = -1$, $C(1) = 0$, and $A_u(1)' = (0,0)$. The holding return for 2-period Treasury bond as in (A17) can be simplified as

$$\begin{aligned} R_{t+1}^{(2)} &= -A(2)\cdot\beta_t + C(1) - C(2) \\ &\quad -(\rho_r r_t + \sigma_r \varepsilon_{t+1}^r) - A_r(2) r_t + \tfrac{1}{2}\sigma_r^2 - A_u(2)' u_t. \end{aligned} \tag{A42}$$

Next, we set $\tau = 2$ in the iteration equation for A_r in (A39), which leads to

$$A_r(1)\rho_r - A_r(2) - 1 = A_r(1)\gamma\left(\sigma_r^2 A_r(1)\alpha(2)' A_r\right)$$

$$-\rho_r - A_r(2) - 1 = -\gamma\left(-\sigma_r^2(-b, \frac{a}{2})\begin{pmatrix} -1 \\ A_r(2) \end{pmatrix}\right)$$

$$-\rho_r - A_r(2) - 1 = \gamma\sigma_r^2\left(b + \frac{a}{2}A_r(2)\right).$$

Therefore,

$$A_r(2) = -\frac{1 + \rho_r + \gamma\sigma_r^2 b}{1 + \frac{1}{2}\gamma\sigma_r^2 a}.$$

To obtain $A(2)$, we set $\tau = 2$ in the iteration equation for A in (A38),

$$\begin{aligned} -A(2) &= A_r(1)\gamma\left(\sigma_r^2 A_r(1)(\zeta(2) + \alpha(2)' A + \theta(2))\right) \\ &= -\gamma\left(-\sigma_r^2(\zeta(2) + (-b, \frac{a}{2})\begin{pmatrix} 0 \\ A(2) \end{pmatrix} + \theta(2))\right) \\ &= \gamma\sigma_r^2(\frac{a}{2}A(2) + \zeta(2) + \theta(2)). \end{aligned}$$

which leads to

$$A(2) = -\frac{\gamma\sigma_r^2(\theta(2) + \zeta(2))}{1 + \gamma\sigma_r^2\frac{a}{2}}.$$

Next, we solve for A_u. For $\tau = 2$, equation (A40) leads to

$$-A_u(2)' = -\gamma\sigma_r^2\left(-\alpha(2)'\begin{pmatrix} A_u(1)' \\ A_u(2)' \end{pmatrix} - (0, A_r(1))\right)$$

$$-A_u(2)' = -\gamma\sigma_r^2\left(-(-b, \frac{a}{2})\begin{pmatrix} A_u(1)' \\ A_u(2)' \end{pmatrix} - (0, -1)\right)$$

$$A_u(2)' = \gamma\sigma_r^2\left(-\frac{a}{2}A_u(2)' - (0,-1)\right)$$

$$A_u(2)' = \frac{1}{1+\gamma\sigma_r^2\frac{a}{2}}(0,\gamma\sigma_r^2).$$

Consequently, we obtain the 2×2 matrix A_u as

$$A_u = \begin{pmatrix} 0 & 0 \\ 0 & \frac{\gamma\sigma_r^2}{1+\gamma\sigma_r^2\frac{a}{2}} \end{pmatrix}. \tag{A43}$$

Then, we solve for $C(2)$ via setting $\tau = 2$ in equation (A41),

$$\frac{1}{2}\sigma_r^2 + C(1) - C(2) = A_r(1)\gamma\left(\sigma_r^2 A_r(1)(\bar{S}(2) - \theta_0(2) + \alpha(2)'C)\right)$$

$$\frac{1}{2}\sigma_r^2 - C(2) = \gamma\sigma_r^2\left(\bar{S}(2) - \theta_0(2) + (-b,\frac{a}{2})\begin{pmatrix} 0 \\ C(2) \end{pmatrix}\right)$$

$$\begin{aligned} C(2) &= \frac{\frac{1}{2}\sigma_r^2 - \gamma\sigma_r^2\bar{S}(2) + \gamma\sigma_r^2\theta_0(2)}{1+\gamma\sigma_r^2\frac{a}{2}} \\ &= \frac{\frac{1}{2} - \gamma\bar{S}(2) + \gamma\theta_0(2)}{\frac{1}{\sigma_r^2} + \gamma\frac{a}{2}}. \end{aligned}$$

Summarizing all the above, we obtain

$$p_t^{(2)} = -\frac{1+\rho_r+\gamma\sigma_r^2 b}{1+\frac{a}{2}\gamma\sigma_r^2}r_t - \frac{\gamma\sigma_r^2(\zeta(2)+\theta(2))}{1+\frac{a}{2}\gamma\sigma_r^2}\beta_t + \frac{\gamma\sigma_r^2}{1+\frac{a}{2}\gamma\sigma_r^2}u_t(2) + \frac{\frac{1}{2} - \gamma\bar{S}(2) + \gamma\theta_0(2)}{\frac{1}{\sigma_r^2} + \frac{a}{2}\gamma},$$

which is identical to equation (24).

Proof of Proposition 1

To prove Proposition 1, we derive three important sensitivities.

$$\frac{\partial p_t^{(2)}}{\partial \beta_t} = -\frac{\gamma\sigma_r^2(\zeta(2)+\theta(2))}{1+\frac{a}{2}\gamma\sigma_r^2}$$

$$\frac{\partial p_t^{(2)}}{\partial u_t} = \frac{\gamma\sigma_r^2}{1+\frac{a}{2}\gamma\sigma_r^2}$$

$$\frac{\partial p_t^{(2)}}{\partial \theta_0(2)} = \frac{\gamma\sigma_r^2}{1+\frac{a}{2}\gamma\sigma_r^2}$$

The magnitudes of these three sensitivities clearly all increase with γ.

Proof of Proposition 2

The expectation component of the long-term Treasury yield is

$$\bar{y}_t^{(2)} = \frac{1+\rho_r}{2} r_t$$

Using $y_t^{(2)} = -p_t^{(2)}/2$ and equation (24), we get the term premium expression

$$\begin{aligned}
& y_t^{(2)} - \bar{y}_t^{(2)} \\
=& \left(\frac{1+\rho_r+\gamma\sigma_r^2 b}{2+a\gamma\sigma_r^2} - \frac{1}{2}(1+\rho_r) \right) r_t + \frac{\gamma\sigma_r^2(\zeta(2)+\theta(2))}{2+a\gamma\sigma_r^2}\beta_t - \frac{\gamma\sigma_r^2}{2+a\gamma\sigma_r^2}u_t(2) - \frac{\frac{1}{2}-\gamma\bar{S}(2)+\gamma\theta_0(2)}{2+a\gamma\sigma_r^2} \\
=& \frac{b-\frac{1}{2}(1+\rho_r)a}{\frac{2}{\gamma\sigma_r^2}+a} r_t + \frac{\gamma\sigma_r^2(\zeta(2)+\theta(2))}{2+a\gamma\sigma_r^2}\beta_t - \frac{\gamma\sigma_r^2}{2+a\gamma\sigma_r^2}u_t(2) - \frac{\frac{1}{2}-\gamma\bar{S}(2)+\gamma\theta_0(2)}{2+a\gamma\sigma_r^2}
\end{aligned}$$

As a result,

$$\frac{\partial(y_t^{(2)} - \bar{y}_t^{(2)})}{\partial r_t} = \frac{b-\frac{1}{2}(1+\rho_r)a}{\frac{2}{\gamma\sigma_r^2}+a}$$

while the baseline response according to the expectation hypothesis is

$$\frac{\partial \bar{y}_t^{(2)}}{\partial r_t} = \frac{1+\rho_r}{2} > 0$$

Consequently, the full response is

$$\frac{\partial y_t^{(2)}}{\partial r_t} = \underbrace{\frac{1+\rho_r}{2}}_{\text{expectation hypothesis}} + \underbrace{\frac{b-\frac{1}{2}(1+\rho_r)a}{\frac{2}{\gamma\sigma_r^2}+a}}_{\text{change of term premium}}$$

When $2b > (1+\rho_r)a$, the term premium component is positive, so that the long-term Treasury yield over-reacts to a monetary policy shock compared to the expectation hypothesis. When $2b < (1+\rho_r)a$, the term premium component is negative, so that the long-term Treasury yield under-reacts to a monetary policy shock compared to the expectation hypothesis.

