How the Financing of Balance-Sheet Policy Shapes Its Effects*

Dawis Kim†

November 18, 2026

Click here for the latest version

Abstract

I ask whether the financing of Federal Reserve balance-sheet policy shapes its effects on output and the price level. In a monthly U.S. SVAR I identify two balance-sheet shocks, one financed by reserves and one by overnight reverse repurchases (ON RRP). A reserve-financed contraction raises the risk premium in corporate bond spreads and lowers output and the price level. An ON RRP-financed contraction moves neither the premium nor the price level, and its effect on output is less certain and shorter-lived. The reserve-financed shock is the main source of the unexpected movements in output and the price level over the quantitative easing and tightening that reserves financed. Which liability finances a balance-sheet policy shapes what that policy does.

Keywords: quantitative tightening, quantitative easing, central bank liabilities, overnight reverse repurchases, excess bond premium, sign and zero restrictions, shock heteroskedasticity

JEL codes: C11, C32, E51, E52, E58, G21

*I am deeply grateful to my advisors, Eric M. Leeper, Eric R. Young, and Leland E. Farmer, and I owe special thanks to Tao Zha. For comments at the Federal Reserve Bank of Atlanta I thank R. Anton Braun, Jonathan L. Willis, Salomé Baslandze, Larry D. Wall, and all seminar participants there. For comments at the University of Virginia, Duke University, and Federal Reserve Bank of Richmond Joint Research Workshop I thank Huberto M. Ennis and Alexander L. Wolman. For comments at the University of Virginia practice job talk I thank Andrew Glover and everyone else who commented. All errors are my own.

†University of Virginia. cbe2zn@virginia.edu.

1 Introduction

Do quantitative easing (QE) and quantitative tightening (QT) move output and the price level? This paper finds that the answer depends on which liability finances them. A balance-sheet operation moves long-term assets, and the liability side follows by the same amount, mostly in reserves or overnight reverse repurchases (ON RRP). Since the financial crisis the Federal Reserve (Fed) has run two rounds of QE and two of QT. The liability that financed them changed from one episode to the next. Reserves financed the 2008 to 2014 purchases, and QT1 drained them. Both liabilities financed the COVID purchases. QT2 left the reserves nearly intact and drained the ON RRP balances of money market funds (MMFs) instead (Figure 1).

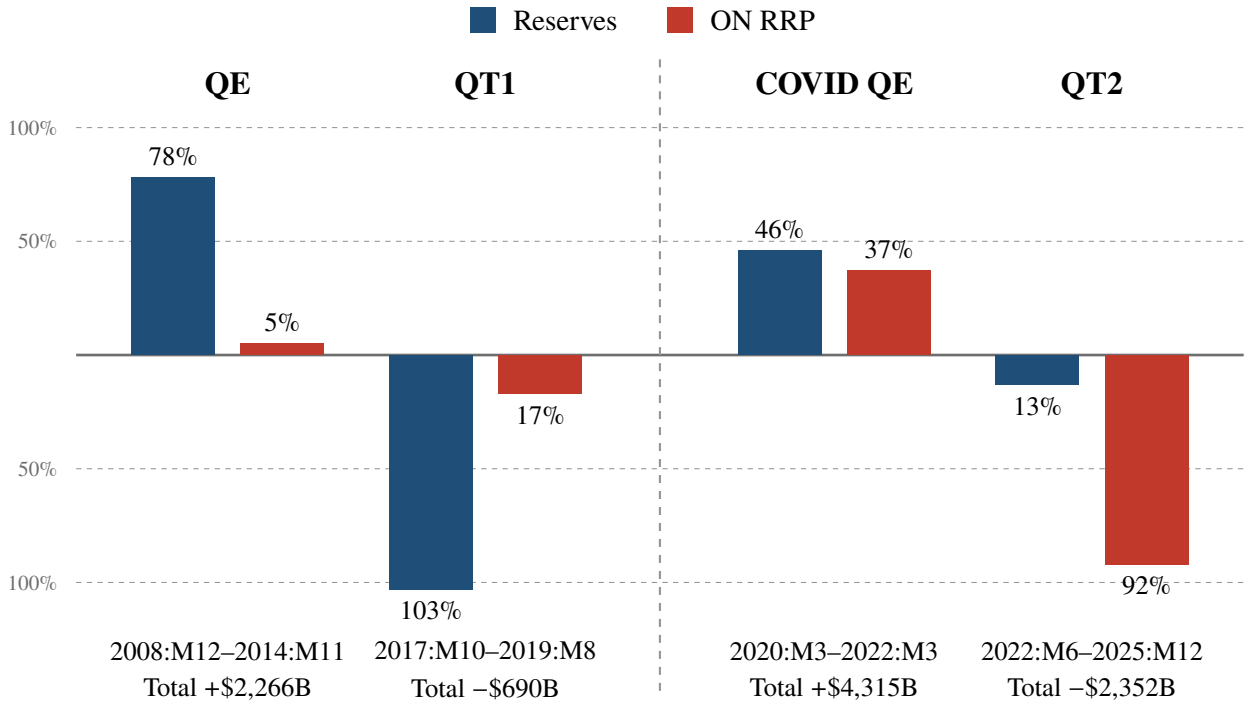


Figure 1: Liability transition across balance-sheet policy.

Note. Bars show the change in reserves and in ON RRP over the episode windows of Table A.1 as a percent of each episode's total asset change. The number above each bar is that share. QE episodes point upward and QT episodes downward. Table A.1 reports the changes in USD billions. The two shares do not sum to 100 percent, and the remainder is the change in the TGA, currency, and the residual liabilities, which is net positive in every episode. Reserves fell by more than total assets over QT1. Inside the COVID QE window the ON RRP buildup followed the TGA drawdown that began in March 2021, as Section 2 documents. Source: Federal Reserve H.4.1 and NY Fed SOMA holdings.

The Fed pays for every purchase with reserves. An operation is reserve-financed if the cash the Fed creates stays in the banking system (the blue bars of Figure 1). It is ON RRP-financed if the

cash ends up with MMFs that lend it back to the Fed overnight (the red bars). MMFs move the cash from month to month. The large buildup and drain of ON RRP happened within terms the Fed set. Within those terms, changes in Treasury issuance and in bank regulation moved the cash. The liability that finances the operation determines who receives the interest the Fed pays. Banks earn it on their reserves, and MMFs earn it on their ON RRP balances.

The two financings differ in how they could reach the cost of credit for firms. Reserve financing could reach it through banks in two ways. Banks earn the margin of interest on reserves over their funding costs, and that income would add to bank capital. Banks with more capital would require a smaller risk premium to lend. Reserves also make banks more liquid. Lending creates deposits that can move to other banks, and more liquid banks would bear that risk at a smaller premium.

ON RRP financing does not reach banks. MMFs pass the ON RRP rate to their shareholders as yield and keep none of it as capital. Unlike banks, MMFs choose where to place their cash and move it to whichever of ON RRP, Treasury securities, and private repo pays more. Over QT2 MMFs moved the cash into short-term Treasury securities and private repo, and hedge funds borrowed in private repo to hold more Treasuries (Section 2.2). That chain stays in the Treasury market, and none of it builds capital or buys firms' debt.[1] If the transmission differs with the liability, the difference should appear first in the risk premium that moves the cost of credit for firms.

Given that the two financings could transmit differently, I identify a reserve-financed shock (RFBS) and an ON RRP-financed shock (OFBS). I estimate them in a nine-variable monthly U.S. SVAR that also identifies demand, supply, monetary policy (MP), and excess bond premium (EBP) shocks. The system includes interest on reserves (IOR) less the one-month Treasury bill rate, which controls for the relative return that governs where MMFs place their cash. The EBP is a risk premium in corporate bond spreads beyond the compensation for expected default. A reserve-financed contraction raises the EBP and lowers GDP and the GDP deflator. A counterfactual following Brunnermeier et al. (2021) holds the EBP at zero and attributes those declines to the rise in the EBP.

[1] Government funds must invest at least 99.5 percent of their total assets in cash, government securities, and fully collateralized repo (U.S. Securities and Exchange Commission, 2026). Commercial paper and other debt that firms issue are not on that list (Section 2.3).

An ON RRP-financed contraction moves neither the EBP nor the deflator, and its effect on GDP is less certain and shorter-lived.

The two balance-sheet shocks together account for about 12% of the forecast variance of GDP and of the deflator at five years, and their share grows with the horizon. The RFBS share of the EBP grows faster and reaches 21% at two years, larger than the share of the EBP shock itself. Over QE and QT1 the RFBS shock is the largest contributor to the GDP and deflator forecast errors. It pushes GDP and the deflator above their forecasts over QE and below over QT1, and it moves the EBP the other way. During COVID, at the zero lower bound, the RFBS shock pushes GDP and the deflator above their forecasts, and the OFBS shock contributes little to either. During QT2 the OFBS contributions to GDP and the deflator have bands that include zero.

I estimate the SVAR with heteroskedastic and heavy-tailed shocks. The two decades since 2008 contain the financial crisis, the COVID collapse, and the post-2021 inflation surge, and the months around COVID are extreme within their phase. I let the variance of each shock take a different value in each of four regimes, which I set at the balance-sheet episodes. Within a regime the shocks are Student-t with heavy tails, as in Brunnermeier et al. (2021), to fit the extreme months. Because the shocks change their variances in different proportions across regimes, the regime variances help separate the shocks from one another.

I label the shocks with sign, zero, and narrative restrictions. The financing is part of Fed policy and responds to the state of the economy (Section 2). In the VAR the balance sheet and its financing follow systematic relations to the current and past values of the variables, as the policy rate follows a rule. The two balance-sheet shocks are the movements those relations do not explain, and the restrictions distinguish them from the other four shocks. The zero restrictions state that real activity and prices do not respond to a balance-sheet shock within the month, the timing assumption of Leeper et al. (1996) and Christiano et al. (2005). The balance sheet changes gradually after an announcement (Nakashima et al., 2024), and real activity and prices respond as the change in holdings accumulates.

I know of no sampler that imposes sign, zero, and narrative restrictions together on a het-

eroskedastic model. This paper develops the heteroskedastic absorb-and-rotate sampler with zero restrictions (HARS-Z). HARS-Z extends the heteroskedastic absorb-and-rotate sampler (HARS) of Kim and Zha (2026), which combines the heteroskedastic likelihood of Brunnermeier et al. (2021) with sign restrictions and with the narrative restrictions of Antolín-Díaz and Rubio-Ramírez (2018). HARS-Z keeps the rotation construction and the elliptical slice sampling of HARS and updates the structural parameters by multiple-try Metropolis (Liu et al., 2000).

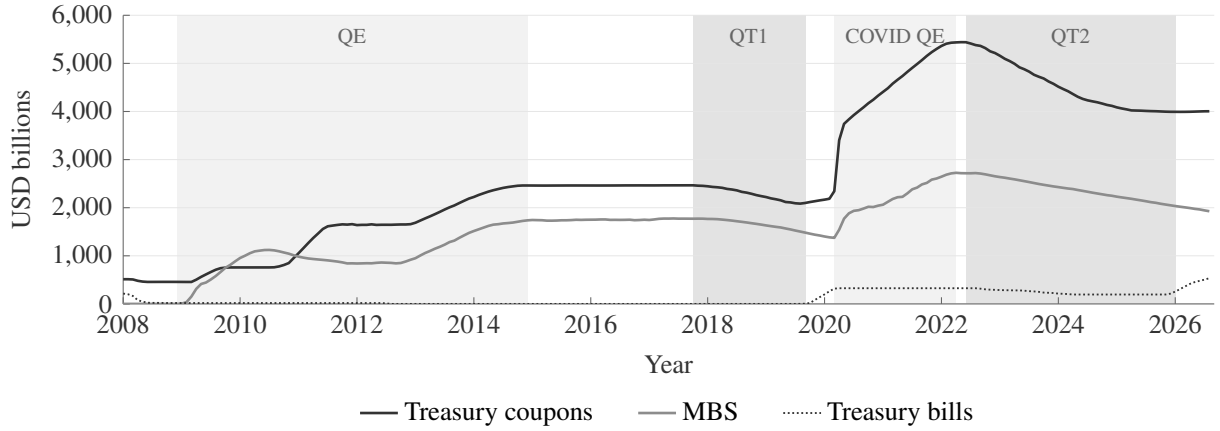
To my knowledge, no existing paper estimates the real effects of balance-sheet policy separately by the liability that finances it. Most of the empirical literature measures the financial-market effects of balance-sheet policy in event studies around announcements, for QE (Krishnamurthy and Vissing-Jorgensen, 2012; Swanson, 2021) and for QT (Smith and Valcarcel, 2023; Lloyd and Ostry, 2024; Lu and Valcarcel, 2024). Theoretical work models the transmission of balance-sheet policy (Chen et al., 2012; d'Avernas et al., 2024; Kumhof and Salgado-Moreno, 2024; Adrian et al., 2025). In these models reserves, or the short-term debt of a consolidated government, finance the purchases. Nakashima et al. (2024) estimate the real effects of the Bank of Japan's balance-sheet policy with a different identification. Their shocks separate the size of the balance sheet from its asset composition, not the liability side.

The rest of the paper proceeds as follows. Section 2 documents the balance-sheet evidence. Section 3 presents the empirical model, the data, and the identification. Section 4 develops HARS-Z and establishes its validity. Section 5 reports the impulse responses, their robustness, and the decompositions. Section 6 concludes. The appendices contain the balance-sheet data, the formal statement of the narrative restrictions, the review of HARS and the theory of HARS-Z with its proofs, and the estimation diagnostics. They also report the responses to the other labeled shocks, the posterior probabilities of the response signs, the counterfactual for the other variables, the robustness figures, and the full decompositions.

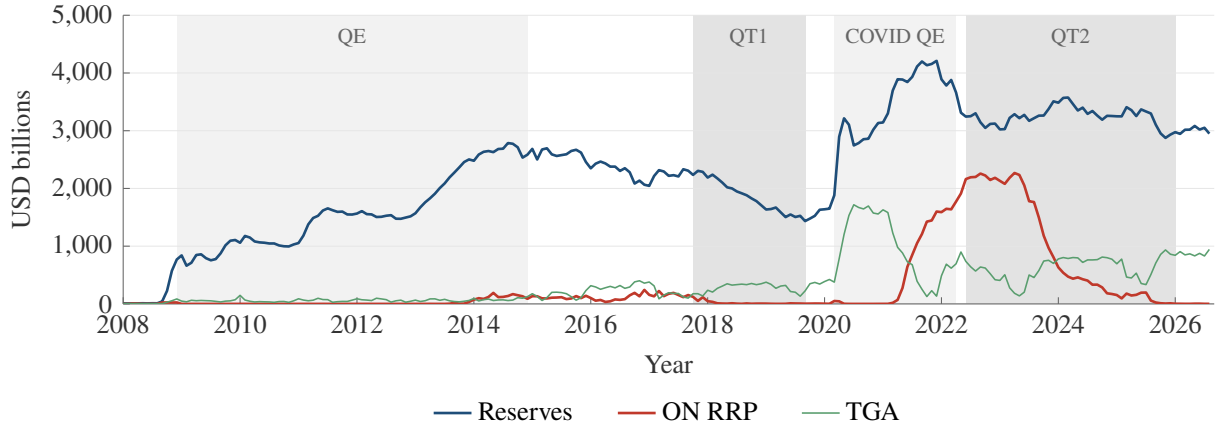
2 Two Ways of Financing the Balance Sheet

2.1 The Two Sides of the Balance Sheet

QE expands the Fed's balance sheet by buying securities, and QT shrinks it by letting securities run off at maturity without reinvestment. Figure 2 plots the assets and liabilities. Every QE raises both sides and every QT lowers them together. The balance-sheet identity fixes the totals and leaves open which liabilities absorb the change.



(a) Fed asset holdings.



(b) Fed liabilities.

Figure 2: The Fed's balance sheet.

Note. Panel (a) plots the holdings of Treasury coupons, Treasury bills, and MBS, and panel (b) plots the reserve, ON RRP, and TGA balances, in USD billions. Each line is a monthly average, and the shaded spans are the episode windows of Table A.1. Source: Federal Reserve H.4.1.

The asset side. In all four episodes Treasury and MBS holdings moved in the same direction. Both expansions raised them together, and both contractions lowered them together (Figure 2a). The MBS share of that change was larger in QE and QT1 than in the COVID QE and QT2 (Table A.1). The COVID QE raised total assets by $4.3 trillion, roughly twice the $2.3 trillion of the 2008 to 2014 expansion. QT2 ran longer and larger than QT1, with coupon runoff of $1.45 trillion against $0.38 trillion (Appendix A).

The liability side. The episodes diverge in which liability they moved. The Fed buys a security from a dealer and pays the dealer's bank in reserves. The bank credits the dealer with a deposit. The reserves stay where they are unless the deposit moves into an MMF that places the cash in the ON RRP facility, which turns the reserves into ON RRP.[2] When the Fed lets a security run off, the Treasury repays it from its account at the Fed and refills that account by issuing new debt. Reserves fall when banks or their depositors buy the new debt, and ON RRP falls when MMFs buy it with cash they take out of ON RRP. Whether the Fed buys or lets a security run off, the holders of the cash move it between the two liabilities (link 1 of Figure 3). They compare the ON RRP rate, which the Fed sets, with bill yields, which Treasury issuance moves (Section 2.2).

Each share below is percent of the episode's total asset change. Reserves absorbed 78% of the 2008 to 2014 expansion. QT1 reversed less than a third of that expansion and drained essentially the entire contraction from reserves. The COVID QE split between reserves (46%) and ON RRP (37%). QT2 drew 92% of the contraction from ON RRP and 13% from reserves (Appendix A).

2.2 The ON RRP Facility from Buildup to Drain

The ON RRP buildup. The ON RRP balance was $154 billion when QT1 began and near $2 trillion when QT2 began. The cash from the COVID purchases stayed in reserves at first, and

[2]Suppose the Fed buys $100 of securities from a dealer. The dealer's bank receives $100 of reserves, and the dealer receives a $100 deposit. The dealer had borrowed from an MMF in repo to hold the security, and it uses the deposit to repay the MMF. The $100 of reserves now sit at the MMF's bank instead of the dealer's bank. If the MMF places the $100 at the ON RRP facility, the reserves leave the banking system and become ON RRP. If the MMF buys a bill from another investor instead, the reserves stay in the banking system.

reserves kept rising into late 2021 as the buying continued (Figure 2b). The balance grew after March 2021, when the Treasury cut bill issuance and drew down the Treasury General Account (TGA) and the released cash flowed through MMFs into ON RRP (Wong, 2021). The expiry of the temporary Supplementary Leverage Ratio (SLR) relief in March 2021 added to it.[3]

The QT2 drain and the ON RRP facility. The ON RRP drain ran from mid-2023, a year after the runoff began. Heavy Treasury bill issuance pushed T-bill yields above the ON RRP rate. MMFs moved their cash out of ON RRP into short-term Treasury securities and private repo. Government funds left first, and their share of the balance fell from 80 percent in 2023:M7 to 35 percent in 2025:M9 (Figure A.1).

The Committee's May 2022 plans state that it would slow and then stop the runoff when reserves are somewhat above the level it judges consistent with ample reserves (Federal Open Market Committee, 2022a). While the runoff drained ON RRP, reserves stayed above that level. Members of the Federal Open Market Committee watched the ON RRP facility as they set the pace of the runoff. Waller (2024) read the cash remaining in ON RRP as excess liquidity and as room to reduce securities holdings further. Logan (2024) attributed the decline in ON RRP to Treasury bill issuance and warned that further runoff would fall on reserves once ON RRP reached a low level. The Committee saw the same risk. In June 2024 it lowered the monthly Treasury redemption cap from $60 billion to $25 billion so that the runoff would not fall sharply on reserves (Federal Open Market Committee, 2024).

ON RRP financing also moves duration. When reserves finance a purchase, the Fed takes duration out of private hands and long-term yields fall (Krishnamurthy and Vissing-Jorgensen, 2012; Chen et al., 2012). When it lets those securities run off, private investors take the duration back

[3]The SLR requires large banks to hold tier 1 capital against total leverage exposure without risk weights. In April 2020 the Fed temporarily excluded Treasuries and reserve balances from the exposure measure to ease strains in the Treasury market. The exclusion expired as scheduled on March 31, 2021, and reserves and Treasuries again counted toward bank leverage. Banks discouraged large deposits, the cash moved to MMFs, and they found few bills to buy (Afonso et al., 2025). Wong (2021) attributes the largest share of the ON RRP buildup to the fall in bill supply during the TGA drawdown, with the SLR expiry and the June 2021 rate adjustment adding to it.

(Crawley et al., 2022; Wei, 2022). A reader may think that the ON RRP-financed episodes move no duration, because MMFs only placed cash in ON RRP over the buildup and took it out over the drain.

Duration moved in both ON RRP episodes. The COVID purchases took coupons and MBS out of private hands whether the cash stayed in reserves or moved into ON RRP. Over QT2 the Fed let $1.45 trillion of Treasury coupons run off (Table A.1). Private investors hold those coupons instead, as they hold the coupons that ran off under QT1. Table A.2 shows who took them on. MMFs added short-term Treasury securities and private repo lending over QT2, and hedge funds borrowed in private repo to hold more Treasuries. Both financings move duration the same way, and Section 3 imposes the same term-spread restriction on both balance-sheet shocks for that reason.

2.3 Why the Financing Should Matter

Each financing pays a different holder. The Fed pays interest on both liabilities, and the two financings differ in who receives it and in what that interest does. Only banks can hold reserves, and interest on reserves goes to them (link 2 of Figure 3). In the aggregate, banks do not choose the quantity of reserves they hold. A bank that lends or buys a security passes its reserves to another bank. The total changes only when the Fed's assets or its other liabilities, such as currency, the TGA, and ON RRP, change.

Banks earn the rate on reserves and pay the cost of the deposits or wholesale borrowing that fund them, and the margin is bank income.[4] A change in reserves that banks expect to persist changes the margin they expect to earn, and with it the value of their capital.[5] Banks with more capital

[4]When the European Central Bank raised its deposit facility rate in 2022, euro area banks with more reserves gained net interest income and book equity relative to other banks (Fricke et al., 2026). Their lending to firms fell less after the hikes, most of all at banks with low equity ratios. That study varies the rate on reserves the banks already held, not the quantity of reserves.

[5]The interest a change in reserves earns within an episode is small. At the rate on reserves, the change in reserves earned at most 1.2 percent of the Tier 1 capital of insured banks over QE and 0.9 percent over QT1 (FDIC Quarterly Banking Profile). A drain that banks expect to persist lowers every future margin, and the value of bank capital prices all of them. Over QT1 the rate on reserves exceeded the average funding cost of banks by 1.1 percentage points. The drain of 709 billion dollars in reserves cut bank income by about 8 billion dollars a year. If banks expected that loss to last five years, its present value at a 5 percent discount rate is 2 percent of their Tier 1 capital. If they expected it to last indefinitely, the present value is 10 percent. Over QE the average funding cost exceeded the rate on reserves, and the income route does not apply.

