--- Disclosure of Interests and Relationships for Sitong Ding ---
Sitong Ding is a PhD Candidate in Economics at LSE and is currently serving as an academic visitor at the Bank of England. The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. The results reported in this paper have been reviewed by the Bank of England, which has approved its publication. There are no financial relationships or other potential conflicts of interest that affected the content of this research in any way.

--- Disclosure of Interests and Relationships for Robert Czech ---
Robert Czech is a Senior Research Manager at the Bank of England, which provided the regulatory data employed for the empirical analysis. The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. The results reported in this paper have been reviewed by the Bank of England, which agreed on its publication. There are no financial relationships or other potential conflicts of interest that affected the content of this research in any way.

Who Moves the Long End? The Dynamics of Bond Market Segmentation[†]

Sitong Ding
London School of Economics

Robert Czech
Bank of England

Preliminary—do not quote without the authors' permission
Click here for the latest version

November 16, 2025

Abstract

Monetary policy moves long-term real interest rates. Using regulatory transaction-level data from the UK bond market, this paper provides new causal estimates of how financial intermediaries adjust their balance-sheet holdings of long-term bonds in response to monetary policy as it transmits into long-term interest rates. The key result is that dealer banks reduce their long-term bond holdings following a monetary tightening, and yet, long-term forward rates rise persistently. These price patterns are inconsistent with an account of term premia arising from intermediaries being exposed to more interest rate risk. We rationalise our empirical findings using a New Keynesian model with segmented bond markets, where arbitrageurs trade against downward-sloping demand for long-term nominal and real bonds. Countercyclical risk aversion arising from portfolio wealth and adjustment costs generate persistent, hump-shaped responses of both prices and quantities that are consistent with the signs observed in the data. We calibrate the model using bond transaction data and quantify the counterfactual effects of market segmentation on aggregate demand and inflation.

Keywords: monetary policy, market segmentation, long-term bond holdings, high-frequency identification.

JEL Codes: E43, E44, G12, G23.

[†]The authors would like to thank Ricardo Reis, Saleem Bahaj and Dmitry Mukhin for their comments, as well as seminar participants at the LSE. E-mails: `s.ding5@lse.ac.uk` and `robert.czech@bankofengland.co.uk`. The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees.

1 Introduction

How does monetary policy work? The conventional account is that it moves real rates temporarily for only as long as prices remain fixed in the short run, during which real aggregates adjust to a new equilibrium. By this account, long-term forward real rates—the real rate expected to prevail only in the far future—should not respond to monetary policy, reflecting its neutrality over the long run. Yet, there is accumulating evidence over the last twenty years that long-term real interest rates exhibit excess sensitivity to macroeconomic news and to monetary policy.[1] Since these empirical patterns are at odds with the New Keynesian paradigm, recent explanations focus on term premia, focussing particularly on market segmentation as a source of friction driving the behaviour of long-term interest rates (Vayanos and Vila, 2021; Hanson, Lucca, and Wright, 2021; Kekre, Lenel, and Mainardi, 2024). According to this view, financial intermediaries demand compensation for bearing risk when integrating markets and providing liquidity, which amplifies the response of long-term yields relative to its rate path when monetary policy changes. This ties the excess sensitivity of long-term interest rates to the balance sheet holdings of these intermediaries. However, there are challenges to testing this view empirically that have often left the causal responses of intermediary holdings unobserved.

To test this view, this paper brings forth regulatory bond transaction data from the UK to identify the balance sheet responses of intermediaries to monetary policy. Using MiFID II transaction data that covers the universe of all trading activities in the UK bond market, we merge this highly granular dataset with yield curve data and identify their joint responses to monetary policy at the high-frequency using a structural vector autoregression. That is, we present new causal estimates of the effects of monetary policy on both long-term forward rates *and* intermediary bond holdings in the long-term segment of the UK bond market, where the latter have frequently been shrouded from view. We use these new estimates to shed new light and distinguish apart different mechanisms—that come with predictions for how prices and quantities in these markets respond to monetary policy—for how monetary policy generates term premia in the bond market. What moves long-term interest rates, and who influences them? What is the economic force behind this mechanism and how does it influence the transmission of monetary policy and shocks? This paper presents new facts and evidences to answer these questions.

We first show that changes in monetary policy result in a persistent overreaction of long-term forward rates following an initial underreaction relative to the rate path. When monetary policy tightens, both nominal and real long-term forward rates initially decline before they persistently rise above zero for more than a year. In contrast, expected inflation falls but recovers quickly within a year. The fact that long forward real rates remain elevated even when the underlying shock seems to have dissipated casts a deeper puzzle on the phenomenon of excess sensitivity. From a macroeconomic standpoint, this is at greater odds with half-life estimates of price-adjustment frequencies.

[1] Cook and Hahn (1989), Kuttner (2001), Cochrane and Piazzesi (2002), Bernanke and Kuttner (2005), Gürkaynak, Sack, and Swanson (2005), Gilchrist, López-Salido, and Zakrajšek (2015), Gertler and Karadi (2015), Hanson and Stein (2015), Abrahams et al. (2016), Nakamura and Steinsson (2018), Hanson, Lucca, and Wright (2021).

From a finance angle, the fact that long rates do not revert immediately can either suggest slow learning by market participants and limited information acquisition, or that the underlying market frictions are long-lasting.

Our second finding is that these price responses in fact come with substantial shifts in bond holdings across intermediaries in the longer-term segment of the market that satisfy a set of sign restrictions. In response to an increase in the 1-year yield by 100 basis points, financial intermediaries—consisting of gilt-edged market makers (GEMMs) and dealer banks—reduce their bond holdings by approximately £20 billion, thus lowering their exposure to duration risk, while the long-term investors that trade this segment—such as pension funds and insurers—increase their long-term bond holdings by buying them from the market. We thus recover a downward-sloping demand curve for these institutional investors who tilt their portfolio allocation towards long-term bonds when monetary policy tightens. Additionally, we find that these intermediaries do not fully respond to unexpected changes in monetary policy immediately. On the contrary, the full adjustment of their bond holdings takes about a full month after the change in monetary policy, and the unwinding of these positions thereafter also occurs at a much slower rate. An unexpected change in monetary policy therefore leads to persistent shifts in intermediary demand.

As such, our first contribution is to show that behind the price movements from long-term interest rates in response to monetary policy are large and persistent shifts in bond holdings across intermediaries that trade in the long-term segment of the bond market.

The second contribution of our paper is to use these new empirical results to discipline theoretical mechanisms that link the widening of market segmentation with either an underreaction or overreaction of long-term interest rates, relative to the rate path of monetary policy or the expectations channel. The evidence that we present in this paper suggests that mechanisms based on intermediary net worth such as Kekre, Lenel, and Mainardi (2024) can be a better characterisation of the data—at least for the UK gilt market—as compared to channels that rely on reach-for-yield (Hanson and Stein, 2015) or rate-amplifying demand (Hanson, Lucca, and Wright, 2021). In the latter two papers, including the modern characterisation of preferred-habitat theory by Vayanos and Vila (2021), there is an underreaction of yields relative to the expectations theory in response to a monetary tightening because the intermediary—or *arbitrageur*—bears less risk and thus term premia falls. In our empirical estimates, we show the opposite: intermediaries bear less quantities of interest rate risk, yet there is an overreaction of long-term interest rates.

As such, our third contribution is to reconcile our empirical results in a general equilibrium New Keynesian model featuring both nominal and inflation-linked long-term perpetuities, whereby expected inflation is endogenous, and aggregate fluctuations lead to a widening of bond market segmentation both *across* the two types of bond markets and across maturities *within* each market. We adopt the mechanism of Kekre, Lenel, and Mainardi (2024), whereby the risk-bearing capacity of arbitrageurs is endogenous to portfolio wealth that fluctuates in response to macroeconomic shocks. Our application differs, however, in three ways. First, we include portfolio adjustment costs to the arbitrageur and the persistence of preferred-habitat demand to motivate slow-moving

capital, such that the model can match the humped-shape dynamics of price and quantity responses in our VAR estimates. We show that adjustment costs lead to an initial under-reaction of long-term interest rates when short rates rise, but amplifies the overreaction thereafter as it slows down the recovery of the arbitrageur's risk-bearing capacity. Secondly, we set the model in a New Keynesian general equilibrium to quantify the feedback between nominal rigidities and the capacity for the arbitrageur to intermediate across both nominal and inflation-linked bond markets. Our model not only predicts the excess sensitivity of long-term interest rates when there is a reduction in risk-bearing capacity of the arbitrageur, but also generates inflation risk premia in response to macroeconomic shocks that raise inflation. As such, the general equilibrium of the model is similar to Ray et al. (2019) and Ray, Droste, and Gorodnichenko (2024), but the model differs in how markets price expected inflation. We then calibrate the model to match the empirical impulse response functions, and use the model to provide the counterfactual effects of market segmentation on aggregate demand.

Connections to the literature. Our paper tightly connects the high-frequency literature on excess sensitivity of long-term interest rates to the quantities behind the prices that populate theories of segmented markets and imperfect arbitrage in finance.

First, there is a large literature on the high-frequency identification of monetary surprises, pioneered by Cook and Hahn (1989) and later built upon by Gürkaynak, Sack, and Swanson (2005), Hanson and Stein (2015) and Nakamura and Steinsson (2018), that has found that long-term real and nominal rates display excess sensitivity to shocks, and move approximately one-for-one several years out into the term structure. Our contribution to this literature, which has focussed exclusively on the price responses, is to link this excess sensitivity to underlying shifts in bond holdings across intermediaries. Our approach also builds on high-frequency identification, but we use the monetary surprises as an external instrument to identify an asset price VAR (Stock and Watson, 2012; Mertens and Ravn, 2013) at the daily frequency in order to jointly identify the comovement and dynamics of both long-term forward rates and intermediary bond holdings. Additionally, we also illustrate the use of a proxy instrument constructed from inflation swap surprises from a tight window surrounding monetary policy announcements that is equally as valid as an interest rate surprise, but can sometimes be superior because of power. We use this as an instrument for breakeven inflation measured from index-linked gilts in the first stage, and find that the Stock and Yogo (2005) F-statistic is frequently larger than 200 in our application. Further, the literature has frequently estimated responses of breakeven inflation that is approximately null. Our methodology highlights a qualifier to this result: the response of breakeven inflation across the term structure is highly transitory across both monetary and non-monetary shocks, and returns to steady state within a year on average. Our interpretation is that UK inflation expectations are well-anchored in the data, and therefore seek an alternative explanation in comparison to Gürkaynak, Sack, and Swanson (2005).

Second, we focus on a mechanism centred on intermediary asset pricing (He and Krishnamurthy, 2013; Brunnermeier and Sannikov, 2014; Haddad and Sraer, 2020) nested within a segmented

markets model (Vayanos and Vila, 2021) in the footsteps of He, Nagel, and Song (2022) and Kekre, Lenel, and Mainardi (2024) in order to explain price discrepancies. Where we differ from many papers in this tradition is in our application of high-frequency identification to both long-term forward rates *and* the quantities behind these prices in a unified macroeconometric framework.

In particular, we test the predictions of such intermediary asset pricing channels by identifying the sign responses of both prices and quantities to a well-studied shock in the data. This distinguishes our paper from related work that ties risk-bearing capacity to estimates of balance sheet wealth and net worth. Our novelty here is to construct daily time series measures of quantities traded by different financial institutions—such as broker-dealers, hedge funds, and pension funds—whose objectives is well known and often populate preferred-habitat term structure models, to identify how they move, and against whom, in response to a high-frequency identified monetary shock, therefore subjecting them to the same treatment effect alongside long-term interest rates that has been traditionally been the focus up until now. Indeed, our empirical estimates suggest that mechanisms linking portfolio duration and net worth to intermediary risk-bearing capacity seem to be a better explanation of why long-term interest rates respond so strongly to monetary policy.

Third, and relatedly, our macroeconometric setup draws from a recent set of papers that rely on a joint system of financial market prices and quantities to identify shocks. We also estimate a VAR, following Hanson, Malkhozov, and Venter (2024) and Bahaj, Czech, Ding, and Reis (2023), but differ methodologically in that we partially identify the system using an external instrument rather than fully identifying the system using sign restrictions, and let the sign restrictions that arise naturally from the estimates to discipline theories of term premia.

Lastly, we add to numerous papers such as Czech et al. (2021b), Pinter (2023) and Giese et al. (2024), that have characterised the microstructure and segmentation of the UK gilt market. As with these papers, we find that pension funds and liability-driven investment (LDI) funds can be well described as a preferred-habitat investor, occupying nearly the entirety of the index-linked gilt market and the long-end (above 15-years maturity) of the conventional gilt market, while asset managers and hedge funds trade most actively in the shorter-term segment (less than 15-years maturity) of the conventional gilt market. Gilt-edged market makers, inter-dealer brokers and banks trade across both markets, mainly serving the role of intermediaries to provide liquidity. We demonstrate the use of MiFID II gilt trading data in a VAR for identification, and illustrate a widening of segmentation across these economic actors in response to shocks. To the best of our knowledge, this is the first paper that identifies the reduction in risk-taking by intermediaries in response to monetary shocks using granular transaction-level data. To complement our inference, we additionally cross-verify using regulatory information from UK banks to show that they reduced their inventories of long-term gilt holdings and duration risk held during the recent rate-hiking cycle of the Bank of England between 2022 to 2024.

As such, a broad reading of this paper should interpret the persistent overreaction of long-term interest rates as reflecting a persistent increase in the price of risk, and the reduction in quantities of

risk borne by intermediaries as reflecting a widening of market segmentation that is slow to recover, for which it then presents aggregate implications for the business cycle. A narrow reading of this paper sees its purpose to distinguish apart different models from the preferred-habitat tradition and identify how is it that monetary policy moves the bond market through this lens.

Organisation of the paper. We start by explaining our main data sources and facts that they reveal about the segmentation of the UK gilt market in Section 2. In Section 3, we discuss our empirical strategy for combining both prices and quantities in a single framework for identification. Section 4 presents our main empirical results. We rationalise these estimates in a structural model in Section 5, and illustrate its quantitative properties in Section 6. Section 7 concludes.

2 Data, Empirics and Facts: the UK Gilt Market

We start with an explanation of our main data sources, starting first with the prices, and then with the quantities, followed by a description of market segmentation. Finally, we establish that intermediaries in the UK gilt market are likely to have reduced duration risk and the inventories of gilts they hold on their balance sheets between 2021 to 2024 as the Bank of England raised interest rates.

2.1 The UK Yield Curve

We use daily estimates of the UK real and nominal yield curve based on UK government bonds (gilts), constructed based on the methodology by Anderson and Sleath (2001), that is made publicly available by the Bank of England.[2] For our baseline results, we focus on the instantaneous forward rates across the term structure implied by these estimates. While the literature has traditionally focussed on the US treasury curve, we focus on the UK yield curve as we can identify the economic agents who trade these prices and bring their transaction data to bear on the phenomenon of excess sensitivity of long-term interest rates.

2.2 The MiFID II Database and Gilt Holdings Data

The MiFID II database. Our main analyses rely on the MiFID II database, which is a regulatory data source maintained by the UK Financial Conduct Authority containing detailed transaction-level data for the UK government bond (gilt) market. It provides highly granular and detailed information of all secondary-market trades of UK-regulated firms and branches of UK firms regulated in the European Economic Area (EEA). Since all gilt dealers are UK-domiciled, our dataset covers the near universe of all transactions in the market. Each transaction-level report contains information such as the exact time of transaction, the international securities identification number

[2] `https://www.bankofengland.co.uk/statistics/details/further-details-about-yields-data`. Date of last access: 30 August 2025.

(ISIN), the notional amount, price, buyer/seller flags indicating the direction of trade, as well as the legal entity identifiers (LEI) of the counterparties to the transaction.

With this information, we are able to categorise various counterparties into broader investor groupings (such as hedge funds, asset managers, insurers etc.) based on a best endeavour classification scheme. We further aggregate across all transactions entered into by these financial institutions, taking into account the direction of the trade and with whom, within each of these groupings on a daily frequency. By this measure, a positive value from an investor group on a given date indicates a collective net buying of UK gilts from the group (against a counterparty) and vice versa.

Despite its richness, the dataset has the limitation that we only observe secondary market transactions. Consequently, this dataset should be interpreted as providing *flow* measures that serve as a proxy for changes in net holdings (i.e., the *stock*), as primary market trades between the UK Debt Management Office and market participants, as well as investors' pre-existing and maturing holdings, are unobserved.

Three alternative measures of gilt holdings. To address these limitations, we further complement the gilt trading data with information on the asset holdings of UK banks, insurers and pension funds. We focus on this subset of economic agents partly because of data availability, and partly because they fit well into preferred-habitat models that have precise predictions for how they would shift their asset holdings in response to shocks. To measure UK banks' gilt inventories, we employ the Monetary Financial Institutions' Balance Sheet dataset, publicly available from the Bank of England. This dataset provides quarterly data on UK-regulated banks and building societies, detailing gross long and short positions in UK government bonds and treasury bills across different maturity buckets. To complement this aggregate view, we draw on the granular exposures dataset described by Covi, Brookes, and Raja (2022), which reports security-level holdings for 35 institutions at a quarterly frequency. This dataset enables us to examine how banks adjust security-specific positions on their balance sheets over time.

For insurers, we rely on the quarterly submissions of returns mandated under the Solvency II Directive for firms regulated by the UK Prudential Regulation Authority (PRA), with the exception of smaller firms granted reporting waivers. This dataset covers 83 insurers with total assets exceeding £2tn as of 2022 Q1. These reports include detailed information on the holdings of a given insurer, such as the ISIN of the asset, asset category and quantity. For pension funds, we use publicly available data from the Office for National Statistics (ONS), which report quarterly aggregated gilt holdings for defined benefit (DB) schemes. As of 2022 Q1, these schemes had total assets exceeding £2.4tn. We further highlight that pension fund holdings are marked-to-market, whereas insurer and bank holdings exclude revaluations.

2.3 The UK Gilt Market and Segmentation of Investors

In this section, we briefly describe the UK gilt market through the lens of our data and categorise the nature of its segmentation using a set of three facts.

The UK gilt market is highly liquid, with an average daily trading volume of approximately £36.4bn, of which £12.5bn occurs in the dealer–client segment. On average, around 4,000 trades are executed daily. The market also features a diverse investor base, comprising more than 18,000 unique legal entities, that can be mapped to more than 15 investor types ranging from standard classifications of end-demand—such as hedge funds, asset managers, pension funds and insurers—to providers of liquidity—such as gilt-edged market makers, proprietary trading firms and trading services.

First fact: segmentation across conventional and index-linked gilt markets. Pension funds and liability-driven investment (LDI) funds (or pension funds, for short) dominate the index-linked gilt market in terms of their aggregate holdings, and fit well the description of a preferred-habitat clientele within this market segment. They hold roughly twice the amount of index-linked gilts as compared to conventional gilts (see Figure 1a), consistent with their liability-matching investment strategies. In contrast, other types of demand are scarce in this segment of the market. In the conventional gilt market, we find hedge funds and asset managers as the most active investors. For comparison, their average daily trading volume across all maturity segments combined is roughly £9bn in the conventional gilt market, and less than £1bn in the index-linked market (see Figure 1b).[3] At a broad level, we thus observe a segmentation of the UK gilt market across the two types of gilts traded: pension funds primarily trade in the index-linked market, while hedge funds and asset managers primarily trade in the conventional gilt market.

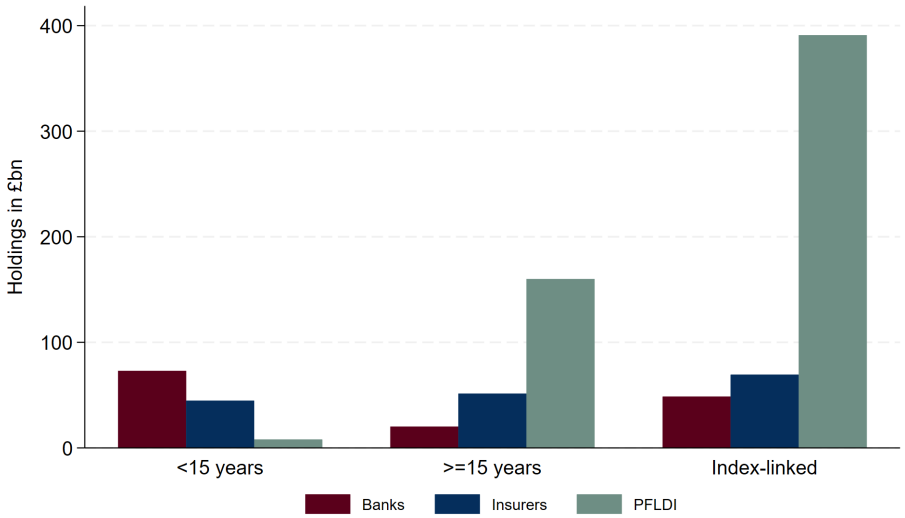
Second fact: segmentation across maturities. The UK gilt market is also segmented by the maturity of gilts traded. This segmentation is clearest in the conventional gilt market. Within this market, hedge funds and asset managers actively trade gilts with maturity less than 15 years, and their trading volume is approximately double the amount in the longer-term segment (see Figure 1b). The longer-term segment of this market is instead more frequently traded by insurers and pension funds, although we highlight that the size of their holdings in this segment is smaller in comparison to those they hold in the index-linked market segment. Their presence in the conventional gilt market is mostly driven by demand for longer-term gilts with maturity above 15-years, consistent with the findings of Giese, Joyce, Meaning, and Worlidge (2024).[4] This segmentation is further evident in the time dimension. We find that the weighted-average maturity of gilt holdings is approximately 18 years for insurers, sampled across the last 6 years for which data is available (see Figure 7). This pattern closely mirrors the average time-to-maturity of their trades in the gilt market that we observe at higher frequencies (see Figure A.1), where we find that this is approximately 12 years for hedge funds and asset managers across the same time period.

[3] In Figure A.4, we further illustrate the cumulative sum of the net trading volume of hedge funds in the index-linked gilt market. This measure is a proxy of the change in their net position in this market segment. Data is sparse and this cumulative sum is approximately zero throughout the sample. We interpret this as an illustration that hedge funds do not actively trade in the index-linked gilt market.

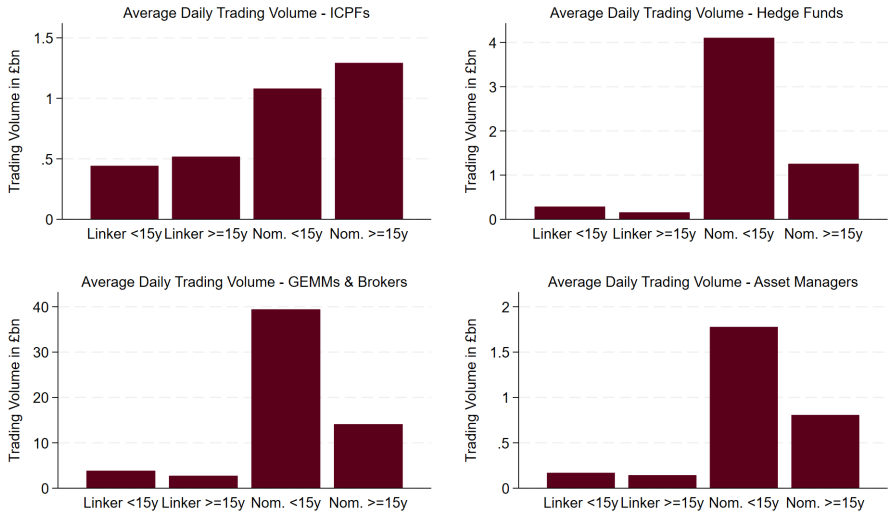
[4] See Figure 1a. Through the lens of gilt holdings data, we observe that pension funds maintain nearly negligible holdings in shorter-term conventional gilts. Note that while insurers and pension funds also trade in the long-end of the conventional gilt market, these positions are smaller in comparison to their holdings in the index-linked gilt market.

Figure 1 Measures of market segmentation

(a) By aggregate holdings



(b) By trading volume

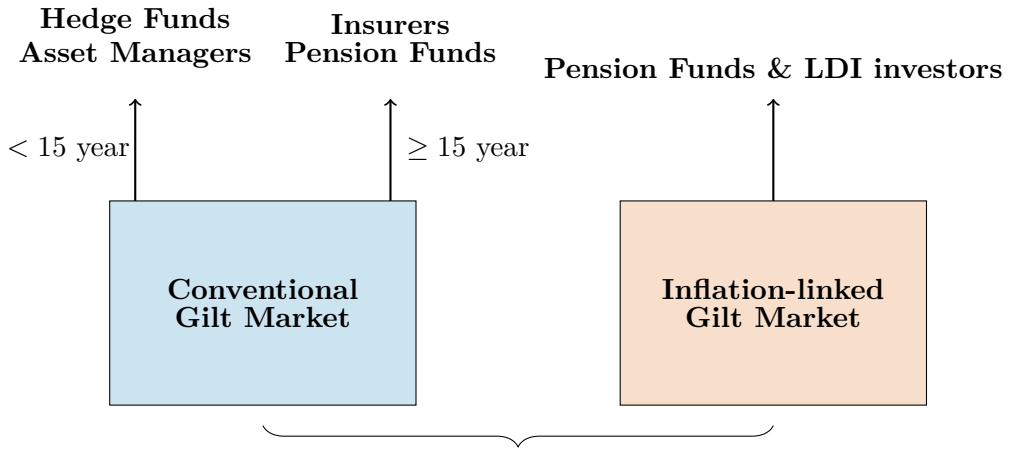


Note: In panel (a), "PFLDI" refers to pension funds and liability-driven investment funds. Source: To measure UK banks' gilt inventories, we employ the Monetary Financial Institutions' Balance Sheet dataset, publicly available from the Bank of England. For insurers, holdings data is from the quarterly submissions mandated under the Solvency II Directive. For pension funds, we rely on publicly available data from the Office for National Statistics (ONS). In panel (b), "ICPF" refers to insurance companies, pension funds and liability-driven investment funds. Source: MiFID II database on UK gilt transactions, Bank of England.

Third fact: an intermediary trading in all markets. Gilt-edged market makers (GEMMs), inter-dealer brokers and other dealer banks provide liquidity across all segments (see Figure 1b), with trading motives distinct from the different types of preferred-habitat clienteles that we have described up to now. For instance, while they account for a large share of the trading volume within the conventional gilt market, the stock of gilts that they hold on their balance sheets is smaller in

comparison to that held by pension funds and insurers who, in fact, account for a much smaller share of the trading volume and thus less frequently trade (see Figure 1). This feature defines their role as an intermediary between different clienteles, rather than a source of end-demand.