E.4. Why Arbitrageur Holdings Alone Are Not Sufficient

A natural question, in light of the simple-model proofs in the previous subsection, is the following: given that arbitrageur Treasury holdings $X_t(\tau)$ are directly observable, is it enough to read the term-premium response off these holdings, skipping the granular demand estimation? Intuitively, a higher monetary policy rate induces more arbitrageur long-term Treasury holdings and thus raises the term premium, so one might think the holdings response alone suffices.

The answer is no, for two reasons. First, mapping observed arbitrageur Treasury holdings to the term-premium response requires both arbitrageur risk aversion γ and the outside-asset loading Λ_r, which cannot be disentangled from yields and observed holdings alone; the Treasury-specific latent demand variation recovered by sector-level demand estimation is the separating moment (see Section 5.1).

Second, even the holdings response itself is an equilibrium composite consistent with different demand systems, so the same observed holdings path can imply opposite-sign term-premium responses, as we will show below.

The holdings response is an equilibrium composite. Using the equilibrium arbitrageur holdings in (A27) for the two-period bond, with $\alpha(2)' = (-b, a/2)$ and $y_t(\tau) = -p_t(\tau)/\tau$, the long-bond demand can be written as $Z_t(2) = \theta_0(2) + a\, y_t(2) - b\, y_t(1) - \theta(2)'\beta_t + u_t(2)$. Holding β_t and $u_t(2)$ fixed for the monetary-policy experiment and using $y_t(1) = r_t$,

$$X_t(2) = \text{const} + \big[\zeta_r(2) + b\big] r_t - a\, y_t(2). \tag{A44}$$

Differentiating gives

$$\frac{\partial X_t(2)}{\partial r_t} = \underbrace{\zeta_r(2)}_{\text{supply response}} + \underbrace{b}_{\text{cross-maturity substitution}} - \underbrace{a\frac{\partial y_t(2)}{\partial r_t}}_{\text{own-yield response}}. \tag{A45}$$

Different combinations of b, $\zeta_r(2)$, and a can generate the same observed holdings response, so identifying b requires the sector-level demand data that the granular demand system provides.

Outside exposure breaks the link from holdings to term premia. The simple-model proofs above abstract from arbitrageurs' outside-portfolio exposure. Adding it back, let the outside portfolio contribute a scalar loading $\Lambda_r r_t$ to the price of short-rate risk, consistent with the definition

of Λ_r in (A32). The two-period FOC then takes the form

$$p_t(2) = -(1+\rho_r)r_t - \gamma\sigma_r^2 X_t(2) + \gamma\Lambda_r r_t + \text{const}, \tag{A46}$$

which with $y_t(2) = -p_t(2)/2$ gives

$$y_t(2) = \frac{1+\rho_r}{2} r_t + \frac{\gamma\sigma_r^2}{2} X_t(2) - \frac{\gamma\Lambda_r}{2} r_t + \text{const}. \tag{A47}$$

Subtracting the expectations component $\frac{1+\rho_r}{2} r_t$ yields the term premium

$$TP_t(2) = \frac{\gamma\sigma_r^2}{2} X_t(2) - \frac{\gamma\Lambda_r}{2} r_t + \text{const}, \tag{A48}$$

so that

$$\frac{\partial TP_t(2)}{\partial r_t} = \frac{\gamma}{2}\left[\sigma_r^2 \frac{\partial X_t(2)}{\partial r_t} - \Lambda_r\right]. \tag{A49}$$

A positive Treasury-holdings response is therefore not sufficient for a positive term-premium response: even if $\partial X_t(2)/\partial r_t > 0$, term premia fall whenever $\Lambda_r > \sigma_r^2 \partial X_t(2)/\partial r_t$. Observed Treasury holdings reveal how much long Treasury risk arbitrageurs hold, but not their total priced exposure after netting the outside portfolio. Moreover, Λ_r itself is not separately identified from γ without the granular demand system: macro-yield covariation pins down only the composite price-of-risk loading in (A32), and the additional moment that separates them is the Treasury-specific latent demand variation recovered by sector-level demand estimation.

Empirical counterpart. Section 6.3 confirms this equivalence in the full model: Cases (2) and (3) of the model comparison match the same observed arbitrageur Treasury holdings under a common outside-asset structure, yet produce opposite-sign term-premium responses because only Case (3) estimates the cross-maturity demand elasticities from sector-level data. Thus, observing arbitrageur holdings alone is not sufficient to infer the term-premium response to monetary policy shocks. The demand elasticities that determine the response are crucial inputs and must be identified by directly estimating the demand system.

E.5. Setting Model Parameters

The model is quite flexible, accounting for the rich dependence of investor demand on macroeconomic factors and Treasury prices, as well as dynamics in the state variables. In this subsection, we provide details of how we use data to directly inform model parameters.

We take the average duration as the maturity for each maturity bucket, obtaining $\tau_1 = 2$,

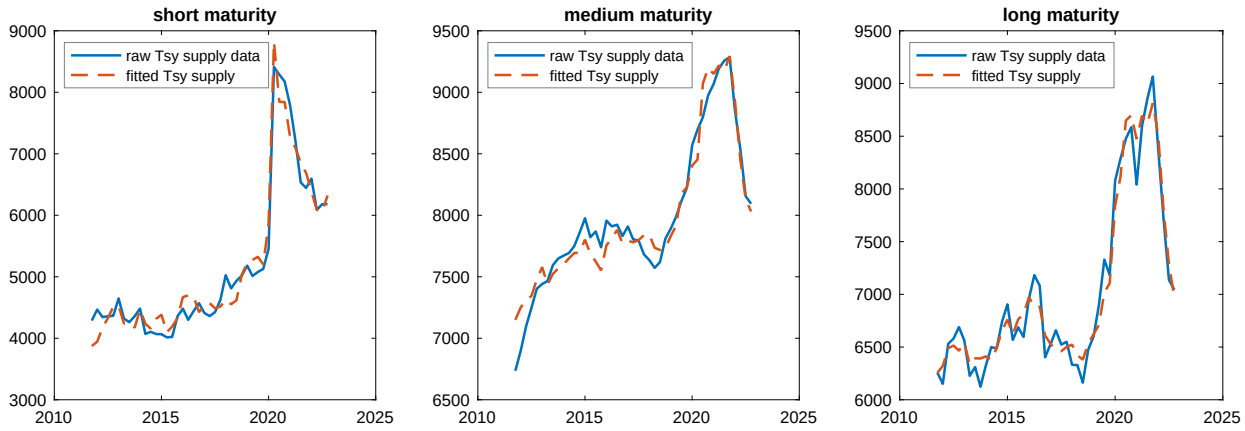


Figure A6. Treasury Supply: Data versus Model Fitting.