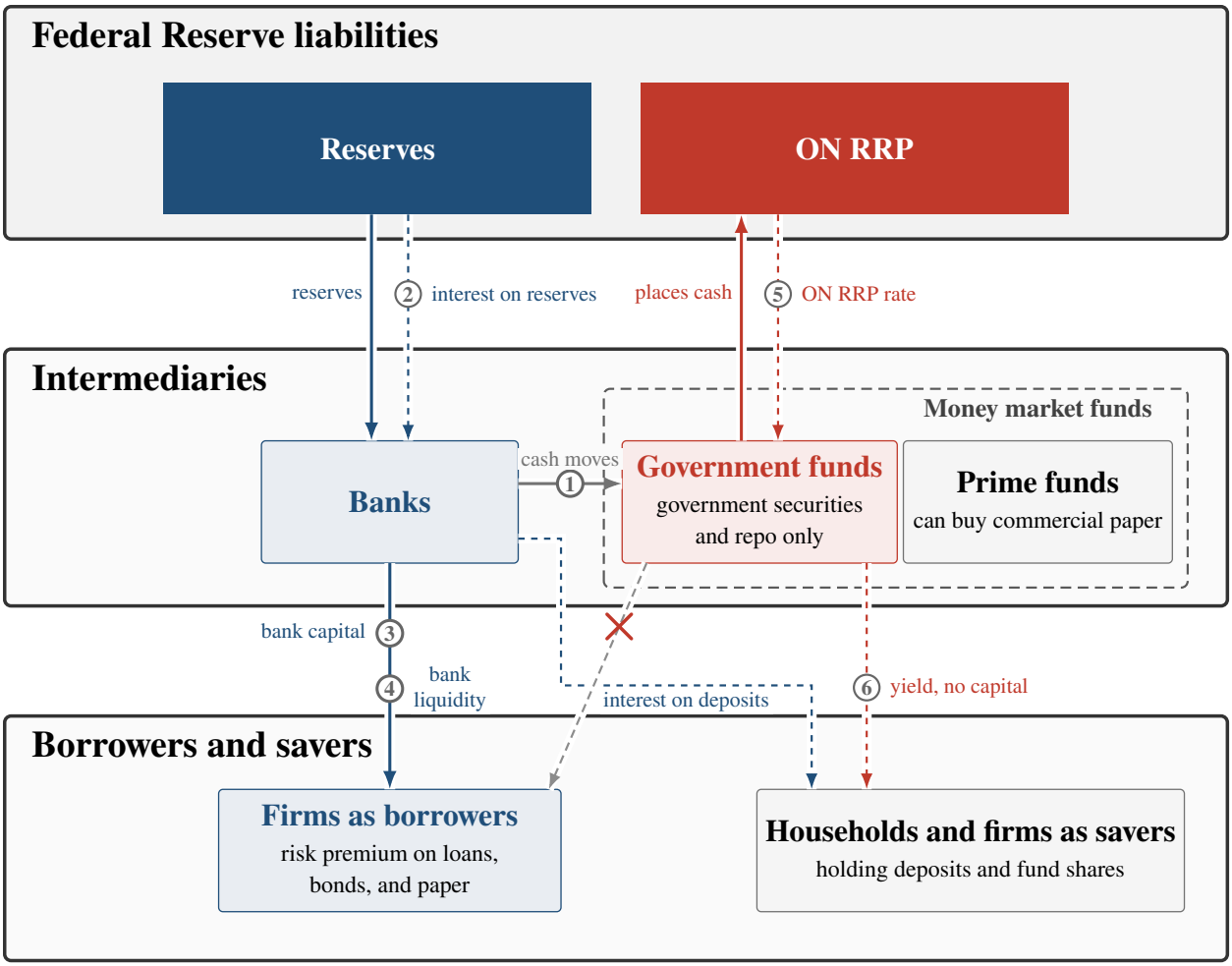


Figure 3: Cash and interest under the two financings.

Note. A purchase of securities creates reserves. The cash stays in reserves or moves into ON RRP as MMFs compare the ON RRP rate with bill yields. The figure leaves out the asset side, which moves the same way under both financings. Solid arrows carry cash in a purchase, and dashed arrows carry interest. Blue marks the reserve-financed chain and red the ON RRP-financed chain. The circled numbers mark the six links the text follows. The crossed gray arrow marks a purchase that government funds cannot make. They hold only government securities and repo.

would carry more credit risk, the risk that borrowers do not repay. They would require a smaller risk premium to lend to the same borrower (link 3, He and Krishnamurthy, 2013). Reserves also make banks more liquid (Copeland et al., 2021). Lending creates deposits, and banks pay out reserves when those deposits move to other banks. Banks hold reserves as a precaution against those outflows (Bianchi and Bigio, 2022). Banks with more reserves would require a smaller premium over the rate on reserves to lend (link 4). Over QE, banks that received more reserves expanded lending and shifted toward riskier loans (Kandrac and Schlusche, 2021).

The interest on ON RRP goes almost entirely to MMFs, mostly to government funds (link 5).[6] MMFs keep none of that interest as capital and pass it to their shareholders as yield (link 6). Unlike banks, MMFs choose where to place their cash and move it to whichever of ON RRP, Treasury securities, and private repo pays more.[7] The yield reaches households and firms as interest income, of the same kind that depositors receive under reserve financing.[8] Wherever MMFs place the cash, it stays in the Treasury market (Section 2.2). The chain neither builds capital nor buys firms' debt.

Only the reserve-financed chain reaches the risk premium. The difference between the two financings should appear first in the risk premium that moves the cost of credit for firms.[9] Section 5 asks whether it shows in the risk premium and in GDP and the price level.

The terms and the realized split. The financing is part of Fed policy, as the policy rate is. The large buildup and drain happened within terms the Fed set. It set the operating framework, the ON RRP award rate, the per-counterparty caps, and the runoff caps (Federal Open Market Committee, 2022b,a). Within those terms, changes in Treasury issuance and in bank regulation moved the realized split (Section 2.2, Afonso et al., 2025). If the Fed had not wanted the ON RRP drain during QT2, it could have raised the ON RRP rate within its target range and kept the cash in ON RRP. It did not, and it slowed the runoff to keep the drain off reserves (Section 2.2). I do not model how the Fed chose those terms.

Like the policy rate, the balance sheet and its financing respond to the state of the economy. The Fed expanded the balance sheet in the financial crisis and in the COVID collapse, and it shrank

[6] At the December 2022 peak MMFs held 92 percent of the ON RRP balance, and government funds held 77 percent. Government funds must invest at least 99.5 percent of their total assets in cash, government securities, and fully collateralized repo (U.S. Securities and Exchange Commission, 2026). Commercial paper, the short-term unsecured debt that firms issue, is not on that list. Prime funds, which can buy commercial paper, never held more than 17 percent (Figure A.1).

[7] Government MMFs exchanged their ON RRP balance for other assets. From 2023:M5 to 2024:M5 they reduced their ON RRP holdings by $1.39 trillion and added $1.36 trillion of Treasury securities and $0.58 trillion of repo with private counterparties. Their total assets rose by $0.44 trillion.

[8] In the Financial Accounts, households and nonprofit organizations held 57 percent of money market fund shares at the end of 2022, and nonfinancial businesses held 17 percent (table F.521) (Board of Governors of the Federal Reserve System, 2026).

[9] Twenty-one of the twenty-six primary dealers are bank-affiliated (Federal Reserve Bank of New York, 2026). Bank capital limits the corporate bonds these dealers can hold. Banks are also the largest lenders to firms, and firms move from loans to bonds when banks tighten their terms (Becker and Ivashina, 2014). Borrowers who move to the bond market raise the spread there.

the balance sheet in the expansions that followed. Which liability finances an operation responds as well, through Treasury issuance, bank regulation, and the bill yields that MMFs compare with the ON RRP rate. The two balance-sheet shocks are the movements in the balance sheet and its financing that the current and past values of the variables do not explain (Section 3.3).

The gap in existing work. Existing work prices balance-sheet policy without separating the liability that finances it. Crawley et al. (2022) in an FRB/US simulation and Wei (2022) in a preferred-habitat model convert a runoff into an equivalent policy-rate increase. Both price it through the duration risk that private investors absorb. Wei (2022) traces which liability a runoff drains, reserves when banks buy the reissued coupon securities and ON RRP when MMFs buy bills. He does not estimate the two cases.

3 Empirical Model

Let y_t be the $n \times 1$ vector of endogenous variables and x_t the $r \times 1$ vector of predetermined terms, including the constant and p lags of y_t. The structural equation is

$$y_t' A_0 = x_t' A_+ + \eta_t', \tag{1}$$

where A_0 is the $n \times n$ nonsingular structural impact matrix and A_+ is the $r \times n$ matrix stacking the constant and lag coefficients. The vector η_t collects the model shocks, independent across equations and time conditional on the regime.[10]

Following the deterministic variance-regime specification of Brunnermeier et al. (2021), let

$$s_t : \{1, \ldots, T\} \to \{1, \ldots, h\}$$

be a known date-to-regime map that assigns each date to one of h volatility regimes. The shocks are

[10] Under this right-multiplication convention, which (1) shares with Kim and Zha (2026) and Arias et al. (2018), columns of A_0 correspond to equations.

heteroskedastic across regimes and heavy-tailed within them,

$$\eta_{i,t} \mid \nu_{i,t},\, s_t = k \;\sim\; \mathcal{N}\big(0,\; \lambda_{i,k}\, \nu_{i,t}\big), \qquad i = 1, \ldots, n, \tag{2}$$

where $\lambda_{i,k} > 0$ is the relative variance of shock i in regime k and $\nu_{i,t}$ is a latent scale with an inverse-gamma distribution, as in Brunnermeier et al. (2021). The inverse-gamma parameters make each standardized shock a Student-t with unit scale, and the variance of shock i in regime k is proportional to $\lambda_{i,k}$. I normalize the regime-specific variances to satisfy

$$\frac{1}{h} \sum_{k=1}^{h} \lambda_{i,k} = 1, \qquad i = 1, \ldots, n,$$

which fixes the scale of each shock variance path across regimes.

Post-multiplying (1) by A_0^{-1} gives the reduced form

$$y_t' = x_t' B + u_t', \qquad B = A_+ A_0^{-1}, \qquad u_t' = \eta_t' A_0^{-1}, \tag{3}$$

where the regime-specific matrices

$$\Sigma_k \;=\; (A_0^{-1})'\, \Lambda_k\, A_0^{-1}, \qquad \Lambda_k := \operatorname{diag}(\lambda_{1,k}, \ldots, \lambda_{n,k}), \tag{4}$$

are proportional to the reduced-form covariances $\operatorname{Var}(u_t \mid s_t = k)$, with a factor that the Student-t degrees of freedom set and that is the same in every regime. These matrices identify the columns of A_0 up to signed permutation whenever the variance ratios across regimes are not mutually proportional (Rigobon, 2003; Lanne et al., 2010). The covariances identify (A_0, Λ) and do not label the shocks. In this paper, the economic labeling comes from the restricted rotation, and the two roles stay separate throughout.

3.1 Data and Priors

I estimate the model of Section 4 on a monthly U.S. system with nine variables ($n = 9$) and $p = 6$ lags. The sample is November 2008 through June 2026.[11]

The system contains nine variables. The first four are the two-year Treasury yield, the log of the holdings of the System Open Market Account (SOMA), log(Reserves + ON RRP), and the ON RRP ratio ON RRP/(Reserves + ON RRP). SOMA holdings cover Treasury notes and bonds plus agency MBS and exclude bills.[12] The next four are the log of real GDP, the log of the GDP deflator (GDPDEF), the term spread, and the excess bond premium (EBP). The ninth is interest on reserves (IOR) less the one-month Treasury bill rate. The term spread is the 10-year Treasury yield less the 3-month bill rate. I map quarterly GDP and the deflator to monthly frequency by Chow–Lin interpolation and aggregate every series to monthly means.

The two-year Treasury yield stands in for the policy rate. At the zero lower bound the funds rate carries no innovation. Gertler and Karadi (2015) take a short-maturity government yield as the policy indicator, on the ground that announcements about future policy operate over a horizon of about two years. The two-year yield spans that horizon.

I include the EBP for the reason Section 2 gives. The two financings differ in how they move the risk premium intermediaries require to bear credit risk. The EBP measures that premium in the corporate bond market. It is the part of corporate bond spreads not attributable to expected default risk. Gilchrist and Zakrajšek (2012) read a rise in the EBP as a fall in the financial sector's capacity to bear that risk.[13] In their estimates, a rise in the EBP that is orthogonal to the state of the economy lowers economic activity and, over time, inflation. Glover and Oliyide (2024) read the EBP as the willingness of lenders to lend beyond what the riskiness of the loans explains. Gertler and Karadi

[11] The November 2008 start coincides with the financial-crisis interventions that preceded QE1. It drops the pre-crisis regime, where reserves stayed near zero and reserve variance was orders of magnitude smaller.

[12] The Fed treats bills separately from the programs. It purchased bills from October 2019 into 2020 to keep reserves ample. It described those purchases as technical measures that do not represent a change in the stance of monetary policy (Federal Open Market Committee, 2019). Under the 2022 runoff the decline in Treasury holdings covers coupon securities, and bills fall only to the extent that coupon maturities run below the monthly cap (Federal Open Market Committee, 2022a). Bill holdings track the maturity schedule and the reserve-management purchases rather than the purchase and runoff decisions.

[13] The part of the spread they remove moves with the borrowers, and what remains moves with the lenders.

(2015) include it in a system like this one.

Section 2 implies a prediction for the EBP. Suppose the reserve-financed chain adds to the income and capital of banks and the ON RRP-financed chain does not. Intermediaries with more capital require a smaller risk premium, and a reserve-financed expansion would lower the EBP while a reserve-financed contraction would raise it. An ON RRP-financed operation would move the EBP in neither direction. The EBP carries no restriction under either balance-sheet shock, and the data decide whether that pattern appears. Section 5.1 reports how each balance-sheet shock moves it.

Two liability variables enter the system. The size variable log(Reserves + ON RRP) sums the two liabilities that absorbed the QE and QT episodes of Section 2. It treats reserves and ON RRP as close substitutes, two forms of overnight central-bank money that commercial banks and MMFs hold, respectively (Diamond et al., 2024; Copeland et al., 2021). The ON RRP ratio ON RRP/(Reserves + ON RRP) measures the sector split of that total. The ratio distinguishes the two contractions. A reserve drain and an ON RRP drain reduce the size variable by the same amount, but they move the ratio in opposite directions. The tables abbreviate the two variables as log(R + RRP) and ON RRP/(R + RRP).[14]

The ninth variable controls for the relative return that governs where MMFs place their cash. MMFs move their cash to whichever of ON RRP and Treasury bills pays more, and those choices move the ON RRP ratio that separates the two balance-sheet shocks. Section 5.2.2 reports the estimates without it. The variable is the monthly mean of the daily difference between IOR and the four-week Treasury bill rate.[15] It carries no restriction under any shock, and the figures report it in basis points.

The priors follow Sims and Zha (1998) and Brunnermeier et al. (2021). Each element of A_0 has an independent Gaussian prior with mean 100 on the diagonal, 0 off the diagonal, and standard deviation 200. At the prior mean $A_0 = 100I$ the reduced-form residual of each equation

[14]The ratio lies between zero and one and equals zero before the ON RRP facility became active. It enters in levels, with no log or inverse-hyperbolic transform.

[15]IOR is the interest rate on excess reserves through July 28, 2021, and the interest rate on reserve balances from July 29, 2021. I use IOR in place of the ON RRP rate because the ON RRP rate starts in September 2013. A spread on the ON RRP rate would cut the sample at October 2013 and drop most of the first variance regime, over which reserves financed the balance sheet. IOR and the one-month bill rate cover the whole sample.

has a standard deviation of 0.01, and the standard deviation of 200 keeps the prior diffuse. A Dirichlet($\alpha = 2$) prior on the normalized variance shares centers the prior on equal variances across regimes. It enforces the normalization that the relative variances of each shock average to one.

Sims–Zha dummy observations center each equation on an independent random walk. The exceptions are the ON RRP ratio, the term spread, the EBP, and IOR less the bill rate, which I center on white noise.[16] Following Brunnermeier et al. (2021), the latent scales are i.i.d. inverse-gamma and make each standardized shock Student-t with 5.46 degrees of freedom. Section 3.2 reports the residual diagnostics that motivate the regime structure and the Student-t errors, including the two-step fit of the degrees of freedom.

3.2 Volatility Regimes and Heavy-Tailed Shocks

For the heteroskedasticity identification I use $h = 4$ volatility regimes. Table 1 reports the dates. Each regime boundary is a boundary of one of the four balance-sheet episodes of Table A.1. The first break is the end of the QE purchases in 2014:M11. The second is the start of the COVID purchases in 2020:M3, and the third is the start of the QT2 runoff in 2022:M6. Each of the four regimes contains one balance-sheet episode, and none contains two.

Table 1: Volatility regimes.

Regime	Start	End	Description
1	2008:M11	2014:M11	QE1 through QE3.
2	2014:M12	2020:M2	Between the QE purchases and the COVID purchases, and QT1.
3	2020:M3	2022:M5	COVID-19 QE.
4	2022:M6	2026:M6	QT2 and the months after it ended.

Note: End dates are inclusive. Each break is a date on the balance sheet, the end of the QE purchases, the start of the COVID purchases, and the start of the QT2 runoff. Each regime contains one balance-sheet episode of Table A.1. The QT2 runoff ended in 2025:M12, and the six months that follow are too few for a regime of their own.

The sample spans the financial crisis, COVID, and the post-2021 tightening. Figure 4 plots the

[16] The hyperparameters are overall tightness 3, lag decay 0.5, sum-of-coefficients (μ) weight 1, and co-persistence (λ) weight 5. The dummy scales use each variable's univariate AR(p) residual standard deviation, and the prior precision stays invariant to units.

reduced-form residuals u_t in (3) from estimating (1) with a single variance regime and Gaussian errors, without the identification restrictions, under the prior of Section 3.1. Each panel shows the posterior median residual of one equation in its own units. The four subsamples follow the dates of Table 1 and partition the residual sample for description only. The number above each subsample is its residual standard deviation.

The standard deviations vary across subsamples, and they do not move together across equations. SOMA is more than twice as volatile in the third subsample as in the first. IOR less the one-month bill rate is most volatile in the fourth subsample, three times as volatile as in the first. Non-proportional patterns of this kind motivate the regime variances. Equation (4) identifies the columns of A_0 when no two structural shocks change their variances in the same proportion across regimes. At the posterior mode of the full model and across the retained draws, the estimated variance ratios are far from proportional.[17]

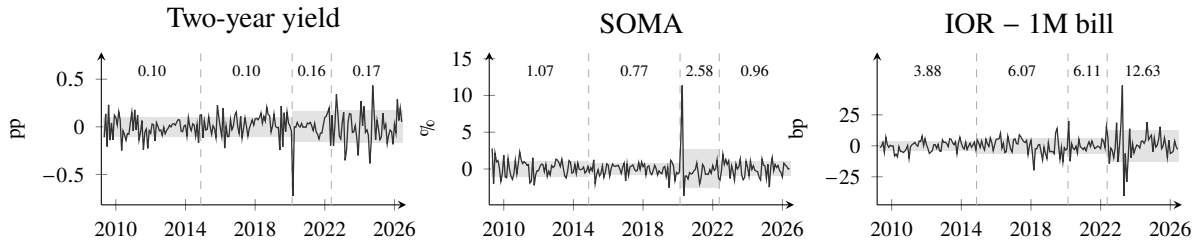


Figure 4: Reduced-form residuals under a single variance regime, three equations.
Note. Each panel plots the posterior median of the reduced-form residual of one equation from the estimation of (1) with a single variance regime, Gaussian errors, no identification restrictions, and the prior of Section 3.1. Units are percent for SOMA, percentage points for the two-year yield, and basis points for IOR less the one-month bill rate. Figure D.1 plots the other six equations. Dashed lines mark the regime breaks of Table 1. The shaded band in each subsample is plus and minus one subsample standard deviation of the series, and the number above each subsample is that standard deviation in the units of the panel. The residual sample runs from 2009:M5 to 2026:M6.

I measure each residual against the typical size of a shock in its own regime, and a few months still stand far outside the rest. The SOMA residual of April 2020 is 9.1 times that size, and the term-spread residual of March 2020 is 5.8 times. Those are the months of the emergency purchases at the onset of COVID and of the emergency cuts to the zero lower bound.

Without the latent scales $\nu_{i,t}$ of (4) the estimated variance of a regime that holds one of these months reflects that month rather than the regime. The heteroskedasticity identification weakens

[17] Table D.2 in Appendix D reports them.

with it. The latent scales give each shock in each month its own scale. The two months lose weight in the likelihood, and the regime variances keep their meaning. The two-step procedure of Brunnermeier et al. (2021) fits 5.46 Student-t degrees of freedom to these residuals, and the estimation uses that value. Section D reports the diagnostics behind these numbers.

3.3 Identification

I identify the shocks with three sets of restrictions. Sign restrictions label six of the nine shocks. Zero restrictions hold at impact. The zeros bind under the average transmission, the responses evaluated at the average of the relative variances across regimes, and need not hold in the regime-specific responses (Appendix C.1). Narrative restrictions tie RFBS and OFBS to the episodes that generated them. All sign restrictions bind from impact through horizon five and must hold both under the average transmission and in every regime. I describe the two balance-sheet shocks first and the four remaining labeled shocks second. Table 2 collects the restrictions on the balance-sheet shocks, and Table 3 collects the restrictions on the demand, supply, monetary policy (MP), and EBP shocks.

The balance-sheet shocks. Table 2 collects the restrictions on RFBS and OFBS. The two shocks correspond to the two financing channels. I normalize the positive direction of both shocks as an expansion. A positive RFBS shock raises SOMA, raises log(Reserves + ON RRP), and compresses the term spread, and the cash the expansion creates stays in reserves. A positive OFBS shock moves the same three variables in the same direction, and the cash the expansion creates ends up in ON RRP. The ON RRP ratio separates the two columns. A reserve-financed balance-sheet expansion lowers the ratio (reserves raise the total while ON RRP is unchanged). An ON RRP-financed expansion raises the ratio directly. Without the ratio restriction the two columns would satisfy identical restrictions. The restrictions carry no episode label. QE and QT1 enter as positive and negative realizations of RFBS, and the QT2 drain enters as negative OFBS realizations.

The Fed's terms for ON RRP, Treasury issuance, and bank regulation make the financing respond to the state of the economy. The VAR includes the two-year yield and IOR less the one-month bill

Table 2: Restrictions on the two balance-sheet shocks.

Response variable	RFBS	OFBS
Panel A: sign and zero restrictions		
Two-year yield		
SOMA	+	+
log(Reserves + ON RRP)	+	+
ON RRP/(Reserves + ON RRP)	−	+
GDP	Z	Z
GDPDEF	Z	Z
Term spread	−	−
EBP		
IOR − 1M bill		
Panel B: narrative restrictions (window W, regime average)		
Window W	2020:M4–2021:M2	2023:M9–2024:M7
Mean shock realization, eq. (B.1)	+	−
|HD| on log(R + RRP), eq. (B.2)	dominant	dominant

Note: A + or − is a sign restriction on the response of the row variable to the column shock, binding from impact through horizon five $[0, 5]$. Z is a zero at impact. Blank cells are unrestricted. Panel B ties each balance-sheet shock to its window. Over the window the mean realization of RFBS must be positive and the mean realization of OFBS must be negative, and the shock's historical-decomposition contribution to log(Reserves + ON RRP) must exceed the sum of the contributions of all other labeled shocks.

rate, and it labels an MP shock. RFBS and OFBS are the movements in the balance sheet and its financing that the systematic relations to the current and past values of the variables do not explain.

One economic assumption generates the zero restrictions in Panel A of Table 2.