Figure 2 The segmentation of the UK gilt market



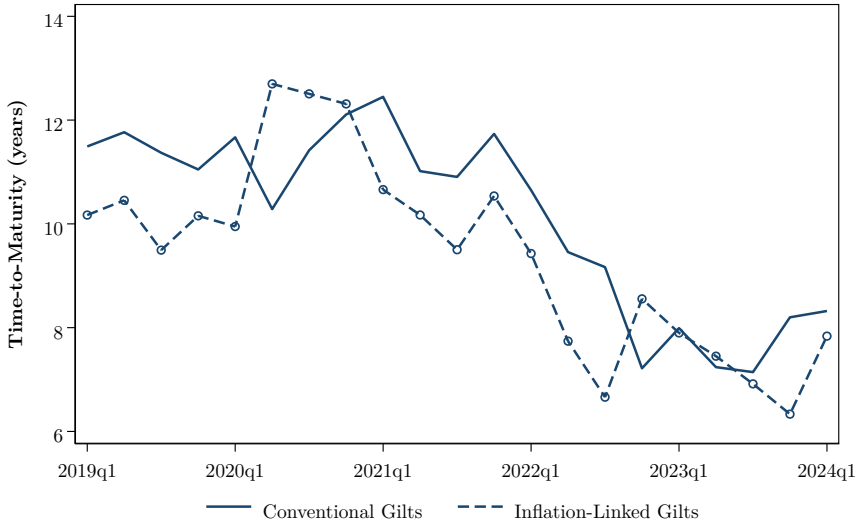
In summary, the UK gilt market is segmented across two dimensions (i) between the types of gilts (nominal v.s. index-linked gilts), and (ii) across maturities (shorter-term v.s. longer-term gilts with maturity exceeding 15 years). Each segment attracts distinct clientele. Hedge funds and asset managers fit well into a description of a preferred-habitat investor mostly trading shorter-term conventional gilts, while pension funds and insurers demonstrate preferred demand in the index-linked segment and mostly for longer-term gilts. GEMMs, inter-dealer brokers and banks serve as intermediaries to these clients, trading across all segments of the gilt market. Figure 2 provides a stylised depiction of this market.

2.4 Banks and the Reduction in Gilt Inventories: A Prelude

Having described the segmentation of the gilt market, we now turn to the dynamics of gilt holdings in the recent UK experience of monetary policy tightening, with a particular emphasis on banks as an intermediary. We focus on this pattern in the data as it can provide a preliminary view of how intermediaries respond to a monetary tightening cycle, and to what extent they fulfil the role of a marginal trader and account for the behaviour of long-term interest rates.

The fact most relevant to our subsequent analysis is our finding that banks reduced duration risk unambiguously as the Bank of England raised the Bank Rate between 2021 to 2024. Figure 3 shows the reduction in the average time-to-maturity of gilts held by banks on a trade-weighted basis, which declined from a peak of about 13 years in 2021 to 8 years in 2024. Further, this reduction in duration risk held is likely to have been driven by a reduction in banks' inventory holdings of longer-term gilts. In Figure 4 below, we highlight that banks cumulatively reduced their longer-term

Figure 3 Banks reduce average duration risk



Note: Figure shows the trade-weighted average time-to-maturity of gilts held by banks across all maturities in the conventional gilt market and the index-linked gilt market. Source: UK bank granular exposures data (Covi, Brookes, and Raja, 2022).

gilt holdings by approximately £38bn across both the conventional and index-linked gilt markets during this period of time.[5]

To cross-validate this pattern across our data sources, we additionally provide a complementary view from a wider group of intermediaries using MiFID II gilt trading data at high-frequency. Figure A.3 shows the cumulative change in net volume traded collectively in the gilt market by GEMMs, banks and hedge funds, since the beginning of January 2018, that provides a proxy measure of how their total holdings have changed to date.[6,7] Between January 2018 and January 2021, this measure was approximately zero on net, reflecting the role of these institutions as intermediaries, buying and selling on average the same quantity with equal probability. However, as UK inflation rose between 2021 to 2024, these intermediaries sold gilts systematically on a net basis. This is qualitatively consistent with the reduction in duration risk that we have calculated for banks, which is but one member of this collective intermediary group.

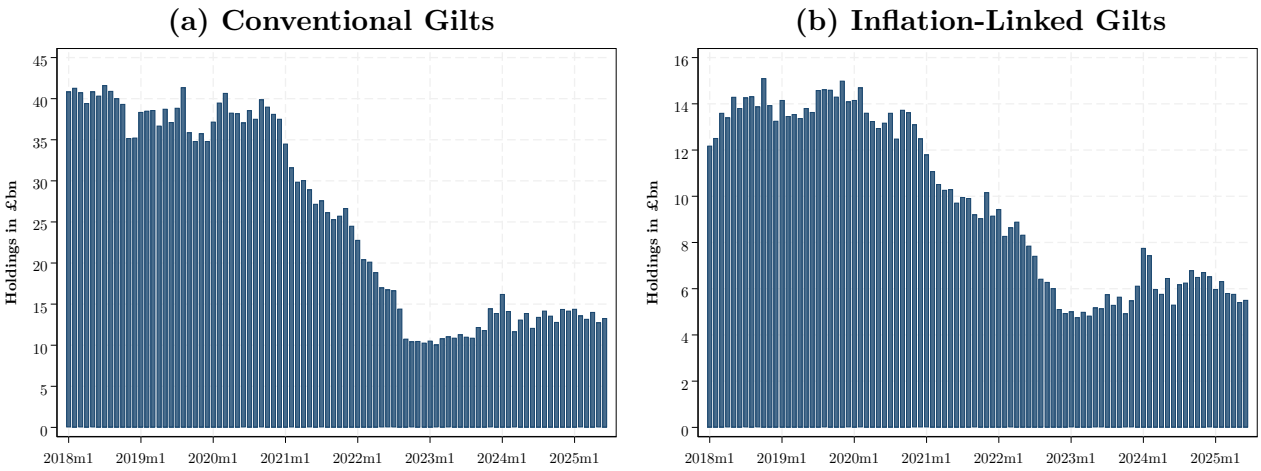
These patterns in the data can be illustrative of two possible effects on long-term interest rates.

[5]The reading is less clear in the shorter-term segment of these markets (see Figure A.2), where banks in fact increased their holdings of shorter-term gilts. Hence, we summarise the result using Figure 3 that provides an aggregate measure accounting for duration risk held by banks across all segments of the market. In our subsequent high-frequency identification of how these quantity measures respond to a monetary surprise, we also find that banks reduce their holdings in the shorter-term segment of the markets. This is consistent with lower-frequency holdings data from this section that indicates a reduction in duration risk held by banks in the aggregate, across all the maturity segments. Our interpretation of the holdings data for the shorter-term segment is that they could possibly be attributed to a response by banks in this segment to different shocks to inflation during this sample period, not only monetary shocks.

[6]Figures A.4 and A.5-A.6 illustrate respectively, this cumulative measure across the conventional gilt market and index-linked gilt market, and within market segments of each market.

[7]We include hedge funds in this calculation because it is sometimes customary within the literature on preferred-habitat models to label them as an arbitrageur. The dynamics from Figure A.3 are preserved even if we leave out hedge funds from this grouping.

Figure 4 Banks reduce inventories of long-term gilts



Note. Figures show the aggregate holdings of UK gilts by banks in the longer-term segment (more than 15 years maturity) of the conventional gilt market and the index-linked gilt market. Souce. The Monetary Financial Institutions' Balance Sheet dataset, Bank of England.

On the one hand, the reduction in duration risk can be interpreted as an equilibrium adjustment by intermediaries in response to a high interest rate environment, reflecting a higher risk aversion and thus a higher price of risk. By this account, term premium should *rise* if they are the marginal trader, generating an over-reaction of long rates. On the other hand, there is also a reduction in the quantity of risk that intermediaries have to warehouse on their balance sheets. Term premium should thus *fall*. For long-term interest rates to over-react to a rise in the short rate due to term premia explanations, the stylised patterns in our data indicate that the price of risk should dominate even as intermediaries reduce the quantities of risk they bear. We return to this discussion subsequently with estimates that allow for a causal interpretation.

3 Empirical Strategy

We now turn to our main empirical strategy. We start with a generalisation of the high-frequency identification method to an alternative instrument based on inflation swap surprises, and argue that this can sometimes be superior to interest rate surprises due to statistical power, followed by its measurement and construction. We then explain our econometric framework, provide our baseline specification of data in estimation, and explain the shock identification.

3.1 A Tale of Two Instruments

What is the response of long forward interest rates to shocks? A common approach to this question is to estimate a regression equation of the form (see e.g., Hanson and Stein (2015); Nakamura and Steinsson (2018)):

$$\Delta y_t^e = \beta \Delta i_t^e + \eta_t, \quad t \in \Theta, \tag{1}$$

where y_t^e is a market-based long-term interest rate (e.g., a 10-year forward real rate) and i_t^e is a measure of a short rate (e.g., yield from a Euribor or Sonia futures contract for delivery 4 quarters ahead), η_t is an error term, and Δ denotes their change within a high-frequency window of a news event. The convention is to measure interest rate surprises (or a proxy) $\Delta i_t^e \approx \varepsilon_{i,t}$ to a monetary shock on a set of dates $\Theta \equiv \Theta_i$ for which monetary policy is announced.[8] Under the identifying assumption that no other sources of news contaminate this measurement, Equation 1 recovers a structural interpretation of β.[9]

Our objective here is to present an extension of this insight to an alternative shock instrument based on surprise changes in *expected inflation*. Depending on the estimation sample and shock identification, this instrument can recover estimates that are more statistically significant as compared to when interest rate-based proxies are used. In Section 4, we present estimation results using both sets of proxies, and highlight when one may work better than the other.

To clarify intuition, we consider a stylised representation of market expectations as follows. We assume away measurement error and limit attention to a demeaned system for exposition:

$$\Delta i_t^e = \delta_1 \Delta \pi_t^e + \varepsilon_{i,t}, \tag{2}$$

$$\Delta \pi_t^e = \delta_2 \Delta i_t^e + \varepsilon_{\pi,t}. \tag{3}$$

In this system, short-term interest rate surprises Δi_t^e arise from both a random component $\varepsilon_{i,t}$ and a systematic component $\Delta \pi_t^e$ via a coefficient δ_1, where $\Delta \pi_t^e$ refers to the market's short-term expected inflation.[10] The systematic component is the response of interest rate expectations to inflation or *non-monetary* shocks $\varepsilon_{\pi,t}$. Expanding the system yields:

$$\Delta i_t^e = \frac{\delta_1}{1 - \delta_1 \delta_2} \varepsilon_{\pi,t} + \frac{1}{1 - \delta_1 \delta_2} \varepsilon_{i,t}, \tag{4}$$

$$\Delta \pi_t^e = \frac{1}{1 - \delta_1 \delta_2} \varepsilon_{\pi,t} + \frac{\delta_2}{1 - \delta_1 \delta_2} \varepsilon_{i,t}. \tag{5}$$

Changes in market expectations are thus driven by linear combinations of both monetary and non-monetary shocks. This system immediately clarifies that the high-frequency approach can isolate each of these shocks not only using an interest rate-based surprise Δi_t^e, but also by an inflation surprise $\Delta \pi_t^e$.[11] That is, around a high-frequency narrow window (-10/+20 minutes) bracketing

[8] We denote the set of non-monetary data release dates as Θ_π.

[9] The estimate that $\beta > 0$ has subsequently motivated questions of whether there is long-run monetary non-neutrality, or whether the underlying response is driven by term premia.

[10] The coefficient δ_1 is a reduced-form characterisation of this systematic response that is not necessarily the Taylor rule coefficient. It can be interpreted as a factor loading on expected inflation, where the unexplained residual is driven by a monetary shock (thus a two-factor representation).

[11] To identify β from Equation (1) from a monetary surprise, the convention in the literature has been to rely on measures of $\Delta i_t^e = (1 - \delta_1 \delta_2)^{-1} \varepsilon_{i,t}$ isolated in response to monetary announcements. To identify β instead from a non-monetary data release, the relevant interest rate surprise requires $\Delta i_t^e = \delta_1 (1 - \delta_1 \delta_2)^{-1} \varepsilon_{\pi,t}$ that are instead measured in response to macroeconomic data releases. See e.g., Gürkaynak, Sack, and Swanson (2005); Gürkaynak, Kısacıkoğlu, and Wright (2020); Zhu (2022).

either a monetary announcement, or a non-monetary data release, the inflation surprises to a monetary and a non-monetary type shock correspond respectively to:

$$\Delta\pi_t^e = \frac{\delta_2}{1-\delta_1\delta_2}\varepsilon_{i,t}, \ \ \forall\, t \in \Theta_i, \ \ \textit{and} \ \ \Delta\pi_t^e = \frac{1}{1-\delta_1\delta_2}\varepsilon_{\pi,t}, \ \ \forall\, t \in \Theta_\pi. \tag{6}$$

Therefore, the exclusion restriction to shocks also applies to inflation surprises, and they are equally valid as interest rate surprises in their use as a proxy for these shocks.

3.2 The Measurement of Inflation Surprises to News

High-frequency identification does not guarantee, however, that the inflation surprise is a relevant instrument even if it satisfies the exclusion restriction. The main methodological difference is to construct an inflation surprise measure as an instrument for exogenous shifts in *breakeven inflation*, a commonly used measure of inflation expectations from the index-linked gilt market, in the first-stage of the 2SLS estimation that we further elaborate in Section 3.3.

A natural candidate is to construct the inflation surprise using inflation swap rates. An inflation swap is a bilateral contract where one counterparty is the inflation receiver, and pays the cumulative increase in the price level over the life of the contract. The other counterparty is an inflation payer, who pays a fixed rate agreed upon at the initiation of the contract. If the fixed rate were to be exactly equal to the cumulative growth of the price level ex post, then no payment exchanges hands. For this reason, the fixed rate traded in the inflation swap market is another measure of expected inflation apart from breakeven inflation measured from linkers. Both measures trade closely and are highly correlated by no-arbitrage, and investors can seek inflation protection from either markets.[12] This is the feature that guarantees the relevance condition when inflation swap surprises are used as an instrument for breakeven inflation.[13]

To construct the inflation surprise measures, we obtain tick-by-tick price quotes of zero coupon UK RPI swaps from the Refinitiv Tick History database. We subsequently measure the surprise in expected inflation using the difference in median transacted quotes of the swap rates within the 10 minutes before and 20 minutes after a narrow window surrounding a news release that is either a monetary policy announcement by the Bank of England, or a non-monetary data release e.g.

[12] In the UK, because the linker market predates the inflation swap market and inflation-linked coupons are calculated based on the retail price index (RPI), the inflation swap market has also adopted the convention to compute cash flows on the floating leg based on the RPI. However, the two measures do not exactly equal due to differences in market liquidity, upfront capital costs associated with holding a linker, and differences in supply of the two assets. See Bahaj et al. (2023) for further details on the UK RPI swap market, and Barria and Pinter (2023) and Fleckenstein, Longstaff, and Lustig (2014) for mispricing in these markets as a violation of the law of one price.

[13] At a more general level, the relevance condition is more difficult to establish as compared to the exclusion restriction of the instrument. This is due to incompleteness of markets: derivative financial markets trading assets with payoffs tied to measurable macroeconomic outcomes seldom exist beyond interest rates and inflation. For example, it is also possible to measure the high-frequency changes in the stock market in response to monetary announcements. This stock market surprise can be a valid instrument to identify a monetary shock because it satisfies the exclusion restriction and is not contaminated by other sources of news. However, its use as an instrument is limited because it is less clear what (macroeconomic) variable or asset price this surprise can be highly correlated with for the relevance condition to hold.

RPI inflation. We adapt the methodology from Braun, Miranda-Agrippino, and Saha (2025) and Altavilla et al. (2019), making appropriate adjustments where necessary to take into consideration cases when market trading is thin during the sampling intervals prior or after the data release as well as instances where the date of data releases and monetary events may coincide.[14] Figure A.7 shows the inflation swap surprises measured in response to MPC announcements.

3.3 Econometric Framework

If one were only interested in the average instantaneous response of long-term interest rates to shocks over high-frequency event windows during which there is a lumpy arrival of news, then estimating an event-based regression such as Equation (1) suffices.

Our novelty, however, is to bring to this literature data on traded quantities that are endogenous to these prices, where both are likely to shift in equilibrium simultaneously in response to a common fundamental such as a monetary shock. If gilt holdings adjust slowly due to possible market frictions, and these are marginal to the equilibrium behaviour of prices, then the excess sensitivity reflected by long rates is also unlikely to be a short-lived anomaly. As such, our objective is not only to estimate the price responses from long-term interest rates to shocks over time, but also the underlying shifts in liquidity, with what magnitude and sign, and across which of these agents in the market, that can all provide useful insights to explain the excess sensitivity of long-term interest rates.

Hence, we depart from the tradition in this literature and identify a structural VAR at the high-frequency that can simultaneously return (i) estimates of the comovement between prices and quantities in response to shocks, and (ii) the dynamics of this comovement across time. Our structural VAR model takes the form:

$$\mathbf{y}_t = \boldsymbol{\alpha} + \sum_{\ell=1}^{p} \boldsymbol{\Phi}_\ell \mathbf{y}_{t-\ell} + \mathbf{u}_t, \;\; \mathbf{u}_t = \mathbf{B}\boldsymbol{\varepsilon}_t, \tag{7}$$

where $\mathbf{y}_t$ is a $N \times 1$ vector of endogenous variables consisting of asset prices and quantities at date t, $\boldsymbol{\alpha}$ is a vector of constants, and $\boldsymbol{\Phi}_\ell$ is a $N \times N$ matrix of estimated coefficients for the ℓ^{th} lag of the endogenous variables, with the total lag length given by p. We partially identify the matrix $\mathbf{B}$ using an external instrument to impose a single restriction, following Stock and Watson (2012), Mertens and Ravn (2013) and Gertler and Karadi (2015). We provide more details in Section 3.3.2.

3.3.1 The Baseline Specification and Estimation Samples

We now clarify the baseline specification that we use to establish the main results in the paper. An important limitation is that we have data on the UK term structure that span a much longer sample period as opposed to the data we have pertaining to their traded quantities. The yield

[14] To avoid over-censoring the data at zero for these instances, we additionally identify the timestamps of the first price quote prior and after the data release and sample the median price quotes in the same 10 minutes that bracket it as long as it does not coincide with a monetary announcement.

curve data that we use in estimation is available from June 2004 to August 2025, while our access to MiFID II gilt transaction data only covers the sample from January 2018 to February 2025. This has important implications for how we approach estimation.

The baseline (subsample) specification with prices and quantities. Guided by necessity, our baseline VAR specification will consist of data on prices and quantities at the daily frequency only from January 2018 to February 2025. The set of prices we include is a minimal term structure representation. They include the 1-year yield as a short rate indicator, 3-year breakeven inflation as a measure of short-term inflation expectations, the 25-year instantaneous forward rates (both nominal and real) and the 25-year inflation swap rate as measures of long-term interest rates and expected inflation respectively.[15] We include a set of two quantity measures to this set of prices. These are, respectively, the daily gilt transactions from pension funds, liability-driven investment (LDI) funds and insurers—the long-term investors—aggregated across the longer-term segment of the entire UK gilt market (i.e., combining the value of their trades across both the conventional and index-linked gilt markets longer than 15 years in maturity), and the equivalent measure aggregated from GEMMs who serve this market segment as an intermediary and as a counterparty to end-demand.[16] We consider these quantity measures to be most relevant for our baseline specification, since our focus is on the transmission of the short rate into the long-end of the yield curve, and the facts on UK gilt market segmentation inform us that these financial institutions are the largest investors that trade in this segment of the market (see Section 2.3). For controls, we include the FTSE All-Share index, UK non-financial corporate bond spreads, and the broad-based nominal effective exchange rate.

The full sample specification with only prices. In principle, our baseline specification is well-suited for testing the data, as it covers a sample period in the UK when inflation was highly volatile and the Bank of England raised the Bank Rate from 0.1% to 5.25% within a short span of less than two years. As such, we should be able to observe significant shifts in bond holdings across different intermediaries in the market as prices move and search for underlying mechanisms to explain these joint comovements. However, the compromise is that our estimates can be fairly imprecise, since this baseline specification is estimated over a shorter data sample. To partially overcome this limitation, we additionally identify the VAR over a longer sample—from June 2004 to August 2025—using only the set of asset prices included in our baseline specification. Although we can no longer draw inference from the variation in quantities, we can still characterise the responses of long-term interest rates to shocks over a longer sample, and use these estimates to cross-validate

[15] We include the 25-year inflation swap rate as our measure of long-term expected inflation instead of the 25-year breakeven rate because we also provide supplementary estimates using the 25-year nominal and real yields in place of the instantaneous forward rates as the relevant long rate. Including the 25-year breakeven rate leads to issues of multicollinearity in estimation. Since these inflation swap rates are only available starting from June 2004, we use data on the yield curve from June 2004 to August 2025 for all estimates in our paper that do not rely on quantity data so that the VAR specifications can be kept consistent across different estimates.

[16] We show in the appendix that our subsequent estimates are robust to adding hedge funds and banks to this definition. Further, the transaction quantities from the collective group of long-term investors and intermediaries we include here do not net out to zero due to the presence of other investors in this market segment that trade in much smaller amounts.

those obtained from our baseline specification.

3.3.2 Shocks, Identification Assumptions and the External Instrument

We focus our inference of prices and quantities in response to a monetary shock within our baseline specification, as existing models in the preferred-habitat tradition have the sharpest predictions for how long-term rates deviate from the expectations channel when responding to this shock, and with what shifts in portfolio holdings between the arbitrageurs and habitat investors these come with. Following in the footsteps of the proxy-IV literature, we identify a monetary shock from our baseline VAR using the target factor from Braun, Miranda-Agrippino, and Saha (2025) as an external instrument for the 1-year yield. The moment restrictions satisfied by these interest rate surprises Δi_t^e are:

$$\mathbb{E}[\Delta i_t^e \widetilde{\boldsymbol{\varepsilon}}_t] = \mathbf{0}_{N-1}, \qquad \text{(Exogeneity)}$$

$$\mathbb{E}[\Delta i_t^e \varepsilon_{i,t}] = \lambda_i \neq 0, \qquad \text{(Relevance)}$$

where $\boldsymbol{\varepsilon}_t = [\varepsilon_{i,t}, \widetilde{\boldsymbol{\varepsilon}}_t']'$ and for $t \in \Theta_i$. An additional, implicit identifying assumption that we rely on is that the gross supply of gilts in the market is exogenous to the monetary shock, which we assume to hold at the high-frequency. Under this assumption, we can identify shifts in gilt holdings across heterogeneous investor types that are endogenous to the monetary shock.

In our full sample estimates based on asset prices alone, we extend this identification additionally to non-monetary shocks. A noteworthy complication, however, is that interest rate-based proxies do not identify significant responses of long-term interest rates to these shocks when we apply our methodology in the full sample. To this end, we demonstrate the use of an alternative external instrument based on inflation swap surprises to identify these shocks.

When identifying a monetary shock, we use the inflation swap surprises $\Delta\pi_t^e$ measured at high-frequencies to MPC announcements for breakeven inflation within the same VAR specification. The moment restrictions they satisfy are:

$$\mathbb{E}[\Delta \pi_t^e \widetilde{\boldsymbol{\varepsilon}}_t] = \mathbf{0}_{N-1}, \qquad \text{(Exogeneity)}$$

$$\mathbb{E}[\Delta \pi_t^e \varepsilon_{i,t}] = \alpha_i \neq 0, \qquad \text{(Relevance)}$$

for $\boldsymbol{\varepsilon}_t = [\varepsilon_{i,t}, \widetilde{\boldsymbol{\varepsilon}}_t']'$ and for $t \in \Theta_i$.

For a type of non-monetary shock identification, we measure a *direct* surprise to expected inflation within high-frequency windows that surround UK RPI inflation data releases for which our instrument is well-designed to capture.[17] Since inflation swap rates are tied to the underlying

[17]The identification of non-monetary structural shocks require more structure and identifying restrictions to be imposed on the data, given that there is arguably a wider set of $N-1$ non-monetary shocks in the VAR system for a set of N endogenous variables. Hence, we propose the identification of an inflation surprise as an alternative to this that is simpler and does not require additional structure to be imposed. The trade-off, however, is that we lose the structural interpretation as to what an inflation surprise/shock is. We return to this point in Section 4.2.

price index and they reflect continuous expectations for how it would evolve, a lumpy release of this price index leads to a discontinuous change in the inflation swap rate if the value of the price index were not exactly equal to the market's forecast. This is, however, precisely what the definition of a shock to expected inflation is, and is what can be measured at high-frequencies (see Equation (6)). The exclusion restriction of this identification is that there are no other sources of news on the day of the inflation release, whether monetary or other types of data releases. It then allows for the identification of how long-term interest rates respond to a surprise change in inflation. Formally, the moment restrictions satisfied by the external instrument are:

$$\mathbb{E}[\Delta\pi_t^e \widetilde{\boldsymbol{\varepsilon}}_t] = \mathbf{0}_{N-1}, \qquad \text{(Exogeneity)}$$

$$\mathbb{E}[\Delta\pi_t^e \varepsilon_{\pi,t}] = \lambda_\pi \neq 0, \qquad \text{(Relevance)}$$

where $\boldsymbol{\varepsilon}_t = [\varepsilon_{\pi,t}, \widetilde{\boldsymbol{\varepsilon}}_t']'$ and for $t \in \Theta_\pi$.

4 Empirical Estimates

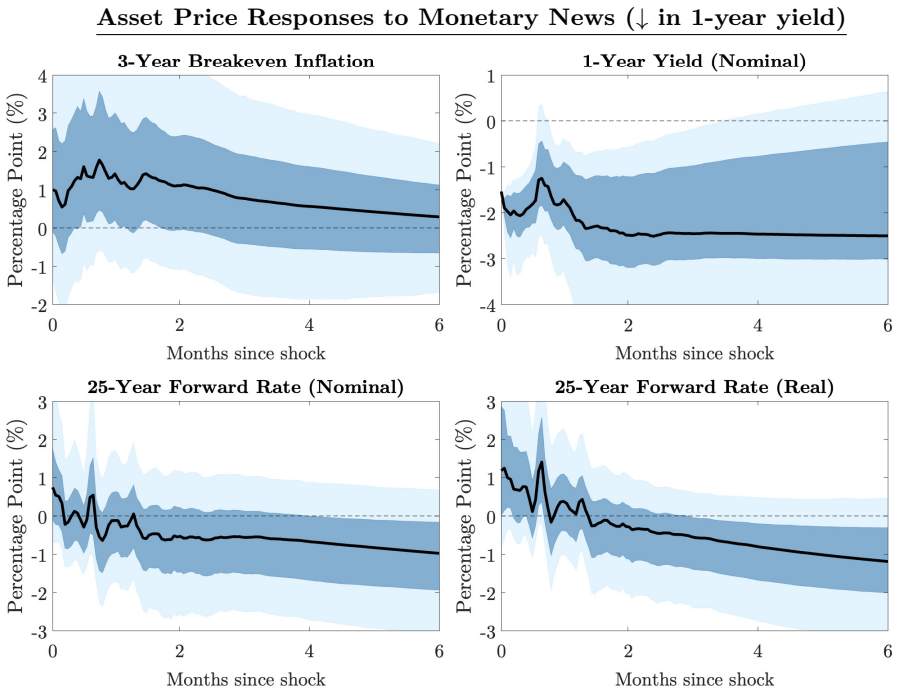
This section shows our main estimates of interest: the dynamic responses of long-term forward rates to shocks and the corresponding movements in bond holdings from segmented investors in the UK.

4.1 Baseline Estimates

We start by presenting our results from the baseline VAR specification consisting of both prices and quantities estimated from the shorter data sample.

For our baseline inference, we restrict the identification of the VAR to a monetary shock because this is a shock well-studied by the literature and there are existing theoretical mechanisms—especially from those belonging to the preferred-habitat tradition—that provide sharp predictions for how these quantities should move in response to this particular shock.[18] As such, the value of these estimated quantity responses is that they can be informative of whether these predictions are satisfied in the data, and in turn allow us to discipline different term premia mechanisms that can all account for the excess sensitivity of long-term interest rates.[19] Our primary inference of interest, therefore, is how the traded quantities respond to a monetary shock. In what follows, we discuss the responses of long-term interest rates and their traded quantities each in turn.