$\tau_2 = 10$, and $\tau_3 = 42$ (all in quarters). For each maturity bucket, we sum up the coefficients of non-arbitrageur demand in Table 2 and 3. To convert regression results to the model format, we express the demand for each maturity bucket separately, and use the intercept term to capture maturity-bucket fixed effects. We then add the maturity-by-maturity bucket estimates of the Fed to the granular-demand investor demand to obtain total non-arbitrageur demand. For simplicity, our model does not capture characteristic-based demand (i.e., loadings on coupon rate and bid-ask spread), so we take the average of these components and add them to the intercept of non-arbitrageur demand.

Moreover, in the model, the demand is expressed as a function of prices, not yields, so we need to convert the yield sensitivity into price sensitivity, using the chain rule,

$$\frac{\partial Z(\tau)}{\partial p^\tau} = \frac{\partial Z(\tau)}{\partial y^\tau}\frac{\partial y^\tau}{\partial p^\tau} = -\frac{1}{\tau}\frac{\partial Z(\tau)}{\partial y^\tau} \tag{A50}$$

Second, we estimate the supply dynamics in Equation (16). We implement a linear regression of the Treasury total supply in each maturity bucket and then recover the loadings on macro factors, the short rate, and the intercept $\bar{S}$. Similar to the demand estimation, we concentrate the supply into three maturities that represent the average duration of three maturity buckets. In Figure A6, we illustrate that the model fits the total supply well. The R^2s of all three regressions are above 95%.

Third, we estimate the monetary policy dynamics in (11). We rewrite the monetary policy equation as

$$r_{t+1} = (\bar{r} - \phi_r'\bar{\beta}) + \phi_r'\beta_{t+1} + \rho_r r_t + \sigma_r \varepsilon_{t+1}^r, \tag{A51}$$

where the intercept term is identified as a whole. To fit the monetary policy rule, we have to use a longer time period, because the monetary policy rate does not have much variation during our

main sample period. In particular, we use the post-Volcker period (1990 to 2024) excluding the zero lower bound (ZLB) period (2008–2015). We start from 1990 because it is when the Fed gained credibility in its fight against inflation. The resulting monetary-policy equation is:

$$\begin{aligned} r_{t+1} = {} & 1.9 - 1.36 * \text{credit spread}_{t+1} + 0.06 * \text{GDP gap}_{t+1} + 0.22 * \text{core inflation}_{t+1} \\ & - 1.13 * \text{debt/GDP}_{t+1} + 0.78 * r_t + 0.75 * \varepsilon^r_{t+1} \end{aligned} \tag{A52}$$

Equation (A52) suggests that the Fed lowers the interest rate if credit spread is high or GDP gap (GDP deviation from potential GDP) is low, and tightens the interest rate if inflation is high. The coefficients on GDP gap and inflation have the same signs as the classical Taylor rule (Taylor 1993) but much smaller coefficients. Moreover, there is a moderate amount of monetary policy inertia reflected by the coefficient of 0.78 on lagged policy rate. This dependence on lagged policy rate generates an impact of the monetary policy rate on long-term yields from the expectation effect and is critical to understanding how the yield curve responds to monetary policy shocks ε^r_{t+1}.

Fourth, we estimate the dynamics of macro factors in Equation (10). It is important to get the long-run average of macroeconomic factors correct. Therefore, we take the sample average of macro factors directly as $\bar{\beta}$. Denote the demeaned macro factors as $\hat{\beta}_t$. Then we recover the coefficients with the following regression:

$$\hat{\beta}_{t+1} = \Phi \hat{\beta}_t + \Sigma^{1/2} \varepsilon_{t+1}. \tag{A53}$$

Alternatively, we could directly run a linear regression with an intercept to uncover $\bar{\beta}$ and Φ simultaneously. We find that the estimates of Φ are similar between the two approaches, but the simultaneous estimation of $\bar{\beta}$ and Φ gives unreasonable long-run averages of macro variables. The matrix Σ is estimated as the covariance matrix of the regression residuals in (A53).

E.6. Model Estimation

As shown in Section E.2 (equation (A28)), the prices of risk $\lambda_{\beta,t}$ and $\lambda_{r,t}$ are affine in (β_t, r_t, u_t) with slope matrices $\lambda_{\beta\beta}$, $\lambda_{\beta r}$, $\lambda_{r\beta}$, λ_{rr}, $\lambda_{u\beta}$, λ_{ur} and constant terms c_β, c_r (equations (A35)–(A36)) that affect only the unconditional yield level. According to equations (A38) and (A39), the iteration equations for A and A_r depend on γ and the outside-portfolio parameters $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$ only through the composite slope matrices $\lambda_{\beta\beta}$, $\lambda_{\beta r}$, $\lambda_{r\beta}$, and λ_{rr}. We can therefore reformulate the estimation problem over the parameter set $\{\gamma, A_u, \lambda_{\beta\beta}, \lambda_{\beta r}, \lambda_{r\beta}, \lambda_{rr}\}$, replacing Ψ, Ψ_r, Λ, and Λ_r with the composite slope matrices. This reparameterization reveals that the yield curve's dynamic responses to β_t and r_t are pinned down by $\lambda \equiv \{\lambda_{\beta\beta}, \lambda_{\beta r}, \lambda_{r\beta}, \lambda_{rr}\}$ alone, regardless of how they decompose into γ and outside-asset exposure. The constant terms c_β and c_r depend

on $(\gamma, \psi, \psi_r, C)$ and affect only the unconditional yield level; in Step 1 they are absorbed by the free maturity-specific intercept $C(\tau)$, and (ψ, ψ_r) are estimated in Step 3. This motivates our orthogonalized estimation procedure, which separates the identification of aggregate prices of risk from the identification of γ.

We estimate the model in three steps. Steps 1 and 2 provide economically motivated starting values by sequentially identifying the price-of-risk parameters and an initial estimate of γ, using a free intercept C to absorb unconditional yield levels without imposing structural constraints. Step 3 is the full structural estimation: it jointly optimizes all parameters $(\lambda, \gamma, \psi, \psi_r)$ with C solved from the market-clearing recursion at every candidate parameter vector, so all estimated objects are mutually consistent with the model's equilibrium conditions. Step 3 also evaluates the full model prediction, including $A_u u_t$, against observed prices and positions. The reported estimates are from Step 3; Steps 1 and 2 make it computationally tractable by providing a good initialization. The outside-portfolio risk loadings $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$ are recovered in closed form after Step 3. We describe each step in detail below.

A. Step 1: Macro-driven yield dynamics

In the first step, we estimate the aggregate price-of-risk parameters from the dominant source of yield curve variation: macroeconomic fluctuations and monetary policy. Specifically, we jointly estimate the price-of-risk parameters and a free maturity-specific intercept $C(\tau)$ by solving

$$\min_{\{\lambda_{\beta\beta}, \lambda_{\beta r}, \lambda_{r\beta}, \lambda_{rr}, C\}} \sum_t \sum_\tau \left(A\beta_t + A_r r_t + C(\tau) - p_t^o(\tau)\right)^2, \tag{A54}$$

subject to the iteration equations on A and A_r (derived in (A38) and (A39)):

$$A(\tau)' = A(\tau-1)'\Phi + A_r(\tau-1)\phi_r'\Phi - \hat{A}(\tau-1)'\lambda_{\beta\beta} - A_r(\tau-1)\lambda_{\beta r} \tag{A55}$$

$$A_r(\tau) = A_r(\tau-1)\rho_r - 1 - \hat{A}(\tau-1)'\lambda_{r\beta} - A_r(\tau-1)\lambda_{rr} \tag{A56}$$

that determine $\{A, A_r\}$ from the price-of-risk parameters, where $\hat{A}(\tau-1)$ is defined in (A18). Throughout the estimation, $C(\tau)$ is treated as a free, maturity-specific intercept estimated jointly with λ by absorbing the average yield at each maturity. As we discuss in Step 3, this free intercept is sufficient, together with Steps 1–2, for identifying the dynamic objects that govern how yields and positions respond to macro shocks, monetary policy, and demand shocks; it does not resolve the separate steady-state identification of (C, ψ, ψ_r). We therefore compute a model-implied steady-state C only as a consistency check under an explicit normalization (Step 3).