Assumption 1 (Balance-sheet shocks do not reach real activity and prices within the month)**.** *The two balance-sheet shocks have no impact effect on real GDP or the GDP deflator.*

Assumption 1 follows the timing of Leeper et al. (1996), whose sluggish private sector responds to financial variables only with a delay. Production and final-goods pricing involve planning processes, and movements in interest rates and in the balance sheet reach output and prices smoothly over time rather than within the month. Christiano et al. (2005) impose the same timing on the MP shock. Assumption 1 applies it to the two balance-sheet shocks instead, and the MP shock stays sign-restricted at impact. The balance sheet changes gradually after each program announcement

(Nakashima et al., 2024), and its real effects plausibly arrive after the quantities accumulate. The timing argument binds more tightly for the quantity channel than for the rate channel.

Beyond impact, the GDP and deflator responses to both balance-sheet shocks stay unrestricted. The zeros bind on the average transmission, the response evaluated at the regime average of the relative variances. That average is the object I report. The regime-specific impact responses stay unrestricted (Section 4). The macroeconomic transmission is data-determined and can differ across the two channels.

I impose two narrative restrictions, one for each balance-sheet shock, and Panel B of Table 2 summarizes them. Each adapts the single-date narrative sign restrictions of Antolín-Díaz and Rubio-Ramírez (2018) to a window average and ties the shock to the episode that generated it through two requirements. The mean realization of the shock over the window must carry the sign of its episode. The shock's contribution to the unexpected movement of log(Reserves + ON RRP) over the window must dominate the contributions of the other labeled shocks.[18] The baseline measures the contribution by the multi-horizon interval historical decomposition of Antolín-Díaz and Rubio-Ramírez (2018). Appendix B states the two requirements formally.

I tie the RFBS shock to the first phase of the COVID purchases, April 2020 through February 2021, when the cash from the purchases stayed mostly in reserves. The window carries a positive sign. From March 2021 the cash moved increasingly through MMFs into the ON RRP facility (Section 2). Including those months would risk misallocating the TGA-driven movement in log(Reserves + ON RRP) to the RFBS shock.

I tie the OFBS shock to the sharp ON RRP drain of September 2023 through July 2024. The window contains eleven months, the same length as the RFBS window. A positive OFBS shock is an expansion, the QT2 drain corresponds to negative OFBS realizations, and the window carries a negative sign. Over that window the ON RRP facility ran down steadily while reserves stayed roughly flat. This window starts more than a year after QT2 began in mid-2022, as Section 2

[18] I impose the dominance on log(Reserves + ON RRP) rather than on SOMA for two reasons. Section 2 defines the two episodes by their effects on the liability side, and log(Reserves + ON RRP) carries no zero restriction. The requirement discriminates the balance-sheet shock against every labeled shock, including the demand, supply, and MP shocks. On SOMA the zero restrictions would leave only the other balance-sheet shock as a competitor.

documents.[19] Tying the restrictions to these phases, rather than to full episodes, keeps each inside the window where its channel was the active margin. Averaging over windows rather than single dates makes the restriction robust to the timing noise of individual months.

The demand, supply, MP, and EBP shocks. Table 3 collects the restrictions on the four remaining labeled shocks. I label them because the dominance requirement compares each balance-sheet shock against the other labeled shocks. The labels put demand, supply, MP, and the EBP shock among the competitors it must dominate over its window. Without the labels the comparison would run against the other balance-sheet shock alone. Demand, supply, and a conventional MP shock follow standard conventions. A demand shock raises the two-year yield, GDP, and the GDP deflator. A supply shock raises GDP and lowers the deflator. An MP shock raises the two-year yield and lowers GDP and the deflator. The EBP shock raises the EBP. Two economic assumptions generate the zero restrictions on these shocks.

Assumption 2 (Balance-sheet policy drives SOMA within the month)**.** *Within the month the Fed moves its SOMA holdings only through its balance-sheet policy. The demand, supply, MP, and EBP shocks have no impact effect on SOMA.*

Assumption 2 sets four zeros on SOMA, and the reasons differ by shock. For the demand and supply shocks the zero follows the policy-equation exclusions of Leeper et al. (1996). In their system the policy instrument does not respond within the month to output and final-goods prices, the counterparts of GDP and the GDP deflator here. They give two grounds. The data for those series arrive only after the month ends, and a shift in the policy stance requires consultation and consensus building. The preannounced schedules and caps of asset purchases and runoff strengthen

[19] Part of the ON RRP decline over the window coincides with the post-debt-ceiling rebuild of the Treasury General Account (TGA). The Treasury refilled the TGA through heavy bill issuance from June through October 2023, and the rebuild drew down ON RRP rather than reserves. This raises the question of whether to attribute the drain to QT2 or to the fiscal rebuild. The two are far apart in magnitude. From September 2023 onward the monthly decline in ON RRP is roughly three times the contemporaneous rise in the TGA. The rebuild does not account for the bulk of the drain, and among the six labeled shocks OFBS is the dominant driver of log(Reserves + ON RRP) over the window. I read the restriction as identifying the dominant balance-sheet shock, while noting that the bill issuance associated with the rebuild plausibly amplified the same channel.

Table 3: Restrictions on the demand, supply, MP, and EBP shocks.

Response variable	Demand	Supply	MP	EBP
Two-year yield	+		+	
SOMA	Z	Z	Z	Z
log(Reserves + ON RRP)				
ON RRP/(Reserves + ON RRP)				
GDP	+	+	−	Z
GDPDEF	+	−	−	Z
Term spread				
EBP				+
IOR − 1M bill				

Note: A + or − is a sign restriction on the response of the row variable to the column shock, binding from impact through horizon five $[0, 5]$. Z is a zero at impact. Blank cells are unrestricted. The regime covariances identify (A_0, Λ) and do not label the shocks, and the three unlabeled shocks carry no restriction.

both grounds, and the within-month path of the balance sheet does not react to macroeconomic news.

For the EBP shock only the need for consultation and consensus building applies, since the corporate bond prices behind the EBP arrive within the month. The preannounced schedules and caps of purchases and runoff keep SOMA from responding to it within the month. The restriction constrains the impact responses of SOMA only, and GDP and the GDP deflator carry zeros under Assumptions 1 and 3 instead.

For the MP shock the zero reflects the separation of the two instruments. Since October 2008 the Fed has implemented the policy rate through administered rates rather than through open market operations. It has paid interest on reserves since October 2008 and has offered the ON RRP rate since September 2013. A rate surprise requires no change in asset holdings within the month. An MP shock that moved SOMA contemporaneously would also conflate the two instruments. The sample begins in November 2008 and lies entirely within this operating framework.

Leeper et al. (1996) find that movements in reserves and M1 in their sample arise mostly from policy accommodating shifts in private demand. An identification that treats reserves or M1 as a policy instrument mislabels those demand shifts as policy shocks. They call such identifications

unreliable. The critique applies to any quantity that private demand can move. SOMA is not such a quantity. Only the Fed trades in its own portfolio, and under administered rates the accommodation of demand runs through the liability side rather than through asset holdings.

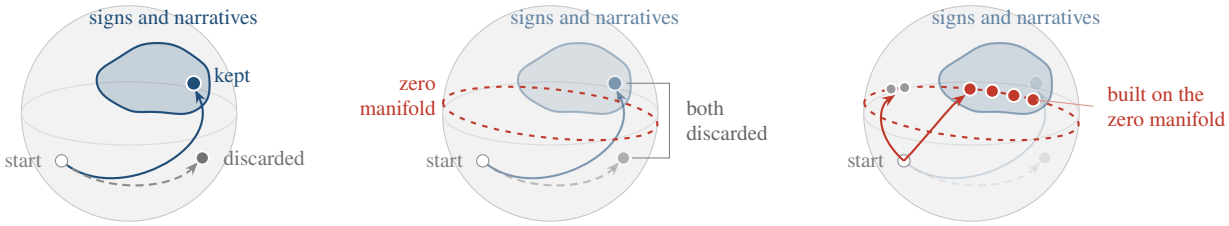
Assumption 3 (Shocks to the EBP do not reach real activity and prices within the month)**.** *The EBP shock has no impact effect on real GDP or the GDP deflator.*

Assumption 3 is the identifying assumption of Gilchrist and Zakrajšek (2012), in whose quarterly VAR shocks to the EBP affect economic activity and inflation with a lag. Over a month the same assumption restricts less.

4 HARS-Z

This section presents the sampler that draws from the posterior of the heteroskedastic model of Section 3 conditional on the sign, zero, and narrative restrictions of Section 3.3. HARS-Z is the heteroskedastic absorb-and-rotate sampler with zero restrictions. I develop it by extending HARS, the heteroskedastic absorb-and-rotate sampler of Kim and Zha (2026, hereafter KZ) that combines shock heteroskedasticity with sign and narrative restrictions, to include zero restrictions. I adopt the notation of KZ throughout. Appendix C contains the full theoretical analysis. Readers focused on the estimates can note Algorithm 1 and continue to Section 5.

Throughout, $\theta = (A_0, \Lambda, A_+)$ collects the structural objects with $\Lambda = \{\Lambda_k\}_{k=1}^{h}$, $Q \in O(n)$ denotes the rotation, and $SR(\theta, Q)$ is the indicator that the sign and narrative restrictions hold. HARS (KZ, Algorithm 1) combines likelihood-preserving posterior draws of the model parameters with rotations drawn from a well-defined probability distribution restricted by the sign and narrative restrictions, adopting a column-wise elliptical slice sampler (ESS; Murray et al., 2010). Sign and narrative restrictions are inequalities. They truncate the conditional distribution of Q to a subset of positive Haar measure, and the posterior target in HARS (Appendix C.1) is well defined. A random rotation lands in the subset with positive probability (Figure 5, panel a). Appendix C.1 reviews the sampler and the repair steps in detail.



(a) HARS under signs and narratives **(b)** HARS with a zero restriction **(c)** HARS-Z

Figure 5: Why zero restrictions break HARS, and what HARS-Z does instead.

Note. Each sphere stands for $O(n)$. The shaded region is the set of rotations that satisfy the sign and narrative restrictions, and the dashed curve is the zero manifold $Q(\theta)$. In panel (a) HARS draws rotations from the starting point and keeps the ones that land in the region. In panel (b) a zero restriction adds the curve, a set of measure zero. A random rotation reaches it with probability zero, and every draw is discarded. In panel (c) HARS-Z builds each rotation on the curve and the region selects among them.

A zero restriction is not a truncation. A zero restriction on the impulse responses is a linear equality constraint on the j-th column q_j of Q. Following the notation of Arias et al. (2018, hereafter ARRW), the $z_j \times n$ matrix $Z_j(\theta)$ stacks the z_j zero restrictions on shock j as rows. The restrictions on shock j read $Z_j(\theta)\, q_j = 0$, the rows depend on the draw through A_0 and the moving-average coefficients, and the sampler must rebuild them at every iteration. The zeros bind under the average transmission only and need not hold in the regime-specific responses, where $\Lambda_k^{1/2}$ reweights the constrained column. Define the zero manifold

$$Q(\theta) \;=\; \left\{ Q \in O(n) : Z_j(\theta)\, q_j = 0 \text{ for every restricted } j \right\}. \tag{5}$$

Whenever some $z_j \geq 1$, $Q(\theta)$ is a lower-dimensional smooth manifold in $O(n)$ and $\mu(Q(\theta)) = 0$, where μ is Haar measure on $O(n)$, as ARRW note in motivating their Algorithm 2 (Figure 5, panel b).

One may be tempted to impose the zeros by adding them to $SR(\theta, Q)$ and running HARS unchanged. The zeros lower the dimension of the admissible set $R(\theta) = \{Q \in O(n) : SR(\theta, Q) > 0\}$ instead of truncating it, and the condition $\mu(R(\theta)) > 0$ fails at every θ with at least one zero restriction. The restricted posterior target is not defined, the Gaussian coordinatization never reaches the zeros, and every step of KZ's Algorithm 1 fails with probability one, including its initialization. The

failure originates in the measure rather than in the algorithm. Appendix C.2 shows it formally. The sampler must parameterize the zero-restricted set directly rather than reach it by sampling. It must build the rotation on $Q(\theta)$, and its slice-sampling updates must never leave it (Figure 5, panel c). The rest of this section constructs exactly such a parameterization.

HARS-Z replaces the QR map γ with a map $\gamma_z(\cdot\,;\theta)$, built on the null-space construction of ARRW's Algorithm 2, that sends a Gaussian matrix to a rotation on $Q(\theta)$ by construction. These are the built rotations of Figure 5, panel c. The state of the chain is the pair (θ, X), where X is an $n \times n$ Gaussian matrix, the coordinates in which HARS runs its column-wise ESS update. The sampler recomputes the rotation $Q = \gamma_z(X;\theta)$ whenever it needs one. A structural update moves the manifold $Q(\theta)$, the recomputed rotation lies on the new manifold, and the zeros hold after every move. Appendix C.3 builds the map γ_z and establishes its properties.

HARS-Z targets a posterior on the zero-restricted set. Let $\mu(\cdot \mid Q(\theta))$ denote the uniform measure on $Q(\theta)$ and let $R(\theta) = \{Q \in Q(\theta) : SR(\theta, Q) > 0\}$ denote the admissible set on it. The target is

$$\pi_v(d\theta, dQ \mid y, SR) \;\propto\; \pi(\theta \mid y)\, \mathbf{1}\{Q \in R(\theta)\}\, \mu\big(dQ \mid Q(\theta)\big)\, d\theta. \tag{6}$$

Given θ, the rotation is uniform on $R(\theta)$. The marginal of θ is proportional to $\pi(\theta \mid y)\, \kappa_u(\theta)$, where $\kappa_u(\theta) = \mu(R(\theta) \mid Q(\theta))$ is the uniform probability of the admissible set. The restrictions weight each draw of θ by the share of $Q(\theta)$ they admit.

The target (6) is the analogue on the zero-restricted set of the posterior of ARRW's Algorithm 1. Their Algorithm 1 draws from the posterior over the orthogonal reduced-form parameterization conditional on the sign restrictions. It keeps the draws of the reduced form and a Haar rotation that satisfy the restrictions. Each reduced-form draw survives in proportion to the Haar probability of its admissible set. In (6) the uniform probability $\kappa_u(\theta)$ plays that role. KZ's Definition 1 defines the same kind of target on $O(n)$, with the marginal of θ proportional to $\pi(\theta \mid y)\, \mu(R(\theta))$. Equation (6) carries it to $Q(\theta)$, with the uniform measure on $Q(\theta)$ in place of Haar measure. The zero-restricted set has Haar measure zero, and a conditional distribution on it depends on the parameterization. I condition with the uniform measure on $Q(\theta)$. ARRW's importance sampler (their Algorithm 3)

conditions in another parameterization, and its bands need not match mine.

Generating the rotation through γ_z induces the null-space construction measure on $\mathcal{Q}(\theta)$, not the uniform measure. The chain draws θ in proportion to $\pi(\theta \mid y)\,\kappa(\theta)$, where $\kappa(\theta)$ is the Gaussian probability of the admissible set under the construction. One weight corrects both differences. Each retained draw receives $v(Q;\theta)/\mathrm{E}_z[v \mid \theta]$, where $v(Q;\theta)$ is the volume element of the construction and $\mathrm{E}_z[v \mid \theta]$ is its mean over unrestricted construction draws at θ.[20] The ratio is the density of the uniform measure with respect to the construction measure on all of $\mathcal{Q}(\theta)$. It corrects the conditional of Q given θ and turns $\kappa(\theta)$ into $\kappa_u(\theta)$ at once. The column-wise ESS update remains valid on $\mathcal{Q}(\theta)$ and leaves the conditional target invariant. Appendix C.4 establishes these results.

The HARS-Z sampler, stated below as Algorithm 1, cycles through three blocks. Following KZ's Theorem 1, partition the parameter collection as $z_0 = (A_0, \Lambda)$ and $z_+ = A_+$, and let s_0 collect everything held fixed in the structural update, the latent Student-t scales included. The structural block draws (z_0, z_+) at fixed X from the target of KZ's Theorem 1, the density $p_0(z_0 \mid s_0)\,p_+(z_+ \mid z_0, s_0)$ restricted to the completed draws that pass $SR\big(\theta, \gamma_z(X;\theta)\big)$. Here p_0 is the collapsed density with A_+ integrated out and p_+ is the Gibbs conditional of A_+. The latent scales move by their Gibbs conditional, as in the sampler of Brunnermeier et al. (2021, hereafter BPSS). The rotation block moves X by the column-wise ESS update of KZ.

HARS-Z uses KZ's rotation construction and ESS and updates the structural block by multiple-try Metropolis (MTM; Liu et al., 2000). A single random-walk proposal of z_0, completed by a draw of A_+, rarely passes the sign and narrative restrictions at the current X. MTM draws K completed candidates in one step and weights each by its collapsed density times the admissibility indicator. It accepts one candidate with a probability that keeps the structural target invariant. With $K = 1$ the step is the Metropolis–Hastings (MH) step of BPSS with the admissibility indicator in the acceptance probability and with A_+ drawn as part of the proposal.

Algorithm 1. (**Heteroskedastic absorb-and-rotate sampler with zero restrictions (HARS-Z).**)

[20] KZ write $v(\theta)$ for the Haar volume $\mu(R(\theta))$ of the admissible set. In this paper $v(Q;\theta)$ is the volume element of ARRW's Theorem 3, a different object. The probability $\kappa_u(\theta)$ of the admissible set under the uniform measure on $\mathcal{Q}(\theta)$ plays the role of KZ's $v(\theta)$ on the zero manifold.

Let the current admissible state be (θ, X) with $Q = \gamma_z(X;\theta)$ and $SR(\theta, Q) > 0$. Fix the number of tries K, the random-walk covariance Σ_{rw}, the number m of ESS updates before the structural update, and the number N_{ess} of retained rotations. For the initialization, construct an admissible rotation by a numerical search at the initial θ and convert it into Gaussian coordinates by Lemma 1(iii). One iteration proceeds as follows.

1. **Rotation update.** Run m column-wise ESS updates of X at fixed θ (Proposition 3).

2. **Structural update by MTM.** Write $w(\theta') = p_0(z_0' \mid s_0)\,\mathbf{1}\{SR(\theta', \gamma_z(X;\theta')) > 0\}$ for the weight of a completed draw $\theta' = (z_0', z_+')$.

   (a) **Candidates.** For $k = 1, \ldots, K$, draw $z_0^{(k)} = z_0 + e_k$ with $e_k \sim N(0, \Sigma_{\mathrm{rw}})$ and $z_+^{(k)} \sim p_+(A_+ \mid z_0^{(k)}, s_0)$, and form $\theta^{(k)} = (z_0^{(k)}, z_+^{(k)})$. Each weight uses the rebuilt rotation $\gamma_z(X;\theta^{(k)})$, which satisfies every zero restriction by construction.

   (b) **Selection.** If every candidate has weight zero, keep θ and go to Step 3. Otherwise select one candidate θ^c with probability proportional to its weight.

   (c) **Reference points.** Draw $K-1$ reference points from θ^c as in Step 2(a), and take the current θ as the K-th.

   (d) **Acceptance.** Set $\theta \leftarrow \theta^c$ with probability $\min\{1, \sum_k w(\theta^{(k)}) / \sum_k w(\tilde{\theta}^{(k)})\}$, where $\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(K)}$ are the reference points. Otherwise keep θ.

3. **Latent scales.** Draw the latent Student-t scales from their Gibbs conditional given the structural residuals of the current θ, with the A_+ that the state carries.

4. **Rotation draw.** Run N_{ess} column-wise ESS updates of X at fixed θ, producing Gaussian states $X^{(1)}, \ldots, X^{(N_{\mathrm{ess}})}$ and rotations $Q^{(r)} = \gamma_z(X^{(r)};\theta)$ with $SR(\theta, Q^{(r)}) > 0$ for every r. Retain all N_{ess} rotations together with θ.

5. **Advance the state.** Set the next state to $(\theta, X^{(N_{\mathrm{ess}})})$ and return to Step 1.

Step 2 never redraws A_+ alone to rescue a candidate that fails the restrictions, the shortcut that KZ's Theorem 1 rules out. A rejected step keeps z_0 and A_+ together. Step 3 draws the latent scales from the residuals of that pair, as BPSS do with the pair their sampler holds. The ESS updates evaluate only the admissibility indicator, not the likelihood. In the estimates $K = 300$, $m = 5$, and $N_{\text{ess}} = 1$, and Appendix D reports the remaining settings. Posterior inference reports medians and bands weighted by $v(Q;\theta)/\mathrm{E}_z[v \mid \theta]$. Appendix C.5 states the estimator, and Appendix C.4 establishes the validity of Algorithm 1. With an empty zero set, γ_z reduces to the QR map γ of KZ, $\mathcal{Q}(\theta) = \mathcal{O}(n)$, and every weight equals one.

5 Results

I estimate the baseline posterior with HARS-Z. Appendix D reports the sampler settings and the diagnostics.

5.1 Baseline Impulse Responses

Figures 6 and 7 report the impulse responses to the two balance-sheet shocks under the baseline identification of Section 3.3, with the relative variances set to their average of one across regimes.[21] The sign and narrative restrictions of Section 3.3 define each shock in its expansionary direction, and the identification is unchanged. Each figure shows one shock in its contractionary direction, the direction of the QT episodes. To make the two shocks directly comparable, I normalize each shock to a one-percent impact decline in SOMA holdings. The decline continues after impact. Figure E.1 in Appendix E reports the responses to the demand, supply, MP, and EBP shocks.

Figure 6 plots the responses to the RFBS shock. The shaded regions are the 68% credible bands. The restricted responses hold by construction. Panel A of Table 2 restricts the expansionary direction. Under the contractionary display, SOMA, log(Reserves + ON RRP), the ON RRP ratio, and the term spread move with the signs of Panel A reversed. The zero restrictions hold GDP and the

[21] The figures report one chain of 10,000 retained draws. Three further chains of the same specification give the same signs.

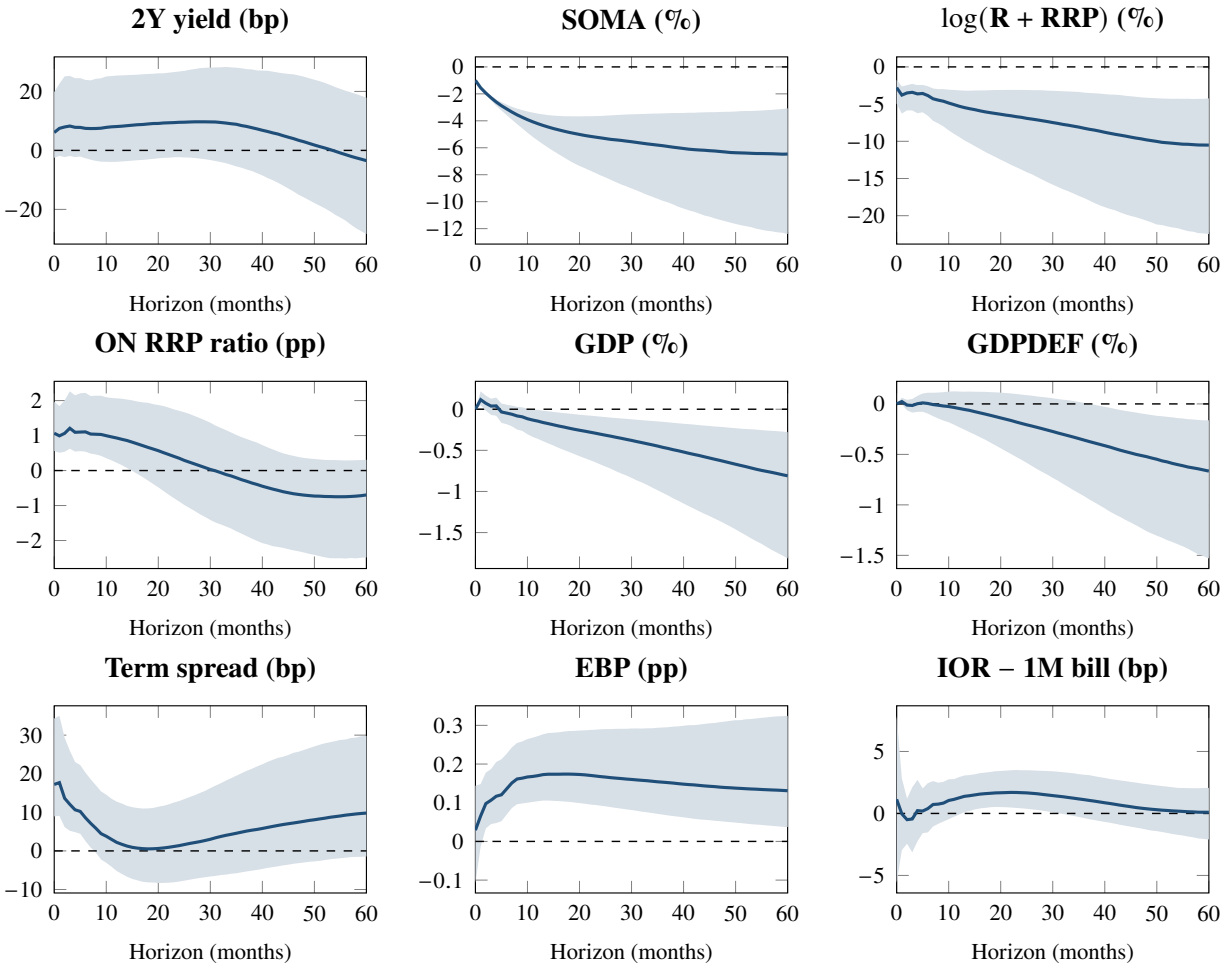


Figure 6: Impulse responses to the RFBS shock.