Figure 5 The persistent overreaction of long-term interest rates



Note: IRFs indicate the responses to a monetary shock that reduces the 1-year yield. Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 9.6. The Lewis & Merten (2025) generalized test statistic is 4.1. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

4.1.1 Price Responses to a Monetary Shock

Figure 5 shows the estimated responses of the 25-year instantaneous forward rates to a monetary policy shock. These impulse response functions are scaled such that the monetary shock leads to a *reduction* of the 1-year yield leading to a rise in the 3-year breakeven inflation rate by 1pp on impact (i.e., a monetary policy easing that leads to a rise in expected inflation). This corresponds to a monetary shock that reduces the 1-year yield by approximately 1.5 percentage points on impact. We observe an initial under-reaction of long-term forward rates relative to the expectations theory. Rather than falling immediately in response to the reduction in the 1-year yield, these long-term interest rates initially rise. However, they thereafter decrease to the monetary policy easing and remain significantly negative along with the 1-year yield at the 68% confidence level. Moreover, these responses are persistent, and capture effects on long-term interest rates rates that are considerable in magnitude: the 25-year forward real rate falls by approximately half as much as the 1-year yield by the sixth-month mark. Taken together, we argue that this reflects a short-lived, initial under-reaction followed by a persistent over-reaction of long-term interest rates to changes

[18] We additionally include a discussion of non-monetary shocks in a subsequent section when presenting our full sample estimates based solely on asset prices.

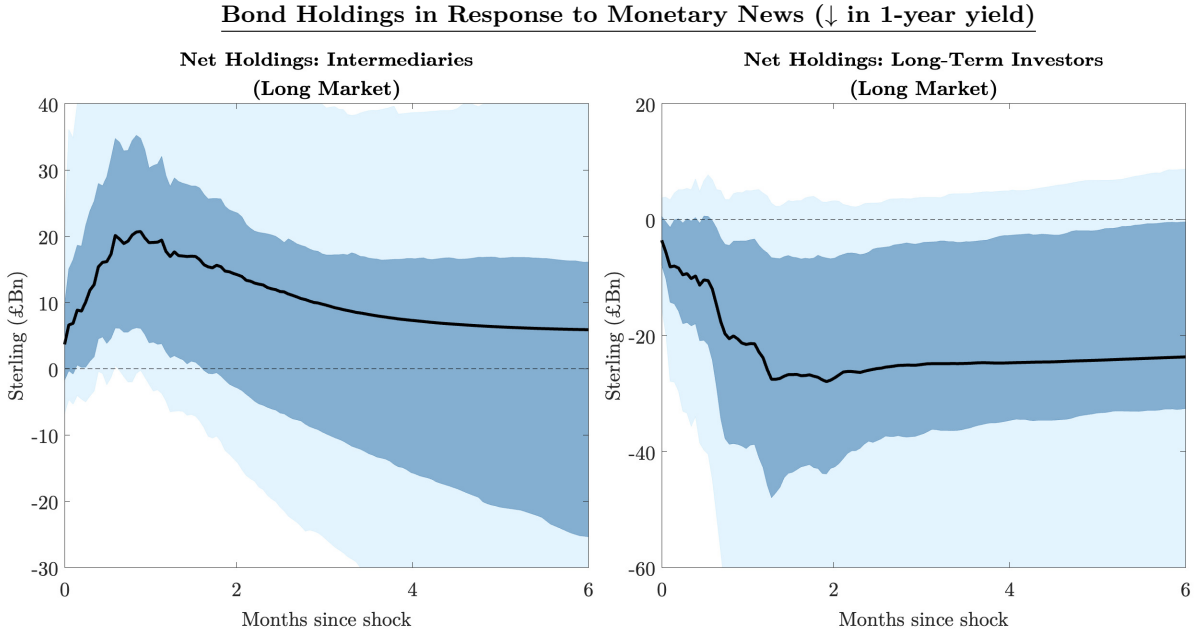
[19] We provide a more elaborate discussion in Section 4.3 after presenting all of our empirical results.

in the short rate. Finally, we highlight that the first stage of this estimation has a F-statistic of 9.6 that is marginally weak. We tackle this point subsequently when presenting our full sample estimates.

4.1.2 Quantity Responses to a Monetary Shock

To understand what may be the main driving force behind these results, we now turn to the quantities behind the prices that have been shrouded from view up to now. Figure 6 presents the quantity-based estimates from the same preceding identification. Since our data on bond transactions are flow measures, we present the cumulative effects of the monetary shock so that these results can be more closely interpreted as a change in total *holdings*.[20]

Figure 6 Intermediaries bear larger quantities of risk as yields decline



Note: IRFs indicate the aggregate response in the long market segment ($\geq$ 15-year) across the conventional gilt and inflation-linked market. Intermediaries are GEMMs and inter-dealer brokers. Long-term investors are asset managers, pension funds and liability-driven investment (LDI) funds. Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 9.6. The Lewis & Merten (2025) generalized test statistic is 4.1. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

We highlight two main findings from this identification. First, our primary inference is that intermediaries, consisting of GEMMs and inter-dealer brokers, respond to a monetary policy easing by bearing *more* interest rate risk on their balance sheets. They cumulatively increase their total

[20] See Figure B.8 for the estimated impulse response functions of aggregate bond *flows* where the same inference applies. The cumulative sum of those responses yield Figure 6.

gilt holdings by a maximum of roughly £20bn within a month of the shock.[21] As bond prices rise, the other side of this market is the response from long-term investors to sell gilts, reducing their total holdings by approximately £25bn. Given our limited data availability, these responses are estimated fairly imprecisely with large confidence intervals. However, we can at least infer the sign of these responses at the 68% confidence level. Second, we find that these estimates are fairly persistent: in the month after intermediaries responded to the monetary easing by increasing their gilt inventories, they begin to partially unwind these positions but this occurs at a much slower pace in comparison to the rate at which they increased bond holdings initially. Long-term investors also start to rebuild their gilt holdings approximately after a month, but they also recover them at a much slower pace. These shifts in gilt holdings occur as long-term forward rates persistently decline.

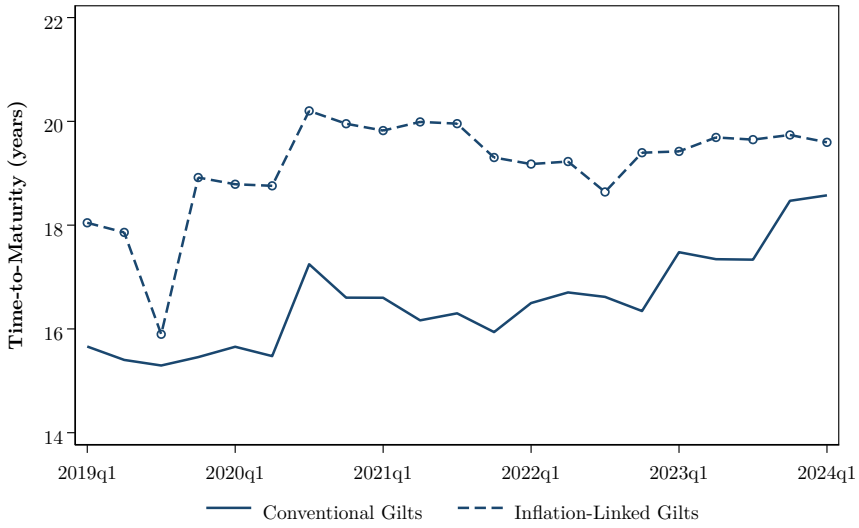
Combining these results, we learn that when monetary policy tightens, intermediaries become less willing to hold gilts on their balance sheets, and sell them to long-term investors. Our interpretation of this result is that a monetary policy tightening leads to a *widening* of gilt market segmentation, since the reduction in gilt inventories of intermediaries reduces their intermediation capacity to provide liquidity in the different market segments they intermediate across. As such, a supply of assets end up in the holdings of long-term investors, whose objective is not financial intermediation but rather to meet hedging requirements and funding targets. Importantly, our result also indicates that this widening of market segmentation in response to monetary tightening does not recover quickly, as there is a *quasi-permanent* shift in bond holdings across these agents in the market.

Caveats. This inference, however, is subject to the caveat that the cumulative responses that we estimated above can only serve as a proxy measure for the change in gilt holdings, since our MiFID II database does not reveal the stock measure. To partly overcome this limitation, we cross-validate our VAR results with lower-frequency gilt holdings data for banks and insurers—each of whom are members of the investor classifications above—from alternative sources as discussed in Section 2.4. There, our main finding is that banks reduced their inventories of long-term bonds held on their balance sheets between 2021 to 2024 as the Bank of England raised interest rates, while insurers lengthened duration risk (see Figures 3 and 7 respectively). These are dynamics of gilt holdings that are at best correlated with UK monetary policy during this period of time, since there were also several other shocks that hit the economy simultaneously. Through the lens of our VAR estimates, however, these dynamics are qualitatively consistent with a monetary policy tightening. Likewise, if the change in duration risk were to be endogenous to a monetary tightening, then the cumulative estimates from our VAR would indeed be a reliable measure of how the stock of gilt

[21] Within the literature, hedge funds are also sometimes considered as an intermediary with arbitrage capacity (Vayanos and Vila, 2021; He, Nagel, and Song, 2022). We show that hedge funds respond differently to GEMMs and inter-dealer brokers in our sample, at least to a monetary shock. They respond to a monetary easing by instead reducing the quantities of risk they hold. That said, our results are robust to the inclusion of hedge funds as an intermediary along side GEMMs and inter-dealer brokers (see Figure B.9). When we aggregate the measures of their bond transactions together, we find that the same monetary easing still leads to an overall cumulative increase in gilt holdings, only that the maximum attain is smaller in magnitude.

holdings respond. By cross-validating both sources, we are more confident that our VAR estimates provide an accurate view of how holdings adjust, which allows us to better connect our results to existing theories.

Figure 7 Insurers lengthen duration risk



Note: Figure shows the trade-weighted average time-to-maturity of gilt holdings held by insurers across all maturities in the conventional gilt market and the index-linked gilt market. Data is at a quarterly frequency. Source: Solvency II Directive, Bank of England.

Disaggregated responses. We conclude this section by providing estimates from additional VAR specifications that reveal how gilt holdings shift in the individual segments of the market, categorised by bond-type (conventional gilts v.s. index-linked gilts) and maturity segments ($<$ 15-year and $\geq$ 15-year). These specifications include the same set of prices, and differ from our baseline only in the set of traded quantities and their level of aggregation that we elaborate below.

Figure 11a illustrates the individual shifts in gilt holdings from different types of the longer-term gilt market in response to a monetary easing. These are estimates from a specification that replace the baseline set of quantities with an alternative set of four disaggregated quantity measures: the traded quantities of insurers and pension funds in the longer-term conventional gilt market and index-linked gilt market respectively, and the equivalent measures from GEMMs and inter-dealer brokers in each of these market segments. We recover the baseline specification if these gilt transactions are summed across the two types of markets for the same investor classification. These estimates indicate that most of our baseline results (both magnitude and sign of estimates) from Figure 6 can be attributed to responses from within the conventional gilt market, which yields the sharpest inference: we now observe an even slower rate at which intermediaries unwind their positions. These estimates thus suggest a widening of market segmentation that is even more persistent as compared to our baseline if monetary policy were to tighten. In contrast, the responses of intermediaries and long-term investors are flipped in the index-linked market segment and are much more imprecise. One explanation is that the demand for index-linked gilts from pension funds is close to being inelastic, likely due to the discrepancy between the limited supply and pension

funds' desire to hedge their large amount of inflation-linked liabilities. An alternative explanation may attribute this to trading noise: in appendix Figure B.10, we show that our baseline inference is preserved in this market segment when we re-estimate this VAR specification using instead a 20-day moving average of the four disaggregated quantity measures.

For completeness, Figure 11b also extends this analysis to the shorter-term segment ($<$ 15-year) corresponding to the two types of gilt markets. We replace the disaggregated quantities with those traded by the preferred-habitat clientele in these market segments appropriately. These are hedge funds and asset managers who are active in the conventional gilt market, and pension funds that also trade in the short-end of the index-linked market.[22] Our baseline inference also applies, although they are not as sharp as compared to the longer-term segment.

4.2 Full-Sample Estimates (Prices)

So far, we have presented our baseline estimates from a monetary shock that are identified using an external instrument that is marginally weak. This is likely due to the shorter sample period allowed by the inclusion of the traded quantities in our estimation.

In this section, we present full-sample estimates from an alternative VAR specification that *excludes* data on traded quantities from our baseline as a possible workaround. This is closer in tradition to the high-frequency literature that has solely focussed on prices across a long sample, since characterising the excess sensitivity of long-term interest rates to shocks does not require data on quantities per se. Our main inference of interest is therefore the extent to which long-term interest rates also persistently over-react to shocks—including non-monetary ones—over a longer sample. We see these estimates as potentially useful for shedding light on the external validity of our estimated quantity responses beyond the shorter data sample.

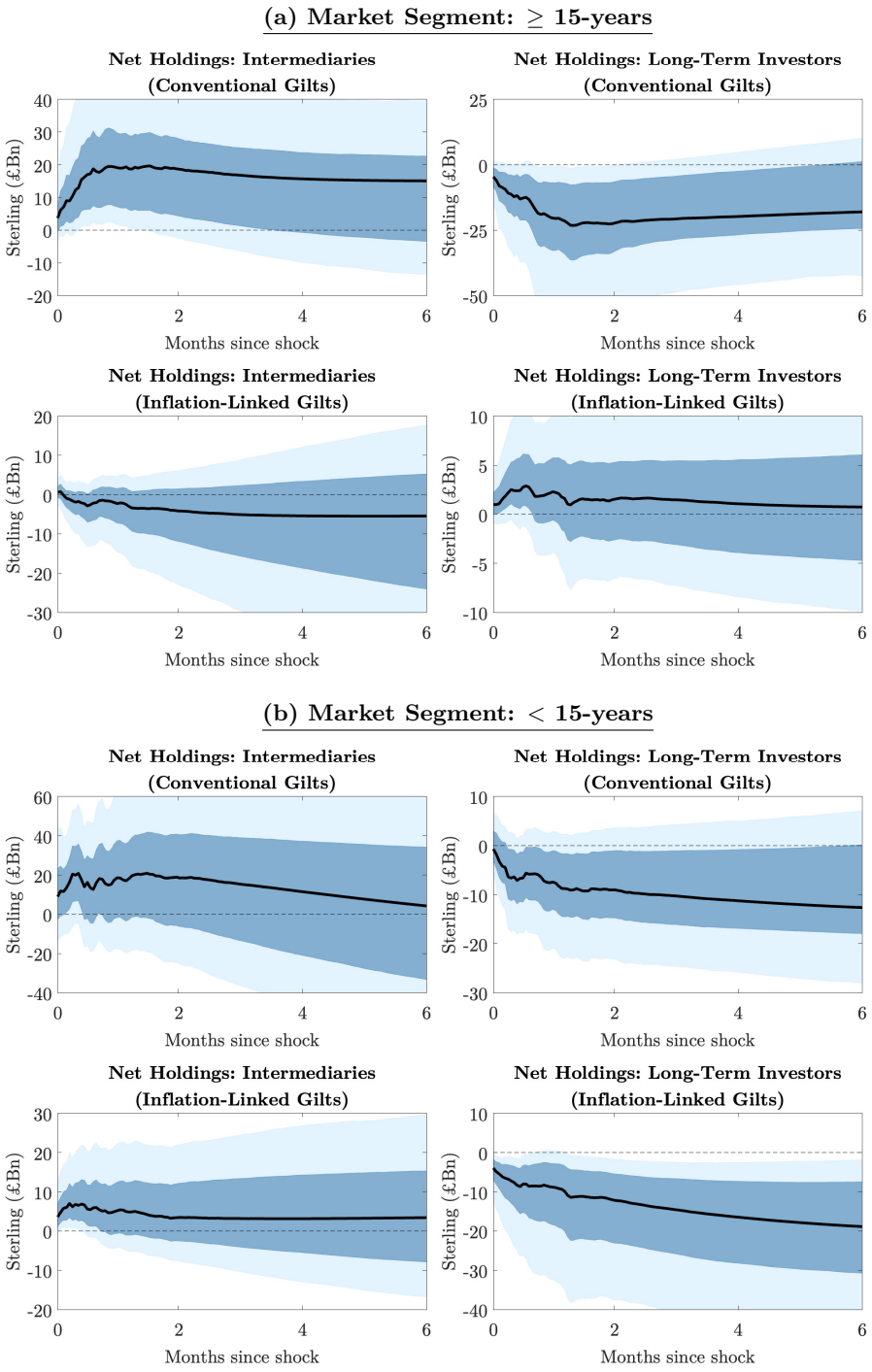
4.2.1 Inference based on Interest Rate Surprises to Shocks

We find, however, that these estimates over a longer sample do not allow us to reject the null of zero response from long-term real interest rates under our methodology. This is the case when we use interest rate proxies for both a monetary shock and a non-monetary inflation shock as external instruments for the 1-year yield when identifying these shocks from the VAR.[23] In Figure B.4, we show there is indeed a persistent rise in the 25-year forward instantaneous real rate in response to an increase in the 1-year yield, but both 68% and 90% confidence intervals are fairly large and include zeros across all horizons. The imprecision of these estimates may be a reason why the literature has primarily relied on estimating the instantaneous responses of asset prices to shocks through the use of event-based regressions at the high-frequency.

[22] We omit the other investor types from our analysis, such as e.g., sovereign wealth funds, non-financial firms, other financial and proprietary trading firms, as their trades are quantitatively small in comparison.

[23] Following convention, we use the target factor from Braun, Miranda-Agrippino, and Saha (2025) as an external instrument for the 1-year yield when identifying the monetary shock, and the high-frequency interest rate surprise in response to UK RPI inflation release dates measured from the SONIA futures contract for delivery in 4-quarters' time as the external instrument for the 1-year yield when identifying the non-monetary inflation shock.

Figure 8 Bond holdings in response to monetary News (↓ in 1-year yield)



Note: IRFs indicate the responses of bond holdings from intermediaries and habitat investors to a monetary policy easing. Intermediaries in each individual segments are GEMMs and inter-dealer brokers. Habitat investors are asset managers (< 15-year) and insurers (≥ 15-year) in the conventional gilt market, and pension funds in the index-linked gilt market (both maturity segments). Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 9.6. The Lewis & Merten (2025) generalized test statistic is 4.1. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

4.2.2 Inference based on Inflation Swap Surprises to Shocks

Given these results, we propose and showcase an alternative instrument that may be well-suited for our application. This alternative relies on the use of inflation swap surprises measured at high-frequencies in response to these news events as an external instrument for *breakeven inflation* within the same VAR specification (see Equation (6)). This proxy instrument also satisfies the exclusion restriction required for identification, but can differ from interest rate surprises in terms of statistical power.[24]

We find that identification of our VAR system based on this external instrument has the property that there is always a gradual and persistent rise of long-term forward rates with humped-shape dynamics in response to shocks, even if these shocks were to only raise expected inflation in a transitory manner: long-term interest rates remain significantly elevated long after expected inflation has reverted, and the result applies to both nominal and real rates. Figure 9 presents these estimates when we measure inflation swap surprises at high-frequencies either to a monetary shock, or a direct shock to inflation that was unexpected by the financial market. To compare the estimation results across the two shocks, we scale the impulse response functions such that both lead to an increase in the 3-year breakeven inflation by 1 percentage point on impact.

Monetary shock. In panel (a), the shock that raises expected inflation on impact is a surprise monetary policy easing, identified using inflation swap surprises measured in response to MPC policy announcements as an external instrument to the VAR. The identified responses are such that monetary policy reacts by immediately raising the 1-year yield to fend off inflationary pressures, to which breakeven inflation responds by falling quickly, but the 25-year forward nominal and real rates rise gradually with humped-shapes, and they remain persistently elevated even after inflation expectations have reverted. As the rise in expected inflation is only transitory, there are no discernible differences in how forward nominal and real rates respond.

To cross-validate the use of these inflation swap surprises, we apply them to our baseline identification over a shorter data sample in Section 4.1 and find that we recover estimates of both prices and quantities that are similar in sign and magnitude as compared to the baseline. This suggests that they identify the same monetary shock.[25] These estimates are presented across Figure B.11 and Figure B.12.[26] This opens the door for our baseline inference—particularly the responses of quantities—to possibly hold over a longer data sample.

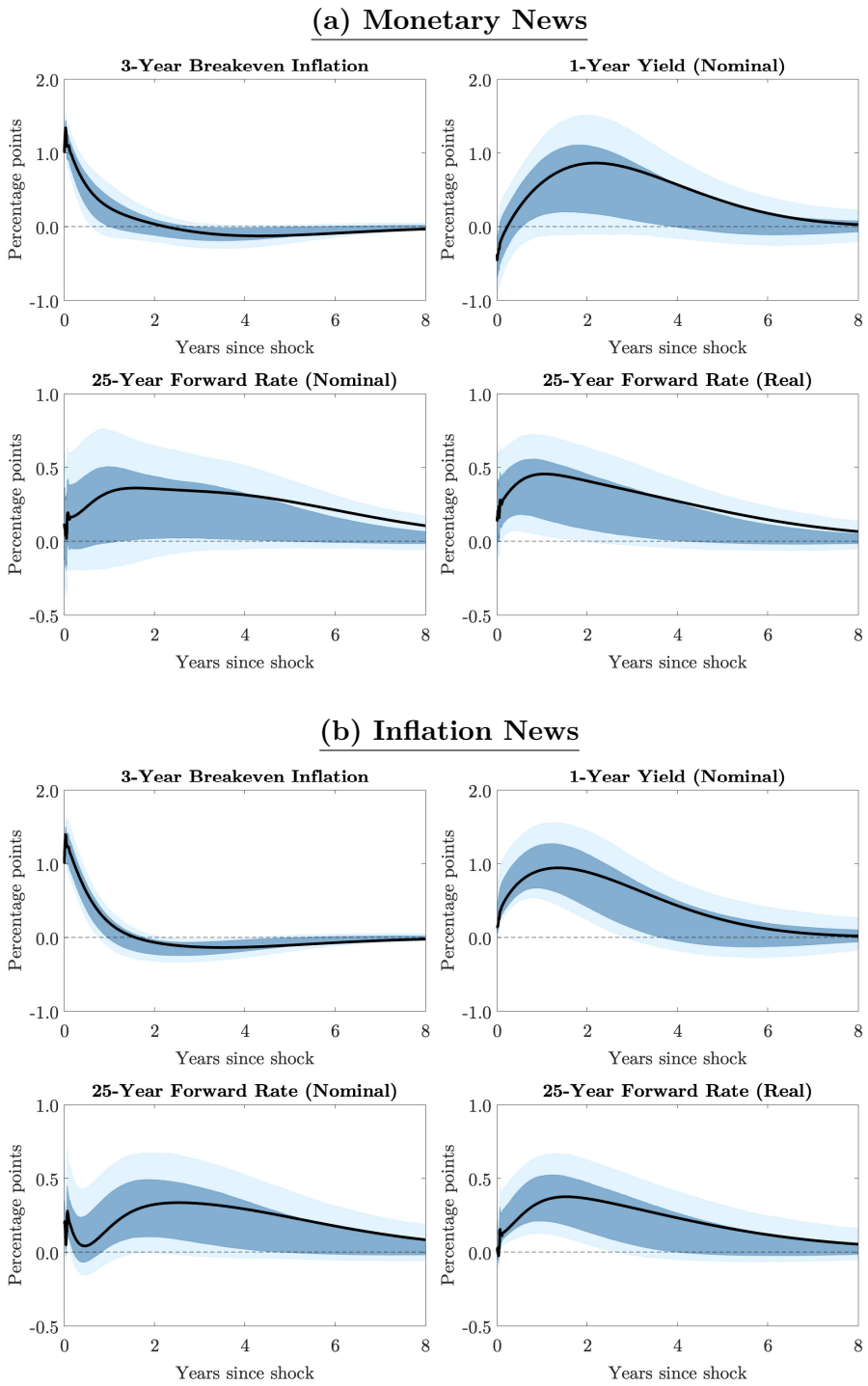
Non-monetary inflation shock. In panel (b), the shock that raises expected inflation is simply a direct shock to inflation itself, measured from a surprise inflation print that arises from the lumpy release of inflation data. Similar to the surprise monetary easing, expected inflation rises on impact but falls relatively quickly as soon as the 1-year yield begins to rise. Thus, our VAR estimates suggest that financial markets do indeed anticipate that the Bank of England will

[24] These instruments equally satisfy the exclusion restriction by design, refer to Section 3.1 for more details.

[25] Note that in the shorter sample, the first-stage F-statistics from both sets of monetary proxies are similar: they are 9.8 for the first-stage identified by inflation swap surprises and 9.6 for the target factor.

[26] Figure B.12 additionally illustrates the responses of estimated gilt *flows* to the monetary shock. The cumulative sum of these responses give rise to the quantity responses presented in Figure B.11.

Figure 9 The Persistent Overreaction of Long-Term Interest Rates



Note: Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 20.2 for monetary news, and 223.7 for inflation news. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

raise interest rates quickly when inflation expectations rise: the 1-year yield reaches its peak within approximately a year of the shock. Long-term forward rates rise persistently, and remain elevated long after expected inflation has reverted.

In appendix B.1, we provide a possible interpretation of what the direct shock to inflation estimated above is, which is arguably nonstandard. We argue that it resembles a transitory business cycle shock that only shifts short-term—not longer-term—inflation expectations (i.e., a slope factor), and is likely to be demand-driven. By this view, we would interpret our estimation results above as the response of long-term forward rates to exogenous demand shocks that are transitory. In fact, our estimation results are even sharper when we focus on this factor, which explains approximately 70% of the UK RPI swap curve variance on inflation release dates.[27] We also show that a supply-driven shift in expectations, such as that arising from imported oil shocks, lead to similar responses of expected inflation and long-term interest rates except these are quantitatively much larger in magnitude (even when the shock is scaled comparably).

Instrument strength. While both inflation swap and interest rate surprises satisfy the exclusion restriction to the same release of information, they can differ in statistical power largely for the identification of an inflation surprise. For example, the first-stage F-statistic is 224 when we use these inflation swap surprises to a UK RPI data release as an instrument for breakeven inflation to identify a non-monetary inflation shock in Figure 9b. This is much larger than the F-statistic of 40 provided by an interest rate-based proxy, possibly contributing to the statistically significant estimates.[28]

To ensure that our inference by the use of inflation swap surprises as an external instrument for the VAR is robust, we additionally implement an alternative, heteroskedasticity-based identification scheme that, while similar in that it also relies on the lumpy release of news, differs from the proxy-IV approach in that it does not rely on these inflation swap surprises. It relies instead on variation from shifts in the variance of the yield curve on days when there is a particular news release. Consider, for instance, the identification of a direct shock to inflation. Identification by heteroskedasticity imposes the structure that the variance of the reduced-form $\mathbf{\Sigma}_u = \mathbb{V}\text{ar}[\mathbf{u}_t]$ can be partitioned into a low (L) and a high (H) variance regime, with the high variance regime representing the subset of inflation data release dates (a total of 221 UK RPI data release dates in our sample):

$$\mathbf{\Sigma}_u = I_t \cdot \mathbf{\Sigma}_H + (1 - I_t) \cdot \mathbf{\Sigma}_L, \tag{8}$$

where $\mathbf{\Sigma}_\ell = \mathbf{B}\mathbf{\Omega}_\ell\mathbf{B}'$, and $\mathbf{\Omega}_\ell = \mathbb{V}\text{ar}[\boldsymbol{\varepsilon}_{\ell,t}]$ is the covariance matrix of the structural shocks that differ depending on the regime $\ell \in [L, H]$. By assuming that those direct shocks to expected inflation on these dates drive the shift in variances across the two regimes, we can identify this shock up

[27] The first two factors together explain approximately 80% of the UK RPI swap curve variance on inflation data release dates. We interpret the second factor as a level factor which shifts longer-term inflation expectations. News shocks, perceptions of monetary policy credibility and inflation uncertainly would likely belong to this factor.