In the estimation, the dimension of β is $K = 4$, and the dimension of r_t is 1. The price-of-

risk matrices $\lambda_{\beta\beta}$, $\lambda_{\beta r}$, $\lambda_{r\beta}$, λ_{rr} contain $4\times4+4\times1+1\times4+1\times1=25$ parameters. Although this is a constrained nonlinear optimization, we leverage an important insight from the affine term structure literature: regression-based initialization of the coefficient matrices provides a reliable starting point for the algorithm.

Initialization. Our initialization follows a key intuition from the term structure literature. We start with a linear regression:

$$\min_{A,A_r,C}\sum_t\sum_\tau(A\beta_t+A_r r_t+C-p_t^o(\tau))^2,$$

which is equivalent to regressing the log-price vector

$$p_t^o=-(y_t^o(1),2y_t^o(2),\cdots,Ny_t^o(N))$$

on β_t and r_t, where C serves as the intercept term. Next, knowing the values of the matrices A, A_r, we can view the iteration equations in (A55) and (A56) as another set of regressions. Rewriting them in regression form,

$$\begin{aligned}\underbrace{A(\tau)'-A(\tau-1)'\Phi}_{\text{left hand side}}&=\underbrace{A_r(\tau-1)}_{\text{dep var}}(\phi_r'\Phi-\lambda_{\beta r})-\underbrace{\hat{A}(\tau-1)'}_{\text{dep var}}\lambda_{\beta\beta},\\ \underbrace{-A_r(\tau)+A_r(\tau-1)\rho_r-1}_{\text{left hand side}}&=\underbrace{\hat{A}(\tau-1)'}_{\text{dep var}}\lambda_{r\beta}+\underbrace{A_r(\tau-1)}_{\text{dep var}}\lambda_{rr},\end{aligned}\tag{A57}$$

where the regression coefficients are $\phi_r'\Phi-\lambda_{\beta r}$, $\lambda_{\beta\beta}$, $\lambda_{r\beta}$, and λ_{rr}, and ϕ_r and Φ are directly estimated from data. This provides initial values for all four price-of-risk matrices.

Yield residuals (non-systematic components). After solving (A54), the yield residuals

$$\eta_t^o(\tau)\equiv p_t^o(\tau)-A\beta_t-A_r r_t-C(\tau)\tag{A58}$$

capture the component of yield variation unexplained by macroeconomic factors, monetary policy, and the steady-state yield level. According to the model, these residuals reflect the price impact of latent demand shocks: $\eta_t^o(\tau)=A_u(\gamma,\tau,\cdot)\hat{u}_t+\text{noise}$, where $\hat{u}_t$ denotes the estimated latent demand shocks from the demand regressions in Section 3.

B. Step 2: Identifying arbitrageur risk aversion γ

Given the price-of-risk parameters and the resulting $\{A, A_r, \hat{A}\}$ from Step 1, the latent-demand price impact matrix A_u is uniquely determined by γ through a linear system. That is, conditional on the aggregate prices of risk, γ is the sole free parameter governing A_u.

Computing A_u from γ. For any given γ and the solved matrices A, A_r, and $\hat{A}$ from Step 1, we can uniquely pin down A_u. Denote

$$\begin{aligned} \hat{A}^{\text{shift}} &= \big(0, \hat{A}(1), ..., \hat{A}(N-1)\big)' \\ A_r^{\text{shift}} &= (0, A_r(1), ..., A_r(N-1))' \end{aligned} \tag{A59}$$

which are "shifts" of the original $\hat{A}$ and A_r matrices. Stacking all maturities τ in equation (A40) yields the linear system

$$\begin{aligned} &\left(\hat{A}^{\text{shift}} \gamma \Sigma \sum_{\hat{\tau}=2}^{N} \hat{A}(\hat{\tau}-1)\alpha(\hat{\tau})' + A_r^{\text{shift}} \gamma \sigma_r^2 \sum_{\hat{\tau}=2}^{N} A_r(\hat{\tau}-1)\alpha(\hat{\tau})' + I \right) A_u \\ =&\hat{A}^{\text{shift}} \gamma \Sigma (\hat{A}^{\text{shift}})' + A_r^{\text{shift}} \gamma \sigma_r^2 (A_r^{\text{shift}})'. \end{aligned} \tag{A60}$$

This is a linear equation for A_u that can be solved immediately for any candidate value of γ. The key observation is that γ enters multiplicatively on both sides and no free parameter remains to absorb it, so A_u is a unique, nonlinear function of γ alone given the Step 1 estimates of $\{A, A_r, \hat{A}\}$. Moreover, A_u is monotonically increasing in γ.

In the simplified model of Section E.3, the 2×2 matrix A_u in equation (A43) has the single non-zero entry $A_u(2,2) = \gamma\sigma_r^2/(1+\gamma\sigma_r^2 a/2)$, which is strictly increasing in γ. This reflects the fact that a more risk-averse arbitrageur (equation (19)) demands a larger price concession (equation (27)) to absorb each unit of latent demand shock (equation (21)).

In the general model of equation (A60), writing $(\gamma M + I)A_u = \gamma R$ gives $A_u(\gamma) = \gamma(\gamma M + I)^{-1} R$, so that $dA_u/d\gamma = [(\gamma M + I)^{-1}]^2 R$. Since R is positive semidefinite by construction, the diagonal elements of $dA_u/d\gamma$ are positive whenever the symmetric part of M is positive definite, an economic stability condition requiring that own-elasticity effects dominate cross-elasticity feedback loops. This condition is satisfied under our estimated demand parameters, and we verify numerically that all diagonal elements of A_u are monotonically increasing in γ over the relevant range. This monotonicity ensures that the Step 2 objective has a unique minimum: the data pins down a single value of γ that rationalizes the observed price impact of demand shocks.

Joint pricing and quantity objective. We exploit this mapping to identify γ from two complementary moments: the comovement of yield residuals with latent demand shocks, and the corresponding response of arbitrageur positions.