Note. Impulse responses to the RFBS shock, normalized to a one-percent impact decline in SOMA holdings. I evaluate the responses at the regime average of the relative variances. Solid lines are posterior medians and shaded regions are 68% credible bands over the retained draws. Vertical axes are in basis points for the two-year yield, the term spread, and IOR less the one-month bill rate. They are in percent for SOMA, log(Reserves + ON RRP), GDP, and GDPDEF, and in percentage points for the ON RRP ratio and the EBP.

GDP deflator at zero on impact. The economic content lies in the GDP and GDP-deflator responses beyond impact and in the responses of the EBP and of IOR less the bill rate. No restriction binds on any of them.

A negative RFBS shock lowers GDP persistently. The GDP response turns negative at five months and falls through the horizon, and the deflator response is negative from seven months on. The 68% bands exclude zero from eleven months onward for GDP and from thirty-eight months onward for the deflator. The same shock raises the EBP. The response rises over the first seventeen months and stays positive through the horizon, and the 68% bands exclude zero from two months

onward. The shock also raises IOR relative to the bill rate, and those bands exclude zero from thirteen to thirty-one months.

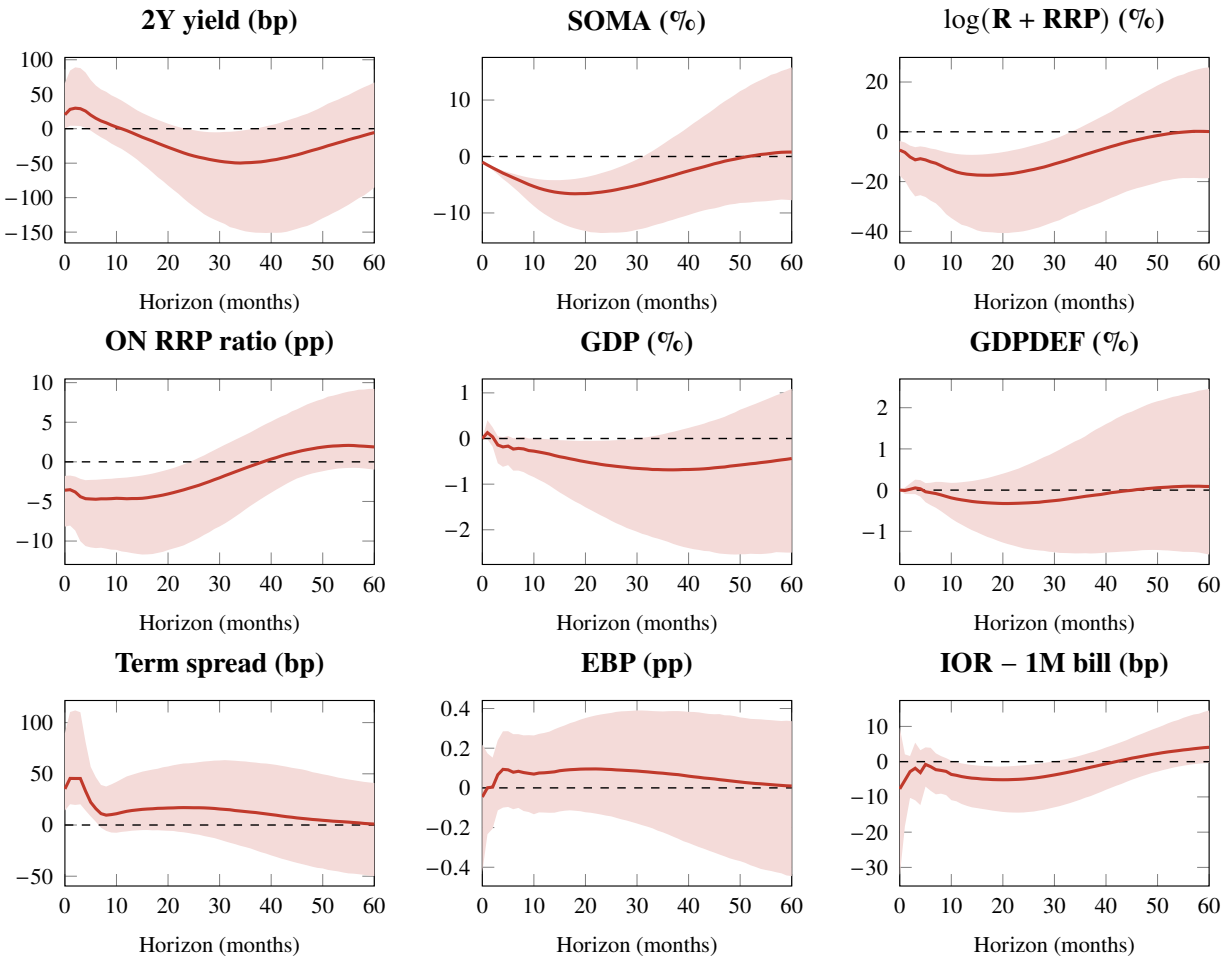


Figure 7: Impulse responses to the OFBS shock.

Note. Impulse responses to the OFBS shock, normalized to a one-percent impact decline in SOMA holdings. The figure shows the shock in the direction of the QT2 drain. I evaluate the responses at the regime average of the relative variances. Solid lines are posterior medians and shaded regions are 68% credible bands over the retained draws. Vertical axes are in basis points for the two-year yield, the term spread, and IOR less the one-month bill rate. They are in percent for SOMA, log(Reserves + ON RRP), GDP, and GDPDEF, and in percentage points for the ON RRP ratio and the EBP.

Figure 7 plots the responses to the OFBS shock. The restricted responses again hold by construction and reverse the signs of the OFBS column in Panel A of Table 2. The economic content again lies in the GDP and GDP-deflator responses beyond impact and in the responses of the EBP and of IOR less the bill rate. No restriction binds on any of them.

A negative OFBS shock does not move the deflator or the EBP. Their 68% bands include zero at every horizon. The median response of GDP falls, and its 68% band excludes zero from eleven to

twenty-nine months with the upper edge close to zero. This decline is less certain and shorter-lived than the one after the RFBS shock. The posterior probability that GDP falls after the OFBS shock never exceeds 0.87 and drops to 0.64 at five years. After the RFBS shock it exceeds 0.9 from seventeen months on (Table F.1 in Appendix F). The shock lowers IOR relative to the bill rate, with the bands excluding zero from eleven to twenty-nine months.

Interpreting the two shocks. Neither figure describes a single episode. The linear model has one response function for each shock, and realizations of both signs enter its estimation. QE and QT1 enter the RFBS shock as positive and negative realizations, and the ON RRP buildup and the QT2 drain enter the OFBS shock the same way (Section 3.3). Realizations of both signs also fall outside those episodes. The contraction display is a sign convention. I read each plotted response as the average over both directions and not as the response to the QT episode alone.

5.2 Testing the Two Responses

5.2.1 The Reserve-Financed Contraction and the EBP

Section 2.3 links a reserve-financed operation to GDP and the price level through the risk premium in the cost of credit for firms, which the EBP measures. If the declines in GDP and the deflator after a reserve-financed contraction come from the rise in the EBP, they should disappear when the EBP does not rise. I hold the EBP response to the RFBS shock at zero in every month from 0 to 60 and compare the GDP and deflator responses with and without it. Following BPSS, who keep the federal funds rate constant with a sequence of monetary policy shocks, I use a sequence of EBP shocks to keep it at zero.[22]

With the EBP response held at zero, the declines in GDP and the deflator after a reserve-financed

[22] In 594 of the 10,000 draws the required shocks grow without bound. I drop those draws and renormalize the weights of the other 9,406. In each posterior draw, let $r_x(h)$ be the response of variable x to the RFBS shock at month h. Let $b_x(h)$ be its response to an EBP shock of one standard deviation. Both responses use the average relative variances. Adding EBP shocks $\zeta(0), \ldots, \zeta(60)$ to the RFBS shock gives the counterfactual response $c_x(h) = r_x(h) + \sum_{j=0}^{h} b_x(h-j)\,\zeta(j)$. I choose $\zeta(h)$ to set $c_{\text{EBP}}(h) = 0$, one month at a time from $h = 0$. The penalized offset of Figure G.2 instead chooses $\zeta(0), \ldots, \zeta(60)$ to minimize $\sum_{h=0}^{60} c_{\text{EBP}}(h)^2 + \rho \sum_{h=0}^{60} \zeta(h)^2$, with $\rho = 10^{-3} \max_h r_{\text{EBP}}(h)^2$, and keeps every draw.

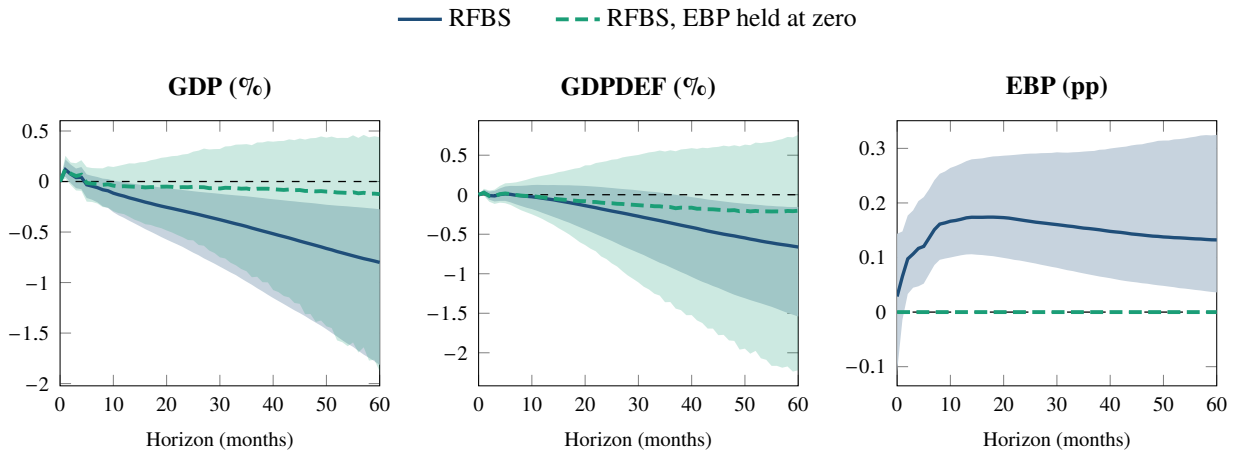


Figure 8: Responses to the RFBS shock with the EBP held at zero.
Note. Responses to the RFBS shock, normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid blue lines and blue regions show the RFBS shock alone. Dashed green lines and green regions show the RFBS shock combined with the EBP shocks that keep the EBP response at zero in every month. Lines are posterior medians and shaded regions are 68% credible bands over the 9,406 draws in which those shocks stay finite. Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

contraction are no longer present (Figure 8). Neither band lies below zero at any horizon. A penalized offset that keeps all 10,000 draws gives the same result (Figure G.2 in Appendix G). Figure G.1 reports the other six variables.

The counterfactual attributes the effects of a reserve-financed contraction on GDP and the price level to the rise in the EBP. In the estimates of Gilchrist and Zakrajšek (2012), a rise in the EBP lowers economic activity and, over time, inflation. Section 2.3 offers two channels for that rise. A reserve-financed contraction would lower bank income and with it bank capital (link 3 of Figure 3), and it would make banks less liquid (link 4). Either channel would raise the risk premium banks require to bear credit risk. The estimates are consistent with both channels and do not separate them.

5.2.2 The ON RRP-Financed Contraction and Bill Yields

The realized financing includes the part that the portfolio choices of MMFs produce (Section 2.3). MMFs move their cash to whichever of ON RRP and bills pays more. Those choices move the ON RRP ratio, the one variable on which the sign restrictions of the two balance-sheet shocks differ (Table 2). When bills pay more than ON RRP, MMFs move cash out of ON RRP and the ratio falls, as under a reserve-financed expansion. When bills pay less, MMFs move cash into ON RRP and the

ratio rises, as under an ON RRP-financed expansion. If a movement in bill rates coincides with an operation, the estimated realizations of either balance-sheet shock can include it. Without a control for bill rates, the responses of GDP and the price level to that shock would mix the effect of the operation with the effect of the movement in bill rates. The two-year yield carries the expected path of the policy rate, which also moves bill yields.

The baseline carries IOR less the one-month bill rate for this reason (Section 3.1). Figure 9 compares it with the eight-variable system without that variable, under the same restrictions.

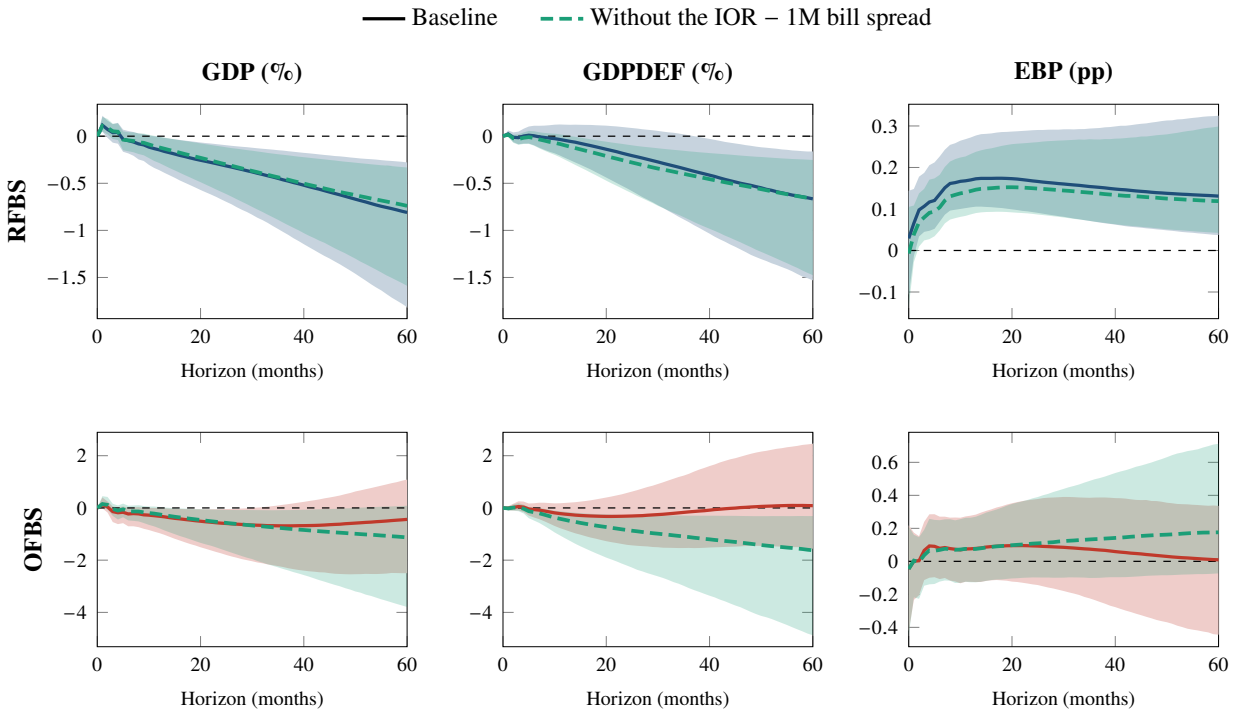


Figure 9: Responses to the two balance-sheet shocks without IOR less the one-month bill rate.
Note. Rows report the RFBS and OFBS shocks, each normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid lines and shaded regions in the color of each shock are the baseline posterior medians and 68% credible bands of Figures 6 and 7. Dashed green lines and green regions are the posterior medians and 68% credible bands of the eight-variable system without IOR less the one-month bill rate, under the same restrictions. Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

Without IOR less the bill rate, the ON RRP-financed contraction lowers GDP and the deflator (Figure 9). The GDP bands lie below zero from fifteen to forty-nine months and the deflator bands from seven months onward. The EBP bands include zero at every horizon. The reserve-financed contraction still raises the EBP and lowers GDP and the deflator. Its decline in the deflator appears from fourteen months onward, earlier than in the baseline, and its decline in GDP from twelve

months, one month later.

I do not read the declines in GDP and the deflator in the eight-variable system as an effect of a balance-sheet operation financed by ON RRP. In that system the estimated realizations of the OFBS shock can include movements in the ON RRP ratio that MMFs make when bill yields change relative to the Fed's administered rates. With IOR less the bill rate in the system, the model attributes those movements to other shocks. The decline in the deflator disappears, and the decline in GDP becomes less certain and shorter-lived.

5.3 Robustness

I examine whether the responses to both balance-sheet shocks survive further changes to the system.

5.3.1 The TGA and the Supply of Bills

I also add the log of the Treasury General Account (TGA) to the baseline as a tenth variable, with no restriction on it.[23] The TGA controls for the Treasury's cash at the Fed. When that cash rises, it drains reserves and ON RRP without any change in SOMA. With the TGA, the reserve-financed contraction still lowers GDP and the deflator and raises the EBP. The ON RRP-financed contraction again moves neither the EBP nor the deflator, and its effect on GDP is again less certain and shorter-lived (Figure H.1 in Appendix H).

Where the baseline controls for bill rates through IOR less the bill rate (Section 5.2.2), a bill supply shock labels one source of their movements directly. Heavy bill issuance drew down ON RRP over the drain (Section 2), and a shock to the supply of bills could move the ON RRP ratio without any balance-sheet operation. I label a Treasury bill supply shock in the ten-variable system with the TGA. An increase in bill supply raises the TGA and lowers both log(Reserves + ON RRP) and IOR relative to the bill rate from impact through horizon five. It has no impact effect on GDP or the deflator. The bill supply shock enters the dominance requirement of the narrative restrictions on

[23] Figure H.1 reports one chain of 10,000 draws. The ten-variable system uses the degrees of freedom that the same two-step procedure fits to its own residuals.

weaker terms. Each balance-sheet shock must dominate the sum of the other labeled shocks without the bill supply shock, and it must exceed the bill supply shock on its own.[24] With the bill supply shock, the reserve-financed contraction still lowers GDP and the deflator and raises the EBP. The ON RRP-financed contraction again moves neither the EBP nor the deflator, and its effect on GDP is again less certain and shorter-lived (Figure H.2 in Appendix H).

5.3.2 Changes to the Specification

I estimate five specifications. Each changes one element of the baseline and keeps all other elements and restrictions the same. The first drops every zero restriction, those of Assumptions 1, 2, and 3. The second puts industrial production and the CPI in place of GDP and the deflator. The third sets the regime breaks at the exits from and the return to the zero lower bound (2015:M11, 2020:M3, and 2022:M2) in place of the balance-sheet episodes. The fourth uses one variance regime (homoskedasticity) in place of four variance regimes (heteroskedasticity). The fifth uses ten lags in place of six. Appendix H plots those responses against the baseline.

In every specification a reserve-financed contraction raises the EBP and lowers output. It lowers the price level in three of them, all but the one without zero restrictions and the one with the CPI. The ON RRP-financed contraction moves none of output, the price level, and the EBP in four of them. With ten lags it raises the EBP and lowers output.

Without zero restrictions, a reserve-financed contraction raises GDP in the first four months and the deflator in the first eight (Figure H.3).[25] I read that rise as the state of the economy around the balance-sheet actions, which the impact zeros of Assumption 1 separate from transmission. The contraction still raises the EBP and lowers GDP, from forty months onward. The decline in the deflator does not appear without the zero restrictions. The ON RRP-financed contraction moves none of GDP, the deflator, and the EBP.

With industrial production and the CPI in place of GDP and the deflator, a reserve-financed contraction still raises the EBP and lowers industrial production (Figure H.4). The decline in

[24] Figure H.2 reports one chain of 10,000 draws.
[25] Figure H.3 reports one chain of 10,000 draws.

industrial production appears from fourteen months onward, later than the decline in GDP in the baseline. The CPI band includes zero at every horizon. The ON RRP-financed contraction moves none of industrial production, the CPI, and the EBP.

With the regime breaks at the lower bound, a reserve-financed contraction raises the EBP and lowers GDP and the deflator (Figure H.5). The decline in the deflator appears later than in the baseline, from fifty-three months onward. The ON RRP-financed contraction moves none of GDP, the deflator, and the EBP.

With one variance regime, a reserve-financed contraction raises the EBP and lowers GDP and the deflator (Figure H.6). The declines in GDP and the deflator both appear from thirty-five months onward, later than the baseline in GDP and earlier in the deflator. The ON RRP-financed contraction moves none of GDP, the deflator, and the EBP.

With ten lags, a reserve-financed contraction raises the EBP and lowers GDP and the deflator (Figure H.7). The declines in GDP and the deflator both appear from sixteen months onward, later than the baseline in GDP and earlier in the deflator. The ON RRP-financed contraction also raises the EBP, from three to thirty-three months, and lowers GDP from twelve months onward and the deflator from ten to twenty-two months. It no longer lowers IOR relative to the bill rate.[26]

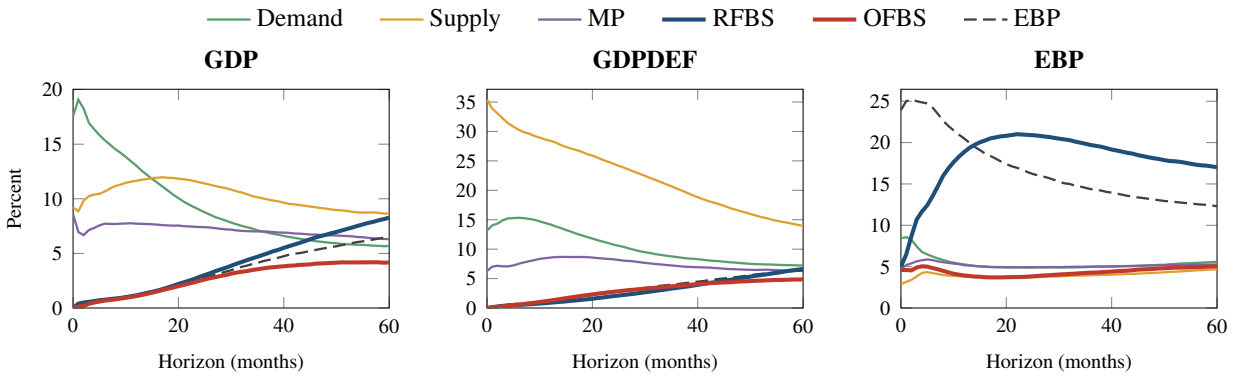
5.4 Variance and Historical Decompositions

The impulse responses show how the effects of a balance-sheet policy of a given size differ with its financing. They do not show how much of the variation in GDP and the deflator the two balance-sheet shocks account for. This subsection reports two decompositions that answer that question. The forecast-error variance decomposition gives the share of the forecast variance at a given horizon. The historical decomposition gives the contribution of each shock to the unexpected change in a variable over an episode. Both use the baseline draws of Section 5.1 and weight the draws as the impulse responses do.