[28] This is not the case, however, for the identification of monetary news. The relevant F-statistic is 20.2, as compared to 43.4 provided by the target factor of Braun, Miranda-Agrippino, and Saha (2025).

to sign and scale.[29] Figure B.5 compares the identification by heteroskedasticity of a direct shock to inflation to estimates from Figure 9 across the term structure: both estimates are extremely similar, with the same gradual and persistent rise in long-term forward rates.

In summary, we highlight two distinct features of the UK yield curve that describe how it responds to macroeconomic shocks. First, the speed at which expected inflation reverts suggests that inflation expectations are well-anchored, and shocks do not generate any long-lasting effects on long-run inflation expectations. Second, long forward rates rise gradually with humped-shapes and remain persistently elevated, even after the shock to expected inflation has dissipated. We argue that this likely reflects a *persistent* overreaction of long-term interest rates to shocks relative to the expectations benchmark.

4.2.3 The Measurement of Excess Sensitivity

Why do our estimates imply a persistent overreaction of long-term interest rates? How severe is this overreaction, and is it quantitatively large? These are questions that cannot be answered in the absence of a credible benchmark. The consensus within the literature seems to be that the expectations theory of the term structure is a natural choice for this benchmark.[30] Yet, most empirical estimates of excess sensitivity, such as ours above, do not come with alternative measures of what those responses should be for them to be consistent with economic theory.

We attempt to overcome this limitation by providing a model-free measurement approach that is particularly well-suited to our VAR model, and show that this can be a useful method to measure the sign and magnitude of excess sensitivity.[31] This method is model-free in that it does not require any calibrated parameters, but it still relies on the theoretical underpinnings of the expectations hypothesis. Its main feature is to solely use the estimates of the entire term structure response to a shock to internally cross-examine one another within the system.

We start by imposing two assumptions: (1) markets are frictionless, and (2) there is a shock ε_t with property $\mathbb{E}_t[\varepsilon_{t+1}] = 0$. According to the expectations theory, no-arbitrage pricing under assumption (1) implies that the one-period forward rate $f_t^{(\tau-1,\tau)}$ known in time t to prevail in the future between periods $t+\tau-1$ and $t+\tau$ is defined as:[32]

$$f_t^{(\tau-1,\tau)} = \mathbb{E}_t[i_{t+\tau-1}], \tag{9}$$

where i_t refers to the period t short-rate.

By the law of iterated expectations and assumption (2), the dynamic responses to a shock at

[29] Formally, the eigenvector associated with the largest eigenvalue of $\mathbf{\Sigma}_L^{-1}\mathbf{\Sigma}_H$ identifies this shock.

[30] See e.g., Greenwood, Hanson, and Vayanos (2024). The authors provide a risk-based, term premia explanation of excess sensitivity of long-term interest rates based on the preferred-habitat view of financial markets, and provide conditions when there is over- and under-reaction of these interest rates relative to the expectations hypothesis.

[31] We thank Ricardo Reis for this suggestion.

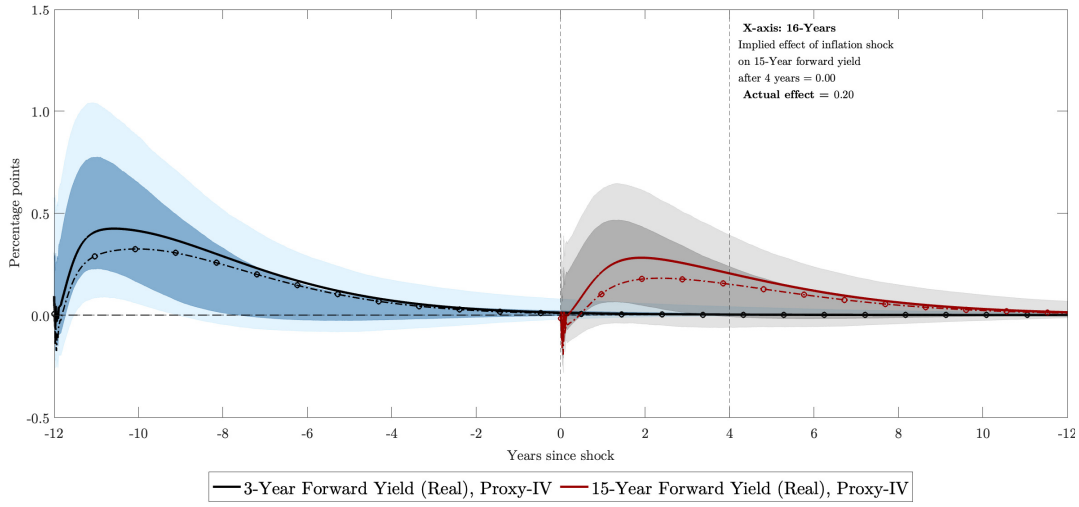
[32] For example, $f_t^{(1,2)} = \mathbb{E}_t[i_{t+1}]$ is the 1-year (ahead) forward rate. Our estimates rely on τ-year ahead instantaneous forward rates, defined as $\lim_{\Delta t \to 0} f_t^{(\tau-\Delta t,\tau)} = \mathbb{E}_t[i_{t+\tau}]$.

horizon h (i.e., after h periods) must satisfy:

$$\underbrace{\frac{df_{t+h}^{(1,2)}}{d\varepsilon_t}}_{\substack{IRF\ of\ f_t^{(1,2)} \\ h\ periods\ after\ shock}} \equiv \frac{d\,\mathbb{E}_{t+h}[i_{t+h+1}]}{d\varepsilon_t} \overset{L.I.E.}{=} \frac{d\,\mathbb{E}_t[i_{t+h+1}]}{d\varepsilon_t} \equiv \underbrace{\frac{df_t^{(h+1,h+2)}}{d\varepsilon_t}}_{\substack{Response\ of\ f_t^{(h+1,h+2)} \\ on\ impact}} . \tag{10}$$

This suggests a VAR system restriction for how forward rates across the term structure should respond to the shock under the expectations theory. In practice, this implies that the effect of the shock on the τ-period instantaneous forward rate *after* h periods is exactly its immediate impact effect on the $(\tau + h)$-period instantaneous forward rate, and so on, for a value of τ that spans the term structure. For instance, the response of 3-year forward rates to a shock after 12 years would equal the impact response of the shock on 15-year forward rates.[33] We cross-examine these estimated impulse response functions across different market segments in Figure 10. The confidence intervals estimated from responses of 3-year forward rates include zero after 12 years. These intervals estimated from 15-year forward rates also include zero on impact, but they rise soon after. In particular, the 68% confidence interval does not include zero for about two years. According to this measure, there is an overreaction of the 15-year forward rate for roughly 2 years. Even after 4 years, the point estimate implies that this overreaction is by roughly 20 basis points, highlighting the persistence of the excess sensitivity.[34]

Figure 10 The measurement of excess sensitivity



Note: The x-axis has been shifted by 12 years to the right when interpreting the responses of the 3-year forward rate. Impulse responses are identified from a direct shock to expected inflation, using inflation swap surprises as an external instrument for breakeven inflation. Dashed lines are heteroskedasticity-based estimates. The First-Stage F-statistic is 219.6. The Lewis & Merten (2025) generalized test statistic is 47.9. Shaded intervals are 68% and 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

[33] See Figure B.6 for an illustration of this result.

[34] In Figure B.7, we show that this argument can be made even sharper with a rotated inflation factor that only shifts short-term and not longer-term inflation expectations.

4.3 Discussion of Results

Up to this juncture, we have documented new facts that characterise the excess sensitivity of long-term interest rates. We have shown that this excess sensitivity to shocks—whether monetary or non-monetary—also lies in the time-series, where long-term interest rates persistently overreact to the 1-year yield with humped-shape dynamics (characterised by an initial underreaction). Our estimates of the monetary shock further reveal that long-term investors exhibit an *downward-sloping* demand curve, such that intermediaries in the gilt market increase their long-term bond holdings in response to a monetary policy easing, and the subsequent change in the *level* of these holdings seems to be quasi-permanent. How can we make sense of these results in the light of existing theories?

A view from the preferred-habitat literature (Vayanos and Vila, 2021) suggests segmented markets populated by heterogeneous agents, whom, because of their different trading motives, capacity to bear risk and preferred habitats, exhibit varying elasticities of demand for the underlying asset (gilt) traded. As a result, a shock to a common fundamental—such as a monetary shock—leads to both demand and supply imbalances within these segmented markets that can push its market-clearing prices away from economic fundamentals. In particular, one sharp prediction of models in this tradition is to tie the direction of this disconnect to the change in the net amount of risk borne by the marginal trader in this market segment. As the marginal trader is typically considered to be risk averse, she requires greater compensation for bearing larger quantities of risk. As such, an overreaction of yields in the long-end in response to shocks to the short rate can arise due to positive term premia that the marginal trader demands, as a result of shifts in the common fundamental that leads her to bear more quantities of this risk in equilibrium.[35]

Hence, the literature so far has explained the excess sensitivity of long-term interest rates by relying on versions of the demand curve of habitat investors to be *decreasing* in bond yields (and increasing in prices). Hanson, Lucca, and Wright (2021) refers this as a form of *rate-amplifying* demand, that, for instance, describes the behaviour of long-term investors who are either biased and over-extrapolate changes in the short rate (Cieslak, 2018; Giglio and Kelly, 2018; d'Arienzo, 2020), or those who demonstrate a reach-for-yield behaviour (Hanson and Stein, 2015).[36] According to this view, the net supply of long-term bonds that arbitrageurs have to warehouse on their balance sheet is *positively* correlated with shocks to the short rate, conditional on the gross supply of long-

[35] See Greenwood, Hanson, and Vayanos (2024) for an in-depth discussion.

[36] Hanson, Lucca, and Wright (2021) also raise mortgage refinancing as a possible explanation for rate-amplifying demand, which is more relevant in countries such as the US.

term bonds being fixed (Jansen, Li, and Schmid, 2024).[37] As a result, these mechanisms typically require changes in the stock of bond holdings of arbitrageurs to be in the same sign as the changes in the short rate in order to generate an overreaction of long-term interest rates.

Our estimates, however, point to a different set of sign restrictions that is satisfied by the data. We find that pension funds and insurers who occupy the long-end of the gilt market behave differently in contrast to the predictions of the abovementioned mechanisms. In response to an unexpected decline in the short rate, these long-term investors reduce their holdings of long-term gilts—not increase them. The overall duration risk held by UK insurers, as we have shown in Figure 7, also rose during the period when the Bank of England raised the Bank Rate. These are outcomes inconsistent with a rate-amplifying demand characterisation of long-term investors. Therefore, these long-term investors seem to populate a supply curve that is upward-sloping—not downward-sloping—in prices, at least for the UK. As a consequence, we also observe the net supply of gilts held by GEMMs as intermediaries in this market to be *negatively* correlated with the change in the short rate.[38] According to a standard Vayanos-Vila environment, this would generate an under-reaction of long-term interest rates that is inconsistent with our earlier findings that they persistently over-react to changes in the short rate.

We stress, however, that we are not objecting to rate-amplifying demand or the existence of downward-sloping supply from long-term investors. We instead argue that these mechanisms may be shock-dependent. In particular, our estimates above are based solely on the identification of an unexpected change in the short rate that arises from a *monetary* shock: it is only conditional on a monetary shock that we identify the set of signs satisfied by our price and quantity estimates above. Many models in this tradition, however, are partial equilibrium models that carry a representation of the short rate as an exogenous process where the short rate shock is a forecast error component. These models do not cast a strict distinction between monetary and non-monetary shocks to the short rate, which may be important for how intermediaries and long-term investors respond. There is little evidence that non-monetary shocks impose the same sign restrictions on the responses of prices and quantities as would a monetary shock. By this reasoning, there could be intrinsic differences in how investors respond to non-monetary shocks that could render the rate-amplifying demand channel to be more consistent with non-monetary driven short rate changes than monetary-

[37] According to the view of extrapolative investors, they overestimate the rise of short rates when they rise unexpectedly. These investors then respond by reducing their demand for long-term bonds. The quantity of bonds that must be held by unbiased investors—who serve the capacity of an arbitrageur—therefore rises, leading to a rise in term premium and generating an excess sensitivity of long-term yields in response to a rise in the short rate. Under a reach-for-yield type of investor behaviour, investors who seek to attain an overall target return in their portfolio demonstrate yield-seeking behaviour by increasing their demand for long-term bonds when there is an unexpected decline in short rates. Under the assumption that gross supply is bonds in the market is fixed, arbitrageurs clear the market by selling bonds to them and thus hold less interest rate risk as a result. In equilibrium, this reduces the term premia and thus reduces long-term interest rates (excess sensitivity) as short rates decline.

[38] Our identification relies on the same assumption imposed by the rate-amplifying demand view of the market, which is that the gross supply of UK gilts in the market remains fixed. Our identification assumption is that this gross supply does not change in response to a monetary policy shock, which we assume to hold at the high-frequency. Our estimation results suggest that this assumption holds ex post, since the cumulative amount of bond holdings that intermediaries acquire onto their balance sheet is approximately equal to the cumulative amount sold by the long-term investors.

based changes. We return to this point again towards the end of the paper.

In the light of all these, we highlight an alternative mechanism from Kekre, Lenel, and Mainardi (2024) that does not restrict long-term investors to a downward-sloping supply curve, but rather, rely on intermediary wealth effects to explain the excess sensitivity of long-term interest rates in response to a monetary shock within the same Vayanos-Vila environment. According to this view, intermediaries reduce the quantities of interest rate risk they bear on their balance sheets in response to an unexpected rise in the short rate. However, the subsequent reduction in portfolio wealth additionally raises intermediary risk aversion and thus the price of risk, conditional on their portfolio featuring positive duration. If this channel dominates, an increase in the price of risk leading to an overreaction of the long-term rate can equally be observed with a *decline* in the stock of bond holdings held by these intermediaries.

Our estimates satisfy these counterfactual sign restrictions that this mechanism imposes on a monetary shock, and we view them as offering new evidence in direct support of such intermediary wealth effects at play. While we do not provide any measures of intermediary net worth that would be informative of such channels, we let our data speak by identifying the sign in which traded quantities respond to a monetary shock, which would be a stronger and more direct validation of the proposed mechanism.

Through the lens of our results, they suggest that the persistent overreaction of long-term interest rates in the UK gilt market in response to an unexpected increase in the short rate is more likely to be driven by intermediaries whose risk-bearing capacity falls as yields rise, rather than long-term investors with a downward-sloping supply. There is a quasi-permanent widening of market segmentation with a slow recovery of bond inventories from these financial intermediaries, leading to a persistent disconnect of long-term interest rates in this market segment from economic fundamentals.

5 Model

Having discussed our main empirical results, we now focus our attention on a model that can rationalise them and provide counterfactuals for their effects on the aggregate economy.

In summary, our model has to match three features that we have identified from the data. First, long-term interest rates exhibit excess sensitivity and over-react persistently to shocks with humped-shape dynamics. Second, these interest rates are traded by long-term investors against intermediaries in segmented bond markets. Third, an unexpected rise in the short rate as a result of monetary tightening leads to a quasi-permanent shift in bond holdings from intermediaries to long-term investors: intermediaries reduce the stock of their long-term bond holdings and aggregate duration risk as short rates rise, only recovering them at a much slower pace thereafter.

Motivated by these facts, we present a New Keynesian model featuring a segmented bond market in the Vayanos-Vila tradition where investors trade a set of nominal and inflation-linked perpetuities. In this model, aggregate shocks lead to demand and supply imbalances within these

market segments that lead to sustained deviations of prices from economic fundamentals, reconciling the excess sensitivity of long-term interest rates with the property of long-run monetary neutrality. The main actors in our model are a set of arbitrageurs and habitat investors, that respectively represent GEMMs and long-term investors such as pension funds and insurers. Given our sign estimates of how these arbitrageurs respond to a monetary shock, we adopt the mechanism of Kekre, Lenel, and Mainardi (2024) but additionally interact it with portfolio transaction costs to slow down the recovery of arbitrage capacity after a monetary tightening. We show that in the presence of persistent demand from long-term investors, this can generate humped-shaped dynamics of both prices and quantities that are qualitatively consistent with our estimates. We further nest this mechanism in a New Keynesian general equilibrium following the footsteps of Gourinchas, Ray, and Vayanos (2022); Ray, Droste, and Gorodnichenko (2024), so that expected inflation is endogenous to the dynamics of long-term nominal and real rates, and non-monetary shocks to the short rate also invite the same overreaction as illustrated by our estimates.

In the subsequent section, we calibrate the abovementioned mechanism using our MiFID II data on bond transactions, and subsequently use this model as a laboratory to provide a series of counterfactual analyses for the transmission of monetary policy and the design of open-market operations. We return to this point at a later part in this paper.

5.1 The Bond Market

We start with the financial market, where the heart of the model's mechanism lies. We first describe the assets traded to define their payoffs and excess returns, before defining the segmentation of this market. Then, we describe where demand and supply of these assets arise, and provide intuition for how the model mechanisms collectively work to qualitatively match our empirical results.

5.1.1 Assets and Market Segmentation

We assume that the financial market trades a set of risky, long-term bonds that are in net zero supply. These long-term bonds differ by their type: they are either nominal bonds or inflation-linked bonds. We model them as perpetuities with declining coupon payments that decay at a rate $\kappa_d \in [0, 1)$, where κ_d is the key parameter that parameterises the duration of the asset, $d = (1 - \overline{\beta}\kappa_d)^{-1}$. In the special case where $\kappa_d = 0$, the asset replicates a payoff profile identical to a one-period bond. By assigning different values of κ_d to these perpetuities, we are able to recover a term structure representation for these long-term bonds. We denote this representation using $d \in \Xi$. As such, long-term bonds traded in this financial market are heterogeneous in two dimensions: by duration and bond-type.

Motivated by our empirical facts in Section 2.3, we adopt the feature of bond market segmentation, where each segment of the market is differentiated according to the duration and bond-type of the underlying bond traded. In equilibrium, bond prices are determined by demand and supply in each of these market segments that we will shortly describe. We let the price of the nominal

bond and inflation-linked bond of duration d be denoted by $Q_t^{(d)}$ and $J_t^{(d)}$ respectively, and their yields to be inversely defined as:

$$\exp(y_t^{(d)}) = \frac{1}{Q_t^{(d)}} + \kappa_d, \;\; \exp(\widetilde{y}_t^{(d)}) = \frac{1}{J_t^{(d)}} + \kappa_d. \tag{11}$$

The ex post *excess* returns from these perpetuities over the gross real rate R_t (and net real rate r_t) from period t to $t+1$ are respectively defined as:

$$\mu_{t+1}^{(d)} = \left(\frac{P_t}{Q_t^{(d)}} \frac{1 + \kappa_d Q_{t+1}^{(d)}}{P_{t+1}} \right) - R_t = R_t^{(d)} - R_t, \tag{12}$$

$$\widetilde{\mu}_{t+1}^{(d)} = \left(\frac{1 + \kappa_d J_{t+1}^{(d)}}{J_t^{(d)}} \right) - R_t = \widetilde{R}_t^{(d)} - R_t. \tag{13}$$

Under the assumption of risk neutrality and no limits to arbitrage, there will be zero excess returns from investing in the long-term bond and financing this position via the short rate. Then, Equations (12) and (13) characterise the transmission of the short rate into the long solely via the expectations channel. However, away from these assumptions, and given that our bond market is segmented, this transmission now depends on the adjustment of demand and supply in these market segments in response to shocks and changes in the short rate.

5.1.2 The Blanchard-Yaari Arbitrageur

We now specify the arbitrageur's environment. The arbitrageur is broadly defined as an investor who trades across all market segments, with the objective of arbitraging away excess returns through a series of yield curve carry trades financed by the riskless rate. In the data, the arbitrageur represents the collection of GEMMs and dealer banks, who share the characteristic of maintaining a bond inventory on their balance sheets to provide liquidity to end-demand.

Following the footsteps of Kekre, Lenel, and Mainardi (2024), we represent these arbitrageurs as a continuum of Blanchard-Yaari perpetual youths indexed by $i \in [0,1]$, who derives utility from consumption only upon death. Death occurs at a probability $\zeta \in (0,1)$, and the unit mass is preserved by birth at the same rate. As such, there is a stationary measure of arbitrageurs at any point in time. Their period utility takes the CRRA form with relative risk-aversion φ. These assumptions imply that her utility satisfies the Inada condition in portfolio wealth. An accumulation of wealth reduces her risk aversion, and vice versa. As such, the dynamics of portfolio wealth generate time-varying risk aversion in this environment, such that their demand for long-term bonds across market segments is not perfectly elastic. We additionally assume that arbitrageurs face portfolio adjustment costs when adjusting their positions, which gives rise to a form of slow-moving capital across market segments to slow down the recovery of holdings after a shock. This allows the model to better match the dynamics in the data by generating humped-shape responses in both bond yields and the stock of bond holdings. Limits to arbitrage in our environment thus

arise from both risk aversion and slow-moving capital with cross-amplification effects in equilibrium.

The Bellman equation of the arbitrageur is thus given by:

$$V(W_{i,t}, \mathbf{q}_{i,t-1}) = \max_{\mathbf{q}_{i,t}} \overline{\beta}\, \mathbb{E}_t \left(\zeta \frac{W_{i,t+1}^{1-\varphi} - 1}{1-\varphi} + (1-\zeta) V(W_{i,t+1}, \mathbf{q}_{i,t}) \right), \;\; i \in [0,1], \tag{14}$$

where $\overline{\beta}$ is the arbitrageur's time discount factor. We use $\mathbf{q}_{i,t}$ to denote a vector of positions optimally chosen by arbitrageur i in period t. This vector is a quantity measure of the *stock* of bonds held by the arbitrageur across the market segments that differ by both duration and bond-type. Due to the presence of portfolio adjustment costs, $\mathbf{q}_{i,t-1}$ is a state variable to the arbitrageur in period t. The other state variable to this problem is portfolio wealth, denoted in real terms by $W_{i,t}$. Portfolio wealth is assumed to accumulate according to:

$$W_{i,t+1} = \exp(r_t) W_{i,t} + (\mathbf{q}_{i,t} \circ \mathbf{p}_t)' \boldsymbol{\mu}_{t+1} - C(\mathbf{q}_{i,t}, \mathbf{p}_t, \mathbf{q}_{i,t-1}, \mathbf{p}_{t-1}; \boldsymbol{\Gamma}), \;\; W_{i,0} \;\; \textit{given}, \tag{15}$$

and aggregate wealth follows the law of motion:

$$W_{t+1}^a = (1-\zeta) \sum_i W_{i,t+1} + \zeta \overline{W}, \tag{16}$$

where $\overline{W}$ denotes an exogenous endowment received by newly born arbitrageurs.

This accumulation equation states that the arbitrageur can grow her portfolio wealth through a set of yield curve carry trades, where she borrows at the riskless real rate r_t to finance her positions in perpetuities of longer duration risk.[39] Stacking Equations (11), (12) and (13) across the set of durations $d \in \Xi$, we denote the vector of excess returns from these positions by $\boldsymbol{\mu}_{t+1}$ and their associated bond prices by $\mathbf{p}_t$.[40] We denote portfolio adjustment cost function by C and impose a set of three assumptions on this function: (i) it penalises the real dollar value of portfolio wealth; (ii) the arbitrageur incurs costs in changing the value of her position; and (iii) the function C is convex and parameterised by a quadratic form, given by:

$$C(\mathbf{q}_{i,t}, \mathbf{p}_t, \mathbf{q}_{i,t-1}, \mathbf{p}_{t-1}; \boldsymbol{\Gamma}) = \Lambda(\mathbf{q}_{i,t}, \mathbf{p}_t, \mathbf{q}_{i,t-1}, \mathbf{p}_{t-1})' \boldsymbol{\Gamma} \Lambda(\mathbf{q}_{i,t}, \mathbf{p}_t, \mathbf{q}_{i,t-1}, \mathbf{p}_{t-1}), \tag{17}$$

where $\Lambda(\mathbf{q}_{i,t}, \mathbf{p}_t, \mathbf{q}_{i,t-1}, \mathbf{p}_{t-1}) = (\mathbf{q}_{i,t} \circ \mathbf{p}_t) - (\mathbf{q}_{i,t-1} \circ \mathbf{p}_{t-1})$. The matrix $\boldsymbol{\Gamma}$ contains fixed coefficients that serves as a catch-all for frictions to liquidity, balance sheet capacity or behaviour costs to adjustment. Importantly, it is used in our environment to regulate direct spillovers across different market segments. For instance, a diagonal $\boldsymbol{\Gamma}$ imposes the assumption that, apart from their indirect effects on wealth and thus risk-bearing capacity, adjustment frictions in one segment of the market

[39] This assumption is without loss of generality. Alternatively, the arbitrageur can also borrow at the riskless nominal rate instead, but that requires specifying the arbitrageur's preferences in nominal terms rather than in real terms.

[40] This is a modelling assumption that allows a quantity measure that is closer to our data on MiFID II bond transactions, since these are denoted by notional values.

do not directly impede the capacity of the arbitrageur to shift capital across other segments of the market.

We emphasize one other feature that distinguishes our environment from other Vayanos-Vila type models: our arbitrageur holds an optimal portfolio allocation not only across different durations, but also across both nominal and inflation-linked perpetuities. We rely on this feature to explain our empirical finding that breakeven inflation always reacts to an inflationary shock in a transitory manner, in contrast to the persistent responses of long-term rates. Through the lens of our model, this is because there is an implicit inflation curve trade that arises in equilibrium. An inflationary shock that raises long-term nominal rates encourages the arbitrageur to substitute away from inflation-linked perpetuities as the nominal yield curve steepens, but this exerts an equilibrating force that raises long-term real rates, holding as constant the supply of inflation-linked perpetuities. Our model thus predicts that inflation risk-bearing capacity of the financial market can be tied to the stock of real assets available. We return to this point again at a later part of the paper.

5.1.3 Optimality and Nesting Benchmarks

By the envelope condition, it is straightforward to show that the vector of positions $\mathbf{q}_t$ chosen by the arbitrageur must satisfy the following First-Order Condition in equilibrium:

$$\mathbb{E}_t\left[\left(\zeta W_{i,t+1}^{-\varphi}+(1-\zeta)\frac{dV(W_{i,t+1},\mathbf{q}_{i,t})}{dW_{i,t+1}}\right)\frac{dW_{i,t+1}}{d\mathbf{q}_{i,t}}+(1-\zeta)\frac{dV(W_{i,t+1},\mathbf{q}_{i,t})}{d\mathbf{q}_{i,t}}\right]=\mathbf{0}, \tag{18}$$

where the marginal change in wealth of the arbitrageur in response to $\mathbf{q}_{i,t}$ satisfies:

$$\frac{dW_{i,t+1}}{d\mathbf{q}_{i,t}}=\Big[\boldsymbol{\mu}_{t+1}-\boldsymbol{\Gamma}\Lambda(\mathbf{q}_{i,t+1},\mathbf{p}_{t+1},\mathbf{q}_{i,t},\mathbf{p}_t)\Big]\circ\mathbf{p}_t. \tag{19}$$

This is informative of how portfolio adjustment costs interact with excess returns and thus bond prices in equilibrium. Given a vector of excess returns $\boldsymbol{\mu}_{t+1}$, the arbitrageur is incentivised to shift her positions across different segments of the market to arbitrage them away. However, the presence of portfolio adjustment costs reduces these excess returns on *net*, and the arbitrageur shifts their holdings by less than desired. If these adjustment costs are sufficiently large to exceed the excess returns, then Equation (19) implies that these excess returns must rise. We show that this enables the model to rationalise humped-shape dynamics of bond yields in response to shocks.