To construct the quantity residual, we use observed Treasury supply $S_t^o(\tau)$ directly rather than its parametric approximation. The equilibrium arbitrageur holding is

$$X_t(\tau) = S_t^o(\tau) - \big(\theta_0(\tau) - \alpha(\tau)' p_t - \theta(\tau)'\beta_t + u_t(\tau)\big).$$

Substituting the affine price equation $p_t = A\beta_t + A_r r_t + A_u u_t + C$ from Step 1 and collecting terms by state variable,

$$X_t(\tau) = \underbrace{S_t^o(\tau) - \theta_0(\tau) + \big(\theta(\tau)' + \alpha(\tau)'A\big)\beta_t + \alpha(\tau)'A_r \cdot r_t + \alpha(\tau)'C}_{X_t^{sys}(\tau)} + \big(\alpha(\tau)'A_u - e_\tau'\big)u_t,$$

where $X_t^{sys}(\tau)$ is the component of arbitrageur holdings explained by (β_t, r_t) and the Step 1 estimates, and e_τ is the unit vector selecting the τ-th element of u_t. Because $S_t^o(\tau)$ is taken directly from data, the supply loading $\zeta(\tau)$ need not appear explicitly. The arbitrageur holding residual in the model then is

$$X_t^{resid}(\tau) = X_t(\tau) - X_t^{sys}(\tau) = (\alpha(\tau)'A_u(\gamma) - e_\tau')\hat{u}_t \tag{A61}$$

directly related to γ through $A_u(\gamma)$. The counterpart from the data is defined as

$$X_t^{o,resid}(\tau) \equiv X_t^o(\tau) - X_t^{sys}(\tau), \tag{A62}$$

where $X_t^o(\tau)$ is the arbitrageur's Treasury holdings of maturity τ in the data. We then estimate γ by minimizing

$$\min_\gamma \underbrace{\sum_t \sum_\tau \left(\frac{\eta_t^o(\tau) - e_\tau' A_u(\gamma)\hat{u}_t}{\bar{\eta}^o(\tau)} \right)^2}_{\mathscr{L}^{\text{price}}(\gamma)} + \underbrace{\sum_t \sum_m \left(\frac{X_t^{o,resid}(m) - X_t^{resid}(m)}{\bar{X}_{abs}^{o,resid}(m)} \right)^2}_{\mathscr{L}^{\text{qty}}(\gamma)}, \tag{A63}$$

where e_τ selects the τ-th element of $\hat{u}_t$; $A_u(\gamma)$ is computed from equation (A60) for each candidate γ; n_f is the number of fitted Treasury maturities; n_b is the number of maturity buckets; $\bar{\eta}^o(\tau) \equiv T^{-1}\sum_t |\eta_t^o(\tau)|$ is the mean absolute yield residual at maturity τ; $\bar{X}_{abs}^{o,resid}(m) \equiv T^{-1}\sum_t |X_t^{o,resid}(m)|$ is the mean absolute holding residual in bucket m; and m indexes all maturity buckets. As in Step 3, each sum aggregates squared normalized residuals, so the relative emphasis on pricing versus quantity scales with the number of observations in each panel (Tn_f versus Tn_b). In the pricing component, τ ranges over the fitted Treasury maturities ($n_f = 30$). The quantity component

covers all maturity buckets ($n_b = 3$).

Figure 5 (in the main text) reports the Step 2 profile of $\mathscr{L}^{\text{price}}(\gamma)$, $\mathscr{L}^{\text{qty}}(\gamma)$, and their sum from equation (A63) in levels. All three profiles are convex over the search range and attain a unique minimum near the baseline estimate, which makes the minimization problem for γ well defined.

Computation. Since Step 2 involves a single parameter γ, the optimization is straightforward. For each candidate γ, we solve the linear system (A60) for $A_u(\gamma)$ and evaluate the objective (A63), then minimize over $\gamma \geq 0$. This avoids the high-dimensional joint optimization and associated initialization challenges of alternative approaches.

C. Step 3: Joint refinement of structural parameters

Using the Step 1 and 2 estimates as starting values, Step 3 jointly minimizes over $(\lambda, \gamma, \psi, \psi_r)$:

$$\min_{\lambda,\gamma,\psi,\psi_r} \underbrace{\sum_t \sum_\tau \left(\frac{A\beta_t + A_r r_t + A_u u_t + C(\tau) - p_t^o(\tau)}{\bar{p}^o(\tau)} \right)^2}_{\mathscr{L}^{\text{price,full}}} + \underbrace{\sum_t \sum_m \left(\frac{X_t(m) - X_t^o(m)}{\bar{X}^o(m)} \right)^2}_{\mathscr{L}^{\text{qty,full}}}, \tag{A64}$$

where $\bar{p}^o(\tau) \equiv T^{-1} \sum_t |p_t^o(\tau)|$ and $\bar{X}^o(m) \equiv T^{-1} \sum_t |X_t^o(m)|$ scale residuals so that maturities and buckets are comparable in units despite level differences. Within each evaluation, A and A_r are determined by the recursions (A55)–(A56); A_u is solved from (A60); and C is solved from the market-clearing recursion (A41) given (ψ, ψ_r), so C is structurally consistent at every candidate parameter vector rather than a free parameter.

Aggregation across Step 2 and Step 3. Step 2 in equation (A63) and Step 3 in equation (A64) both use *sums* of squared normalized residuals in the implementation, so the implied relative weight on pricing versus quantity scales with the number of observations in each panel (Tn_f versus Tn_b). The normalizers differ: Step 2 uses $(\bar{\eta}^o, \bar{X}^{o,\text{resid}}_{abs})$ from Step 1 residuals, whereas Step 3 uses $(\bar{p}^o, \bar{X}^o)$ from the data series used in the joint objective. The estimates reported in the paper come from Step 3.

Outside-portfolio risk loadings. Given the jointly estimated (λ, γ) from Step 3, the outside-asset risk loadings $\{\Psi, \Psi_r, \Lambda, \Lambda_r\}$ follow in closed form from the definitions in equations (A38) and (A39):

$$\Psi = \frac{1}{\gamma}\lambda_{\beta\beta} - \sum_{\hat{\tau}=2}^{N} \Sigma \hat{A}(\hat{\tau}-1) \left(\zeta(\hat{\tau})' + \alpha(\hat{\tau})' A + \theta(\hat{\tau})' \right)$$

$$\Psi_r = \frac{1}{\gamma}\lambda_{\beta r} - \sum_{\hat{\tau}=2}^{N} \sigma_r^2 A_r(\hat{\tau}-1)\left(\zeta(\hat{\tau})' + \alpha(\hat{\tau})' A + \theta(\hat{\tau})'\right)$$

$$\Lambda = \frac{1}{\gamma}\lambda_{r\beta} - \sum_{\hat{\tau}=2}^{N} \Sigma\hat{A}(\hat{\tau}-1)\left(\zeta_r(\hat{\tau}) + \alpha(\hat{\tau})' A_r\right)$$

$$\Lambda_r = \frac{1}{\gamma}\lambda_{rr} - \sum_{\hat{\tau}=2}^{N} \sigma_r^2 A_r(\hat{\tau}-1)\left(\zeta_r(\hat{\tau}) + \alpha(\hat{\tau})' A_r\right).$$

These expressions depend only on the estimated λ and γ and do not require separate knowledge of ψ or ψ_r. This completes the estimation of the dynamic parameters governing the model's response to macroeconomic, monetary policy, and demand shocks.

E.7. Mapping γ to Relative Risk Aversion

We follow Vayanos and Vila (2021) in mapping the estimated γ to a coefficient of relative risk aversion (RRA). For a CRRA arbitrageur with wealth W, the Arrow-Pratt absolute risk aversion equals RRA$/W$. The mean-variance objective $\mathbb{E}_t[W_{t+1}] - \frac{\gamma}{2}\mathbb{V}_t(W_{t+1})$ is a second-order approximation to expected utility, giving $\gamma = \text{RRA}/W$ and therefore

$$\text{RRA} = \gamma \times W, \tag{A65}$$

where W is the equity capital devoted to Treasury intermediation. Since model quantities are expressed in 2022Q4 nominal potential GDP dollars, γ has units of $1/(\text{billion 2022Q4 dollars})$ and W must be measured in the same units.