[26] Figure H.7 pools three chains of 10,000 draws each, and each chain gives the same signs. In each chain the structural parameters move in fewer than 400 of the 10,000 sweeps.

Forecast-error variance decomposition. Figure 10 reports the median share of each labeled shock at every horizon from 0 to 60 months. The shares of both balance-sheet shocks in the forecast variance of GDP and the deflator rise with the horizon. Both shares start at zero and stay below 2% over the first year. They rise steadily after it. At five years the RFBS shock accounts for 8% of GDP and 7% of the deflator, and the OFBS shock for 4% of GDP and 5% of the deflator. The RFBS share of the EBP rises until about two years and declines slowly after that. From two months on it is the largest share after that of the EBP shock itself, and from fourteen months on it exceeds that share as well.

Figure 10: Forecast-error variance decomposition.



Note. Posterior median share of each labeled shock in the forecast-error variance of the variable, in percent, over the retained draws. The shares are under the average transmission, with the relative variances at their average of one across regimes. Horizons are in months. Table I.1 reports the shares at 36 months with their 68% intervals.

Historical decomposition. The historical decomposition splits the accumulated forecast error of a variable over a window among the shocks that the sample realizes inside it. Let $W = \{t_0 + 1, \ldots, t_1\}$ and let $\mathcal{I}_{t_0}$ collect the observables through t_0. The contribution of labeled shock j to variable i is

$$\mathrm{HD}_{j\to i}(W) \;=\; \sum_{\ell=0}^{t_1-t_0-1} \left[\Theta_{s_{t_1-\ell},\ell}(Q)\right]_{ji} \tilde{\varepsilon}_{j,t_1-\ell}, \qquad \sum_{j=1}^{n} \mathrm{HD}_{j\to i}(W) \;=\; y_{i,t_1} - \mathbb{E}\left[y_{i,t_1} \mid \mathcal{I}_{t_0}\right], \tag{7}$$

where the regime index follows the date of the shock. Figure 11 reports (7) over the four volatility regimes, as posterior medians with 68% intervals, for the six labeled shocks. Figure J.1 in Appendix J reports the same decomposition for the other six variables.

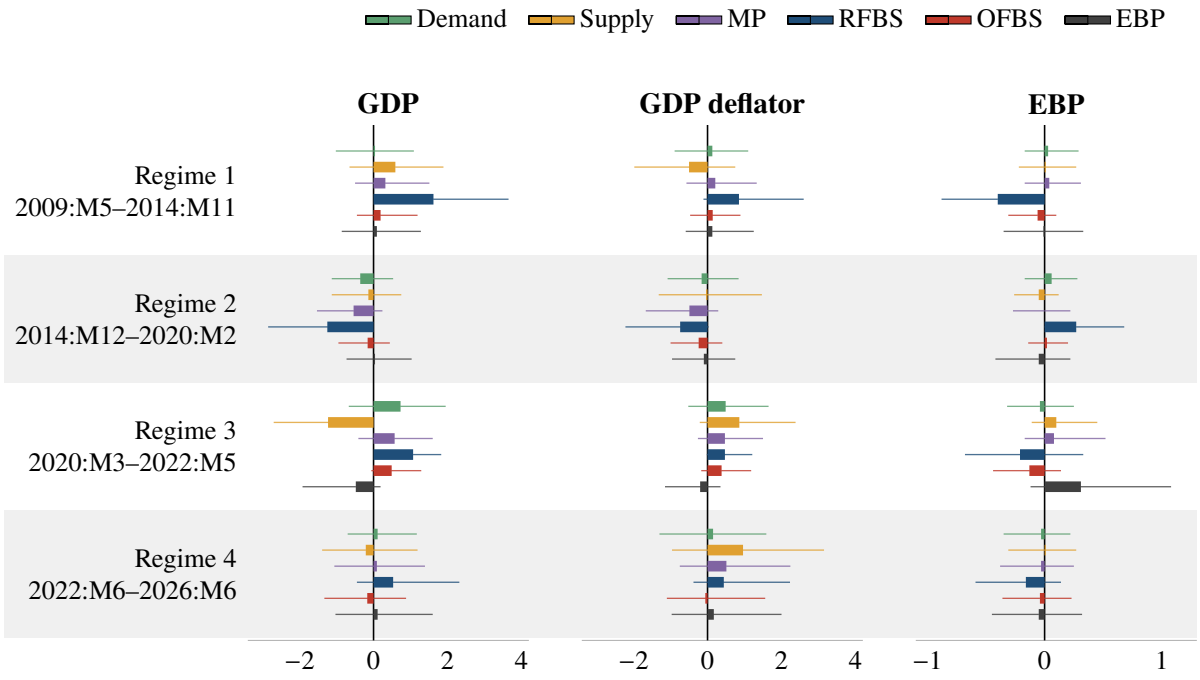


Figure 11: Historical decomposition by volatility regime.

Note. Contribution of each labeled shock to the accumulated forecast error of the variable over the window. Units are percent for GDP and the deflator and percentage points for the EBP. Bars are posterior medians and whiskers are 68% credible intervals, both over the retained draws. The windows are the volatility regimes of Table 1. Regime 1 opens at 2009:M5, the first estimable month under the six estimation lags. The three unlabeled shocks account for the remainder.

Reserves mainly financed the balance sheet over the first two regimes (Figure 1), and the RFBS shock is the largest contributor to the GDP and deflator forecast errors in both. Over Regime 1, the QE years, it pushes GDP and the deflator above their forecasts and the EBP below. Over Regime 2, which holds QT1, it pulls GDP and the deflator below their forecasts and pushes the EBP above. It is also the largest contributor to the EBP forecast error in both regimes. The GDP band excludes zero in both regimes and the EBP band in Regime 1, and the other three bands include it. The OFBS bands include zero in both regimes.

Over Regime 3, the COVID years, the ON RRP balances build up to near $2 trillion (Figure 1) while the COVID purchases add reserves as well, and neither liability mainly financed them. The supply shock is the largest contributor to the GDP and deflator forecast errors. It pulls GDP below its forecast, and its GDP band excludes zero. It pushes the deflator above its forecast. The RFBS shock is the second largest contributor to the GDP forecast error. It pushes GDP and the deflator above their forecasts. Its GDP band excludes zero, and its deflator band includes it. Among the

shocks that push GDP above its forecast, the RFBS shock is the largest. The EBP shock is the largest contributor to the EBP forecast error, and the RFBS shock pulls the EBP below its forecast. The OFBS bands include zero for all three variables.

Over Regime 4, which holds QT2, no labeled shock has a band that excludes zero for GDP or the deflator. The RFBS shock is the largest contributor to the GDP forecast error and the supply shock to the deflator forecast error. The OFBS shock ranks third for GDP and last for the deflator.

Interpreting the decompositions. Reserves mainly financed QE and QT1, and the RFBS shock is the main source of the unexpected movements in GDP and the deflator over both. It moved GDP, the deflator, and the EBP together. GDP and the deflator rose when the EBP fell and fell when it rose, the pattern the counterfactual of Section 5.2.1 attributes to the EBP.

ON RRP mainly financed QT2, and over Regime 4 the OFBS contributions to the GDP and deflator forecast errors have bands that include zero. The responses of Section 5.1 give no reason to read its contributions as an effect of the financing.

The shares in Figure 10 say how much the financing matters on average. The two balance-sheet shocks account for little of the forecast variance of GDP and the deflator in the first year. Their shares grow with the horizon and together reach about 12% for both GDP and the deflator at five years. The RFBS share of the EBP grows faster and reaches 21% at two years, larger than the share of the EBP shock itself. The balance-sheet shocks move the risk premium more than output and prices, and they move it through reserve financing.

6 Conclusion

Since 2008 the Fed has financed its balance-sheet operations with two overnight liabilities. Reserves financed the purchases after the financial crisis and the runoff of QT1. Both liabilities financed the COVID purchases, and ON RRP financed most of the runoff of QT2, which left reserves nearly intact (Figure 1). A reserve-financed operation reaches bank capital and the liquidity of banks. An ON RRP-financed operation reaches the shareholders of MMFs and builds no capital. This paper

asks whether the transmission of balance-sheet policy to real GDP and the price level depends on which liability finances it.

A reserve-financed contraction raises the EBP and lowers GDP and the deflator. Banks with less capital or fewer reserves would require a larger risk premium to lend, and that premium would raise the cost of credit for firms. The rise in the EBP and the decline in output hold in each of the five changes to the specification of Section 5.3.2. The decline in the price level holds in three of them. An ON RRP-financed contraction moves neither the EBP nor the deflator, and its effect on GDP is less certain and shorter-lived. With ten lags in place of six, it raises the EBP and lowers GDP. Without IOR less the bill rate in the system it lowers GDP and the deflator, and I do not read those declines as an effect of the financing. At three years the two balance-sheet shocks account for 8.5% of the forecast variance of GDP and 7% of that of the deflator. They account for about a quarter of the forecast variance of the EBP, mostly through the RFBS shock.

In the sample the RFBS shock is the largest contributor to the GDP and deflator forecast errors over QE and QT1, the episodes that reserves financed. It pushed GDP and the deflator above their forecasts over QE and pulled both below over QT1. During COVID both liabilities financed the purchases at the zero lower bound. The RFBS shock pushed GDP and the deflator above their forecasts, and the OFBS shock contributed little to either. During QT2, which ON RRP mainly financed, the OFBS contributions to the GDP and deflator forecast errors have bands that include zero.

The empirical design combines heteroskedasticity, sign restrictions, narrative restrictions, and zero restrictions. The heteroskedastic likelihood separates the model shocks, and the sign and narrative restrictions label the two balance-sheet shocks. HARS-Z imposes the impact-zero restrictions on that combination, and the rotation satisfies every zero restriction at every draw.

Which liability finances an operation shapes what the operation does to output and the price level. In the recent tightening, the Federal Open Market Committee kept the policy rate as its main tool (Federal Open Market Committee, 2022b) and in 2022 planned to end the runoff with reserves still ample. In 2024 it slowed the runoff to keep them ample (Section 2.2). My results give a rationale for

both. A change that falls on reserves raises the EBP and lowers GDP and the deflator, and holding the EBP at zero removes those declines. A change that drains ON RRP has no clear effect on the economy. The second runoff, which ON RRP mainly financed, did what the Committee intended and shrank the balance sheet without moving the economy. The size of the balance sheet does not summarize what a balance-sheet policy does.

References

Adrian, T., Erceg, C. J., Kolasa, M., Linde, J. L., and Zabczyk, P. (2025). Macroeconomic and fiscal consequences of quantitative easing.

Afonso, G., Cipriani, M., and La Spada, G. (2025). Bank regulation and monetary policy: The role of non-bank financial institutions. Staff Reports 1041, Federal Reserve Bank of New York. First issued December 2022, revised December 2025.

Antolín-Díaz, J. and Rubio-Ramírez, J. F. (2018). Narrative sign restrictions for SVARs. *American Economic Review*, 108(10):2802–29.

Arias, J. E., Rubio-Ramírez, J. F., and Shin, M. (2026). Are fiscal transfers inflationary? Working paper, Federal Reserve Bank of Philadelphia.

Arias, J. E., Rubio-Ramírez, J. F., and Waggoner, D. F. (2018). Inference based on structural vector autoregressions identified with sign and zero restrictions: Theory and applications. *Econometrica*, 86(2):685–720.

Becker, B. and Ivashina, V. (2014). Cyclicality of credit supply: Firm level evidence. *Journal of Monetary Economics*, 62:76–93.

Bianchi, J. and Bigio, S. (2022). Banks, liquidity management, and monetary policy. *Econometrica*, 90(1):391–454.

Board of Governors of the Federal Reserve System (2026). Financial accounts of the united states: Flow of funds, balance sheets, and integrated macroeconomic accounts, second quarter 2026. Statistical Release Z.1, https://www.federalreserve.gov/releases/z1/. Release of September 11, 2026; tables F.521 and S.123.

Brunnermeier, M., Palia, D., Sastry, K. A., and Sims, C. A. (2021). Feedbacks: Financial markets and economic activity. *American Economic Review*, 111(6):1845–79.

Chen, H., Cúrdia, V., and Ferrero, A. (2012). The macroeconomic effects of large-scale asset purchase programmes. *The Economic Journal*, 122(564):F289–F315.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy*, 113(1):1–45.

Copeland, A., Duffie, D., and Yang, Y. D. (2021). Reserves were not so ample after all. Staff Reports 974, Federal Reserve Bank of New York.

Crawley, E., Gagnon, E., Hebden, J., and Trevino, J. (2022). Substitutability between balance sheet reductions and policy rate hikes: Some illustrations and a discussion. FEDS notes, Board of Governors of the Federal Reserve System.

d'Avernas, A., Hubert de Fraisse, A., Ning, L., and Vandeweyer, Q. (2024). The fiscal cost of quantitative easing. Working paper, Social Science Research Network.

Diamond, W. F., Jiang, Z., and Ma, Y. (2024). The reserve supply channel of unconventional monetary policy. *Journal of Financial Economics*, 159:103581.

Federal Open Market Committee (2019). Statement regarding monetary policy implementation. Board of Governors of the Federal Reserve System, October 11. Available at https://www.federalreserve.gov/newsevents/pressreleases/monetary20191011a.htm.

Federal Open Market Committee (2022a). Plans for reducing the size of the federal reserve's balance sheet. Board of Governors of the Federal Reserve System, May 4. Available at https://www.federalreserve.gov/newsevents/pressreleases/monetary20220504b.htm.

Federal Open Market Committee (2022b). Principles for reducing the size of the federal reserve's balance sheet. Board of Governors of the Federal Reserve System, January 26. Available at https://www.federalreserve.gov/newsevents/pressreleases/monetary20220126c.htm.

Federal Open Market Committee (2024). Statement regarding reinvestment of principal payments from treasury securities, agency debt, and agency mortgage-backed securities. Federal Reserve

Bank of New York, Operating Policy, May 1. Available at https://www.newyorkfed.org/markets/opolicy/operating_policy_240501.

Federal Reserve Bank of New York (2026). Primary dealers list. https://resources.newyorkfed.org/newsevents/news/markets/2026/an260115p. Announcement of January 15, 2026.

Fricke, D., Greppmair, S., and Paludkiewicz, K. (2026). Excess reserves and monetary policy tightening. *Review of Finance*, 30(5):1777–1816.

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

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

Glover, A. and Oliyide, J. (2024). Why haven't recent rate increases slowed the economy more? Look to unusually low private lending spreads. Economic bulletin, Federal Reserve Bank of Kansas City.

He, Z. and Krishnamurthy, A. (2013). Intermediary asset pricing. *American Economic Review*, 103(2):732–770.

Kandrac, J. and Schlusche, B. (2021). Quantitative easing and bank risk taking: Evidence from lending. *Journal of Money, Credit and Banking*, 53(4):635–676.

Kim, D. and Zha, T. (2026). Sharpening economic interpretation with HARS. Working Paper 35483, National Bureau of Economic Research.

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

Kumhof, M. and Salgado-Moreno, M. (2024). Quantitative easing and quantitative tightening: The money channel. Staff Working Paper 1090, Bank of England.

Lanne, M., Lütkepohl, H., and Maciejowska, K. (2010). Structural vector autoregressions with markov switching. *Journal of Economic Dynamics and Control*, 34(2):121–131.

Leeper, E. M., Sims, C. A., and Zha, T. (1996). What does monetary policy do? *Brookings Papers on Economic Activity*, 1996(2):1–78.

Liu, J. S., Liang, F., and Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134.

Lloyd, S. and Ostry, D. (2024). The asymmetric effects of quantitative tightening and easing on financial markets. *Economics Letters*, 238:111722.

Logan, L. K. (2024). Opening remarks at panel on market monitoring and the implementation of monetary policy. Annual meetings of the International Banking, Economics and Finance Association and the American Economic Association, San Antonio, January 6. Available at https://www.dallasfed.org/news/speeches/logan/2024/lkl240106.

Lu, Y. and Valcarcel, V. J. (2024). A tale of two tightenings. *Journal of Economic Dynamics and Control*, 166:104906.

Murray, I., Adams, R. P., and MacKay, D. J. C. (2010). Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *JMLR Workshop and Conference Proceedings*, pages 541–548.

Nakashima, K., Shibamoto, M., and Takahashi, K. (2024). Identifying quantitative and qualitative monetary policy shocks. *International Journal of Central Banking*, 20(3):1–61.

Rigobon, R. (2003). Identification through heteroskedasticity. *Review of Economics and Statistics*, 85(4):777 – 792. Cited by: 459.

Rubio-Ramírez, J. F., Waggoner, D. F., and Zha, T. (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies*, 77(2):665–696.

Sims, C. A. and Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, 39(4):949–968.

Smith, A. L. and Valcarcel, V. J. (2023). The financial market effects of unwinding the federal reserve's balance sheet. *Journal of Economic Dynamics and Control*, 146:104582.

Swanson, E. T. (2021). Measuring the effects of federal reserve forward guidance and asset purchases on financial markets. *Journal of Monetary Economics*, 118:32–53.

U.S. Securities and Exchange Commission (2026). Rule 2a-7, money market funds, 17 CFR 270.2a-7. Code of Federal Regulations. Paragraph (a)(14), definition of a government money market fund.

Waller, C. J. (2024). Thoughts on quantitative tightening. Speech at the 2024 U.S. Monetary Policy Forum, New York, March 1. Available at https://www.federalreserve.gov/newsevents/speech/waller20240301a.htm.

Wei, B. (2022). Quantifying "quantitative tightening" (QT): How many rate hikes is QT equivalent to? Working Paper 2022-8, Federal Reserve Bank of Atlanta.

Wong, R. (2021). The borrower of last resort: What explains the rise of ON RRP facility usage? Economic Brief 21-43, Federal Reserve Bank of Richmond.

A Balance-Sheet Data

This appendix collects the episode-level balance-sheet evidence discussed in Section 2. Table A.1 reports the Fed's flows by episode. Table A.2 reports the MMF and hedge fund balance sheets over the same windows, and Figure A.1 reports who held the ON RRP balance.

Table A.1: Fed balance-sheet flows by episode, USD billions.

	QE 2008-12 → 2014-11	QT1 2017-10 → 2019-08	COVID QE 2020-03 → 2022-03	QT2 2022-06 → 2025-12
Asset side				
Total assets	+2,266	−690	+4,315	−2,352
MBS	+1,725	−267	+1,340	−666
SOMA Treasury	+1,985	−379	+3,095	−1,559
coupons	+2,004	−381	+3,087	−1,447
T-bills	−18	+2	+8	−112
Liability side				
Total liabilities	+2,266	−690	+4,315	−2,352
Reserves	+1,765	−709	+1,998	−312
ON RRP	+107	−115	+1,588	−2,152
TGA	+12	−43	+242	+124
Other	+381	+177	+487	−11
Absorption ratios (% of total asset change)				
Reserves	77.9	102.8	46.3	13.3
ON RRP	4.7	16.7	36.8	91.5

Notes: Each value is the monthly average in the last month of the window less the monthly average in the first month. The asset side uses the Federal Reserve H.4.1 series WALCL, WSHOMCB, TREAST, and WSHOBL. Coupons are the Treasury securities other than bills, including TIPS and FRNs. The liability side uses the H.4.1 series WRESBAL and WTREGEN and the daily ON RRP series RRPONTSYD. "Other" is total assets less reserves, ON RRP, and the TGA, and it holds currency in circulation and residual Fed liabilities (capital, foreign and other deposits). Total liabilities equal total assets by accounting identity.

Table A.2 records the ON RRP route. When MMFs take cash out of ON RRP, the offsetting item sits on their own asset side, in Treasury securities and in repo with private borrowers. Total assets need not fall. The hedge fund rows record who borrowed the repo cash and what they held. The issuance rows record the coupons and bills that the Treasury sold over each window, which is the supply MMFs and hedge funds absorbed.

Figure A.1 records who held the ON RRP balance. The MMF categories are the MMFs' own

Table A.2: Money market fund and hedge fund balance sheets by episode, USD billions.

	QE 2008-12 → 2014-11	QT1 2017-10 → 2019-08	COVID QE 2020-03 → 2022-03	QT2 2022-06 → 2025-12
ON RRP facility				
All counterparties	+149	−304	+1,870	−1,873
of which MMFs		−286	+1,656	−1,681
Money market fund assets				
Treasury securities		+222	+719	+1,979
Private repo		+609	−605	+2,251
of which through FICC		+173	−106	+1,210
Total investments		+803	+1,058	+3,316
Hedge funds				
Government debt exposure		+953	−493	+2,497
Repo borrowing		+610	−212	+2,289
Treasury issuance				
Net bill issuance	−564	+530	+1,364	+2,874
Net coupon issuance	+7,163	+1,419	+5,006	+4,092

Notes: Changes over the episode windows of Table A.1. Money market fund, facility, and issuance rows are month-end levels at the last month of the window less the level at the last month-end before it. The base of the QT1 window, 2017:M9, is a quarter end, and ON RRP carries its quarter-end spike on that date, which is why the QT1 ON RRP cell is larger than the change in monthly means that Table A.1 reports. Form N-MFP is a month-end snapshot. Hedge fund rows are quarter-ends that bracket the window. Blank cells have no observation at the base date. Form N-MFP begins in 2010:M11 and Form PF in 2013, both after the QE window opens. MMFs are money market funds. Private repo is their repo with every counterparty other than the Fed. Total investments are their portfolio holdings and exclude cash and other assets. FICC is the Fixed Income Clearing Corporation, where hedge funds borrow through sponsored repo. Government debt exposure is hedge funds' gross exposure to U.S. government debt, and repo borrowing is the cash they borrow in repo, both from Form PF. Treasury securities include coupons with remaining maturity under thirteen months. Sources: Federal Reserve daily ON RRP series (RRPONTSYD) and SEC Form N-MFP through the OFR Money Market Fund Monitor. Hedge fund rows come from SEC Form PF through the OFR Hedge Fund Monitor, and issuance rows from the Monthly Statement of the Public Debt.

Form N-MFP categories, pooled as the Office of Financial Research pools them, with Treasury, Government/Agency, and Exempt Government funds as government funds. The non-MMF remainder is the amount the New York Fed reports for government-sponsored enterprises, banks, and primary dealers. The series starts in 2016:M1. The Fed ran term reverse repurchase operations at quarter ends in 2014 and 2015, and MMFs report those alongside the overnight facility while the overnight balance excludes them.