The marginal values of the arbitrageur with respect to the state variables are further characterised by the following equations:

$$\frac{dV(W_{i,t+1},\mathbf{q}_{i,t})}{dW_{i,t+1}}=\overline{\beta}\,\mathbb{E}_{t+1}\left(\zeta W_{i,t+2}^{-\varphi}+(1-\zeta)\frac{dV(W_{i,t+2},\mathbf{q}_{i,t+1})}{dW_{i,t+2}}\right)\exp(r_{t+1}), \tag{20}$$

$$\frac{dV(W_{i,t+1},\mathbf{q}_{i,t})}{d\mathbf{q}_{i,t}} = \overline{\beta}\,\mathbb{E}_{t+1}\left(\zeta W_{i,t+2}^{-\varphi} + (1-\zeta)\frac{dV(W_{i,t+2},\mathbf{q}_{i,t+1})}{dW_{i,t+2}}\right)\boldsymbol{\Gamma}\boldsymbol{\Lambda}(\mathbf{q}_{i,t+1},\mathbf{p}_{t+1},\mathbf{q}_{i,t},\mathbf{p}_t)\circ\mathbf{p}_t. \quad (21)$$

Equation (20) is the marginal value of wealth. By investing a marginal unit of wealth at the riskless real rate, the arbitrageur receives a gross real return in the next period that either delivers marginal consumption utility in the instance of death, or the continuation value that is its recursive representation. Equation (21) refers to the marginal value from portfolio adjustment. This is small when the arbitrageur chooses a current position that lies in the proximity of her future position, and thus represents another forward-looking mechanism in the model—apart from the expected marginal value of wealth—that pins down current bond prices. Taken together with Equation (19), we find that the arbitrageur faces an additional inter-temporal trade-off. By adjusting her portfolio today, she bears the direct cost of this adjustment through a tax on the excess returns. However, she is additionally compensated by the marginal value of this adjustment, which is the marginal value of wealth she receives through a reduction of these future costs. Hence, the convexity of the cost function implies that our arbitrageur prefers inherently to make smaller portfolio adjustments in each period. This is the key mechanism that slows down the recovery of bond holdings after a shock.

To provide further intuition for how our model is able to match the empirical moments in the data, we collapse Equation (18) into three nesting benchmarks to isolate the mechanisms at work:

Proposition 1. *(Frictionless equiilibrium) Suppose that* $\boldsymbol{\Gamma}$ *is a null matrix, and steady state wealth* $\overline{W} \to \infty$*. Then, the yield curve is pinned down solely by the expectations hypothesis:*

$$\mathbb{E}_t[\boldsymbol{\mu}_{t+1}] = \mathbf{0}. \quad (22)$$

Proposition 1 thus corresponds to the special case where there are no limits to arbitrage. Since there are no portfolio adjustment costs, and the marginal value of wealth is equal to zero in the limit, the arbitrageur behaves as a risk-neutral investor with an infinitely elastic demand for perpetuities. She therefore trades freely and arbitrages away all excess returns, such that in equilibrium Equation (18) recovers an excess return vector equal to zero. In this limiting benchmark, the presence of the arbitrageur allows for full integration of segmented markets, and the transmission of shocks into the long rate occurs solely through the expectations channel. As such, there is no excess sensitivity of long rates, and bond holdings are irrelevant in equilibrium.

Proposition 2. *(Risk premia effects) Suppose that* $\boldsymbol{\Gamma}$ *is a null matrix, and steady state wealth* $\overline{W} \to 0$*. The marginal value of wealth becomes a relevant state variable for equilibrium bond pricing:*

$$\mathbb{E}_t\left(\zeta W_{i,t+1}^{-\varphi} + (1-\zeta)\frac{dV(W_{i,t+1},\mathbf{q}_{i,t})}{dW_{i,t+1}}\right)\boldsymbol{\mu}_{t+1} = \mathbf{0}. \quad (23)$$

Proposition 2 characterises the effect of intermediary wealth on equilibrium bond pricing. Assuming that the arbitrageur's portfolio features positive duration, a rise in the short rate leads to

a reduction in portfolio wealth, which increases the marginal value of wealth and risk aversion of the arbitrageur, given her CRRA utility-form. Expected returns must therefore rise with the price of risk, generating an overreaction in long rates relative to the expectations channel. If this effect is sufficiently large, it can qualitatively match the sign responses estimated in our data, even as arbitrageurs bear less quantities of risk. However, because Equation (23) is entirely forward-looking, it is unable to generate humped-shape dynamics from the data.

Proposition 3. *(Portfolio adjustment effects) Suppose that* $\mathbf{\Gamma}$ *is a positive-semidefinite matrix, and steady state wealth* $\overline{W} \to \infty$*. In equilibrium, the vector of excess returns must equal the marginal portfolio adjustment costs. Bond holdings become backward-looking, leading to a slower adjustment of bond prices:*

$$\mathbb{E}_t[\boldsymbol{\mu}_{t+1}] = \mathbf{\Gamma}\Big[(\mathbf{q}_{i,t} \circ \mathbf{p}_t) - (\mathbf{q}_{i,t-1} \circ \mathbf{p}_{t-1})\Big]. \tag{24}$$

Proposition 3 thus characterises a limiting case whereby portfolio adjustment costs lead to deviations from the expectations benchmark. In particular, we highlight that this can result in an under-reaction of long rates that initially offsets the over-reaction effect from risk premia from Proposition 2. For intuition, consider an unexpected rise in the short rate, which reduces the net supply of perpetuities across all markets that arbitrageurs have to bear (taking as given an upward-sloping supply curve). Due to the convexity of the adjustment cost function, the arbitrageur does not deviate too far from her holdings. As such, this exerts an equilibrating force that raises bond prices in the opposite direction, resulting in an under-reaction of long-term yields relative to the expectations channel under risk-neutrality. Our quantitative results highlight that this mechanism allows the model to produce humped-shape dynamics of bond yields, but not the overreaction observed from the data.

Outside of these nesting benchmarks, we highlight that the equilibrium characterised by Equation (18) is a combination of all three with further cross-amplification. If the arbitrageur were to be risk averse due to a reduction in portfolio wealth, the presence of portfolio adjustment costs implies that risk-bearing capacity is slow to recover, and therefore, long-term interest rates can persistently overreact even to shocks that are transitory. Further, if the arbitrageur is to be reluctant in shifting her capital across markets due to the convexity of portfolio adjustment costs, this is even more so when these shifts come with risks and she is risk averse. Thus, amplifying the initial under-reaction of long rates to give rise to humped-shape dynamics that qualitatively matches our empirical estimates.

5.1.4 Long-Term Investors

On the other side of the market are long-term investors, whose net demand for perpetuities represent the net supply that arbitrageurs have to bear in equilibrium.[41] These long-term investors are distinguished by their preferred-habitat demand (Vayanos and Vila, 2021). We adopt the convention

[41] This is by our assumption that the market trades perpetuities that are in net zero supply.

of exogenous, downward-sloping demand functions that differ by both duration ($d \in \Xi$) and bond-type (nominal v.s. inflation-linked). For a perpetuity of duration d, the demand functions are characterised respectively by:

$$Z_t(d) = \rho(d) Z_{t-1}(d) - \alpha(d) \log\left(Q_t^{(d)}\right) - \theta_0(d), \tag{25}$$

$$\widetilde{Z}_t(d) = \widetilde{\rho}(d) \widetilde{Z}_{t-1}(d) - \widetilde{\alpha}(d) \log\left(J_t^{(d)}\right) - \widetilde{\theta}_0(d). \tag{26}$$

For instance, Equation (25) states the (log) demand $Z_t(d)$ for a nominal perpetuity of duration d that is downward-sloping in its price $Q_t^{(d)}$, with $\alpha(d) > 0$ representing the elasticity of demand and $\theta_0(d)$ parameterises the steady state level of demand. We further augment this demand function with a AR(1) component with coefficient $\rho(d) \in [0, 1)$, so that the model has an additional degree of freedom to match the dynamics from the data.[42] Equation (26) states the demand for an inflation-linked perpetuity of duration d. Given our facts on market segmentation, Equation (25) describes the demand of asset managers for $d < \overline{d}$ and the demand from insurers for $d > \overline{d}$, where $\overline{d}$ corresponds to the duration of a 15-year gilt, while Equation (26) describes the demand of pension funds and liability-driven investment (LDI) funds $\forall\, d \in \Xi$. For ease of notation, we stack Equations (25) and (26) into $\mathbf{z}_t$, $\forall\, d \in \Xi$.

Lastly, we specify the intermediary that connects households to the financial markets indirectly via these habitat investors. We impose a set of three assumptions on this intermediary for exposition. First, we assume that this competitive intermediary owns the abovementioned set of investors. In each period t, the competitive intermediary obtains household savings A_{t+1} as deployable capital from households, and allocates this capital to each habitat investor according to a set of predetermined portfolio weights:[43]

$$A_{t+1}(d) = \eta(d) A_{t+1}, \quad \widetilde{A}_{t+1}(d) = \widetilde{\eta}(d) A_{t+1}, \quad \forall\, d \in \Xi. \tag{27}$$

Given the ownership structure that we have assumed, the competitive intermediary has full access to the (excess) returns generated by these investors in period $t+1$ given their positions $\{Z_t(d), \widetilde{Z}_t(d)\}$. Second, we assume a competitive intermediary sector. Hence, the competitive intermediary operates a zero-spread franchise and transfers these investment returns fully to the household at the beginning of $t + 1$. Lastly, we assume that this intermediary sector has full government backing. That is, in the instance where the amount transferred to the household leads to negative balances, this competitive intermediary can borrow from the government at the riskless real rate. There are no limits to this borrowing.

Under these assumptions, the portfolio wealth of the competitive intermediary in period t is

[42] In the absence of a persistent AR(1) long-term demand, large portfolio adjustment may be required to generate humped-shape dynamics of bond holdings and yields in response to shocks. However, if long-term demand is persistent, then only tiny adjustment costs are required.

[43] We use notation A_{t+1} to denote that optimal amount of savings chosen by the household in period t.

given by:

$$\Upsilon_t = \sum_{d \in \Xi} \Big(W_t(d) + \widetilde{W}_t(d) \Big), \tag{28}$$

where $W_t(d)$ and $\widetilde{W}_t(d)$ satisfy:

$$W_t(d) = \exp(r_{t-1}) W_{t-1}(d) + Z_{t-1}(d) Q_{t-1}^{(d)} \mu_t^{(d)} - A_t(d) \mu_t^{(d)}, \tag{29}$$

$$\widetilde{W}_t(d) = \exp(r_{t-1}) \widetilde{W}_{t-1}(d) + \widetilde{Z}_{t-1}(d) J_{t-1}^{(d)} \widetilde{\mu}_t^{(d)} - \widetilde{A}_t(d) \widetilde{\mu}_t^{(d)}. \tag{30}$$

for a given $\{W_0(d), \widetilde{W}_0(d)\}$.

5.2 The New Keynesian Block

The New Keynesian block consists of households and firms, whose optimising behaviour leads to the standard IS equation and the Phillips curve. We specify each in turn.

5.2.1 Households and Aggregate Demand

We work with a representative household block. Following the footsteps of Gourinchas, Ray, and Vayanos (2022); Ray, Droste, and Gorodnichenko (2024), the only nonstandard feature of the household's problem is that they do not directly optimise over the set of risky assets in this economy. Rather, they indirectly access asset returns—including a riskless deposit franchise—through a competitive intermediary, that serves only to allocate the savings of households to long-term investment funds (to be specified subsequently) according to a set of fixed portfolio shares, and thereafter return the gross returns earned by these investment funds to the household. As such, the household only optimises over the external margin of how much to save. We assume that this intermediation process is frictionless, and does not tax away the household's asset returns in any way.

These households maximise over their discounted lifetime expected utility by choosing a consumption bundle C_t, labour supply L_t and how much to save in the abovementioned intermediary A_{t+1}. The Bellman equation of the household is stated as:

$$V(A_t, C_{t-1}) = \max_{C_t, L_t, A_{t+1}} U(C_t, C_{t-1}, L_t) + \beta \, \mathbb{E}_t \Big[V(A_{t+1}, C_t) \Big], \tag{31}$$

where $U(\cdot, \cdot, \cdot)$ is the period utility function that is assumed to satisfy the Inada condition over each of its arguments, and β is the household time discount factor. We assume that this function is represented by the CRRA utility class:

$$U(C_t, C_{t-1}, L_t) = \frac{1}{1-\sigma} \Big(C_t - h C_{t-1} \Big)^{1-\sigma} - \chi \frac{1}{1+\psi} L_t^{1+\psi}, \tag{32}$$

where the parameters σ is the relative risk aversion, ψ^{-1} is the Frisch elasticity of labour supply, h represents the degree of habit persistence in aggregate consumption and χ is a scaling parameter.

The household maximises Equation (31) subject to a flow budget constraint:

$$P_t C_t + P_t A_{t+1} \leq (1 + r_t^A) P_t A_t + W_t L_t + \Pi_t + P_t T_t, \;\; A_0 \;\; \textit{given}, \tag{33}$$

and a no-Ponzi-game condition $\lim_{T\to\infty} \beta^T A_{t+T} = 0$, taking the set of aggregate prices as given. P_t is the aggregate price index, W_t is the aggregate wage, Π_t represents a dividend transfer of profits from firms owned by the households, and T_t represents the real value of taxes paid to the government in each period.

The rebate from the competitive intermediary occurs at the beginning of each period t, whereby the household receives a gross real return $1+r_t^A$ on her total savings A_t invested in the intermediary in the preceding period. Since the intermediation between households and the financial market is competitive and frictionless, the condition that r_t^A satisfies in equilibrium is:

$$r_t^A = \sum_{d\in\Xi} \Big[\eta(d) r_t^{(d)} + \widetilde{\eta}(d) \widetilde{r}_t^{(d)}\Big], \tag{34}$$

where d represents a specific duration of the asset. In particular, the financial market in this model trades a set of assets that is distinguished both by its duration $d \in \Xi$, and by its type, defined as either a nominal or a real asset (whose returns are indicated by a tilde). For instance, $r_t^{(d)}$ refers to the net one-period return from a nominal asset that has a duration d. Importantly, we highlight that the returns from a riskless deposit franchise are also represented in this notation: they correspond to assets where its duration $d \to 1$, and therefore, the household accesses both the nominal and real short-term interest rate via the intermediary above. We revisit this subsequently when defining the assets and their payoffs in the financial market. Finally, $\eta(\cdot)$ is a portfolio weight that is assumed to be fixed by the competitive intermediary, satisfying:

$$1 = \sum_{d\in\Xi} \Big[\eta(d) + \widetilde{\eta}(d)\Big]. \tag{35}$$

In equilibrium, the set of optimality conditions are:

$$\Big(C_t - hC_{t-1}\Big)^{-\sigma} = h\beta\, \mathbb{E}_t \left[\Big(C_{t+1} - hC_t\Big)^{-\sigma}\right] + P_t \lambda_t, \tag{36}$$

$$\chi L_t^{\psi} = \lambda_t W_t, \tag{37}$$

$$P_t \lambda_t = \beta\, \mathbb{E}_t \left[P_{t+1} \lambda_{t+1} \Big(1 + r_{t+1}^A\Big)\right], \tag{38}$$

where $\lambda_t > 0$ is the Lagrange multiplier on household's flow budget constraint. Equations (37) describes the household's labour supply decision on the margin as a function of both substitution (through W_t) and income effects (through λ_t). Combining Equations (36) and (38) yields the consumption Euler equation that ties the real economy with our segmented financial markets that we will shortly describe. In short, demand and supply imbalances from investors that trade in these markets lead to excess volatility in these returns through risk and term premia, resulting in a wedge

that can strengthen the household's savings motives. At the same time, however, these demand and supply imbalances must also adjust in response to the equilibrium set of asset prices that satisfy the household's stochastic discount factor. Hence, there is a general equilibrium feedback from the real economy to the financial market that we later explore in greater detail.

5.2.2 Firms and Aggregate Supply

The production block in the economy is standard with two hierarchies of production, consisting of an intermediate-good sector and a final-good sector. Intermediate-good retailers hire labour inputs from households competitively and produce differentiated varieties indexed by $\iota \in [0, 1]$. These retailers thus compete monopolistically, and they face nominal frictions to price adjustment. The intermediate goods are in turn demanded across the sector by a competitive final-good producer in the production of a final good. We discuss each in turn.

The final consumption good Y_t is produced by the final-good producer through combining a unit mass of intermediate inputs $\{y_{\iota,t}\}_{\iota=0}^{1}$ using a Dixit-Stiglitz CES technology:

$$Y_t = \left(\int_0^1 y_{\iota,t}^{\frac{1}{1+\lambda_p}} d\iota \right)^{1+\lambda_p}, \tag{39}$$

where $y_{\iota,t}$ is the final-good producer's demand for intermediate good ι. The elasticity of substitution across these intermediate goods is given by $\epsilon = \lambda_p^{-1}(1 + \lambda_p)$, satisfying $1 < \epsilon < \infty$, where λ_p represents the steady state markup charged by the intermediate-good producer that is common across all these intermediate varieties. As the final-good producer is competitive and she takes both her input and output prices as given, profit maximisation and zero profits in equilibrium yield the standard demand schedule for the intermediate goods and the aggregate price index:

$$y_{\iota,t} = \left(\frac{p_{\iota,t}}{P_t} \right)^{-\frac{1+\lambda_p}{\lambda_p}} Y_t, \tag{40}$$

$$P_t = \left(\int_0^1 p_{\iota,t}^{-\frac{1}{\lambda_p}} d\iota \right)^{-\lambda_p}. \tag{41}$$

Each intermediate-good retailer ι demands a measure of labour input in production, denoted by $\eta_{\iota,t}$, hiring them competitively in the labour market taking the real wage $w_t = W_t/P_t$ as given. They employ a linear production technology with constant returns to scale that is identical across the sector:[44]

$$y_{\iota,t} = n_{\iota,t}. \tag{42}$$

It immediately follows from the cost-minimisation of intermediate-good retailers that the marginal cost of production is also common across the sector, and is given by $\gamma_t = w_t$. Since the production technology employed by the intermediate retailers is linear, the scale of production is determined

[44] The specificity of the constant returns to scale production function and a single production input is to simplify exposition.

by the labour market clearing condition in equilibrium, where the aggregate demand for labour across all intermediate-good retailers equals the total labour supplied by the household:

$$\int_0^1 n_{\iota,t} d\iota = L_t. \tag{43}$$

Each intermediate-good retailer subsequently chooses the profit-maximising price, given monopolistic competition in the output market. We assume that each retailer sets prices according to Calvo contracts that expire with probability ξ each period. When the Calvo contract expires in period t, the firm chooses a reset price $p^*_{\iota,t}$ flexibly in a forward-looking manner such that it is a weighted-average markup over all current and future marginal costs. Since marginal costs are identical across retailers, we can restrict attention to a symmetric equilibrium $p^*_{\iota,t} = p^*_t$ as is the convention in the literature. We further introduce indexation to lagged inflation whenever the retailer is unable to reoptimise. That is, an unoptimised price in period $t+j$ will be equal to $p^*_t(\frac{P_{t+j-1}}{P_{t-1}})^{\vartheta_p}$, where ϑ_p is a parameter that governs the degree of price indexation. Therefore, for each of these periods when the retailer is unable to reoptimise its price, it supplies the required quantities to meet demand $y_{\iota,t+j}$ from final goods producers at the price $p^*_t(\frac{P_{t+j-1}}{P_{t-1}})^{\vartheta_p}$ by hiring $n_{\iota,t+j}$ units of labour competitively at the prevailing real wage w_{t+j}. The economy thus becomes partially demand-driven. The retailer's problem is thus characterised by:

$$\max_{p^*_t} \mathbb{E}_t \sum_{\tau=t}^{\infty} \Lambda_{t,\tau}(1-\xi)^{\tau-t} \left(\frac{p^*_t(\frac{P_{\tau-1}}{P_{t-1}})^{\vartheta_p}}{P_\tau} - \gamma_\tau \right) y_{\iota,\tau}, \tag{44}$$

subject to the constraints from her own production frontier in Equation (42) and the downward-sloping demand curve from the final-good producer:

$$y_{\iota,\tau} = \left(\frac{p^*_t(\frac{P_{\tau-1}}{P_{t-1}})^{\vartheta_p}}{P_\tau} \right)^{-\frac{1+\lambda_p}{\lambda_p}} Y_\tau. \tag{45}$$

Under the assumption that these retailers are owned by households, all future profits between periods t and τ are discounted by the household's stochastic discount factor $\Lambda_{t,\tau}$.

In equilibrium, the first-order condition characterising the retailer's optimal reset price satisfies:

$$\frac{p^*_t}{P_t} = (1+\lambda_p) \frac{\mathbb{E}_t[\sum_{\tau=t}^{\infty} \Lambda_{t,\tau}(1-\xi)^{\tau-t} y_{\iota,\tau}\gamma_\tau]}{[\mathbb{E}_t \sum_{\tau=t}^{\infty} \Lambda_{t,\tau}(1-\xi)^{\tau-t}(\frac{P_{\tau-1}}{P_{t-1}})^{\vartheta_p} y_{\iota,\tau} \frac{P_t}{P_\tau}]}. \tag{46}$$

Given this reset price equation, and applying the properties of Calvo pricing to the price index equation in (41), it is straightforward to show that inflation $\Pi_t = P_t/P_{t-1}$ evolves according to:

$$1 = (1-\xi)\left(\frac{P_{t-1}}{P_{t-2}}\right)^{-\frac{\vartheta_p}{\lambda_p}} \left(\frac{P_t}{P_{t-1}}\right)^{\frac{1}{\lambda_p}} + \xi \left(\frac{p^*_t}{P_t}\right)^{-\frac{1}{\lambda_p}}. \tag{47}$$

5.3 Government Policy

We now close the model by specifying the set of government policies, which consist of fiscal and monetary policy.

Fiscal policy is Ricardian. It consists of taxes T_t and the real issuance of short-term debt B_t, taking monetary policy as given. In particular, the issuance of short-term debt is endogenous to total private sector demand:

$$B_t + W_t^a + \Upsilon_t + A_{t+1} = 0. \tag{48}$$

The fiscal authority subsequently adjust taxes endogenously to maintain a balanced-budget every period. Under our assumption that the fiscal authority does not issue long-term perpetuities, the flow budget constraint is given by:[45]

$$T_t + B_t = B_{t-1} \exp(r_{t-1}) + C(\mathbf{q}_{i,t-1}, \mathbf{p}_{t-1}, \mathbf{q}_{i,t-2}, \mathbf{p}_{t-2}; \mathbf{\Gamma}). \tag{49}$$

Monetary policy sets the nominal rate i_t via a Taylor rule:

$$i_t = \bar{i} + \phi_i i_{t-1} + (1-\phi_i)\phi_\pi \, \mathbb{E}_{t-1} \log\left(\frac{P_t}{P_{t-1}}\right) + \sigma_i \varepsilon_{i,t}, \;\; \varepsilon_{i,t} \sim N(0,1), \tag{50}$$

where $\bar{i} = (1-\phi_i)\log(\beta^{-1})$ and $\varepsilon_{i,t}$ is a monetary shock.

5.4 Equilibrium

Finally, we conclude the model by providing the definition of equilibrium:

Definition 1. *(General equilibrium). The general equilibrium of this model is a sequence of allocations consisting of prices $\mathcal{P}$ and quantities $\mathcal{Q}$ given by:*

$$\mathcal{P} = \left\{ r_t^a, W_t, \lambda_t, P_t, p_t^*, \{Q_t^{(d)}, J_t^{(d)}\}_{d \in \Xi} \right\}_{t=0}^{\infty}, \tag{51}$$

$$\mathcal{Q} = \left\{ C_t, L_t, Y_t, \mathbf{q}_t, \mathbf{z}_t \right\}_{t=0}^{\infty}. \tag{52}$$

such that the prices satisfy the household's optimality conditions (36), (38) and (37), the firm's profit-maximisation condition (46), and the stock of bond holdings held by the arbitrageur satisfies (18), (19), (20) and (21). Given these quantities, the prices are such that all markets clear:[46]

$$C_t = Y_t \tag{53}$$

$$\mathbf{q}_t + \mathbf{z}_t = \mathbf{0}. \tag{54}$$

[45] $C(\cdot, \cdot; \mathbf{\Gamma})$ acts as friction to govt. borrowing.

[46] A complete description also includes the labour market in Equation (43), the market for intermediate-goods from Equation (42) and (40), and the market for riskless short-term debt from Equation (48).

6 Quantitative Results

This section demonstrates the quantitative properties of our model. We present our calibration strategy, before turning to the model's predicted responses for both long-term interest rates and intermediary bond holdings in response to changes in monetary policy. Finally, we use the calibrated model to present counterfactual effects of market segmentation on the aggregate economy.

6.1 Model Calibration

Since our main empirical results are the impulse response functions of long-term interest rates and intermediary bond holdings to a monetary policy shock, we use these estimated impulse responses as natural targets for the model to match, focussing on the estimates from the longer-term segment of the bond market given that these are the sharpest.

However, as these impulse response functions are simultaneously parameterised by many deep structural parameters in the model, there is not a single parameter that can discipline each of them alone. As such, we proceed with a two-pronged calibration approach. We first fix the parameters within the New Keynesian block using standard values from the literature as much as possible. Having fixed those parameters, we estimate a subset of the model's remaining parameters using a minimum-distance approach to match the model's impulse response functions against those estimated from the data. Specifically, we implement a root-finding algorithm to minimise the sum of squared deviations between the model and estimated impulse response functions from the longer-term gilt market. This procedure disciplines the key financial market parameters while leaving the remaining New Keynesian parameters mostly externally calibrated.

Table 1 provides the full overview of the model's calibrated parameters. The time unit in our model is defined as monthly. We calibrate the household's subjective discount factor β to 0.9984, which corresponds to a 2% annualised real rate. Arbitrageurs are assumed to discount at the same rate as households. We set both the elasticity of intertemporal substitution σ^{-1} and Frisch elasticity of labour supply ψ^{-1} equal to 1, consistent with standard calibration choices within the business cycle literature. The scale parameter χ governing the disutility of labour is internally calibrated to 9.5602, consistent with a steady state fraction of time dedicated to remunerated labour equal to 1/3. The steady state markup is 0.2, corresponding to a constant elasticity of substitution equal to 6. We set the habit persistence parameter h to 0.9283, in line with Smets and Wouters (2007) and Schmitt-Grohé and Uribe (2012), and consistent with DSGE models that do not feature hand-to-mouth households. The price indexation parameter ϑ_p is set equal to 0.5848, in line with estimated values for the UK economy (Burgess et al., 2013). For the remaining parameters within the New Keynesian block, we selectively adopt estimates from Albuquerque et al. (2025) where appropriate.[47]

[47]The reset probability of prices, ξ, is set to be 0.052, following their estimates of the final output price Calvo parameter. This corresponds to an average duration between price changes of roughly seven quarters, and the model thus exhibits a considerable amount of price stickiness. Similarly, the Taylor coefficients for inflation and interest rate smoothing in the monetary policy rule are set to equal 1.49 and 0.91 respectively.