The relevant W is not the total equity of the broker-dealer and hedge fund sectors, since these institutions allocate only a fraction of their capital to Treasury intermediation. Instead, we back out the implied capital from arbitrageur Treasury positions and the balance-sheet leverage of the intermediary: $W = \bar{X}/\ell$, where $\bar{X}$ is the mean gross Treasury exposure and ℓ is the ratio of total assets to equity. Balance-sheet leverage is the appropriate concept here because, while Treasury positions can be repo-financed at very high ratios, intermediaries must still hold equity capital to absorb losses and meet margin calls. It is this equity buffer that W represents, and balance-sheet leverage directly measures how much of it backs the overall balance sheet. From the baseline solution, the time-series average of $\sum_\tau |X_t(\tau)|$ is \$1,536 billion in 2022Q4 dollars. Banegas and Monin (2023) report that the top 50 Treasury-holding hedge funds have average balance-sheet leverage of 6.9-to-1, and primary dealer holding companies typically operate at 10–15-to-1. Table A18 reports the implied W and RRA across this empirically relevant range. At $\ell = 7$, the implied RRA is 6.6; at $\ell =$ 10–15 it ranges from 3.1 to 4.6. All values fall within

the conventional macro-finance range of 2–10 (Mehra and Prescott 1985), confirming that our estimated γ corresponds to economically plausible risk aversion. Reading in the other direction, RRA $= 5$ requires $\ell = 9.2$-to-1 and RRA $= 3$ requires $\ell = 15.4$-to-1, both within the observed range.

Table A18. Implied Relative Risk Aversion at Different Leverage Ratios

The leverage ratio ℓ is the balance-sheet leverage (total assets to equity) of the arbitrageur. Implied wealth $W = \bar{X}/\ell$ where $\bar{X} =$ \$1,536 billion is the sample mean of gross arbitrageur Treasury exposure $\sum_\tau |X_t(\tau)|$ in 2022Q4 dollars. RRA $= \gamma \times W$ with $\gamma = 0.03$. The dagger marks the empirical benchmark from Banegas and Monin (2023) for the top 50 Treasury-holding hedge funds. The bottom panel inverts the calculation: given a target RRA, it reports the implied equity W and the leverage ratio consistent with observed gross exposure.

Leverage ℓ	Implied W (2022Q4 \$B)	Implied RRA
$7\times^{\dagger}$	219	6.6
$10\times$	154	4.6
$15\times$	102	3.1

Target RRA	Implied W (2022Q4 \$B)	Implied leverage
5	167	$9.2\times$
3	100	$15.4\times$

E.8. Bootstrap Confidence Intervals

We quantify estimation uncertainty using a pairs bootstrap. The key identification of γ comes from the time-series covariation between prices p_t and latent demand shocks $\hat{u}_t$: the structural estimation recovers γ by matching the empirical covariance of these two objects to the model-implied relationship. A valid bootstrap must therefore resample the joint pairs $(p_t, \hat{u}_t, \beta_t, r_t, X_t)$ together, so that each resampled draw preserves the within-period relationship between prices and demand shocks.

Specifically, for each of the 200 bootstrap draws we sample $T = 45$ time indices with replacement from $\{1, \ldots, T\}$ and apply the same index vector to all time series simultaneously: Treasury prices p_t, latent demand residuals $\hat{u}_t$, macro factors β_t, the short rate r_t, and arbitrageur holdings X_t. The demand system coefficients (own-yield and other-yield elasticities, macro loadings) are held fixed at their baseline IV estimates. Conditional on each resampled dataset, we re-run the full structural estimation routine (Steps 1–3 of Section 5.2) to recover $(\lambda, C, \gamma, \Psi, \Psi_r, \Lambda, \Lambda_r)$. The resulting confidence intervals reflect sampling uncertainty in the structural moment conditions, conditional on the demand system estimates. The 90% confidence interval for γ is $[0.013, 0.055]$ (standard deviation 0.016), well above zero, confirming tight identification.

Figure A7. **Bootstrap Distribution of Estimated Market Elasticity.**

This figure plots the distribution of market elasticity across 200 pairs-bootstrap draws. Each draw resamples $T = 45$ quarterly time periods with replacement, applying the same index to $(p_t, \hat{u}_t, \beta_t, r_t, X_t)$ jointly, and re-runs the structural estimation routine (Steps 1–3 of Section 5.2) with demand system coefficients held fixed at their baseline IV estimates. The dashed red line marks the baseline estimate; the dotted blue line marks the corporate bond market elasticity of $1/3.5 \approx 0.29$ from Chaudhary et al. (2025).

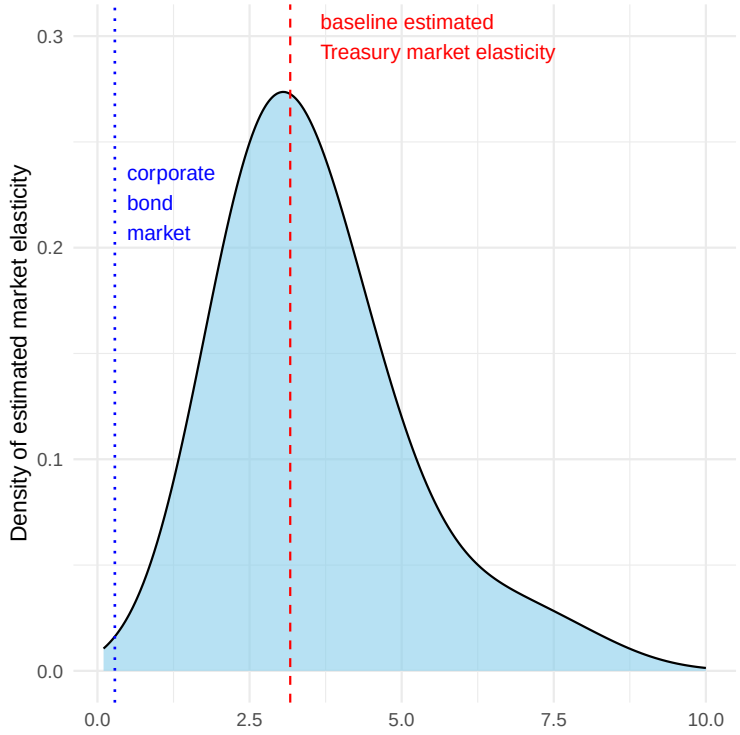


Figure A7 plots the bootstrap distribution of market elasticity. The dotted blue line marks the corporate bond market elasticity of $1/3.5 \approx 0.29$ from Chaudhary et al. (2025). Across all 200 bootstrap draws, the estimated Treasury market elasticity exceeds the corporate bond benchmark. We therefore conclude with extremely high probability that the Treasury market is more elastic than the corporate bond market.

E.9. Kappa Analysis: Derivations

We add a quadratic penalty κ on arbitrageur Treasury holdings to the objective:

$$\max_{\{X_t(\tau)\},\tilde{X}_t} \quad \mathbb{E}_t[W_{t+1}] - \frac{\gamma}{2}\mathbb{V}_t(W_{t+1}) - \frac{\kappa}{2}\sum_{\tau=2}^{N} X_t(\tau)^2. \tag{A66}$$

The parameter κ is not a structural feature of the economy; it is a diagnostic device that forces arbitrageurs toward smaller Treasury positions so we can trace how prices respond to changes in their absorption. The modified first-order condition becomes:

$$\mu_t^{(\tau)} - r_t = \hat{A}(\tau-1)'\lambda_{\beta,t} + A_r(\tau-1)\lambda_{r,t} + \kappa X_t(\tau). \tag{A67}$$

Setting $\kappa = 0$ recovers the baseline FOC exactly.