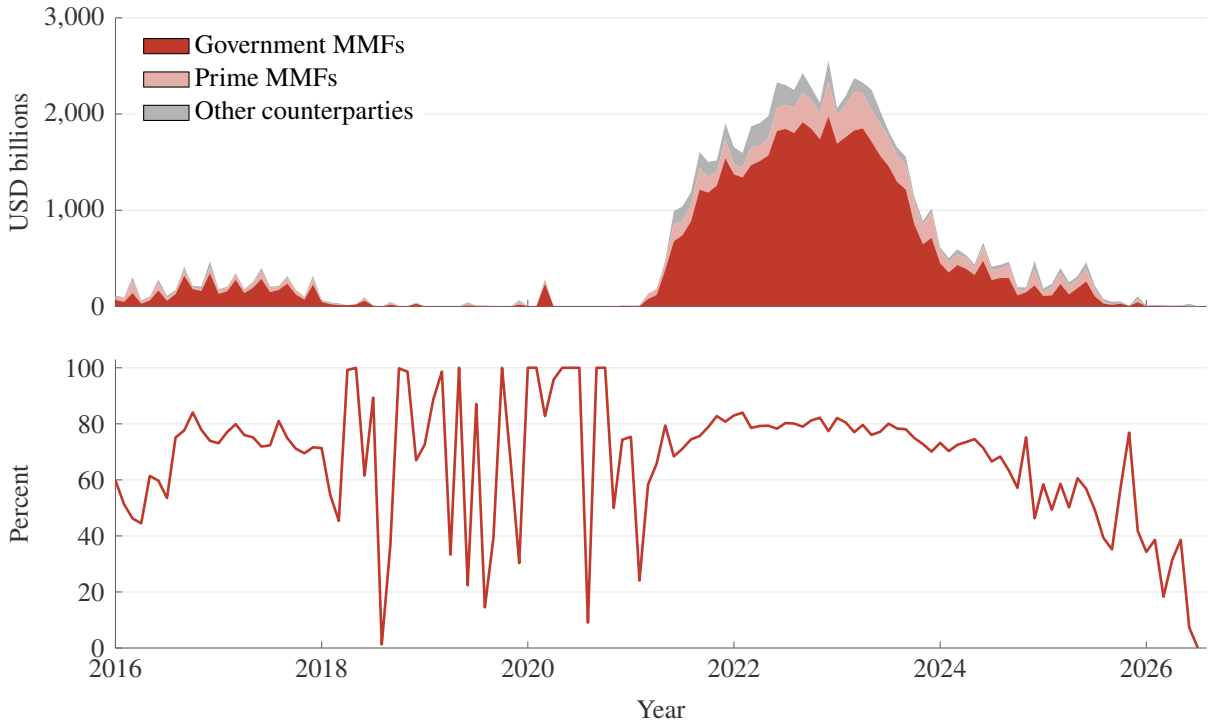


Figure A.1: The ON RRP balance by holder.

Note. Top panel: the month-end ON RRP balance in USD billions, split into government money market funds, prime money market funds, and other counterparties. Bottom panel: the government fund share of the balance in percent. Fund holdings are repurchase agreements with the Federal Reserve Bank of New York reported on Form N-MFP, dated to the last operation of the month; the total and the non-fund remainder are the New York Fed's counterparty-type amounts for that operation. Source: SEC Form N-MFP, Federal Reserve Bank of New York, FRED series RRPONTSYD.

B The Narrative Restrictions

This appendix states the narrative restrictions of Section 3.3. Each restriction ties a labeled shock to the episode that generated it through two requirements. Let W be a window of months inside a regime and j the focal shock. The first is a sign requirement on the shock's realizations. For the RFBS window the mean realization is positive,

$$\frac{1}{|W|}\sum_{t\in W}\tilde{\varepsilon}_{j,t} \;>\; 0. \tag{B.1}$$

For the OFBS window the restriction reverses the inequality in (B.1). The second restricts the historical decomposition. Let $\mathrm{HD}_{j\to L}(W)$ denote the contribution of labeled shock j to $L = \log(\text{Reserves + ON RRP})$ accumulated over the window W. The restriction requires this contribution to dominate those of the other labeled shocks,

$$\left|\mathrm{HD}_{j\to L}(W)\right| \;>\; \sum_{\substack{i \text{ labeled}\\ i\neq j}} \left|\mathrm{HD}_{i\to L}(W)\right|. \tag{B.2}$$

In the baseline, $\mathrm{HD}_{j\to L}(W)$ is the multi-horizon interval historical decomposition of Antolín-Díaz and Rubio-Ramírez (2018), the contribution of shock j to the level of $\log(\text{Reserves + ON RRP})$ at the end of the window. The contributions of the n rotated shocks add up to the unexpected movement of $\log(\text{Reserves + ON RRP})$ over the window, and each restriction assigns a share of that movement to each shock. The sum runs over the labeled shocks only, while Antolín-Díaz and Rubio-Ramírez (2018) compare the focal shock against every other shock.

C HARS and HARS-Z

This appendix expands Section 4. Appendix C.1 reviews HARS in detail. Appendix C.2 derives the zero constraint matrices and states the formal consequences of adding zero restrictions to HARS. Appendix C.3 builds the map γ_z. Appendix C.4 states the results that establish the validity of HARS-Z. Appendix C.5 states the posterior-inference estimator. The next appendix collects the proofs.

C.1 HARS in Detail

The absorbed rotation. HARS (KZ, Algorithm 1) combines likelihood-preserving posterior draws of the model parameters with rotations drawn from a well-defined probability distribution restricted by the sign and narrative restrictions. In homoskedastic SVARs, the rotations of the coefficients, $(A_0, A_+) \mapsto (A_0Q, A_+Q)$ with $Q \in O(n)$, are observationally equivalent. Sign restrictions identify shocks by selecting, among these rotations, those that produce economically sensible impulse responses. Under heteroskedasticity this rotation invariance fails. Rotating the coefficients while holding the regime variances fixed mixes shocks whose relative variances differ across regimes and changes the heteroskedastic likelihood (KZ, Lemma 1 and Proposition 2). The rotation would change the fitted model rather than select an interpretation.

KZ's construction absorbs the regime scale before rotation. Define $\Xi_{s_t} := \Lambda_{s_t}^{-1/2}$ and

$$\varepsilon_t := \Xi_{s_t}\eta_t, \qquad E(\varepsilon_t\varepsilon_t' \mid s_t) \;\propto\; I_n,$$

and the shocks have a common variance in every regime. Post-multiplying (1) by Ξ_{s_t} absorbs the regime scaling into the block $(A_0\Xi_{s_t}, A_+\Xi_{s_t})$. Rotating the entire absorbed block by $Q \in O(n)$ leaves the likelihood unchanged (KZ, Proposition 1),

$$p(y \mid \theta, Q) \;=\; p(y \mid \theta) \qquad \text{for every } Q \in O(n), \tag{C.1}$$

where $\theta = (A_0, \Lambda, A_+)$ collects the structural objects and $\Lambda = \{\Lambda_k\}_{k=1}^h$. Absorbing first keeps the likelihood invariant to Q under heteroskedasticity. I refer to the rotated shocks $\tilde{\varepsilon}_t' = \varepsilon_t' Q$ as the labeled shocks. They are uncorrelated and share one variance in every regime. With the heavy-tailed shocks of (2) they need not be independent. The variance decompositions need only that they are uncorrelated, and the impulse responses and the historical decompositions are linear in them. The dynamic responses of the observables to the labeled shocks in regime k at horizon ℓ are the rows of

$$\Theta_{k,\ell}(Q) \;=\; Q' \Lambda_k^{1/2} A_0^{-1} \, \Phi_\ell(B), \tag{C.2}$$

where $\{\Phi_\ell(B)\}_{\ell \geq 0}$ denotes the moving-average coefficients of the reduced form, with $\Phi_0(B) = I_n$.[27] The regime affects the responses only through the positive diagonal matrix $\Lambda_k^{1/2}$. I also define the average transmission

$$\bar{\Theta}_\ell(Q) \;=\; Q' A_0^{-1} \, \Phi_\ell(B), \tag{C.3}$$

which evaluates the responses at the average of the relative variances across regimes, and that average equals one under the normalization. The average transmission is the object I report as the impulse response. Because Q' mixes the regime-specific scales in $\Lambda_k^{1/2}$ across shocks, the regime-specific responses are not positive rescalings of the average response, and restrictions imposed regime by regime add information.

The restricted posterior target. The likelihood invariance (C.1) separates posterior simulation from identification. Write $SR(\theta, Q)$ for the indicator that the sign and narrative restrictions are satisfied, and define the admissible set $R(\theta) = \{Q \in O(n) : SR(\theta, Q) > 0\}$, as in KZ. When $\int p(\theta \mid y)\, \mu(R(\theta))\, d\theta > 0$, where μ is Haar measure on $O(n)$, KZ's Definition 1 gives the restricted posterior target

$$\pi(d\theta, dQ \mid y, SR) \;\propto\; p(\theta \mid y)\, \mathbf{1}\{Q \in R(\theta)\}\, d\theta\, \mu(dQ). \tag{C.4}$$

[27] Under the right-multiplication convention, $y_{t+\ell}' = \cdots + u_t' \, \Phi_\ell(B)$ and $\Phi_\ell(B) = \sum_{m=1}^{\min(\ell,p)} \Phi_{\ell-m}(B)\, B_m$, where B_m is the m-th $n \times n$ lag block of B. Row j of $\Theta_{k,\ell}(Q)$ collects the horizon-ℓ responses of the n observables to labeled shock j. Unlike KZ, who write $\Theta_{k,\ell}(Q) = \Phi_\ell(B)\, Q' \Lambda_k^{1/2} A_0^{-1}$, this draft places the moving-average matrix on the right, as in (C.2), and all remaining notation is identical.

Conditional on θ, the rotation has Haar measure restricted to $R(\theta)$. Conditional on Q, the model parameters have density proportional to $p(\theta \mid y)\,\mathbf{1}\{SR(\theta, Q) > 0\}$. The marginal density of θ is proportional to $p(\theta \mid y)\,\mu(R(\theta))$ (KZ, Proposition 3). The restrictions reweight θ by the Haar volume of its admissible rotations. HARS never computes this volume. It alternates updates from the two conditional distributions, and its admissibility checks account for the volume.

The rotation in Gaussian coordinates. HARS draws the rotation in Gaussian coordinates. An $n \times n$ matrix X with independent standard-normal entries maps to a Haar-distributed rotation $Q = \gamma(X)$ through the QR map γ. Column-wise updates of X by the elliptical slice sampler (ESS; Murray et al., 2010) leave the conditional distribution of Q in (C.4) invariant (KZ, Proposition 4).

A move in the model parameters can make the current rotation inadmissible, the event KZ call a stale rotation. HARS runs a posterior-draw repair on the completed draw (KZ, Theorem 1(ii)). The repair holds the rotation fixed and starts from the failed candidate. It takes ordinary Metropolis–Hastings steps on (A_0, Λ) and completes each step with a fresh draw of A_+ from its Gibbs conditional. It keeps the first admissible candidate with probability one. If no candidate within the budget is admissible, the chain keeps its current draw. A separate step proposes A_+ from its Gibbs conditional and accepts it if the restrictions hold. Both steps leave the target (C.4) invariant, and neither redraws A_+ alone to rescue a failed candidate.

C.2 The Failure Under Zero Restrictions

Zero restrictions break this construction, and the failure originates in the measure rather than in the algorithm. The zeros invalidate the restricted target (C.4) and the Gaussian coordinatization, while the likelihood invariance (C.1) continues to hold. The obstacle is that a zero restriction is not a truncation. Sign and narrative restrictions truncate the conditional distribution of Q to a subset of positive Haar measure, and the target (C.4) remains well defined. A zero restriction instead imposes a linear equality constraint on a column of Q and lowers the dimension of the admissible set. Adding the zeros to the indicator SR fails, as Remark 1(i) states below.

Under the average transmission (C.3), the impact response of variable i_0 to shock j is $q_j' A_0^{-1} e_{i_0}$, where e_{i_0} is the i_0-th column of I_n. The response is linear in the j-th column q_j of Q, and a zero restriction on this response is a linear constraint on q_j. For a structural draw θ, I describe the zeros on shock j with two objects, following the notation of Arias et al. (2018). The scalar z_j counts the zero restrictions on shock j. The $z_j \times n$ matrix $Z_j(\theta)$ stacks those restrictions as rows. Each row is $(A_0^{-1} e_{i_0})'$ for an impact zero on variable i_0 and $(A_0^{-1}\Phi_\ell(B)\, e_{i_0})'$ for a horizon-ℓ zero, and z_j is the row count of $Z_j(\theta)$.[28] The zero restrictions on shock j read

$$Z_j(\theta)\, q_j = 0,$$

and the sampler must rebuild the constraint matrices at every iteration. The matrices depend on the draw through A_0 and $\Phi_\ell(B)$.[29] Define the zero manifold

$$\mathcal{Q}(\theta) \;=\; \left\{ Q \in \mathcal{O}(n) : Z_j(\theta)\, q_j = 0 \text{ for every restricted } j \right\}.$$

Whenever some $z_j \geq 1$, $\mathcal{Q}(\theta)$ is a lower-dimensional smooth manifold in $\mathcal{O}(n)$ and $\mu(\mathcal{Q}(\theta)) = 0$, as ARRW note in motivating their Algorithm 2.[30] The next remark states the consequences for HARS.

Remark 1 (HARS cannot impose zero restrictions)**.** Fix θ with at least one zero restriction. Then $\mu(\mathcal{Q}(\theta)) = 0$, and the following hold.

[28] ARRW impose the zeros on a transformation $F(A_0, A_+)$ of the structural parameters, writing $Z_j F(A_0, A_+)\, e_j = 0$ with a constant selection matrix Z_j and e_j the jth column of I_n. In their orthogonal reduced-form parameterization the rotation carries the restriction, and their construction evaluates F before the rotation, giving $Z_j F\, q_j = 0$. I fold that product into a single matrix $Z_j(\theta) := Z_j F(\theta)$, which matches KZ's convention $\tilde{\varepsilon}_t' = \varepsilon_t' Q$. The fold is notational. The transformation F appears as the rows of $Z_j(\theta)$, and the count z_j, the dimension bounds, and the null-space construction are identical to theirs.

[29] Unlike the admissibility check $SR(\theta, Q)$, in which the sign restrictions must hold in every regime and Λ enters, the zero constraints do not depend on Λ. The average transmission sets the relative variances to one, and the zeros hold exactly under (C.3). They need not hold in the regime-specific responses (C.2), where $\Lambda_k^{1/2}$ enters the product $Q'\Lambda_k^{1/2} A_0^{-1}\, \Phi_\ell(B)$ and reweights the constrained column.

[30] When $Z_j(\theta)$ has full row rank, the column q_j must lie on the unit sphere of the $(n - z_j)$-dimensional null space of $Z_j(\theta)$. The zero restrictions lower the dimension of $\mathcal{Q}(\theta)$ below the $n(n-1)/2$ dimensions of $\mathcal{O}(n)$, and any lower-dimensional subset has Haar measure zero. ARRW establish the same fact in the structural parameterization. The set of structural parameters satisfying the sign and zero restrictions is of Lebesgue measure zero in the set of all structural parameters. The zero restrictions define a lower-dimensional smooth manifold, and the volume measure of their Theorem 3 serves as the reference measure on it.

(i) The restricted posterior target (C.4) is not defined.

(ii) The Gaussian coordinatization never reaches the zeros, that is, $\Pr\{\gamma(X) \in Q(\theta)\} = 0$.

(iii) Every step of KZ's Algorithm 1 fails with probability one. The failure covers the completed-draw Metropolis–Hastings step (Step 1), the posterior-draw repair (Step 2), the companion update of A_+ (Step 3), and the rotation draw (Step 4). The column-by-column accept–reject initialization fails as well.

Proof of Remark 1. (i) Adding the zeros to the indicator SR forces $R(\theta) \subseteq Q(\theta)$. Monotonicity gives $\mu(R(\theta)) \leq \mu(Q(\theta)) = 0$. The equality $\mu(R(\theta)) = 0$ holds at every θ with at least one zero restriction. The integral $\int p(\theta \mid y)\, \mu(R(\theta))\, d\theta$, which KZ's Definition 1 requires to be positive, is zero.

(ii) The map γ is the QR map that normalizes the R factor to a positive diagonal and returns the orthogonal factor. The rotation $\gamma(X)$ is Haar distributed on $O(n)$ (KZ, Proposition 4). The equality $\mu(Q(\theta)) = 0$ gives $\Pr\{\gamma(X) \in Q(\theta)\} = 0$. The construction projects each Gaussian column only onto the orthogonal complement of the preceding columns, never onto the null spaces of $Z_j(\theta)$.

(iii) The rotation draw, Step 4, fails by (ii). Steps 1 to 3 redraw θ rather than Q and need a separate argument.

Initialization. Each accepted column of the initialization, Algorithm 4 of Arias et al. (2026), is a projected Gaussian draw. A projected Gaussian draw lies in the null space of $Z_j(\theta)$ with probability zero, by (ii). Almost surely no initial admissible pair exists. The chain never occupies a state whose rotation satisfies the zeros.

Steps 1 to 3. All three hold the retained rotation fixed and check completed draws. For a fixed unit column $q_j \neq 0$, the condition $Z_j(\theta)\, q_j = 0$ depends on θ only through A_0 and $\Phi_\ell(B)$. The condition is an exact alignment among functions of A_0^{-1} and $\Phi_\ell(B)$. The set of structural parameters at which the alignment holds is of Lebesgue measure zero, following the argument ARRW use for their set U (their Appendix A.3). Every completed draw and candidate comes from a proposal density and the Gibbs conditional density $p_+(A_+ \mid z_0, s_0)$. Draws with a density avoid the alignment

set almost surely. When the MH step rejects, or when Step 3 keeps z_0, the impact zeros hold at the new draw exactly when they held at the current state. Under HARS they never held at the current state, by the initialization argument, and the rejected step fails as well.

Step 4. The slice step starts from a current admissible point. No such point is ever available, by the initialization argument. Even from a state on $Q(\theta)$, the update fails. Consider the ESS update of a column j that carries zero restrictions. Following KZ's Proposition 4, draw an auxiliary vector $\tau_j \sim \mathcal{N}(0, I_n)$ and consider the ellipse

$$x_j(\vartheta) = x_j \cos\vartheta + \tau_j \sin\vartheta, \qquad \vartheta \in [0, 2\pi].$$

Under γ, the proposed column is the projection of $x_j(\vartheta)$ onto the orthogonal complement of the preceding columns, normalized. Write P for that projection. The projection P is fixed along the ellipse. The preceding columns do not move when the ESS step updates column j. The normalization does not affect the zero condition. The zero restrictions hold at the proposal if and only if $Z_j(\theta)\, P\, x_j(\vartheta) = 0$. Linearity of P gives

$$Z_j(\theta)\, P\, x_j(\vartheta) \;=\; \cos\vartheta\; Z_j(\theta) P\, x_j \;+\; \sin\vartheta\; Z_j(\theta) P\, \tau_j.$$

The coefficient $Z_j(\theta) P\, x_j$ is zero at the current state. The current column satisfies its zeros and is proportional to Px_j. Suppose $Z_j(\theta)P \neq 0$, which fails only when the rows of $Z_j(\theta)$ lie in the span of the preceding columns. The coefficient $Z_j(\theta) P\, \tau_j$ is nonzero almost surely. The auxiliary vector is a fresh Gaussian draw, and a Gaussian vector lies in the null space of a nonzero matrix with probability zero. The zero restrictions hold on the ellipse only where $\sin\vartheta = 0$, at $\vartheta \in \{0, \pi\}$. At these angles the proposal is the current column up to sign. The update never reaches a new rotation on $Q(\theta)$. The slice step retains only angles at which the admissibility indicator passes, and no angle off $\{0, \pi\}$ passes. □

Rotations drawn from the uniform distribution satisfy the zero restrictions with probability

zero.[31] The sampler must parameterize the zero-restricted set directly rather than reach it by sampling. It must build the rotation on $\mathcal{Q}(\theta)$, and its slice-sampling updates must never leave it. Appendix C.3 constructs exactly such a parameterization.

C.3 Extending HARS to Zero Restrictions

HARS-Z extends HARS to include zero restrictions on the rotation. Developing this extension is the paper's methodological contribution. The extension replaces the QR map γ with a map $\gamma_z(\cdot\,;\theta)$ from a Gaussian matrix to a rotation that lies on $\mathcal{Q}(\theta)$ *by construction*. ARRW's Algorithm 2 supplies such a map, and the zero restrictions hold exactly along the chain. The subsection builds the map γ_z and sets the state of the chain and its posterior target. It identifies the measure that the target induces on $\mathcal{Q}(\theta)$ and constructs the rotation draw.

ARRW's Algorithm 2 builds the columns of Q one at a time, following a permutation σ of the shock indices $\{1,\dots,n\}$ that places the most-restricted shocks first. The index $\sigma(j)$ is the shock whose column the construction builds at step j. The permutation must satisfy $z_{\sigma(j)} \le n - j$ for every j, where $z_{\sigma(j)}$ is the zero count of the shock assigned to step j. The j-th column must be orthogonal to the $j-1$ preceding columns and must satisfy its own $z_{\sigma(j)}$ zeros. A candidate direction exists only if $(j-1) + z_{\sigma(j)} < n$.[32] For every restricted shock j, the matrix $Z_j(\theta)$ must also have full row rank, that is, its z_j rows must be non-redundant.[33]

Given a Gaussian matrix $G = [g_1,\dots,g_n]$, the construction proceeds one column at a time. At step j, the admissible directions for column $q_{\sigma(j)}$ form the linear subspace

$$N_j(\theta) \;:=\; \mathrm{null}\big(M_j(\theta)\big), \qquad M_j(\theta) \;:=\; \begin{bmatrix} \big[\, q_{\sigma(1)},\dots,q_{\sigma(j-1)} \,\big]' \\ Z_{\sigma(j)}(\theta) \end{bmatrix}, \tag{C.5}$$

[31] The obstruction is also topological. The elliptical slice sampler of Arias et al. (2026) assumes an admissible set of positive measure (their Assumption 1). Strict-inequality sign restrictions give an open set of positive measure. Zero restrictions give a set that is neither open nor of positive measure.

[32] Later positions have more orthogonality constraints and admit fewer zeros, and the most heavily restricted shocks come first.

[33] ARRW impose this assumption on their selection matrices Z_j when they set their notation (their Section 2.2).

where $M_j(\theta)$ stacks the preceding columns $q_{\sigma(1)}, \ldots, q_{\sigma(j-1)}$ and the zero restrictions of shock $\sigma(j)$. The null space $N_j(\theta)$ collects the vectors orthogonal to every preceding column and consistent with those zeros.

The projection at step j requires the rows of $M_j(\theta)$ to be linearly independent. The null space has the dimension $n - (j-1) - z_{\sigma(j)}$ that the ordering condition above presumes. Each block of $M_j(\theta)$ is independent on its own. The preceding columns are orthonormal, and $Z_{\sigma(j)}(\theta)$ has full row rank. The requirement binds only across the two blocks. I assume that the rows of $Z_{\sigma(j)}(\theta)$ are linearly independent of the preceding columns at every orthonormal collection $q_{\sigma(1)}, \ldots, q_{\sigma(j-1)}$ whose columns satisfy their zero restrictions.[34] ARRW derive this property from a regularity condition on their function F and the full row rank of the Z_j. See their footnote 6 in Section 2.2 and their Appendix A.3, conditions 1 and 2. I assume it directly, at each θ the chain visits rather than over the whole parameter space.