Table 1 CALIBRATED PARAMETERS

	Description	Value	Source/Target
	Externally Calibrated		
β	Household discount factor	0.9984	Target 2% steady state real rate
$\overline{\beta}$	Arbitrageur discount factor	0.9984	Identical to households
σ^{-1}	Intertemporal elasticity of substitution	1	Standard
ψ^{-1}	Frisch elasticity of labor supply	1	Standard
λ_p	Steady state markup	0.2	Standard
ξ	Calvo price adjustment probability	0.054	Albuquerque et al. (2025)
ϕ_i	Taylor coefficient: smoothing	0.91	Albuquerque et al. (2025)
ϕ_π	Taylor coefficient: inflation	1.49	Albuquerque et al. (2025)
h	Consumption habit persistence	0.8	Standard
ϑ_p	Price indexation	0.2	Burgess et al. (2013)
σ_i	Standard deviation of monetary shock	0.0276	Target estimated shock from VAR
	Internally Calibrated		
χ	Labour disutility	9.5752	Target hours worked 0.333
κ_s	Rate of perpetuity coupon decay (short)	0.9738	Target duration 3-years
κ_l	Rate of perpetuity coupon decay (long)	0.9961	Target duration 15-years
	Estimated		
ζ	Probability of death	0.0107	IRF matching
$\overline{W}$	Wealth endowment of new born	5.3542	IRF matching
φ	Arbitrageur risk aversion	3.9973	IRF matching
	Longer-Term Market (corresponding to $\kappa_l = 0.9961$): Nominal Perpetuity		
α_l	Elasticity of habitat demand (nominal)	0.0169	IRF matching
$\theta_{0,l}$	Demand intercept (nominal)	0.4150	IRF matching
ς_l	Portfolio adjustment cost (nominal)	0.0214	IRF matching
ρ_l	AR(1) of habitat demand (nominal)	0.3462	IRF matching
	Longer-Term Market (corresponding to $\kappa_l = 0.9961$): Inflation-Linked Perpetuity		
$\widetilde{\alpha}_l$	Elasticity of habitat demand (real)	0.0008	IRF matching
$\widetilde{\theta}_{0,l}$	Demand intercept (real)	0.0658	IRF matching
$\widetilde{\varsigma}_l$	Portfolio adjustment cost (real)	1.4221	IRF matching
$\widetilde{\rho}_l$	AR(1) of habitat demand (real)	0.0883	IRF matching
	Shorter-Term Market (corresponding to $\kappa_s = 0.9738$): Nominal Perpetuity		
α_s	Elasticity of habitat demand	0.0034	IRF matching
$\theta_{0,s}$	Demand intercept	0.0277	IRF matching
ς_s	Portfolio adjustment cost	0.2500	IRF matching
ρ_s	AR(1) of habitat demand	0.0907	IRF matching
	Shorter-Term Market (corresponding to $\kappa_s = 0.9738$): Inflation-Linked Perpetuity		
$\widetilde{\alpha}_s$	Elasticity of habitat demand	0.0004	IRF matching
$\widetilde{\theta}_{0,s}$	Demand intercept	0.093	IRF matching
$\widetilde{\varsigma}_s$	Portfolio adjustment cost	0.16	IRF matching
$\widetilde{\rho}_s$	AR(1) of habitat demand	0.2663	IRF matching

NOTE: Calibration is done at the monthly frequency.

Turning to the financial market block, we assume that each of the nominal and index-linked bond markets are further segmented across two duration slices $\{s, l\}$, representing a shorter-term and a longer-term duration bond respectively.[48] We internally calibrate κ_s to 0.9738 and κ_l to 0.9961, such that the effective duration of the shorter and longer-term bonds in our model are 3 and 15 years respectively.[49] This matches the average duration of insurers in the longer-term segment of the UK bond market that is approximately 18. We estimate the remaining subset of parameters by minimising the distance between the model's impulse response functions to a monetary policy shock and those we estimate from the data, focussing only on the longer-term segment of the nominal and index-linked market.

We first highlight the parameters governing risk aversion. The arbitrageur's wealth endowment at birth $\overline{W}$ is 5.354. This yields a portfolio duration of 14.5 that is approximately the upper range of the weighted-average time-to-maturity of gilts held by banks prior to the Bank of England's rate hiking cycle in 2021.[50] We recover a coefficient of relative risk aversion φ that is 3.997. In the model, this generates an average index-linked spread of 47 basis points between the shorter-term and longer-term bond, compared to the average 3y/15y index-linked spread of 65 basis points in the data.[51] The probability of death ζ is estimated to be 0.0107, implying that arbitrageur capital dissipates with a half-life of approximately 63 months.

We assume that the adjustment cost matrix $\mathbf{\Gamma} = diag\{\varsigma_s, \widetilde{\varsigma}_s, \varsigma_l, \widetilde{\varsigma}_l\}$, now 4×4 in dimension, is diagonal. Under this assumption, only the asset's own transaction cost is of first-order. The adjustment cost parameters ς_l and $\widetilde{\varsigma}_l$ corresponding to the longer-term segment are estimated to be 0.021 and 1.422 respectively, implying a much larger cost associated with the arbitrageur adjusting her holdings of index-linked bonds.

For the demand functions, we recover elasticities of demand α_l and $\widetilde{\alpha}_l$ by habitat investors equal to 0.0169 and 0.001 respectively in the longer-term market. The model thus requires a relatively inelastic demand for index-linked bonds in order to match their estimated responses in the data, as compared to demand for nominal bonds that are more elastic. The demand intercepts $\theta_{0,l}$ and $\widetilde{\theta}_{0,l}$ are 0.415 and 0.066 respectively. All else equal, a larger intercept amplifies the total adjustment cost associated with the asset on the margin, leading to a slower recovery of bond holdings and prices following a monetary policy shock. The AR(1) coefficients ρ_l and $\widetilde{\rho}_l$ are estimated to be 0.346 and 0.088 respectively. This suggests that much of the persistence in the quantity and price responses arise endogenously from wealth accumulation (i.e., governed by the death rate parameter ζ). The remaining parameters are reported in the bottom panel of Table 1.

Finally, we solve the model by perturbation to the third-order, obtaining impulse response functions via stochastic simulation with 10,000 draws.

[48] For clarity of notation, we further use s and l as subscripts to index parameters that belong to these assets.

[49] The market-wide duration estimates for gilts lie around 8-15 years, with an average maturity around 14.4 years. See Joyce and Tong (2012) and Czech, Huang, Lou, and Wang (2021a).

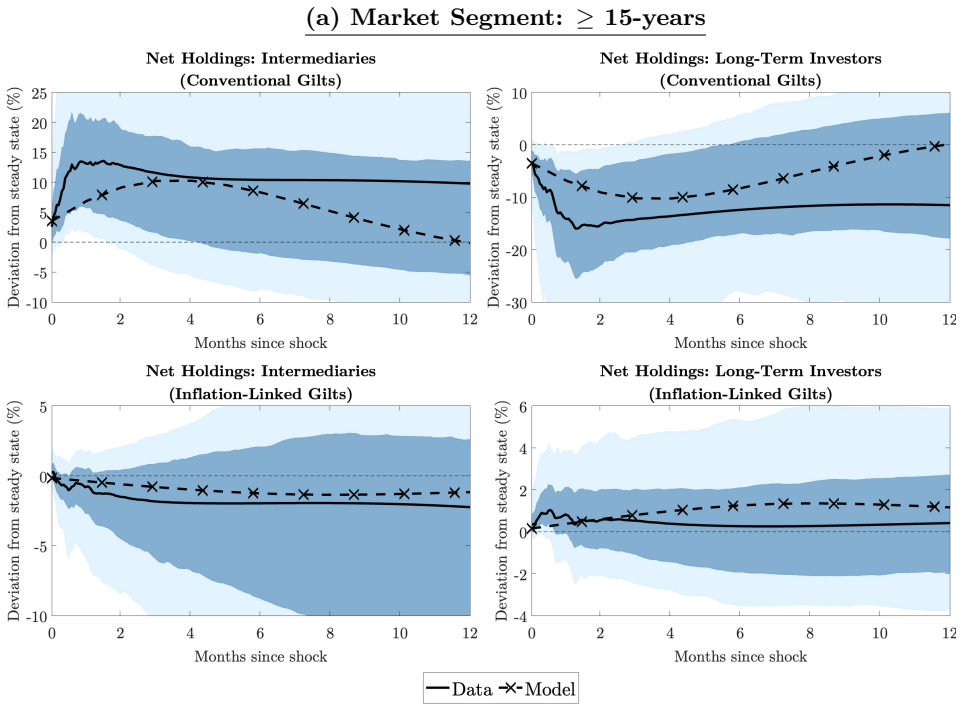
[50] See Figure 3. This portfolio duration is also close to the average time-to-maturity of gilts traded by hedge funds, GEMMs and inter-dealer brokers in the data.

[51] This may be due to the absence of non-monetary shocks in our model, which can lead to larger term premia in the data due to their perceived effects on inflation and thus the volatility of the short rate.

6.2 Model Predictions in Response to Monetary Policy

Figure 11 presents the model's predictions for intermediary holdings in the longer-term bond market in response to an unexpected policy easing. We scale these responses such that the monetary shock *raises* breakeven inflation by 1pp on impact, consistent with same scaling of our empirical estimates. In the model, long-term investors respond to a downward shift of the yield curve by reducing their bond holdings because they have downward-sloping demand curves. This is sharpest in the conventional gilt market, where the elasticity is estimated to be $\alpha_l = 0.0169$ (as opposed to $\widetilde{\alpha}_l = 0.0008$). The model thus attributes the wide confidence bands estimated from the index-linked gilt market to a relatively price inelastic demand that responds to unexpected changes in monetary policy only negligibly. As bonds clear in net zero supply, the equilibrium condition is that arbitrageurs must absorb the additional supply of long-term bonds now sold into the market by these long-term investors, thus bearing duration as yields fall. With portfolio adjustment costs, however, the peak response of holdings does not occur instantaneously on impact of the shock. Rather, the arbitrageurs only partially absorb the additional stock of bonds initially and unwind these positions at a slower pace, replicating the humped-shapes from our empirical impulse responses as we observe in the data.

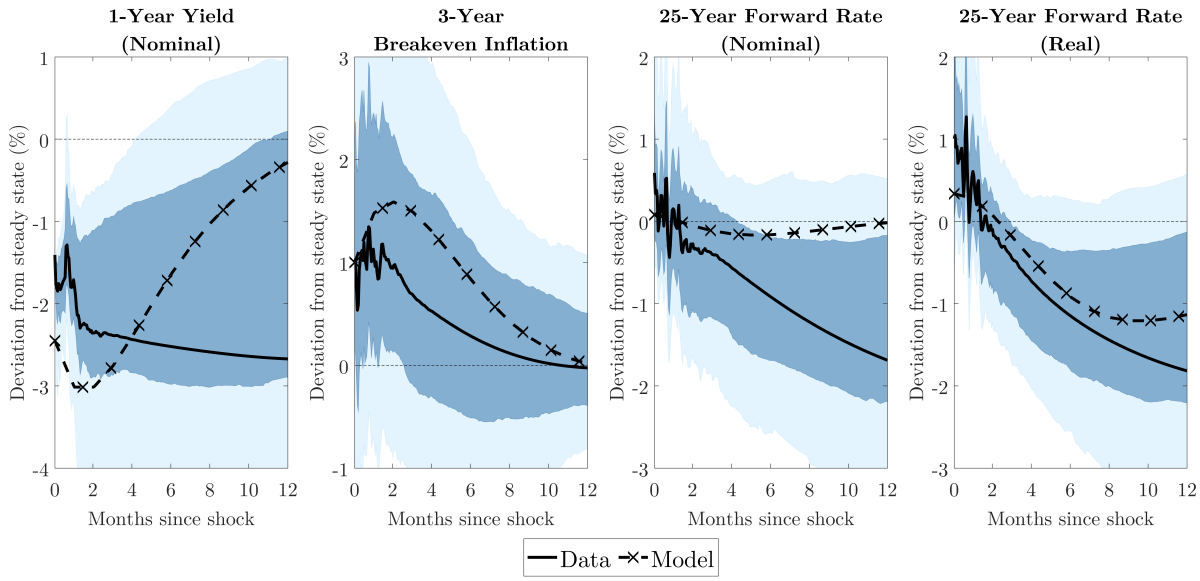
Figure 11 Bond holdings in response to monetary News (↓ in 1-year yield)



Note: Bold lines are empirical IRFs that indicate the responses of bond holdings from intermediaries and habitat investors to a monetary policy easing shock. Intermediaries in each individual segments are GEMMs and inter-dealer brokers. Long-term investors represent pension funds and insurers. Scaling of IRFs is such that 3-year expected inflation raises by 1 percentage point on impact. Shaded intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

The price responses are illustrated in Figure 12. As monetary policy eases, long-term forward rates initially rise on impact and underreact to the unexpected easing of monetary policy: this is due to the interaction between the shifts in long-term demand and the arbitrageur's portfolio adjustment costs. Since the arbitrageur can only partially absorb the increase in supply of long-term bonds due to these adjustment costs, the equilibrium exerts an additional downward force on bond prices to dampen the increase in supply. As such, these forward rates initially rise in response to monetary policy easing. However, this channel is quickly dominated by the subsequent increase in arbitrageur wealth as monetary policy shifts down the yield curve, thus reducing risk aversion substantially. In this instance, equilibrium term premia turn negative, amplifying the response of long-term forward rates and resulting in a persistent overreaction to monetary policy culminating in a hump-shaped response.[52] These responses are persistent precisely because risk-bearing capacity increases for a substantial period of time and intermediaries only unwind their positions gradually. Our baseline calibration is less capable, however, of matching both nominal and real forward rate responses simultaneously due to the existence of inflation risk premia.

Figure 12 Asset response to a monetary shock (↓ in short rate)



Note: Impulse response functions correspond to a monetary policy easing shock that raises 3-year breakeven inflation by 1 percentage point on impact.

6.3 The Aggregate Effects of Market Segmentation

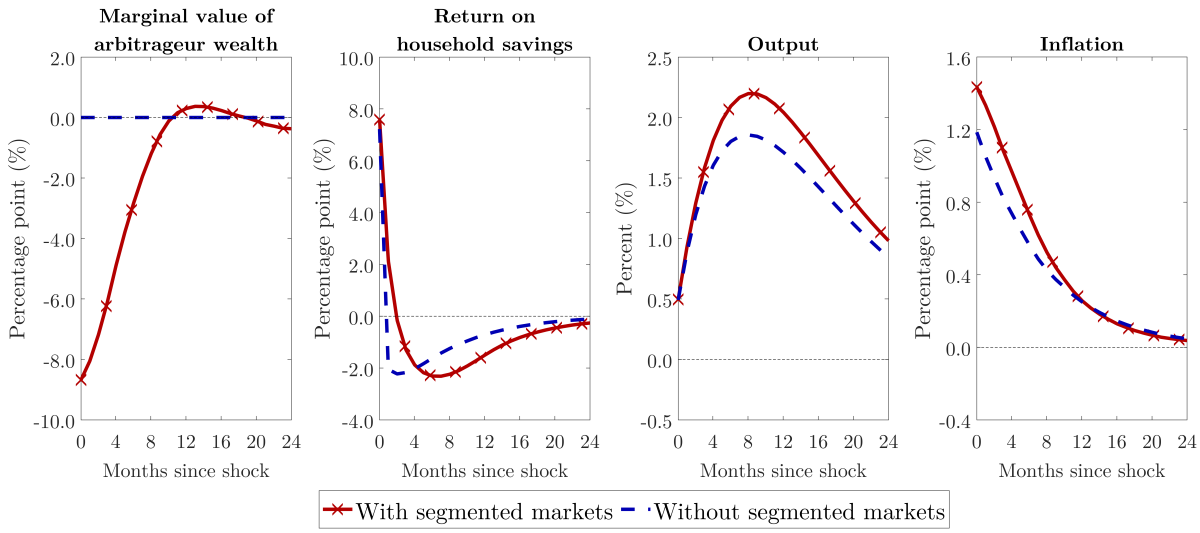
Up to this point, we have focussed on a segmented bond market in which monetary policy generates joint comovements in bond prices and holdings, and we further calibrated the model using both bond transaction and yield curve data from the UK to match these observed empirical patterns. Given the quantitative magnitudes of the model's mechanisms, we now turn to their aggregate implications: what do the shifts in both bond prices and holdings across intermediaries imply for

[52] This is relative to a zero response according to the rate path, see Figure C.1

the broader economy?

In particular, we use the calibrated model to compare the transmission of monetary policy when markets are segmented with the case when markets are fully integrated. Within our framework, market segmentation reflects limits to arbitrage. The arbitrageur is unable to reallocate wealth across different segments of the bond market elastically to arbitrage away excess returns due to a combination of wealth constraints and portfolio adjustment costs. In our counterfactual, we shut down these channels and study the case whereby the bond market is fully integrated in the absence of these intermediary frictions.

Figure 13 Aggregate responses to a monetary shock (↓ in short rate)



Note: Impulse response functions correspond to a monetary policy easing shock that raises 3-year breakeven inflation by 1 percentage point on impact. Bold line indicates the case whereby the arbitrageur faces limits to arbitrage arising from both wealth constraints and portfolio adjustment costs. Dashed line indicates the case when these channels are down shut i.e., arbitrageur demand becomes fully elastic.

Figure 13 presents the counterfactual results. The dashed lines correspond to the case of fully integrated markets—the frictionless benchmark nested within our model that arises under a large steady state wealth injection ($\overline{W} \to \infty$) and zero adjustment costs. While monetary policy raises portfolio wealth, its marginal value to the arbitrageur remains at zero. Consequently, monetary policy does not affect net worth or risk-bearing capacity. Since there are no costs to portfolio adjustment, the arbitrageur has elastic demand and enter markets freely to arbitrage away any excess returns. Monetary policy transmits into the long-end of the yield curve solely via the term structure hypothesis.

Moving away from this frictionless benchmark to the empirically relevant case—consistent with the shifts in both quantities and prices observed in the data—monetary policy has a very different response. Following a monetary easing, the arbitrageur's marginal value of wealth declines, increasing risk-bearing capacity and compressing term premia. This shows up as an additional wedge that reduces the returns households receive from the competitive intermediary on the margin, thus strengthening the inter-temporal substitution channel and amplifying the demand response. In

turn, this leads to a larger increase in labour demand and the firms' wage bill, thus amplifying the response of inflation to monetary policy. In other words, the presence of market segmentation observed in both our model and data suggests that monetary policy has larger real effects.

7 Conclusion: Summary of Findings

Long-term forward rates respond strongly to shocks and particularly to monetary policy. Using regulatory bond transaction data that covers the universe of trading activities in the UK market, this paper sheds new light on the drivers of these observed empirical patterns by identifying the response of intermediaries to monetary policy at the high-frequency. This has led to the following lessons.

First, a monetary shock identified from a proxy-SVAR results in a persistent overreaction of long-term interest rates following an initial underreaction. When monetary policy eases, both nominal and real long-term forward rates initially rise before they persistently fall below zero. We show that this hump-shaped pattern is also robust to identification using inflation swap surprises to monetary policy announcements as an external instrument for breakeven inflation in the VAR.

Second, we find that monetary policy moves the bond market significantly. Behind these price responses are substantial shifts in bond holdings across intermediaries in the longer-term segment of the market that satisfy a set of sign restrictions. In response to a reduction of the 1-year yield by 100 basis points, financial intermediaries—consisting of GEMMs and inter-dealer brokers—increase their bond holdings by approximately £20 billion, thus taking on more duration risk onto their balance sheets, while the long-term investors that trade this segment—such as pension funds and insurers—reduce their bond holdings by selling them in the market. In other words, we recover a downward-sloping demand curve for these institutional investors who tilt their portfolio allocation towards long-term gilts when monetary policy tightens. Moreover, these results are driven nearly entirely by trading in the long-end of the conventional gilt market.

Third, our estimates additionally suggest that these intermediaries do not fully respond to unexpected changes in monetary policy immediately. On the contrary, the full adjustment of their bond holdings takes about a full month after the change in monetary policy, and the unwinding of these positions also occurs at a much slower rate. An unexpected change in monetary policy therefore leads to persistent shifts in intermediary demand.

We rationalise our empirical findings in a New Keynesian model featuring segmented bond markets traded by habitat investors with downward-sloping demand for perpetuities, both nominal and real. Long-term interest rates can deviate from the rate path because these markets are only partially integrated by an arbitrageur whose demand is constrained by both wealth and portfolio adjustment costs. Due to the presence of these costs, the arbitrageur is unable to immediately absorb the additional supply of perpetuities when bond prices rise, and thus the equilibrium exerts a downward force on prices. This generates a hump-shaped response of intermediary holdings and an initial underreaction of forward rates to monetary policy. The countercyclical nature of

risk aversion eventually generates a term premia response, resulting in a persistent overreaction of these forward rates since risk-bearing capacity is slow to recover in the presence of these portfolio adjustment costs. In our baseline calibration, we find that monetary policy is more powerful in a counterfactual economy where a segmented bond market replicates the empirical patterns of both prices and quantities observed in the data.

Overall, this paper has contributed new inferences and techniques through the combination of aggregate prices and micro quantities that, in our application, have proven useful to distinguish apart different accounts of how monetary policy generates term premia and transmits through the bond market into the rest of the wider economy. We hope that this opens up exciting new avenues and applications in other settings and markets where prevailing economic mechanisms cast joint restrictions on prices and quantities.

References

Abrahams, M., T. Adrian, R. K. Crump, E. Moench, and R. Yu (2016): "Decomposing real and nominal yield curves," *Journal of Monetary Economics*, 84, 182–200.

Albuquerque, D., J. Chan, D. Kanngiesser, D. Latto, S. Lloyd, S. Singh, and J. Žáček (2025): "Decompositions, forecasts and scenarios from an estimated DSGE model for the UK economy," .

Altavilla, C., L. Brugnolini, R. S. Gürkaynak, R. Motto, and G. Ragusa (2019): "Measuring euro area monetary policy," *Journal of Monetary Economics*, 108, 162–179.

Anderson, N. and J. Sleath (2001): *New estimates of the UK real and nominal yield curves*, Bank of England.

Bahaj, S., R. Czech, S. Ding, and R. Reis (2023): "The Market for Inflation Risk," *Bank of England Staff Working Paper 1,028*.

Barria, R. and G. Pinter (2023): "Mispricing in inflation markets," .

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

Braun, R., S. Miranda-Agrippino, and T. Saha (2025): "Measuring monetary policy in the UK: The UK monetary policy event-study database," *Journal of Monetary Economics*, 149, 103645.

Brunnermeier, M. K. and Y. Sannikov (2014): "A macroeconomic model with a financial sector," *American Economic Review*, 104, 379–421.

Burgess, S., E. Fernandez-Corugedo, C. Groth, R. Harrison, F. Monti, K. Theodoridis, M. Waldron, et al. (2013): *The Bank of England's forecasting platform: COMPASS, MAPS, EASE and the suite of models*, SSRN.

Cieslak, A. (2018): "Short-rate expectations and unexpected returns in treasury bonds," *The Review of Financial Studies*, 31, 3265–3306.

Cochrane, J. H. and M. Piazzesi (2002): "The fed and interest rates—a high-frequency identification," *American Economic Review*, 92, 90–95.

Cook, T. and T. Hahn (1989): "The effect of changes in the federal funds rate target on market interest rates in the 1970s," *Journal of monetary economics*, 24, 331–351.

Covi, G., J. Brookes, and C. Raja (2022): "Measuring Capital at Risk in the UK banking sector: a microstructural network approach," *Bank of England Staff Working Paper No. 983*.

Czech, R., S. Huang, D. Lou, and T. Wang (2021a): "Informed trading in government bond markets," *Journal of Financial Economics*, 142, 1253–1274.

Czech, R., S. Huang, D. Lou, T. Wang, et al. (2021b): "An unintended consequence of holding dollar assets," Tech. rep., Bank of England.

d'Arienzo, D. (2020): "Maturity increasing overreaction and bond market puzzles," *Available at SSRN 3733056*.

Fleckenstein, M., F. A. Longstaff, and H. Lustig (2014): "The TIPS-treasury bond puzzle," *the Journal of Finance*, 69, 2151–2197.

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

Giese, J., M. Joyce, J. Meaning, and J. Worlidge (2024): "Do preferred habitat investors exist? Evidence from the UK government bond market," *Economics Letters*, 234, 111462.

Giglio, S. and B. Kelly (2018): "Excess volatility: Beyond discount rates," *The Quarterly Journal of Economics*, 133, 71–127.

Gilchrist, S., D. López-Salido, and E. Zakrajšek (2015): "Monetary policy and real borrowing costs at the zero lower bound," *American Economic Journal: Macroeconomics*, 7, 77–109.

Gourinchas, P.-O., W. Ray, and D. Vayanos (2022): "A preferred-habitat model of term premia, exchange rates, and monetary policy spillovers," Tech. rep., National Bureau of Economic Research.

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

Gürkaynak, R. S., B. Kisacikoğlu, and J. H. Wright (2020): "Missing events in event studies: Identifying the effects of partially measured news surprises," *American Economic Review*, 110, 3871–3912.

Gürkaynak, R. S., B. Sack, and E. Swanson (2005): "The sensitivity of long-term interest rates to economic news: Evidence and implications for macroeconomic models," *American Economic Review*, 95, 425–436.

Gürkaynak, R. S., B. P. Sack, and E. T. Swanson (2005a): "Do actions speak louder than words? The response of asset prices to monetary policy actions and statements," *International Journal of Central Banking*.

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

Hanson, S. G., D. O. Lucca, and J. H. Wright (2021): "Rate-amplifying demand and the excess sensitivity of long-term rates," *The Quarterly Journal of Economics*, 136, 1719–1781.

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

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

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

He, Z., S. Nagel, and Z. Song (2022): "Treasury inconvenience yields during the COVID-19 crisis," *Journal of Financial Economics*, 143, 57–79.

Jansen, K. A., W. Li, and L. Schmid (2024): "Granular treasury demand with arbitrageurs," Tech. rep., National Bureau of Economic Research.

Joyce, M. A. and M. Tong (2012): "QE and the gilt market: a disaggregated analysis," *The Economic Journal*, 122, F348–F384.

Känzig, D. R. (2021): "The macroeconomic effects of oil supply news: Evidence from OPEC announcements," *American Economic Review*, 111, 1092–1125.

Kekre, R., M. Lenel, and F. Mainardi (2024): "Monetary policy, segmentation, and the term structure," Tech. rep., National Bureau of Economic Research.

Kuttner, K. N. (2001): "Monetary policy surprises and interest rates: Evidence from the Fed funds futures market," *Journal of Monetary Economics*, 47, 523–544.

Mertens, K. and M. O. Ravn (2013): "The dynamic effects of personal and corporate income tax changes in the United States," *American economic review*, 103, 1212–1247.

Nakamura, E. and J. Steinsson (2018): "High-frequency identification of monetary non-neutrality: the information effect," *The Quarterly Journal of Economics*, 133, 1283–1330.

Pinter, G. (2023): "An anatomy of the 2022 gilt market crisis," *Available at SSRN 4407740*.

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

Ray, W. et al. (2019): "Monetary policy and the limits to arbitrage: Insights from a new keynesian preferred habitat model," in *2019 Meeting Papers*, Society for Economic Dynamics, vol. 692.

Schmitt-Grohé, S. and M. Uribe (2012): "What's news in business cycles," *Econometrica*, 80, 2733–2764.

Smets, F. and R. Wouters (2007): "Shocks and frictions in US business cycles: A Bayesian DSGE approach," *American economic review*, 97, 586–606.

Stock, J. H. and M. W. Watson (2012): "Disentangling the Channels of the 2007-2009 Recession," *Brookings Panel on Economic Activity*, 42, 81–156.

Stock, J. H. and M. Yogo (2005): "Testing for weak instruments in linear IV regression," *Identification and Inference for Econometric Models*, 80–108.