Modified recursions for (A, A_r). Matching β_t and r_t coefficients in the modified FOC, with $\Xi_A(\tau)' \equiv \zeta(\tau)' + \alpha(\tau)'A + \theta(\tau)'$ and $\Xi_{A_r}(\tau) \equiv \zeta_r(\tau) + \alpha(\tau)'A_r$ denoting the market-clearing loadings from $X_t(\tau)$, gives:

$$A(\tau)' = A(\tau-1)'\Phi + A_r(\tau-1)\phi_r'\Phi - \hat{A}(\tau-1)'\lambda_{\beta\beta} - A_r(\tau-1)\lambda_{\beta r} - \kappa\,\Xi_A(\tau)', \tag{A68}$$

$$A_r(\tau) = A_r(\tau-1)\rho_r - 1 - \hat{A}(\tau-1)'\lambda_{r\beta} - A_r(\tau-1)\lambda_{rr} - \kappa\,\Xi_{A_r}(\tau). \tag{A69}$$

For given price-of-risk matrices λ, the correction from κ enters maturity-by-maturity inside the recursion. When the model is re-estimated, λ adjusts through changes in A, A_r, and γ.

Modified A_u system. Matching u_t coefficients adds a term $\kappa X_t(\tau)$ to the right-hand side of the baseline linear system in equation (A60). Letting $J = \mathrm{diag}(0, 1, \ldots, 1)$ restrict the correction to $\tau \geq 2$ (since the one-period bond is pinned by monetary policy) and letting α denote the $N \times K$ matrix whose τ-th row is $\alpha(\tau)'$, the modified system is:

$$\begin{aligned}&\left(\hat{A}^{\text{shift}}\gamma\Sigma\sum_{\hat{\tau}=2}^{N}\hat{A}(\hat{\tau}-1)\alpha(\hat{\tau})' + A_r^{\text{shift}}\gamma\sigma_r^2\sum_{\hat{\tau}=2}^{N}A_r(\hat{\tau}-1)\alpha(\hat{\tau})' + I + \kappa J\alpha\right)A_u \\ &\quad = \hat{A}^{\text{shift}}\gamma\Sigma(\hat{A}^{\text{shift}})' + A_r^{\text{shift}}\gamma\sigma_r^2(A_r^{\text{shift}})' + \kappa J,\end{aligned} \tag{A70}$$

where the $\kappa J\alpha$ and κJ terms are the only additions relative to the baseline.

Modified C system. Matching constant terms, the $\kappa X_t(\tau)$ term contributes $\kappa[\bar{S}(\tau) - \theta_0(\tau) + \alpha(\tau)'C]$ to the equilibrium condition for each maturity $\tau \geq 2$. This modifies the baseline linear system for C by adding κ-proportional corrections from steady-state supply, demand intercepts, and price-sensitivity loadings.

$\kappa \to \infty$ limit. As $\kappa \to \infty$, Treasury holdings vanish and market clearing collapses to $Z_t(\tau) = S_t(\tau)$. Since Z_t depends only on prices and exogenous states, this uniquely pins down Treasury prices

without reference to outside-asset parameters $(\Psi, \Lambda, \psi, \gamma)$. The model collapses exactly to the pure preferred-habitat benchmark.

Simplified model. In the two-maturity simplified model ($N = 2$, $K = 1$, $\tilde{X}_t = 0$), the demand-shock loading is:

$$A_u(2,2;\kappa) = \frac{V+\kappa}{1+\alpha_2(V+\kappa)}, \qquad V \equiv \gamma\sigma_r^2, \tag{A71}$$

which is strictly increasing in κ and converges to $1/\alpha_2$ as $\kappa \to \infty$, the pure habitat price impact with no arbitrageur intermediation. The term-premium loading on r_t is:

$$A_r(2;\kappa) = -\frac{1+\rho_r+(V+\kappa)\,b}{1+\alpha_2(V+\kappa)}. \tag{A72}$$

Under the strong cross-elasticity condition $2b/a > 1+\rho_r$ (which holds in our estimates), the term premium is strictly increasing in κ. Therefore, we expect that the Treasury market becomes less elastic and more sensitive to monetary policy shocks as arbitrageur intermediation is constrained, consistent with the intuition that arbitrageurs provide a stabilizing force that dampens price responses to demand shocks and monetary policy.

F. Additional Quantitative Results

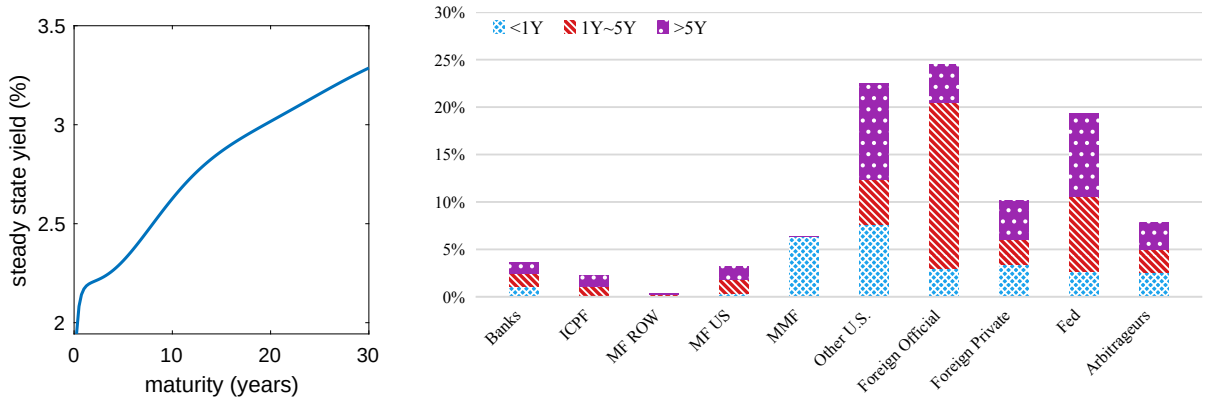
F.1. Steady-State Yields and Holdings

Figure A8 panel (a) shows the steady-state yield curve, which is upward-sloping and reflects the average shape of the yield curve over our estimation period.

Figure A8 panel (b) shows steady-state portfolio allocations across sectors. Foreign investors are the largest holder in the long run, and the Fed plays an important role as well. Insurance and pension funds hold a small share overall but concentrate in long-term Treasuries. As targeted by the calibration, arbitrageurs' holdings of longer-term (>1Y) Treasuries equal 6% of the longer-term Treasury market.