The construction projects g_j orthogonally onto $N_j(\theta)$ and normalizes,

$$q_{\sigma(j)} = \frac{P_j(\theta)\, g_j}{\left\| P_j(\theta)\, g_j \right\|}, \tag{C.6}$$

where $P_j(\theta)$ denotes the orthogonal projection onto $N_j(\theta)$ and the denominator is nonzero for almost every g_j. Write $\gamma_z(\cdot\,;\theta)$ for the map $G \mapsto Q$ that these n projection steps define. The chain below stores this Gaussian input as part of its state and denotes it by X.

Two properties follow. First, each column satisfies its constraints by construction, and after n steps Q is orthogonal and lies on $\mathcal{Q}(\theta)$. The construction involves no accept–reject step. Second, the projected Gaussian column favors no direction within $N_j(\theta)$, and the normalized column is uniform over the admissible directions at step j.[35] This step-by-step distribution matches Steps 2 and 3 of ARRW's Algorithm 2.[36]

[34] I do not derive this condition from the distribution of θ. At a fixed θ, a combination of the rows of $Z_{\sigma(j)}(\theta)$ can fall in the span of the preceding columns for a set of preceding columns of positive measure. The implementation checks the rank of $M_j(\theta)$ at every construction step of every draw and stops if the check fails. No retained draw failed it.

[35] The orthogonal projection of a standard Gaussian vector onto a linear subspace is a standard Gaussian within that subspace, whose distribution is invariant under rotations of the subspace. Its normalized direction is uniform on the subspace's unit sphere.

[36] At each step j, they draw a standard normal vector on $\mathbb{R}^{n+1-j-z_j}$, normalize it to the unit sphere, and map it through

The uniform distribution of each column, conditional on the preceding columns, determines the law of the completed rotation. Write $\mu_z(\cdot\,;\theta)$ for the null-space construction measure on $\mathcal{Q}(\theta)$, the distribution of $\gamma_z(G;\theta)$ under a standard Gaussian matrix G,

$$\mu_z(C;\theta) \;:=\; \Pr\big\{\gamma_z(G;\theta) \in C\big\} \qquad \text{for every measurable } C \subseteq \mathcal{Q}(\theta). \tag{C.7}$$

Proposition 2 characterizes this measure and relates it to the conditional-Haar (uniform) measure on $\mathcal{Q}(\theta)$.

I depart from ARRW in one respect. At step j they draw $x_j \in \mathbb{R}^{n+1-j-z_j}$ from a standard normal distribution and normalize it to the unit sphere. The dimension of their coordinates varies with j and, through z_j, with the restriction pattern. I instead use a single $n \times n$ Gaussian matrix X, the coordinates in which HARS runs its ESS update. Step j takes the j-th column of X as its input. The randomness remains separate across steps, and every column keeps the fixed dimension n. The projection at each step keeps only the components in $N_j(\theta)$, and the projected column has the same distribution as the vector ARRW draw at that step. The fixed $n \times n$ matrix allows me to apply the column-wise ESS update of KZ (Proposition 4) and its validity argument without change.

Lemma 1 (Properties of the map γ_z)**.** *Fix θ such that, for every j, the matrix $Z_{\sigma(j)}(\theta)$ has full row rank and the permutation satisfies $z_{\sigma(j)} \le n - j$. Require also that the matrix $M_j(\theta)$ in (C.5) has full row rank at every orthonormal collection $q_{\sigma(1)}, \ldots, q_{\sigma(j-1)}$ whose columns satisfy their zero restrictions. Then the following hold for almost every G.*

(i) $\gamma_z(G;\theta) \in \mathcal{Q}(\theta)$, and every zero restriction holds exactly, with no rejection.

(ii) Every construction step k depends on G only through $g_1, \ldots, g_k$. Perturbing g_j leaves the columns from steps $1, \ldots, j-1$ unchanged, and the construction re-projects the perturbed and later columns into their own null spaces.

a matrix K_j. The columns of K_j form an orthonormal basis for the null space of M_j, the matrix that stacks the preceding columns and the zero restrictions. The orthogonal projection above induces the same uniform distribution on that null space. The two constructions draw each column from the same distribution.

(iii) If $Q \in \mathcal{Q}(\theta)$ is admissible and $G_Q := [\, q_{\sigma(1)}, \ldots, q_{\sigma(n)} \,]$ collects the columns of Q in build order, then $\gamma_z(G_Q; \theta) = Q$. The construction reproduces its own output, and setting $X = G_Q$ places the chain at an admissible state.

Proof of Lemma 1. (i) At step j the subspace $N_j(\theta)$ in (C.5) is, by construction, orthogonal to every preceding column and contained in $\operatorname{null}\big(Z_{\sigma(j)}(\theta)\big)$. The projected and normalized column $q_{\sigma(j)}$ therefore has unit norm, is orthogonal to its predecessors, and satisfies its zero restrictions exactly. The matrix $M_j(\theta)$ in (C.5) has $(j-1) + z_{\sigma(j)}$ rows. The dimension of $N_j(\theta)$ equals $n + 1 - j - z_{\sigma(j)} \geq 1$ if and only if $M_j(\theta)$ is of full row rank (ARRW). Full row rank holds by hypothesis at every orthonormal collection of preceding columns whose columns satisfy their zero restrictions. The construction completes with Q orthogonal. The projection of a Gaussian column is zero only on a null set. The rotation $\gamma_z(G; \theta)$ lies in $\mathcal{Q}(\theta)$ for almost every G.

(ii) Fix a step k. The subspace $N_k(\theta)$ has two ingredients, the preceding columns and the constraint matrix $Z_{\sigma(k)}(\theta)$. The constraint matrix is fixed at the given θ. The preceding columns come from steps $1, \ldots, k-1$, and step k adds g_k and nothing else. By induction, every construction step k depends on G only through $g_1, \ldots, g_k$. In particular, g_j is not among the inputs of steps $1, \ldots, j-1$ and is among the inputs of steps $j, \ldots, n$. Perturbing g_j leaves steps $1, \ldots, j-1$ untouched. The construction recomputes steps $j, \ldots, n$. Each recomputed column passes through the projection onto its own null space. The zeros hold after the perturbation.

(iii) Let $Q \in \mathcal{Q}(\theta)$ be admissible. Each of its columns already lies in its own subspace $N_j(\theta)$. Orthogonal projection acts as the identity on $N_j(\theta)$. Normalization preserves a unit vector. The construction reproduces Q column by column. The reproduction property is the projection argument ARRW use to establish that their Algorithm 2 can generate every rotation satisfying the zero restrictions. □

Lemma 1 makes ARRW's Algorithm 2 usable inside a Markov chain. Part (i) restates the property of their Algorithm 2 in the fixed $n \times n$ Gaussian parameterization introduced above. Part (ii) is new. ARRW draw each rotation independently and never perturb one afterward. They never need

to ask how the rotation changes when a single Gaussian column moves.[37] The column-wise ESS update moves one column of X along an ellipse and poses exactly that question. Part (ii) supplies the answer. The recomputed rotation stays on $\mathcal{Q}(\theta)$. Part (iii) reuses ARRW's support argument for a new purpose. Setting $X = G_Q$ starts the chain at any admissible rotation, whatever search produced that rotation. Part (iii) supplies the admissible state that Algorithm 1 presupposes.

The state of the chain. A sampler could store the rotation itself, keeping the matrix Q in the state and reusing it after every structural update. Under HARS the two choices are equivalent. KZ carry the Gaussian matrix X between iterations and condition the parameter update on $Q = \gamma(X)$. The QR map γ does not depend on θ, and storing Q or X gives the same chain. On $\mathcal{Q}(\theta)$ the choice matters. The manifold $\mathcal{Q}(\theta)$ moves with θ. A stored Q lies on the old manifold and leaves the new manifold $\mathcal{Q}(\theta^\star)$ with probability one as soon as the structural draw moves. The zero restrictions fail after every structural update. I store the Gaussian coordinates, as KZ do. The state is the pair (θ, X), where X is an $n \times n$ Gaussian matrix, and the sampler recomputes the rotation from the state whenever it needs one,

$$Q \;=\; \gamma_z(X;\theta).$$

The next subsection establishes the validity of the construction. Corollary 1 shows that the Gaussian coordinates are auxiliary, and Proposition 1 shows that the MTM step leaves the structural target invariant. Proposition 2 derives the measure that the construction induces on the admissible set, and Corollary 2 shows that the weighted draws recover conditional-Haar expectations. Proposition 3 gives the column-wise elliptical slice sampling update that the sampler uses. The last paragraph checks each step of Algorithm 1.

[37] ARRW use their coordinates only for independent proposals. The QR map γ of KZ fails for the reason given in Remark 1(ii). Under γ, an elliptical perturbation of a single Gaussian column cannot restore the zeros, and the sign-only ESS never visits the manifold.

C.4 The Validity of HARS-Z

A structural update leaves X unchanged and changes only the rotation $\gamma_z(X;\theta)$. The construction projects the same Gaussian columns onto the null spaces of the new draw, and $\gamma_z(X;\theta^\star)$ lies on $Q(\theta^\star)$ by Lemma 1(i). The zeros hold after every structural move by construction, and the admissibility check reduces to the sign and narrative restrictions. The rotation remains auxiliary, exactly as in KZ. Given (θ, X), the rotated absorbed system of (C.1) evaluated at $Q = \gamma_z(X;\theta)$ generates the data, and the conditional density $p(y \mid \theta, X)$ is well defined. The next definition states the posterior target of HARS-Z, and the corollary that follows establishes its validity.

Definition 1 (The restricted posterior target on (θ, X))**.** Let $\varphi(X)$ denote the standard Gaussian density of the $n \times n$ matrix X and define

$$\kappa(\theta) \;=\; \int \mathbf{1}\big\{SR\big(\theta, \gamma_z(X;\theta)\big) > 0\big\}\, \varphi(X)\, dX,$$

the Gaussian probability of the admissible set at θ. When $\kappa(\theta) > 0$, define the conditional distribution of X given θ,

$$\pi_X(dX \mid \theta, SR) \;:=\; \frac{\mathbf{1}\big\{SR\big(\theta, \gamma_z(X;\theta)\big) > 0\big\}\, \varphi(X)}{\kappa(\theta)}\, dX. \tag{C.8}$$

The restricted posterior target is

$$\pi(d\theta, dX \mid y, SR) \;\propto\; \pi(\theta \mid y)\, \mathbf{1}\big\{SR\big(\theta, \gamma_z(X;\theta)\big) > 0\big\}\, \varphi(X)\, dX\, d\theta \;=\; \pi(\theta \mid y)\, \kappa(\theta)\, \pi_X(dX \mid \theta, SR)\, d\theta. \tag{C.9}$$

Its marginal of θ is proportional to $\pi(\theta \mid y)\, \kappa(\theta)$, and its conditional of X given θ is $\pi_X(\cdot \mid \theta, SR)$.

Definition 1 is the analogue in Gaussian coordinates of the posterior of ARRW's Algorithm 1 and of KZ's Definition 1. KZ's Definition 1 restricts Haar measure on $O(n)$ to $R(\theta)$ without normalizing at each θ, and each θ carries the weight $\mu(R(\theta))$. Definition 1 restricts the joint Gaussian distribution of X in the same way. Each θ carries the weight $\kappa(\theta)$, the Gaussian probability of its admissible set. Each reduced-form draw in ARRW's Algorithm 1 survives in the same way, in proportion to the Haar probability of its admissible set.

A fixed distribution on rotations is unavailable here. A distribution on $Q(\theta)$ assigns no mass to $Q(\theta')$ at any other draw θ'. The Gaussian distribution of X restores a fixed reference distribution. Equation (C.9) restricts the same $\varphi(X)$ at every draw, and θ enters only through the indicator and the map $\gamma_z(\cdot\,;\theta)$. The weight $\kappa(\theta)$ is a probability under the construction, not the uniform probability $\kappa_u(\theta)$ of (6). Appendix C.5 corrects the difference.

Corollary 1 (The Gaussian coordinates are auxiliary)**.** *Fix θ satisfying the conditions of Lemma 1 and write $Q = \gamma_z(X;\theta)$. For almost every X,*

$$p\big(y \mid \theta, X\big) = p\big(y \mid \theta, Q\big) = p\big(y \mid \theta\big).$$

If, in addition, $\kappa(\theta) > 0$ on a set of positive posterior probability, the target (C.9) is the posterior of (θ, X) under the prior that restricts $p(\theta)\,\varphi(X)$ to the admissible set. Drawing X from $\pi_X(\cdot \mid \theta, SR)$ in (C.8) at fixed θ leaves the target invariant.

Proof of Corollary 1. Fix θ and the regime path. Given (θ, X), the rotated absorbed system evaluated at $Q = \gamma_z(X;\theta)$ generates the data,

$$y_t' A_0 \Xi_{s_t} Q = x_t' A_+ \Xi_{s_t} Q + \tilde{\varepsilon}_t', \qquad E(\tilde{\varepsilon}_t \tilde{\varepsilon}_t' \mid s_t) = I_n,$$

where $\tilde{\varepsilon}_t' = \varepsilon_t' Q = \eta_t' \Xi_{s_t} Q$ collects the labeled shocks of Appendix C.1. The conditional density of y given (θ, X) is $p(y \mid \theta, Q)$ evaluated at $Q = \gamma_z(X;\theta)$. The evaluation gives the first equality of the corollary. By Lemma 1(i), $\gamma_z(X;\theta) \in Q(\theta) \subseteq O(n)$ for almost every X. KZ's Proposition 1 gives $p(y \mid \theta, Q) = p(y \mid \theta)$ for every $Q \in O(n)$. Evaluating at $Q = \gamma_z(X;\theta)$ yields the second equality of the corollary. The dependence of the map on θ does not affect the evaluation. The invariance $p(y \mid \theta, Q) = p(y \mid \theta)$ holds at each fixed θ for every $Q \in O(n)$, and $\gamma_z(X;\theta)$ is one such rotation.

For the last claim, specify the prior of the pair as $p(\theta)\,\mathbf{1}\{SR(\theta, \gamma_z(X;\theta)) > 0\}\,\varphi(X)$, the prior $p(\theta)\,\varphi(X)$ restricted to the admissible set. The joint posterior under this prior satisfies

$$\pi(\theta, X \mid y) \;\propto\; p(y \mid \theta, X)\, p(\theta)\, \mathbf{1}\{SR(\theta, \gamma_z(X;\theta)) > 0\}\, \varphi(X)$$

$$
\begin{aligned}
&= p(y \mid \theta)\, p(\theta)\, \mathbf{1}\{SR(\theta, \gamma_z(X;\theta)) > 0\}\, \varphi(X) \\
&\propto \pi(\theta \mid y)\, \mathbf{1}\{SR(\theta, \gamma_z(X;\theta)) > 0\}\, \varphi(X).
\end{aligned}
$$

The first line is Bayes' rule. The second line applies the first part. The third line collects $\pi(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)$. The last expression is the target (C.9). Its normalizer $\int \pi(\theta \mid y)\, \kappa(\theta)\, d\theta$ lies in $(0, 1]$ when $\kappa(\theta) > 0$ on a set of positive posterior probability. The conditional of X given θ under the target is $\pi_X(\cdot \mid \theta, SR)$, and a draw from a full conditional leaves the joint distribution invariant. The argument is the one behind ARRW's Algorithm 1, with Haar measure on $O(n)$ replaced by the Gaussian distribution of X. □

Corollary 1 justifies the block structure of the sampler. The equalities say that the likelihood never involves X, and X enters the target only through the admissibility indicator. At fixed X the target of θ is the posterior restricted to the draws that pass $SR(\theta, \gamma_z(X;\theta))$, the structural target of KZ's Theorem 1. Proposition 1 shows that the MTM step leaves it invariant. The last claim covers the rotation block. The latent scales enter the likelihood but not $SR(\theta, Q)$, and their Gibbs conditional is that of BPSS.

The structural update. Augment the target with the latent scales. The indicator does not involve them, and the augmented target has (C.9) as its marginal in (θ, X). Fix X and the latent scales s_0. Up to a constant, the target (C.9) is a density in $\theta = (z_0, z_+)$,

$$
\pi_{\mathrm{st}}(\theta \mid X, s_0) \;\propto\; p_0(z_0 \mid s_0)\, p_+(z_+ \mid z_0, s_0)\, \mathbf{1}\{SR(\theta, \gamma_z(X;\theta)) > 0\}, \tag{C.10}
$$

the structural target of KZ's Theorem 1 at the rotation $\gamma_z(X;\theta)$. Step 2 of Algorithm 1 proposes θ' from θ by $T(\theta, \theta') = \phi_{\mathrm{rw}}(z_0' - z_0)\, p_+(z_+' \mid z_0', s_0)$, where ϕ_{rw} is the $\mathcal{N}(0, \Sigma_{\mathrm{rw}})$ density.

Proposition 1 (The MTM step leaves the structural target invariant)**.** *Fix X and s_0, and let the conditions of Lemma 1 hold at every θ the chain visits. If the current θ satisfies $SR(\theta, \gamma_z(X;\theta)) > 0$, Step 2 of Algorithm 1 leaves* (C.10) *invariant for every $K \geq 1$.*

Proof of Proposition 1. Liu et al. (2000, Theorem 1) show that MTM satisfies detailed balance with respect to a target π_t when the weight of a candidate y drawn from x is $w(y,x) = \pi_t(y)\,T(y,x)\,\lambda(y,x)$. The function λ must be nonnegative and symmetric with $\lambda(x,y) > 0$ whenever $T(x,y) > 0$, and the proposal must satisfy $T(x,y) > 0$ exactly when $T(y,x) > 0$. Take

$$\lambda(\theta,\theta') \;=\; \frac{1}{\phi_{\mathrm{rw}}(z_0' - z_0)\; p_+(z_+' \mid z_0', s_0)\; p_+(z_+ \mid z_0, s_0)}.$$

The proposal T is positive everywhere, as ϕ_{rw} and the Gibbs conditional p_+ are Gaussian densities, and the support condition of the theorem holds. The random-walk density is symmetric, $\phi_{\mathrm{rw}}(z_0' - z_0) = \phi_{\mathrm{rw}}(z_0 - z_0')$, and λ is symmetric and positive. Substituting (C.10),

$$\begin{aligned} w(\theta',\theta) \;&=\; \frac{p_0(z_0' \mid s_0)\; p_+(z_+' \mid z_0', s_0)\; \mathbf{1}\{SR(\theta', \gamma_z(X;\theta')) > 0\}\; \phi_{\mathrm{rw}}(z_0 - z_0')\; p_+(z_+ \mid z_0, s_0)}{\phi_{\mathrm{rw}}(z_0' - z_0)\; p_+(z_+' \mid z_0', s_0)\; p_+(z_+ \mid z_0, s_0)} \\ &=\; p_0(z_0' \mid s_0)\; \mathbf{1}\{SR(\theta', \gamma_z(X;\theta')) > 0\}, \end{aligned}$$

the weight $w(\theta')$ of Step 2. The weight does not depend on the state that drew the candidate. Steps 2(a) to 2(d) are the MTM step with this weight. They draw K candidates from $T(\theta,\cdot)$ and select one in proportion to the weight. They draw $K-1$ reference points from $T(\theta^c,\cdot)$, take the current θ as the K-th, and accept with probability $\min\{1, \sum_k w(\theta^{(k)}) / \sum_k w(\tilde{\theta}^{(k)})\}$. When every candidate has weight zero, the step keeps θ. The detailed-balance equation for a move from θ to $\theta' \neq \theta$ involves only steps that select a candidate, and keeping θ moves no mass. Step 2 satisfies detailed balance with respect to (C.10) and leaves it invariant. □

The weight omits the Gibbs density of A_+ and the random-walk density. Every candidate pairs z_0 with its own draw of A_+, and a candidate that fails the restrictions has weight zero. Step 2 leaves the joint density of (z_0, z_+) invariant. The marginal of z_0 keeps the admissibility weight of KZ's Theorem 1, the probability that a fresh Gibbs draw of A_+ completes z_0 admissibly. With $K = 1$ the only reference point is the current θ, and the acceptance probability is $\min\{1, w(\theta^c)/w(\theta)\}$. This step is the MH step of BPSS on the collapsed density with the admissibility indicator.

Posterior inference evaluates the impulse responses, functions of (θ, Q). The next definition states the distribution that posterior inference targets.

Definition 2 (The conditional-Haar target on (θ, Q))**.** Fix the conditions of Lemma 1 at every θ with $\kappa(\theta) > 0$. The conditional-Haar target, restated from Section 4, is

$$\pi_v(d\theta, dQ \mid y, SR) \;\propto\; \pi(\theta \mid y)\, \mathbf{1}\{Q \in R(\theta)\}\, \mu\big(dQ \mid \mathcal{Q}(\theta)\big)\, d\theta, \tag{6}$$

where $\mu(\cdot \mid \mathcal{Q}(\theta))$ denotes the conditional-Haar (uniform) measure on $\mathcal{Q}(\theta)$ and $R(\theta) = \{Q \in \mathcal{Q}(\theta) : SR(\theta, Q) > 0\}$ denotes the admissible set on the zero-restricted set. Its marginal of θ is proportional to $\pi(\theta \mid y)\, \kappa_u(\theta)$, and its conditional of Q given θ is $\pi_v(dQ \mid \theta, y, SR) = \mathbf{1}\{Q \in R(\theta)\}\, \mu\big(dQ \mid \mathcal{Q}(\theta)\big)/\kappa_u(\theta)$.

Remark 2. The conditional $\pi_v(dQ \mid \theta, y, SR)$ is well defined whenever $\kappa(\theta) > 0$. The admissible set $R(\theta)$ is relatively open in $\mathcal{Q}(\theta)$, as the restrictions hold with strict inequalities. When $R(\theta)$ is empty, no draw satisfies the indicator in (C.8) and $\kappa(\theta) = 0$. The uniform measure on $\mathcal{Q}(\theta)$ and the construction measure both assign positive mass to every nonempty relatively open subset, and $\kappa_u(\theta)$ and $\kappa(\theta)$ are positive at the same θ.

Definition 2 differs from Definition 1 in the measure on rotations. Both restrict a reference measure to the admissible set without normalizing at each θ, and their marginals of θ are proportional to $\pi(\theta \mid y)\, \kappa_u(\theta)$ and $\pi(\theta \mid y)\, \kappa(\theta)$. Definition 1 states the target on (θ, X), the coordinates in which the chain runs, and Definition 2 states the target on (θ, Q), the coordinates of the impulse responses. Appendix C.5 turns draws from the first into estimates under the second.

Proposition 2 (The induced measure on $\mathcal{Q}(\theta)$)**.** *Fix θ satisfying the conditions of Lemma 1 with $\kappa(\theta) > 0$. The distribution over $\mathcal{Q}(\theta)$ that $\pi_X(\cdot \mid \theta, SR)$ induces through γ_z is the null-space construction measure $\mu_z(\cdot\,; \theta)$ of (C.7), restricted to the admissible set $R(\theta)$. The conditional of (6) given θ is the conditional-Haar measure restricted to $R(\theta)$, written $\pi_v(dQ \mid \theta, y, SR)$. The restricted construction measure has two properties.*

(a) When the zeros restrict a single column of Q and the ordering builds that column first, the restricted construction measure coincides with $\pi_v(\cdot \mid \theta, y, SR)$.

(b) Otherwise the restricted construction measure and $\pi_v(\cdot \mid \theta, y, SR)$ are related through the volume element $v(Q;\theta)$ of the construction (ARRW, Theorem 3). The two coincide when $v(Q;\theta)$ is constant on $R(\theta)$, and weighting each retained draw by $v(Q;\theta)$ recovers $\pi_v(\cdot \mid \theta, y, SR)$ in every case.

Proof of Proposition 2. Two preliminary steps characterize the induced distribution. Parts (a) and (b) then follow in turn.