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

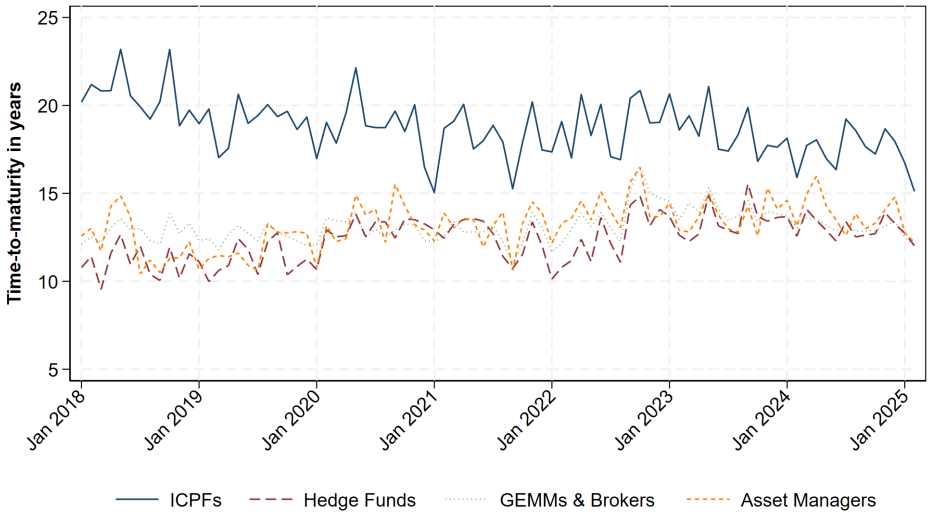
Zhu, L. (2022): "Let the Market Speak: Using Interest Rates to Identify the Fed Information Effect," *Available at SSRN 4035869*.

Appendix

A Additional Data

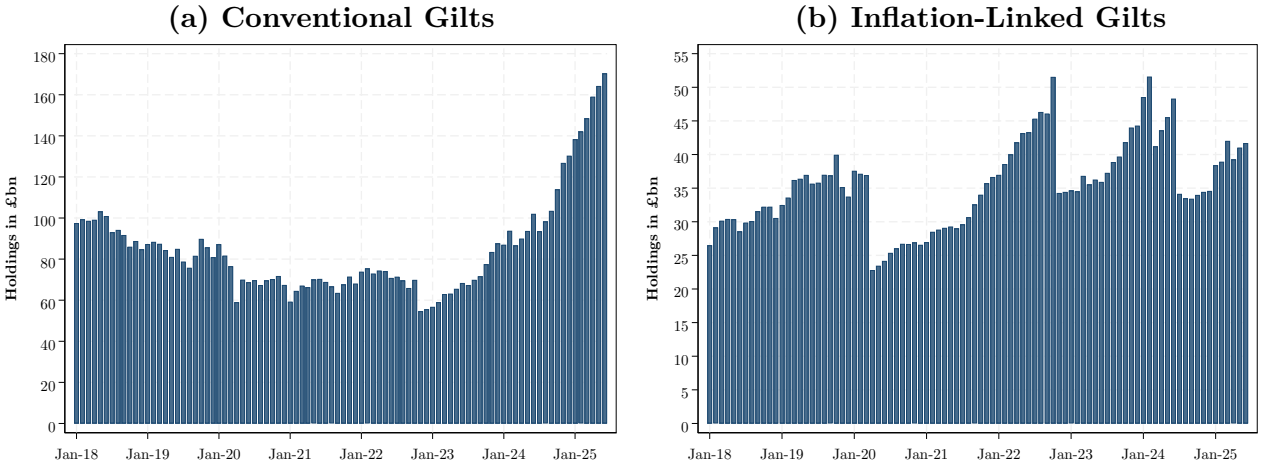
A.1 UK Gilt Market: Further Facts

Figure A.1 Market segmentation: By Maturity Profile of Holdings



Note: "ICPF" refers to insurance companies, pension funds and liability-driven investment funds. Source: MiFID II database on UK gilt transactions, Bank of England.

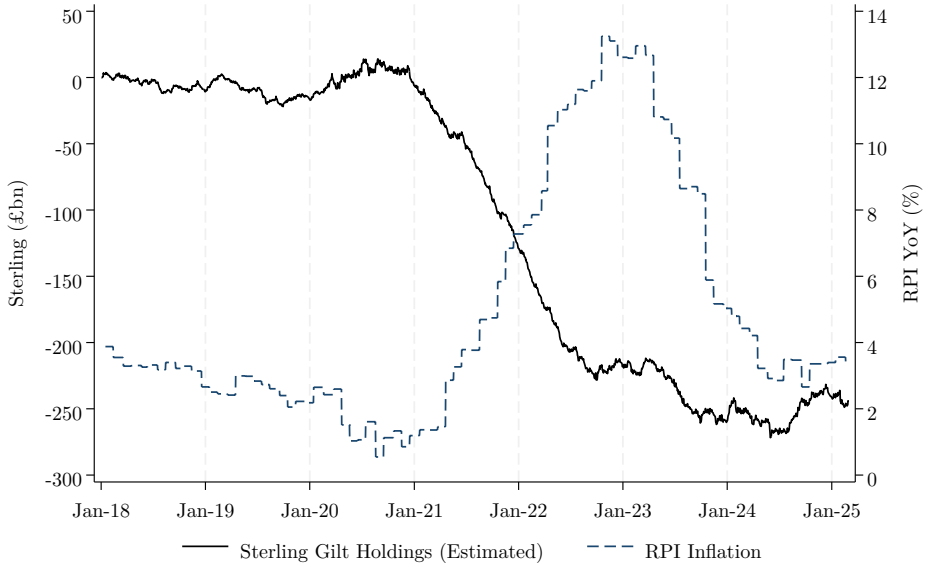
Figure A.2 Aggregate holdings of gilts by banks



Note. Figures show the aggregate holdings of UK gilts by banks in the shorter-term segment (less than 15 years maturity) of the conventional gilt market and the index-linked gilt market. Souce. Monetary Financial Institutions' Balance Sheet dataset, Bank of England.

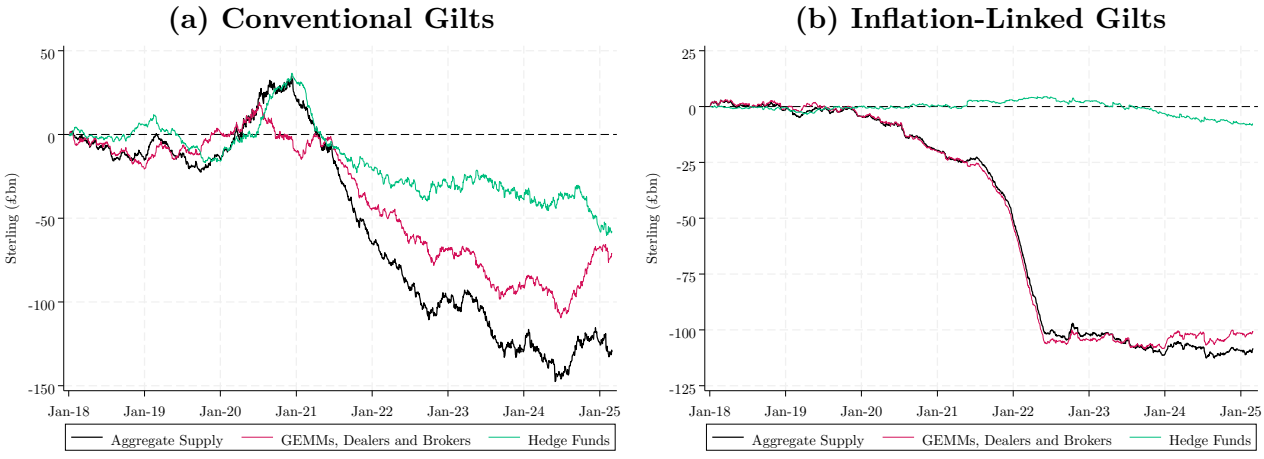
A.2 Construction of Inflation Surprises to News

Figure A.3 UK intermediaries systematically sold gilts as UK inflation rose



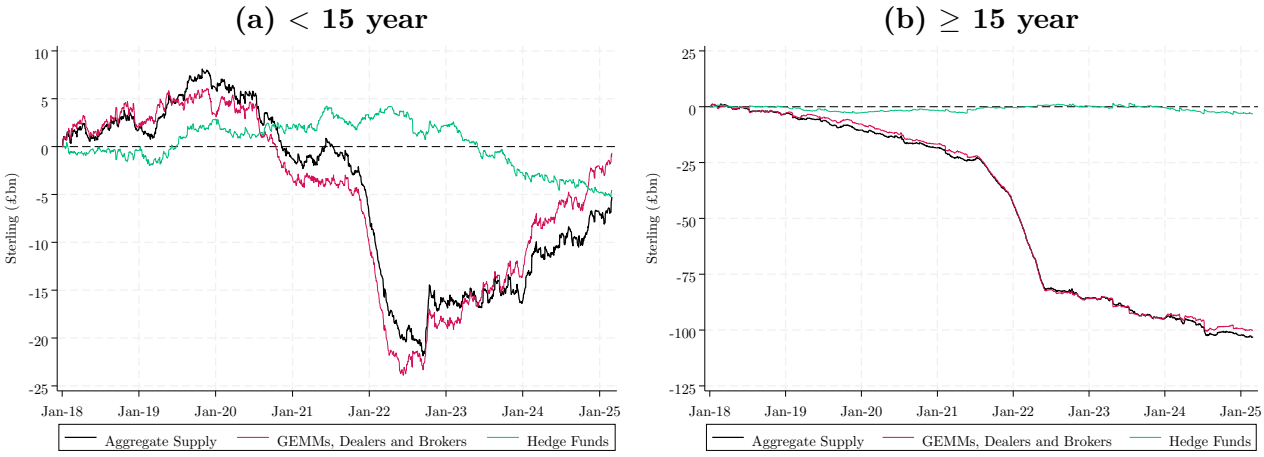
Note: The bold line indicates the cumulative notional value of all secondary market gilt transactions traded by Gilt-Edged Market Makers (GEMMs), Dealer Banks, Inter-Dealer Brokers and Hedge Funds since 1 January 2018. Dashed line indicates year-on-year RPI inflation. Souce: MiFID-II database, Bank of England and the Office for National Statistics.

Figure A.4 UK intermediaries and trading behaviour in the UK gilt market



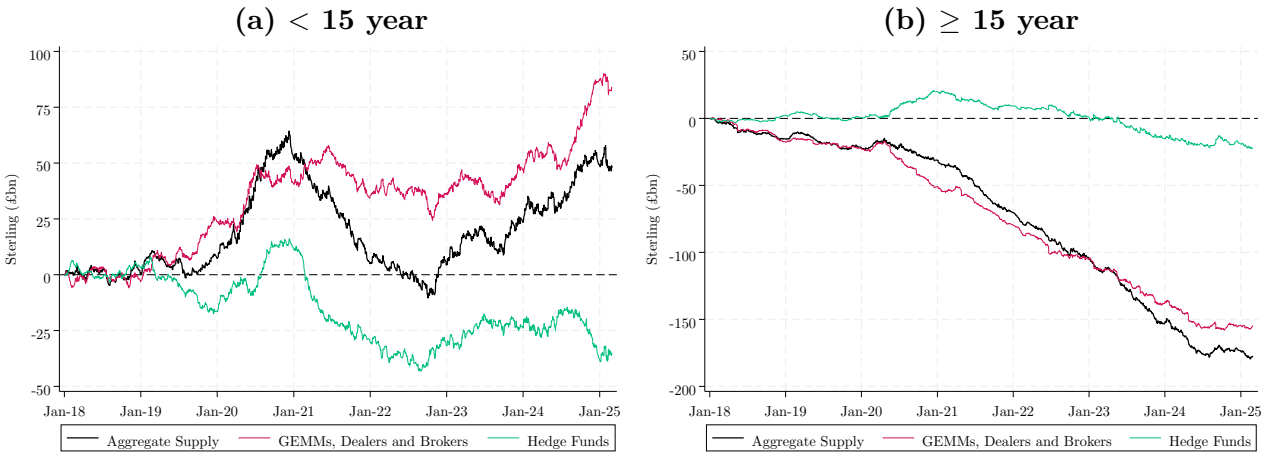
Note. The figures indicate the cumulative notional value of all secondary market gilt transactions traded in the conventional gilt market and the index-linked gilt market by the respective intermediary entities since 1 January 2018. "Aggregate Supply" refers to the sum of these transactions across Gilt-Edged Market Makers (GEMMs), Dealer Banks, Inter-Dealer Brokers and Hedge Funds. Souce: MiFID-II database, Bank of England.

Figure A.5 UK INTERMEDIARIES AND TRADING BEHAVIOUR IN THE INDEX-LINKED GILT MARKET



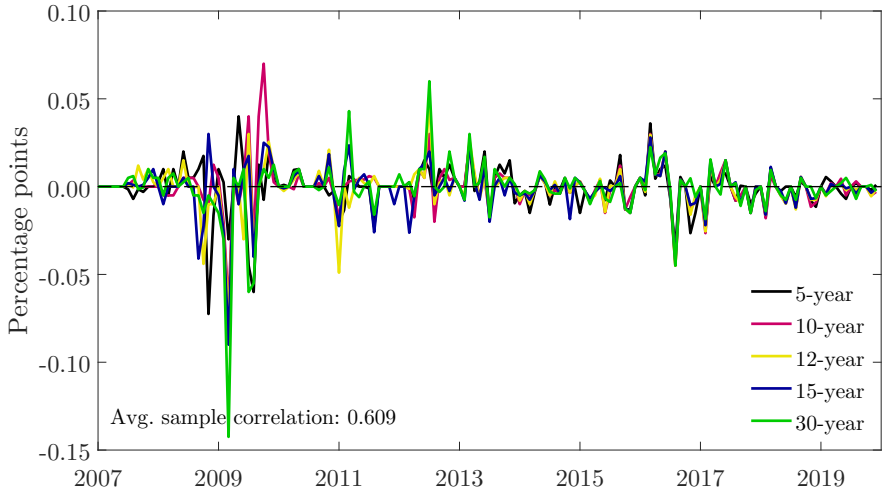
NOTE. The figures indicate the cumulative notional value of all secondary market gilt transactions traded in the index-linked gilt market by the respective intermediary entities since 1 January 2018. "Aggregate Supply" refers to the sum of these transactions across Gilt-Edged Market Makers (GEMMs), Dealer Banks, Inter-Dealer Brokers and Hedge Funds. SOUCE: MiFID-II database, Bank of England.

Figure A.6 UK INTERMEDIARIES AND TRADING BEHAVIOUR IN THE NOMINAL GILT MARKET



NOTE. The figures indicate the cumulative notional value of all secondary market gilt transactions traded in the conventional gilt market by the respective intermediary entities since 1 January 2018. "Aggregate Supply" refers to the sum of these transactions across Gilt-Edged Market Makers (GEMMs), Dealer Banks, Inter-Dealer Brokers and Hedge Funds. SOUCE: MiFID-II database, Bank of England.

Figure A.7 Surprise changes in expected inflation in response to monetary news



Note: Each data point refers to the median change in a zero coupon UK RPI swap rate at intervals -10/+20 minutes around scheduled releases of MPC announcements (excluding the quarterly press conference).

B Additional Empirical Results

B.1 What is an inflation shock?

In the main text, we presented estimates of how long-term interest rates respond to a direct shock to expected inflation. We provide more details in this section as to what this shock may be and how to interpret it in a structural manner. We start by estimating a latent factor model using our inflation surprise measure and argue that the first factor is a transitory business cycle shock that only shifts short-term inflation expectations. We then compare the estimates from a public finance news announcement and an OPEC oil supply announcement, and argue that the first inflation factor has quantitative effects on the long-term forward rates that resembles more to the public finance news shock, which we interpret as a government spending shock.

B.1.1 Latent Inflation Factors

We estimate a latent factor model given by:

$$\mathbf{X} = \mathbf{F}\boldsymbol{\Lambda} + \boldsymbol{\eta}, \tag{B.1}$$

where the matrix $\mathbf{X}$ with dimension $T \times N$ contains the high-frequency inflation swap surprises across maturities $\tau \in [1, ..., N]$, measured in years, in response to UK RPI inflation data release dates (with $N = 10$), the matrix $\mathbf{F}$ with dimension $T \times \kappa$ (with $\kappa < N$) contains the set of latent factors, and the matrix $\boldsymbol{\Lambda}$ with dimension $\kappa \times N$ is the matrix of factor loadings. The residual white noise is contained in the matrix $\boldsymbol{\eta}$. We find that at least two latent factors are present, and they together explain approximately 79% of the swap curve variance (the first factor explains 67%, and the second factor 12%). With three factors, they explain in total 85%.[53] To recover a structural interpretation, we focus on the first two factors and implement an orthogonal rotation via a matrix $\mathbf{Q}$:

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2] = \mathbf{F}\mathbf{Q} \;\; s.t. \;\; \mathbf{Q}\mathbf{Q}' = \mathbf{I}, \;\; \mathbf{z}_1 \perp \mathbf{z}_2. \tag{B.2}$$

where the identifying assumption is that the second factor ($\mathbf{z}_2$) does not move the 1-year inflation swap rate that is the shortest maturity.

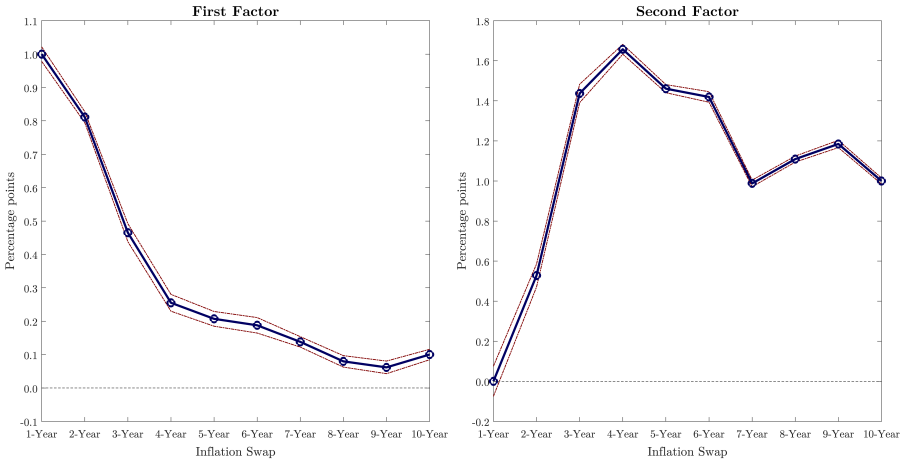
In the monetary policy surprise literature that imposes such an orthogonal rotation (Gürkaynak et al., 2005a; Braun et al., 2025), the first rotated factor is typically interpreted as a shock that changes the policy rate while the second rotated factor is a shock to its path.[54] We draw a parallel interpretation for our application: the first rotated factor is a shock that moves *short-term* inflation expectations (a slope factor),

[53] The Cragg-Donald rank test rejects $H_0 = \kappa_0$ against $H_0 > \kappa_0$ for $\kappa_0 \in \{0, 1\}$, with p-values given respectively by 0.00187 and 0.0345. The third factor explains 6.8% of the swap curve variance.

[54] This is more commonly referred to by the literature as *target* and *path* factors respectively.

while the second rotated factor moves medium to *long-term* expectations (a level factor). By this description, transitory fluctuations of the business cycle belong to the first factor. News shocks, perceptions of monetary policy credibility and inflation uncertainty (or long-run inflation risk) belong to the second factor. Figure B.1 illustrates their factor loadings.

Figure B.1 Factor Loadings



Note: Figures indicate the factor loadings to the first two rotated factors. The identifying assumption is that the second factor does not move the 1-year inflation swap rate. First factor is scaled to raise the 1-year inflation swap rate by 1pp. Second factor is scaled to raise the 10-year inflation swap rate by 1pp.

We use these rotated factors as instruments for breakeven inflation to re-identify the VAR presented in the main text. Given the structural rotation above, we focus on the first factor, which represents a transitory fluctuation (either demand or supply-driven) that raises only short-term inflation expectations and the short-term interest rate.[55] Figure B.2 shows that the overreaction of long-term forward rates and their humped-shape dynamics are robust even to the first factor, and the estimates are sharper, with 90% confidence intervals being further away from zero.

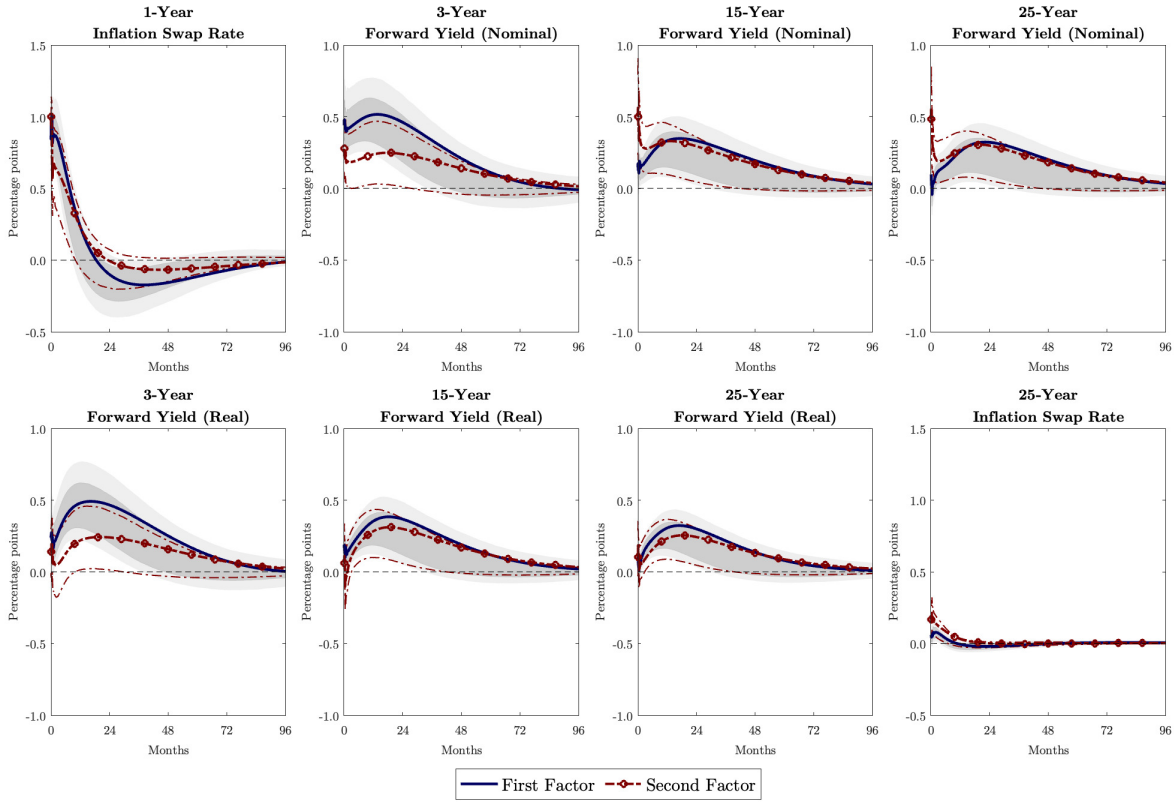
B.1.2 Demand or Supply-Driven Shocks to Expected Inflation?

To further aid interpretation of the inflation factor, we include two additional sources of demand and supply-based variation that each leads to changes in expected inflation, and argue that our inflation factor is likely to resemble an inflationary demand shock more than an inflationary supply shock.

For our supply-driven shock, we identify the joint responses of expected inflation and long-term forward rates in response to a rise in inflation due to an increase in global oil prices. Given that the UK is a small open economy that imports oil for domestic consumption, fluctuations in global oil prices will likely to cause

[55]The two factors differ most in their effects on long-term nominal rates. Consistent with our interpretation of the second factor as capturing long-run inflation risk and monetary policy credibility, the second factor leads to an immediate jump of the long-term nominal forward rates, possibly capturing inflation risk premia. This is, however, transitory, echoing our message in the main text that the UK financial market seems to be capable of bearing inflation risk and inflation expectations are well-anchored.

Figure B.2 Comparison of latent inflation factors



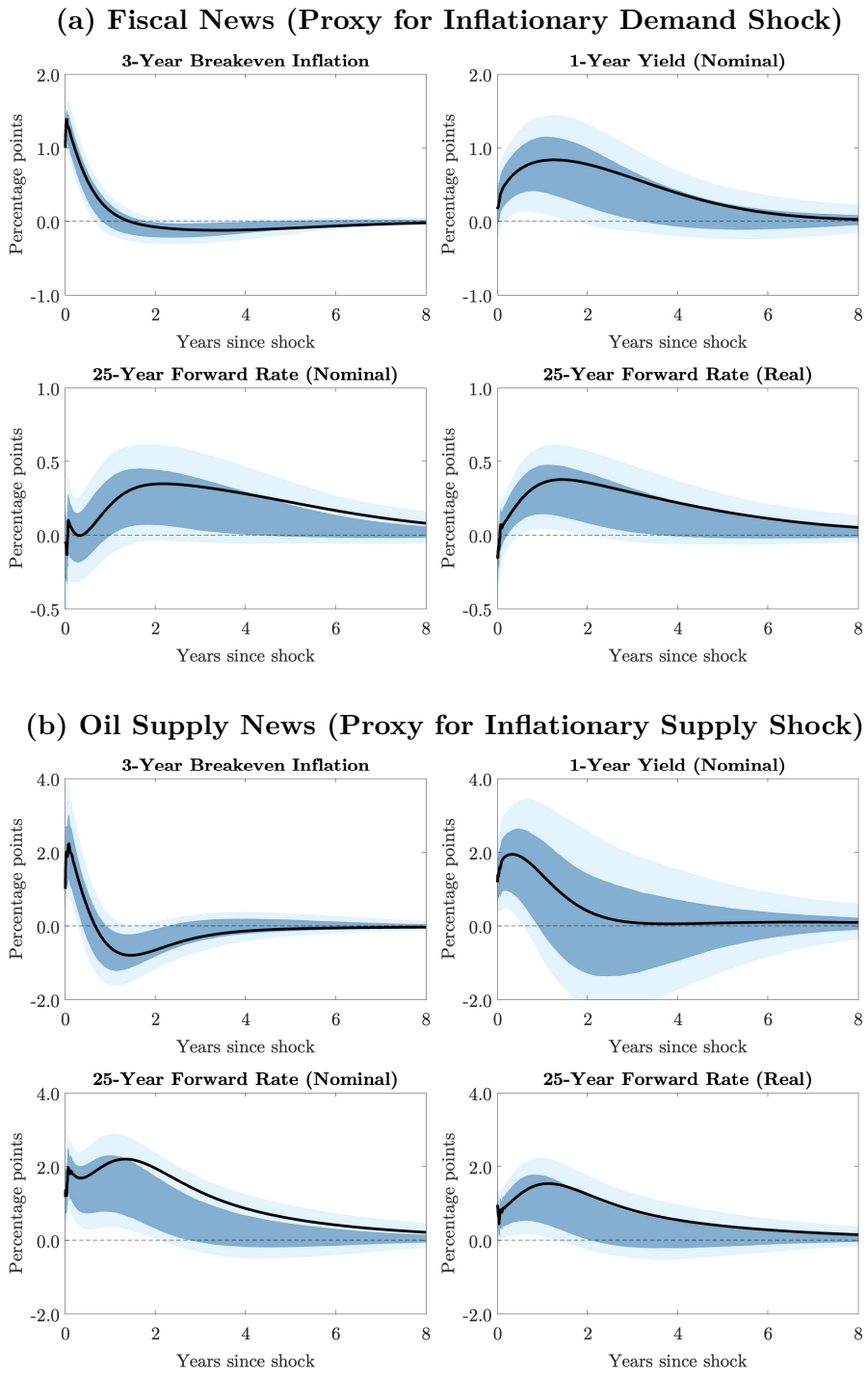
Note: We use the identified factors as instruments for breakeven inflation, as measured by the 1-year inflation swap rate. Scaling of IRFs is such that 1-year expected inflation raises by 1pp on impact. The First-Stage F-statistics are 23.5 and 8.0 for the first and second rotated factors respectively. Shaded (dashed) intervals are 68% and 90% moving block bootstrapped confidence intervals identified using the first (second) factor, sampled with 10,000 bootstrap draws.