F.2. Calculating Treasury Market Multiplier and Elasticity

The total Treasury market multiplier is defined as the percentage valuation change in the whole Treasury market in response to a representative demand shock equal to 1% of total Treasury valuation $S^{total} = \sum_\tau S(\tau)$, where $S(\tau)$ is the steady-state outstanding at maturity τ. The representative



(a) Steady-state yield curve

(b) Steady-state portfolio allocations

Figure A8. **Steady State.** The left panel shows the steady-state yield curve. The right panel shows steady-state portfolio holdings (as % of total outstanding) for each investor group and maturity bucket. The steady state is defined as the state where all shocks are zero.

demand shock is weighted by outstanding shares: define $\omega = S/S^{total}$, so the shock is

$$u = \omega \cdot (S^{total} \cdot 1\%) = S \cdot 1\%.$$

The resulting change in total Treasury valuation as a fraction of S^{total} is

$$\frac{\sum_{\tau'} S(\tau') A_u(\tau', \tau) u(\tau)}{S^{total}}.$$

Dividing by the 1% shock size gives the market multiplier,

$$\mathscr{M} = \frac{1}{S^{total}} S' A_u S = \omega' A_u S. \tag{A73}$$

This is closely related to the bucket-level multiplier in Table 4, where the entry in row τ' and column τ represents the percentage price change at maturity τ' in response to a 1% latent demand shock at maturity τ:

$$\mathscr{M}(\tau', \tau) = \frac{A_u(\tau', \tau) \cdot 1\% \cdot S(\tau)}{1\%} = A_u(\tau', \tau) S(\tau). \tag{A74}$$

The total market multiplier aggregates these bucket-level multipliers as

$$\mathscr{M} = \sum_{\tau'} \sum_{\tau} \omega(\tau') \mathscr{M}(\tau', \tau). \tag{A75}$$

Using equation (A75) and Panel (a) of Table 4, we obtain a total market multiplier of 0.31: a

1% representative demand shock increases total Treasury valuation by 0.31%, or equivalently, a $1 billion demand shock increases Treasury valuation by $0.31 billion. Using Panel (b) of Table 4, the multiplier without arbitrageurs is 3.26. We can also compute the multiplier for a subset $\mathscr{T}$ of maturities:

$$\mathscr{M}(\mathscr{T}) = \sum_{\tau' \in \mathscr{T}} \sum_{\tau \in \mathscr{T}} \frac{S(\tau')}{\sum_{l \in \mathscr{T}} S(l)} \mathscr{M}(\tau', \tau). \tag{A76}$$

Restricting to maturities above one year gives a multiplier of 0.41.

Table A19 reports the price impact of permanent demand shocks, with the same format as Table 4. Using Panel (a), the market multiplier for permanent demand shocks is 0.81, exceeding the latent-shock counterpart because permanent shocks alter the risk premium in addition to current holdings. Restricting to maturities above one year gives 1.07.

Panel (b) of Table A19 is identical to Panel (b) of Table 4. In the absence of arbitrageurs, granular-demand investors treat latent and permanent demand shocks identically, so the market multipliers for the two shock types coincide.

Table A19. Price Impact of Permanent Demand Shocks with and without Arbitrageurs

	Price response (%)		
	short	medium	long
Panel (a): With Arbitrageur			
shock on short maturity	0.000	0.002	0.003
shock on medium maturity	0.013	0.108	0.187
shock on long maturity	0.057	0.503	1.476
Panel (b): Without Arbitrageur			
shock on short maturity	0.057	0.486	1.917
shock on medium maturity	0.191	0.720	3.827
shock on long maturity	0.109	0.555	1.509
Panel (c): Ratio (b)/ (a)			
shock on short maturity	217.607	248.345	620.554
shock on medium maturity	14.274	6.648	20.492
shock on long maturity	1.913	1.103	1.023

We next detail the calculation of the term structure of market elasticity. The market multiplier at maturity τ is the percentage change in total Treasury valuation in response to a demand shock at maturity τ equal to 1% of total Treasury outstanding:

$$\frac{\sum_{\tau'} S(\tau') A_u(\tau', \tau) \cdot (1\% \cdot S^{total})}{S^{total} \cdot 1\%} = \sum_{\tau'} S(\tau') A_u(\tau', \tau). \tag{A77}$$

The market elasticity at maturity τ is then defined as the inverse of this multiplier:

$$\mathscr{E}(\tau) = \frac{1}{\sum_{\tau'} S(\tau') A_u(\tau', \tau)}. \tag{A78}$$

F.3. Subsample Stability

To assess whether the estimated structural parameters are stable across different market regimes, we re-estimate the full pipeline (IV demand system and model estimation) on two alternative subsamples: a pre-COVID period (2011Q4–2019Q4, 33 quarters) that excludes the COVID shock and subsequent quantitative tightening, and a post-ZLB period (2016Q1–2022Q4, 28 quarters) that begins after the zero-lower-bound episode.

Table A20 reports, for each subsample alongside the full-sample baseline, the IV demand system loadings (Panel (a)), the structural parameters (Panel (b)), and the equilibrium price impact matrices (Panels (c) and (d)).

Panel (a) separates the IV demand loadings into non-Fed sectors and the Fed. The non-Fed loadings are relatively stable across subsamples. By contrast, the Fed's own-yield loading swings from 840 pre-COVID to -43 post-ZLB, potentially reflecting the shift from an active, yield-sensitive policy to a general balance-sheet expansion largely insensitive to yields. The Fed is the primary source of variation in the aggregate reduced-form estimates.

Despite this variation, the equilibrium market elasticity $\mathscr{E}$ remains relatively stable (2.72–4.28) across all three samples, and the price impact matrices in Panels (c) and (d) remain comparable in magnitude. The intuition is that the net absorption capacity of non-arbitrageur investors is measured by the sum of own-yield and other-yield coefficients (where the latter are typically negative, reflecting substitution across maturities). When this net capacity is small, arbitrageurs must bear more of the adjustment, and the model estimates a lower γ. Pre-COVID, the Fed's large but offsetting yield coefficients contribute substantial substitution but little net absorption, so arbitrageurs compensate with a low $\hat{\gamma}$. Post-ZLB, the Fed becomes yield-insensitive while non-Fed demand becomes more own-yield elastic; the resulting $\hat{\gamma}$ sits between the full-sample and pre-COVID estimates.

Table A20. **Subsample Stability of Structural Estimates.** Panel (a) reports the IV demand system own-yield and other-yield loadings, supply-weighted across maturity buckets, for all sectors excluding the Fed and for the Fed separately. Panel (b) reports the risk-aversion parameter $\hat{\gamma}$ and market elasticity $\mathscr{E}$. Panels (c) and (d) report the equilibrium bond price impact matrices with arbitrageurs present, measuring the bond price impact (%) of a latent demand shock equal to 1% of outstanding supply in the shocked maturity bucket. All panels are estimated on three samples: the full sample (2011Q4–2022Q4), a pre-COVID subsample (2011Q4–2019Q4), and a post-ZLB subsample (2016Q1–2022Q4).

	Full sample 2011Q4–2022Q4	Pre-COVID 2011Q4–2019Q4	Post-ZLB 2016Q1–2022Q4
Panel (a): IV demand system loadings (supply-weighted average)			
All sectors excl. Fed: own-yield	294	71	284
All sectors excl. Fed: other-yield	-423	-672	-268
Fed only: own-yield	236	840	-43
Fed only: other-yield	-202	-811	127
Panel (b): Structural parameters			
$\hat{\gamma}$	0.0298	0.0117	0.0205
Market elasticity $\mathscr{E}$	3.168	2.722	4.281
Panel (c): Price impact, own-maturity shock (with arbitrageur, % per 1% of supply)			
$\tau < 1$Y	0.001	0.000	0.000
$1 \leq \tau < 5$Y	0.077	0.059	0.033
$\tau \geq 5$Y	0.432	0.630	0.380
Panel (d): Price impact, cross-maturity shock (with arbitrageur, % per 1% of supply)			
Short shock $\rightarrow$ medium	0.006	0.003	0.003
Short shock $\rightarrow$ long	0.015	0.012	0.011
Medium shock $\rightarrow$ short	0.009	0.005	0.004
Medium shock $\rightarrow$ long	0.188	0.183	0.117
Long shock $\rightarrow$ short	0.016	0.014	0.013
Long shock $\rightarrow$ medium	0.142	0.130	0.098