UK inflation to deviate from target, leading to changes in interest rate expectations. We use Känzig (2021)'s oil supply news proxy as an instrument for WTI brent oil prices, including this as an additional variable in our VAR specification in the main text. The interpretation of this identification would thus be the causal effect on long-term forward rates and expected inflation of an exogenous increase in oil prices.

For our demand-driven shock, we isolate exogenous changes in expected inflation from inflation swaps within high-frequency windows that surround the scheduled release of UK public finance data. These releases announce actual UK government spending and deficits, relative to the forecasts from the Office of Budget Responsibility (OBR). In a non-Ricardian economy, there are real effects from a non-neutral budget. Financial markets can interpret this to be a source of inflationary pressure, and interest rate expectations can rise. The interpretation of this identification would thus be the causal effect on long-term forward rates and expected inflation of a surprise increase in government spending (or deficit).

Figure B.3 shows the identification from the two types of news announcements. To merit comparison across the two shocks, we scale the impulse response functions such that each leads to an increase in the 3-year breakeven inflation by 1 percentage point on impact (hence an inflationary demand and supply shock). Across

Figure B.3 The persistent overreaction of long-term interest rates: Demand and supply-driven news



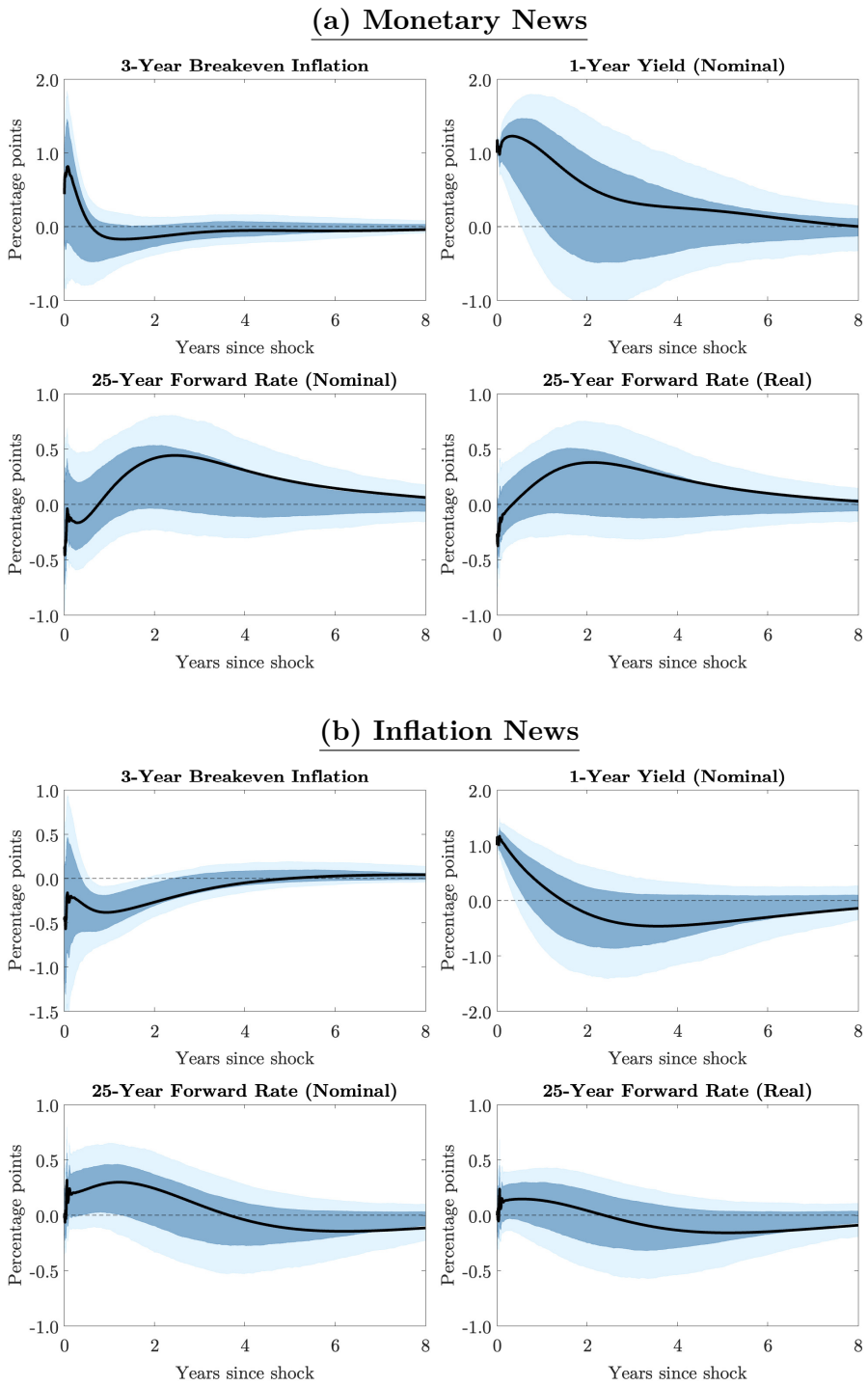
Note: Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 25.8 for fiscal news, and 69.3 for oil supply news. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

both cases, monetary policy responds by raising the 1-year yield. However, this is much more aggressive towards the oil supply shock, where the 1-year yield reaches its peak of approximately 2% within just a quarter of a year. In contrast, monetary policy raises the 1-year yield to its maximum of approximately 1% only after a year in response to the government deficit shock. More over, breakeven inflation responds by falling quickly. Across both cases, there is a persistent rise in the long-term forward rate. For the oil supply shock, it even remains elevated long after the 1-year yield has reverted back to zero.

Overall, we infer across our estimates that long-term interest rates seem to demonstrate the same over-reaction to macroeconomic shocks that are inflationary. Quantitatively, however, the estimates from a government spending shock are closer to those we presented in the main text. Hence, we argue that the first inflation factor resembles an inflationary demand shock that is transitory.

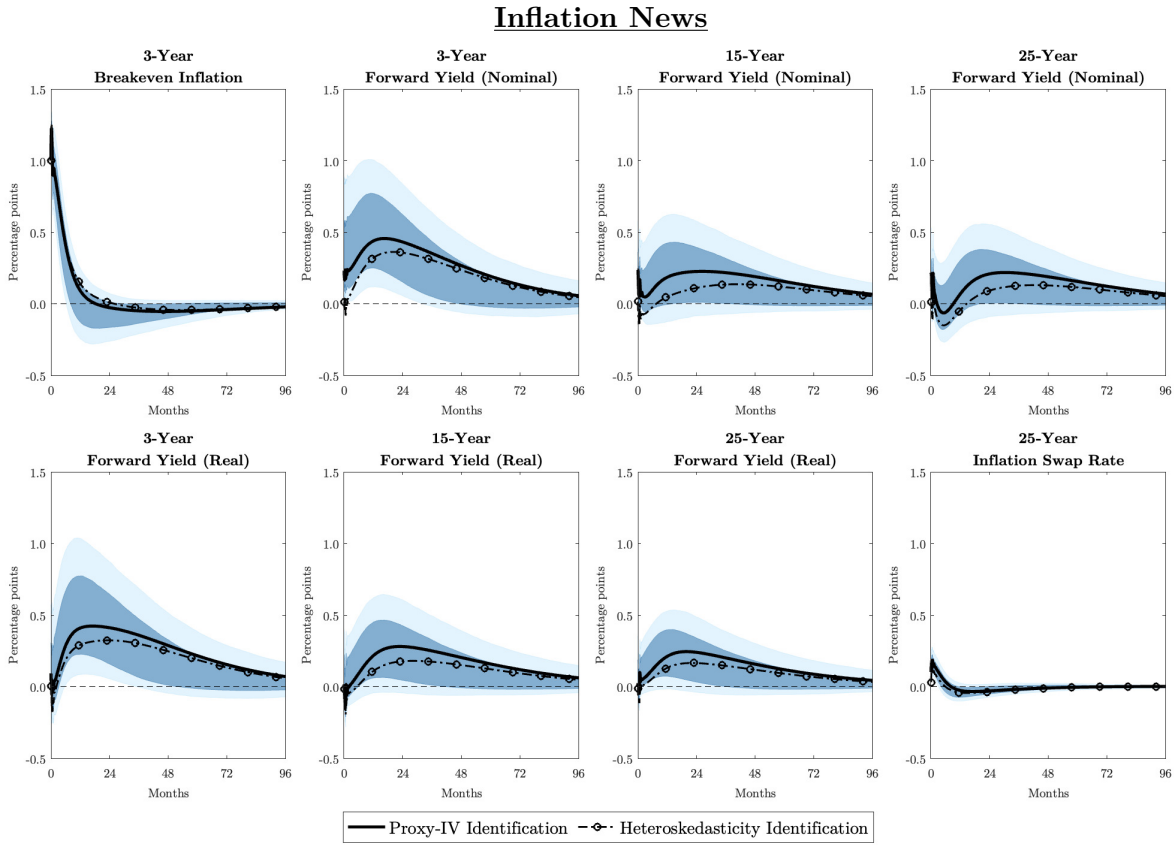
B.2 Additional Estimates

Figure B.4 Identification by interest rate surprises



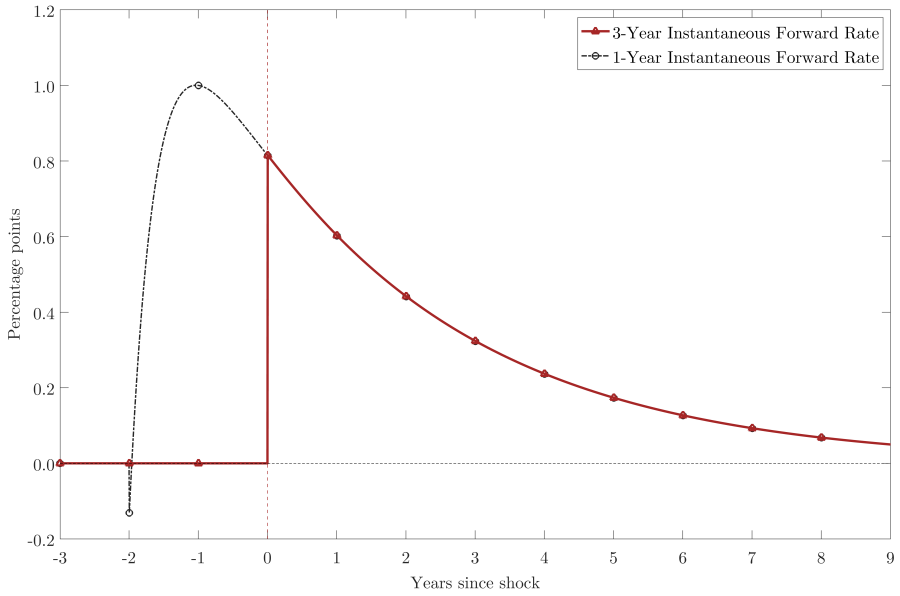
Note: Scaling of IRFs is such that the 1-year yield raises by 1pp on impact. The First-Stage F-statistic is 43.4 for monetary news, and 39.7 for inflation news. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

Figure B.5 Identification by heteroskedasticity



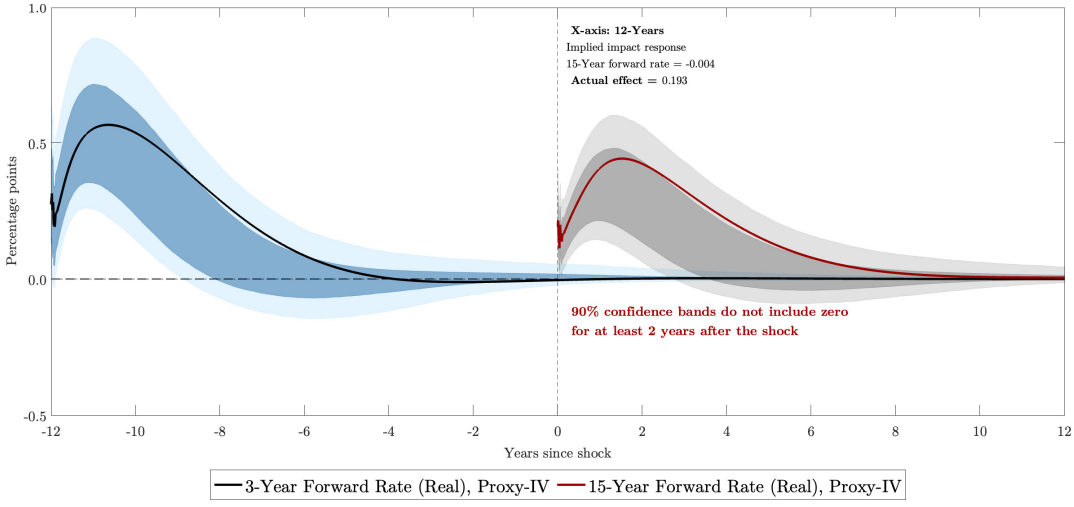
Note: Dashed lines correspond to the point estimates from heteroskedasticity-based identification of the direct shock to inflation. Bold lines and confidence intervals are estimates from proxy-IV identification using inflation swap surprises as an instrument for 3-year breakeven inflation. Scaling of IRFs is such that 3-year breakeven inflation raises by 1pp on impact. The First-Stage F-statistic is 219.6. The Lewis & Merten (2025) generalized test statistic is 47.9. Shaded intervals are 68% and 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

Figure B.6 The measurement of excess sensitivity: An illustration



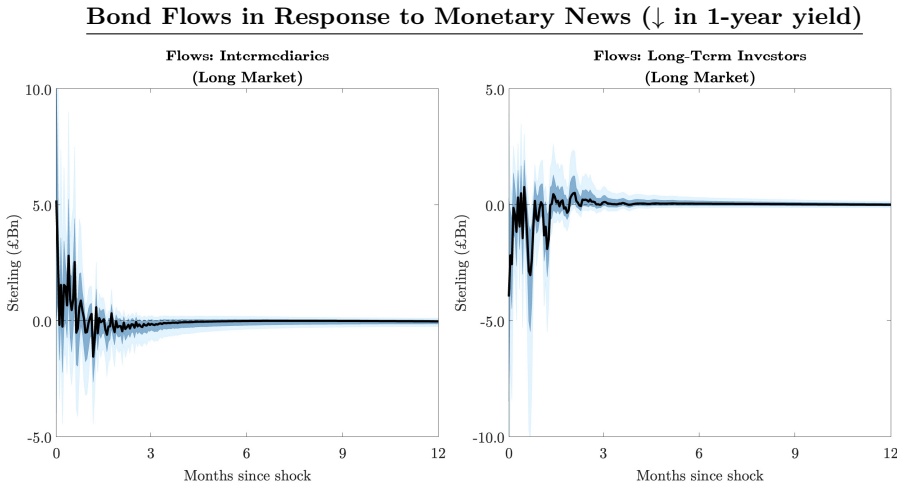
Note: This figure overlays the theoretical impulse response functions of the 1-year and 3-year forward rate that would hold under the expectations theory. For a monetary easing shock that reduces the 1-year forward rate on impact (timing of shock indicated by "-2") followed by monetary tightening, its effect after 2 years should be exactly equal to the immediate effect of the shock on the 3-year forward rate (timing of shock indicated by "0"). In this case, there is no excess sensitivity of the 3-year forward rate. The x-axis has been shifted by 2 years to the right when interpreting the responses of the 3-year forward rate.

Figure B.7 The measurement of excess sensitivity: The first inflation factor



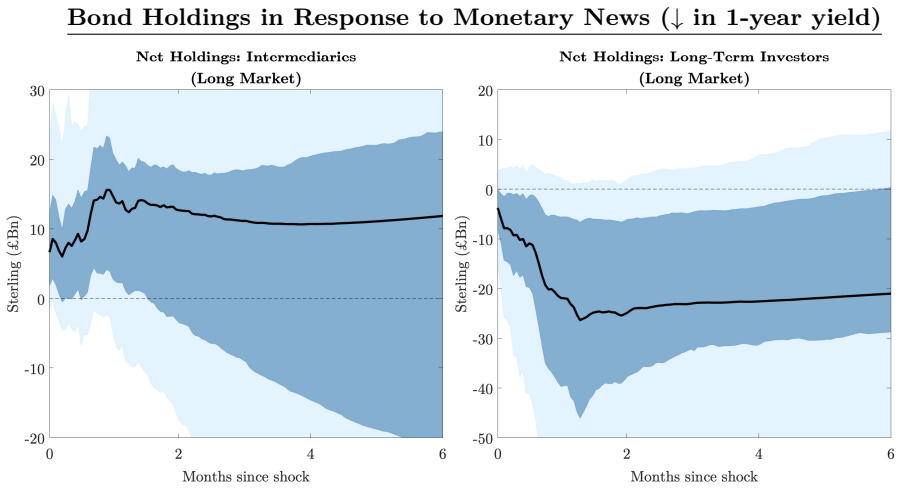
Note: The x-axis has been shifted by 12 years to the right when interpreting the responses of the 3-year forward rate. Impulse responses are interpreted as the response of the yield curve to a direct shock to short-term inflation expectations. The first rotated factor of inflation swap surprises to announcements of inflation data releases is used as an external instrument for breakeven inflation. The First-Stage F-statistic is 23.5. Shaded intervals are 68% and 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

Figure B.8 Intermediaries bear larger quantities of risk as yields decline



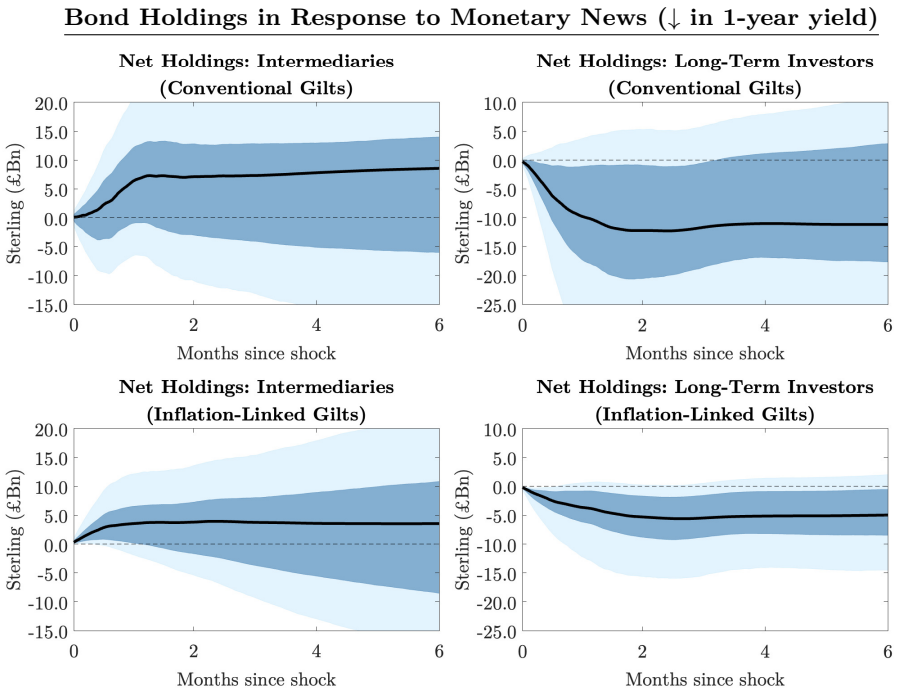
Note: IRFs indicate the aggregate response in the long market segment ($\geq$ 15-year) across the conventional gilt and inflation-linked market. Intermediaries are GEMMs and inter-dealer brokers. Long-term investors are asset managers, pension funds and liability-driven investment (LDI) funds. Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 9.6. The Lewis & Merten (2025) generalized test statistic is 4.1. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

Figure B.9 Intermediaries bear larger quantities of risk as yields decline: Estimates with Hedge Funds as an Intermediary



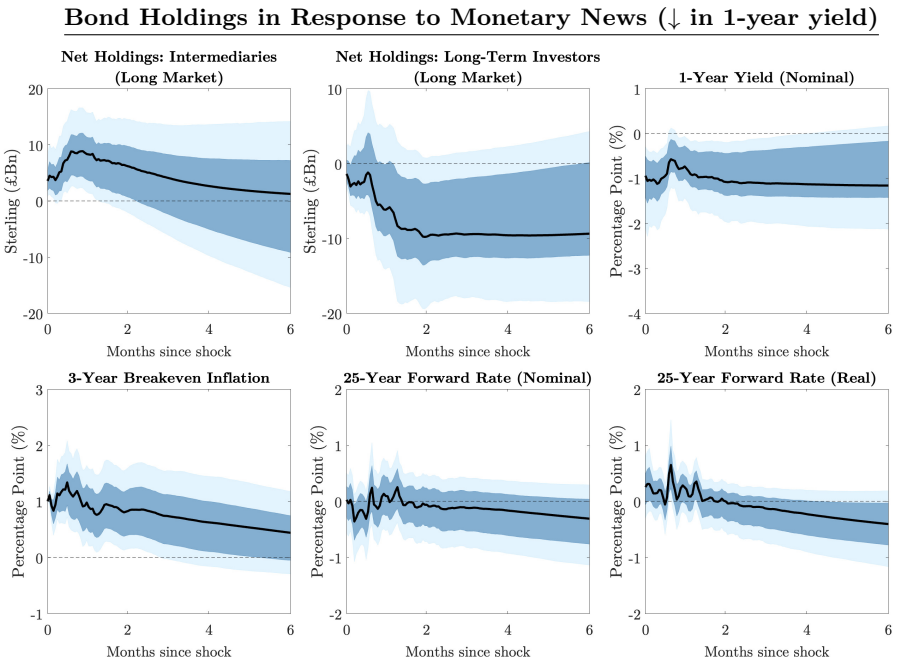
Note: IRFs indicate the aggregate response in the long market segment ($\geq$ 15-year) across the conventional gilt and inflation-linked market. Intermediaries are GEMMs, inter-dealer brokers and hedge funds. Long-term investors are asset managers, pension funds and liability-driven investment (LDI) funds. Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 9.2. The Lewis & Merten (2025) generalized test statistic is 4.1. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

Figure B.10 Intermediaries bear larger quantities of risk as yields decline: Estimates from a 20-day moving average of the raw data



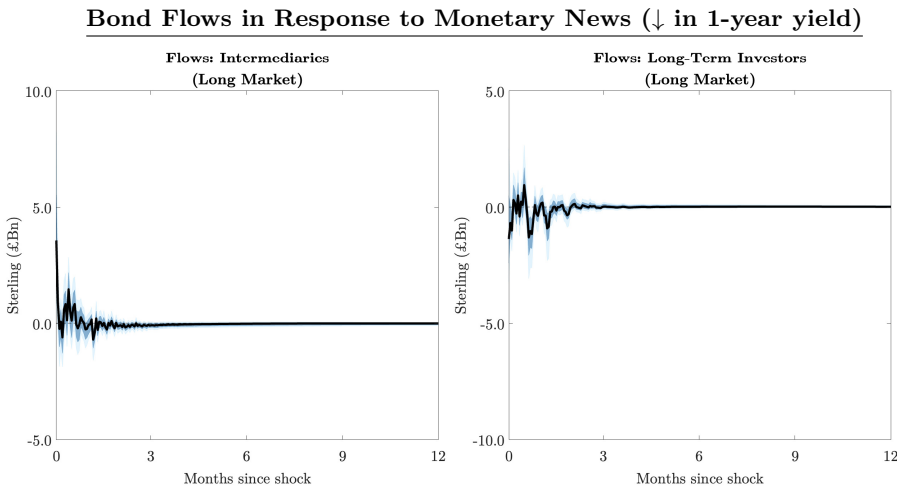
Note: IRFs indicate the aggregate response in the long market segment ($\geq$ 15-year) across the conventional gilt and inflation-linked market. Intermediaries are GEMMs and inter-dealer brokers. Long-term investors are asset managers, pension funds and liability-driven investment (LDI) funds. Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 8.6. The Lewis & Merten (2025) generalized test statistic is 5.1. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

Figure B.11 Intermediaries bear larger quantities of risk as yields decline: Identification using inflation swap surprises as an external instrument



Note: IRFs indicate the aggregate response in the long market segment ($\geq$ 15-year) across the conventional gilt and inflation-linked market. Intermediaries are GEMMs and inter-dealer brokers. Long-term investors are asset managers, pension funds and liability-driven investment (LDI) funds. Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 9.8. The Lewis & Merten (2025) generalized test statistic is 9.6. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

Figure B.12 Intermediaries bear larger quantities of risk as yields decline: Identification using inflation swap surprises as an external instrument

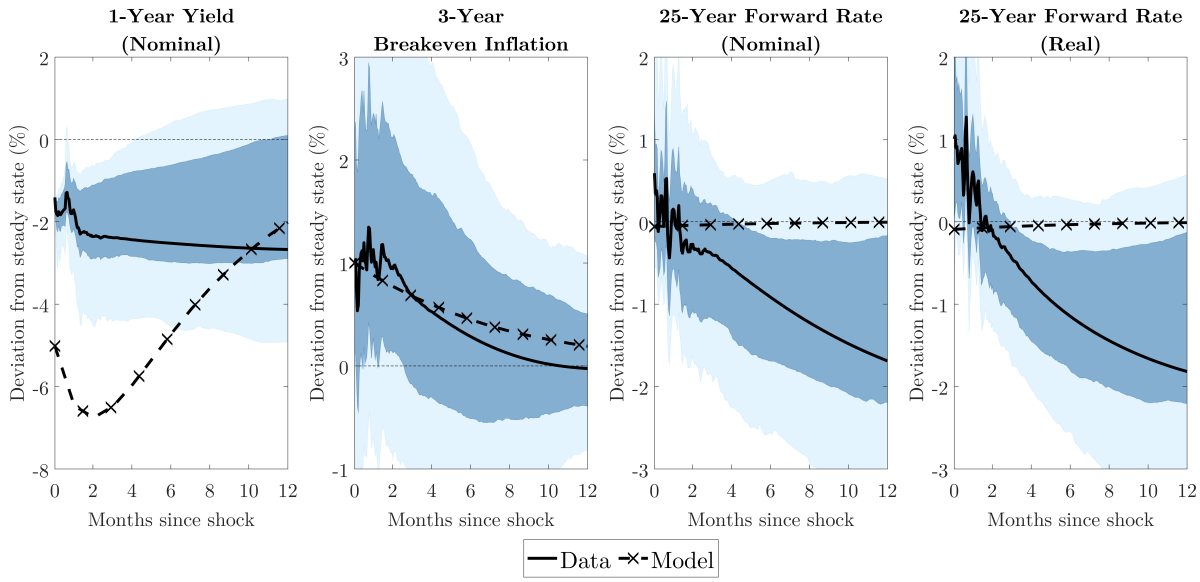


Note: IRFs indicate the aggregate response in the long market segment ($\geq$ 15-year) across the conventional gilt and inflation-linked market. Intermediaries are GEMMs and inter-dealer brokers. Long-term investors are asset managers, pension funds and liability-driven investment (LDI) funds. Scaling of IRFs is such that 3-year expected inflation raises by 1pp on impact. The First-Stage F-statistic is 9.8. The Lewis & Merten (2025) generalized test statistic is 9.6. Dashed intervals are 68% & 90% confidence bands constructed according to the moving block bootstrap approach of Jentsch & Lunsford (2022), sampled with 10,000 bootstrap draws.

C Additional Model Results

C.1 Responses to a Monetary Shock without Market Segmentation

Figure C.1 Short market response to a monetary shock (↓ in short rate)



Note: Impulse response functions correspond to a monetary policy easing shock that raises 3-year breakeven inflation by 1 percentage point on impact.