Step 1: π_X induces the construction measure restricted to the admissible set. The argument follows the image claim of KZ's Proposition 4, with the QR map γ replaced by γ_z and Haar measure replaced by the construction measure $\mu_z(\cdot\,;\theta)$. Recall the conditional (C.8),

$$\pi_X(dX \mid \theta, SR) \;=\; \frac{\mathbf{1}\{SR(\theta, \gamma_z(X;\theta)) > 0\}\,\varphi(X)}{\kappa(\theta)}\,dX.$$

Without the indicator the conditional reduces to the Gaussian density φ. The matrix X then has the same distribution as the Gaussian matrix G in (C.7). The relation $\gamma_z(X;\theta) \sim \mu_z(\cdot\,;\theta)$ holds by definition.

Now restore the indicator. The indicator in (C.8) does not test X directly. The indicator builds the rotation $\gamma_z(X;\theta)$ and tests whether that rotation satisfies the sign and narrative restrictions. The input X passes exactly when its rotation lies in the admissible set $R(\theta) = \{Q \in O(n) : SR(\theta, Q) > 0\}$ of Section 4,

$$\{X : SR(\theta, \gamma_z(X;\theta)) > 0\} \;=\; \{X : \gamma_z(X;\theta) \in R(\theta)\}. \tag{C.11}$$

Take any set C of rotations. Under the conditional $\pi_X(\cdot \mid \theta, SR)$, the probability that the rotation lands in C is

$$\Pr\{\gamma_z(X;\theta) \in C \,|\, SR(\theta, \gamma_z(X;\theta)) > 0\} \;=\; \Pr\{\gamma_z(X;\theta) \in C \,|\, \gamma_z(X;\theta) \in R(\theta)\}$$

$$= \frac{\Pr\{\gamma_z(X;\theta) \in C \cap R(\theta)\}}{\Pr\{\gamma_z(X;\theta) \in R(\theta)\}}$$
$$= \frac{\mu_z(C \cap R(\theta);\theta)}{\mu_z(R(\theta);\theta)}.$$

The first line rewrites the conditioning event. Equation (C.11) shows the two events are the same set of X. The second line applies the definition of conditional probability. The probabilities on the second line are unconditional. On the second line $X \sim \varphi$, and the relation $\gamma_z(X;\theta) \sim \mu_z(\cdot\,;\theta)$ from the start of the proof evaluates each probability as a μ_z-value on the third line.

The set C is arbitrary, and the values on every set determine a distribution. The induced distribution is

$$\frac{\mathbf{1}\{SR(\theta,Q) > 0\}\ \mu_z(dQ;\theta)}{\mu_z(R(\theta);\theta)}. \tag{C.12}$$

The measure (C.12) is the construction measure restricted to the admissible set $R(\theta)$. The restricted construction measure is the counterpart of KZ's Haar measure restricted to $R(\theta)$.

The normalizer of (C.12) coincides with the normalizer $\kappa(\theta)$ of (C.8), $\kappa(\theta) = \mu_z(R(\theta);\theta)$. Both equal the probability of one event, that the rotation is admissible. The normalizer $\kappa(\theta)$ computes this probability in X. The normalizer $\mu_z(R(\theta);\theta)$ computes it on $Q(\theta)$.

Step 2: relation to conditional-Haar. The comparison proceeds column by column. The construction builds the rotation column by column. The construction generates $q_{\sigma(1)}, \ldots, q_{\sigma(n)}$ in the order σ. Each step projects a fresh Gaussian column onto the null space $N_j(\theta)$ that the zero restrictions and the previously built columns determine, then normalizes. The projected Gaussian column favors no direction within $N_j(\theta)$. Conditional on the preceding columns,

$$q_{\sigma(j)} \,\big|\, q_{\sigma(1)}, \ldots, q_{\sigma(j-1)} \ \sim\ \text{uniform on the unit sphere of } N_j(\theta).$$

The uniform law of the projected column is the second property of the construction in Appendix C.3. Write $\mu_j(\cdot \mid q_{\sigma(1)}, \ldots, q_{\sigma(j-1)};\theta)$ for this conditional distribution, the uniform distribution on the unit sphere of $N_j(\theta)$. The chain rule of probability writes the joint law as the product of these

conditionals,

$$\mu_z(dQ;\theta) \;=\; \prod_{j=1}^{n} \mu_j\big(dq_{\sigma(j)} \mid q_{\sigma(1)},\dots,q_{\sigma(j-1)};\theta\big). \tag{C.13}$$

Equation (C.7) defines $\mu_z(\cdot\,;\theta)$ only as the distribution of $\gamma_z(G;\theta)$. The product (C.13) makes the measure explicit. Parts (a) and (b) below use (C.13).

Part (a): the zeros load on a single column. The ordering builds the restricted column first. The first null space $N_1(\theta)$ then equals the null space of the zero restrictions. Every later null space $N_j(\theta)$, $j \geq 2$, is the orthogonal complement of the preceding columns and carries no further constraint.

The comparison involves two distinct measures on $\mathcal{Q}(\theta)$. The construction measure $\mu_z(\cdot\,;\theta)$ is the law the sampler induces. The measure $\mu(\cdot \mid \mathcal{Q}(\theta))$ is the conditional-Haar (uniform) measure on $\mathcal{Q}(\theta)$ and serves as the reference measure of the target. The subscript z denotes the measure that the zero-restricted construction induces. The conditioning on $\mathcal{Q}(\theta)$ denotes Haar measure restricted to $\mathcal{Q}(\theta)$. Factor both measures at the first column,

$$\mu_z(dQ;\theta) \;=\; \mu_1\big(dq_{\sigma(1)};\theta\big)\,\mu_z\big(dq_{\sigma(2)},\dots,dq_{\sigma(n)} \mid q_{\sigma(1)};\theta\big), \tag{C.14}$$

$$\mu\big(dQ \mid \mathcal{Q}(\theta)\big) \;=\; \mu\big(dq_{\sigma(1)} \mid \mathcal{Q}(\theta)\big)\,\mu\big(dq_{\sigma(2)},\dots,dq_{\sigma(n)} \mid q_{\sigma(1)},\mathcal{Q}(\theta)\big), \tag{C.15}$$

and compare factor by factor.

The first factors of (C.14) and (C.15) coincide. The first factor of (C.14) is $\mu_1(\cdot\,;\theta)$, the uniform distribution on the unit sphere of $N_1(\theta)$, by the second property of the construction. The first factor of (C.15) is uniform on the same sphere, by the following symmetry argument. Take any orthogonal matrix R that rotates $N_1(\theta)$ within itself and fixes its orthogonal complement. The map $Q \mapsto RQ$ moves the first column along the sphere and leaves every zero restriction intact. The map $Q \mapsto RQ$ therefore sends $\mathcal{Q}(\theta)$ onto itself. Haar measure is invariant under this map. Conditional-Haar inherits the invariance. Every rotation of $N_1(\theta)$ leaves the marginal distribution of $q_{\sigma(1)}$ unchanged. The only probability distribution on a sphere that every rotation leaves unchanged is the uniform distribution. Both first factors are the same uniform law on the same sphere.

The second factors of (C.14) and (C.15) coincide as well. Fix $q_{\sigma(1)}$. No zero restriction remains after the first column. Every later null space is the orthogonal complement of the preceding columns inside $q_{\sigma(1)}^{\perp}$. Each remaining step projects a fresh Gaussian column onto that complement and normalizes. A projected standard Gaussian column is standard Gaussian within the complement. The remaining steps therefore run the QR map γ of KZ on $q_{\sigma(1)}^{\perp}$. The rotation $\gamma(X)$ of a Gaussian matrix is Haar distributed, by the Gaussian-to-Haar construction of Rubio-Ramírez et al. (2010) in KZ's Proposition 4. The second factor of (C.14) is the uniform distribution over the orthonormal completions of $q_{\sigma(1)}$.

The second factor of (C.15) is the same uniform distribution. In part (a) the zeros constrain $q_{\sigma(1)}$ only. Given $q_{\sigma(1)}$, the conditional of $\mu(\cdot \mid Q(\theta))$ is the conditional of Haar measure given that column. Take any rotation that fixes $q_{\sigma(1)}$ and rotates its orthogonal complement. Left multiplication by it preserves Haar measure and maps the set of orthogonal matrices with first column $q_{\sigma(1)}$ onto itself. The conditional law of the completions is invariant under every such rotation. The uniform distribution is the only law with this invariance. ARRW report constant importance weights for their Algorithm 3 without zero restrictions, consistent with the equality of the two factors.

The factorizations (C.14) and (C.15) coincide factor by factor. The two measures are equal,

$$\mu_z(\cdot\,;\theta) \;=\; \mu\big(\cdot \mid Q(\theta)\big).$$

The restrictions of equal measures to $R(\theta)$ are equal. The volume element is constant. This proves part (a).

Part (b): the zeros load on two or more columns. The construction still draws each column uniformly on the unit sphere of its null space, as in (C.13). The measure $\mu_z(\cdot\,;\theta)$ is the law of the construction that ARRW's Theorem 3 describes. Their change-of-variable theorem for maps into lower-dimensional smooth manifolds gives $\mu_z(\cdot\,;\theta)$ a density proportional to $1/v(Q;\theta)$ with respect to the uniform (volume) measure on $Q(\theta)$. The factor $v(Q;\theta)$ is the volume element of the map

implied by their Algorithm 2 at fixed θ. The ratio of the two densities gives

$$\frac{d\,\mu\big(\cdot \mid Q(\theta)\big)}{d\,\mu_z(\cdot\,;\theta)}(Q) \;\propto\; v(Q;\theta).$$

The density relation holds on all of $Q(\theta)$ and hence on $R(\theta)$. Weighting each retained draw by $v(Q;\theta)$ recovers the conditional-Haar measure restricted to $R(\theta)$, the conditional $\pi_v(\cdot \mid \theta, y, SR)$ of (6). The weighting is necessary. As ARRW document, the distribution that their Algorithm 2 implies is not invariant to a reordering of the shocks. This proves part (b) and the proposition. □

Proposition 2 states that the conditional $\pi_X(\cdot \mid \theta, SR)$ induces the construction measure on $Q(\theta)$ through γ_z, not the uniform (conditional-Haar) measure on $Q(\theta)$. When the zeros load on a single column and the ordering builds that column first, the two coincide and need no correction. With zeros on more than one column, the volume element corrects the conditional of Q given θ, and its mean over unrestricted construction draws corrects the marginal of θ.[38] Proposition 2(b) states the correction as a density, known only up to a constant that depends on θ. The density relation holds on all of $Q(\theta)$, and the density integrates to one under the unrestricted construction measure. The density is $v(Q;\theta)/\mathrm{E}_z[v \mid \theta]$, where $\mathrm{E}_z[v \mid \theta]$ is the mean of v over unrestricted construction draws at θ. The next corollary states the conditional expectations at fixed θ, and Appendix C.5 combines them with the marginal.

Corollary 2 (Conditional-Haar expectations from weighted draws)**.** *Fix θ satisfying the conditions of Lemma 1 with $\kappa(\theta) > 0$. Let $E_z[\,\cdot \mid \theta, SR\,]$ denote the expectation under the restricted construction measure of Proposition 2, the distribution of the retained rotations at θ. For every function h*

[38] The weight is the part of the importance weight of ARRW's Algorithm 3 that survives at fixed θ. Their importance weight is the ratio of the target normal-generalized-normal density to the density of their Algorithm 2, a product of two volume elements. The element v_{f_h} converts between the structural and the orthogonal reduced-form parameterizations (their Proposition 1). The element $v_{(g\circ f_h)|Z}$ corrects their Algorithm 2 on the zero-restricted set (their Theorem 3). Only the second element arises here. The structural update works in the structural parameterization directly. The sampled and target distributions differ only through the construction measure, which moves the conditional of Q given θ and, through $\kappa(\theta)$, the marginal of θ. The volume element $v(Q;\theta)$ of Proposition 2 is their $v_{(g\circ f_h)|Z}$ at fixed θ.

integrable under $\pi_v(\cdot \mid \theta, y, SR)$*,*

$$\int h(\theta, Q)\, \pi_v(dQ \mid \theta, y, SR) \;=\; \frac{E_z\big[v(Q;\theta)\, h(\theta, Q) \mid \theta, SR\big]}{E_z\big[v(Q;\theta) \mid \theta, SR\big]}. \tag{C.16}$$

Proof of Corollary 2. Write $\tilde{\mu}_z(dQ \mid \theta)$ for the restricted construction measure (C.12). By Proposition 2(b), the conditional of (6) has a density proportional to $v(Q;\theta)$,

$$\pi_v(dQ \mid \theta, y, SR) \;=\; c(\theta)\, v(Q;\theta)\, \tilde{\mu}_z(dQ \mid \theta),$$

with a constant $c(\theta) > 0$. For every h integrable under the conditional,

$$\int h(\theta, Q)\, \pi_v(dQ \mid \theta, y, SR) \;=\; c(\theta) \int h(\theta, Q)\, v(Q;\theta)\, \tilde{\mu}_z(dQ \mid \theta) \;=\; c(\theta)\, E_z\big[v(Q;\theta)\, h(\theta, Q) \mid \theta, SR\big].$$

The choice $h = 1$ makes the left side one and pins down the constant, $c(\theta) = 1/E_z[v(Q;\theta) \mid \theta, SR]$. Substituting the constant into the display gives (C.16). □

The rotation draw. With the target of Definition 1 in place, the rotation draw keeps θ fixed and applies the column-wise ESS update of KZ (Proposition 4) through γ_z. The next proposition establishes its validity on $\mathcal{Q}(\theta)$.

Proposition 3 (Column-wise ESS on $\mathcal{Q}(\theta)$)**.** *Fix θ satisfying the conditions of Lemma 1 with $\kappa(\theta) > 0$, and let $\pi_X(\cdot \mid \theta, SR)$ denote the conditional distribution of X given θ in (C.8). The column-wise ESS update proposes*

$$x_j(\vartheta) = x_j \cos\vartheta + \tau_j \sin\vartheta, \qquad \tau_j \sim \mathcal{N}(0, I_n),$$

where x_j denotes the j-th column of the current $X = [x_1, \ldots, x_n]$, and checks admissibility through γ_z. It has the following properties.

(a) Almost surely, it produces proposals whose rotations satisfy every zero restriction exactly

along the entire ellipse, and the zeros never reject a proposal.

(b) Each column update terminates almost surely.

(c) It leaves $\pi_X(\cdot \mid \theta, SR)$ invariant.

Proof of Proposition 3. We prove the three parts in turn.

Part (a): the zeros never reject a proposal. Fix a column j, and let $X^\star(\vartheta)$ denote the current X with column j replaced by $x_j(\vartheta)$. The sampler evaluates admissibility only through the candidate rotation $\gamma_z\big(X^\star(\vartheta);\theta\big)$. By Lemma 1(i), every candidate satisfies the zero restrictions exactly wherever the construction is defined. The construction fails only where some projected column in (C.6) equals the zero vector, an exceptional set of inputs with Gaussian probability zero.

Lemma 1(i) concerns a fresh Gaussian input. The candidates along the ellipse are not fresh Gaussian inputs. A Fubini argument extends the statement to the ellipse. Fix an angle with $\sin\vartheta \neq 0$. Given the current state, the candidate column $x_j(\vartheta) = x_j\cos\vartheta + \tau_j\sin\vartheta$ is a full-support Gaussian draw, and the remaining columns arise from draws with densities. The candidate matrix $X^\star(\vartheta)$ lies in the exceptional set with probability zero. By Fubini's theorem over the uniformly drawn angle, with nonnegative integrands, the exceptional angles form a null set almost surely. The zeros hold along the entire ellipse off this null set, and the zeros never reject a proposal. The slice step retains only angles at which the sign and narrative indicator holds, exactly as in KZ's HARS.

Part (b): almost-sure termination. We argue in three steps.

Step 1: the angle $\vartheta = 0$ is admissible. The bracket of the slice search always contains $\vartheta = 0$. At $\vartheta = 0$ the candidate matrix is the current X. By Lemma 1(iii), γ_z maps the current X to the current rotation. The chain retains only admissible states, and the current rotation is admissible.

Step 2: $\vartheta \mapsto \gamma_z\big(X^\star(\vartheta);\theta\big)$ is continuous at $\vartheta = 0$ almost surely. Each construction step depends on the stacked matrix $M_j(\theta)$ in (C.5), abbreviated M_j, only through the orthogonal projector P_j onto its null space $N_j(\theta)$. The projector P_j does not depend on the choice of orthonormal basis.

Wherever M_j has full row rank,

$$P_j \;=\; I_n - M_j'\big(M_j M_j'\big)^{-1} M_j, \tag{C.17}$$

a continuous function of M_j. The proposition fixes θ under the maintained conditions of Lemma 1. Those conditions give M_j full row rank at every orthonormal collection of preceding columns whose columns satisfy their zero restrictions. Along the ellipse, the construction steps $1, \ldots, j-1$ do not move, and only the construction steps from j onward vary, by Lemma 1(ii). Off the exceptional set of part (a), each construction step is the composition of $M_j \mapsto P_j$ through (C.17) and the normalization in (C.6). Both pieces are continuous at almost every angle. The composition of the n steps is continuous at $\vartheta = 0$ almost surely.

Step 3: the bracket-shrinking rule terminates. The sign and narrative restrictions are strict inequalities. The admissible set of rotations is open. By Steps 1 and 2, the admissible angular set contains an open neighborhood of $\vartheta = 0$ and is nonempty. The bracket-shrinking rule reaches the admissible angular set in finitely many steps almost surely. Steps 1 and 3 make explicit the nonemptiness assertion in the proof of KZ's Proposition 4. That assertion suffices in their setting. The QR map γ is continuous at every full-rank X. Step 2 supplies the analogous continuity for γ_z, the only new ingredient.

Part (c): invariance. The target (C.8) is a standard Gaussian density multiplied by a nonnegative likelihood factor, here the admissibility indicator. The Gaussian density factorizes across columns, $\varphi(X) \propto \exp\big(-\tfrac{1}{2}\sum_{j=1}^{n}\|x_j\|^2\big)$, and the full conditional of a single column is

$$p(x_j \mid x_{-j}, \theta, SR) \;\propto\; \mathbf{1}\big\{SR\big(\theta, \gamma_z(X;\theta)\big) > 0\big\}\exp\big(-\tfrac{1}{2}\|x_j\|^2\big),$$

a Gaussian density times an indicator, as in KZ's Proposition 4. Elliptical slice sampling is exact for targets of this form (Murray et al., 2010). Part (a) reduces the indicator along every ellipse to a sign-and-narrative indicator. Part (b) supplies the almost-sure termination. The argument of KZ's Proposition 4 applies. Each column update samples from the admissible angular set and satisfies

detailed balance for the full conditional $p(x_j \mid x_{-j}, \theta, SR)$. The systematic scan over the columns composes invariant full-conditional updates. The scan leaves $\pi_X(\cdot \mid \theta, SR)$ invariant. The retained states are serially dependent draws from $\pi_X(\cdot \mid \theta, SR)$. This proves the proposition. □

Lemma 1(ii) permits the column-wise ESS update on $\bar{Q}(\theta)$. The zeros hold under every elliptical proposal by construction, and the slice check evaluates only sign and narrative admissibility. The angle-zero proposal is the current admissible state. The admissible angular set is nonempty, and the bracket-shrinking rule terminates (Proposition 3(b)).

The validity of Algorithm 1 step by step. Each step of Algorithm 1 leaves the target (C.9) invariant. By Proposition 3(c), the ESS updates of Steps 1 and 4 leave $\pi_X(\cdot \mid \theta, SR)$ invariant at fixed θ, and an update that leaves the conditional of X invariant leaves the target invariant. By Proposition 1, Step 2 leaves the structural target (C.10) invariant at fixed X and s_0, the conditional of θ under the target. The latent scales do not enter $SR(\theta, Q)$, and Step 3 draws them from their Gibbs conditional under the target, that of BPSS. A composition of updates that each leave the target invariant leaves it invariant. Step 2 never uses the conditional shortcut for z_+, which would have led to the incorrect marginal target of z_0.[39] The retained pairs (θ, X) are serially dependent draws from the target of Definition 1. At each θ the retained rotations follow the restricted construction measure of Proposition 2, and the weight $v(Q;\theta)/\mathrm{E}_z[v \mid \theta]$ of Appendix C.5 turns them into estimates under (6).

C.5 Posterior Inference

Let θ_d denote the d-th retained draw of θ, $Q^{(d,r)} = \gamma_z(X^{(d,r)};\theta_d)$ its r-th retained rotation, and $v_{d,r} = v(Q^{(d,r)};\theta_d)$ the volume element at that rotation. At each retained θ_d, draw M unrestricted Gaussian matrices $G^{(1)}, \ldots, G^{(M)}$ and estimate the mean of the volume element under the unrestricted

[39] The natural alternative holds $z_0 = (A_0, \Lambda)$ fixed at a candidate that fails the check and redraws only $z_+ = A_+$ until the check passes. This targets the wrong distribution. Under the sign-restricted posterior, the marginal of z_0 has the admissibility weight $\omega_0(z_0; s_0, Q)$, the probability that a fresh Gibbs draw of A_+ completes the pair admissibly. Redrawing A_+ until success gives every z_0 a completion. It strips this factor and over-represents structural values whose admissible completions are rare (KZ, Theorem 1).

construction,

$$\widehat{\mathrm{E}}_z[v \mid \theta_d] \;=\; \frac{1}{M}\sum_{m=1}^{M} v\big(\gamma_z(G^{(m)};\theta_d);\theta_d\big).$$

Retained draws that share θ share the estimate. Each retained rotation receives the weight $w_{d,r} = v_{d,r}/\widehat{\mathrm{E}}_z[v \mid \theta_d]$, and the estimator of the posterior mean of a function h of the impulse responses is

$$\widehat{E}[h] \;=\; \frac{\sum_{d=1}^{D}\sum_{r=1}^{N_{\text{ess}}} w_{d,r}\, h\big(\theta_d, Q^{(d,r)}\big)}{\sum_{d=1}^{D}\sum_{r=1}^{N_{\text{ess}}} w_{d,r}}, \tag{C.18}$$

where D is the number of retained draws of θ. The reported quantiles apply the same weights, normalized over all retained draws. The weight is the importance weight of (6) against the distribution the chain samples. Under (C.9) the pair (θ, Q) has the distribution $\pi(\theta \mid y)\,\mathbf{1}\{Q \in R(\theta)\}\,\mu_z(dQ;\theta)\,d\theta$ up to a constant. By Proposition 2(b), $\mu(dQ \mid \mathcal{Q}(\theta)) = \big[v(Q;\theta)/\mathrm{E}_z[v \mid \theta]\big]\,\mu_z(dQ;\theta)$ on all of $\mathcal{Q}(\theta)$, and multiplying the sampled distribution by $v(Q;\theta)/\mathrm{E}_z[v \mid \theta]$ gives (6). The estimator is consistent for expectations under (6) as D and M grow. Any θ-dependent rescaling of the volume element cancels in the weight. In the single-column case of Proposition 2(a), the volume element is constant at each θ, every weight equals one, and (C.18) reduces to the unweighted average of KZ. In the estimates $M = 500$.

D Estimation Diagnostics

The baseline estimates use one chain of 10,000 retained draws after 1,000 burn-in sweeps. Three further chains of the same specification give the same signs (Section 5.1). The MTM step draws $K = 300$ candidates from a random walk with covariance $\Sigma_{\text{rw}} = 0.05H$, where H is the inverse Hessian of the log posterior at its mode. Each sweep runs $m = 5$ ESS updates before the structural update and keeps $N_{\text{ess}} = 1$ rotation. The estimator of Appendix C.5 uses $M = 500$ unrestricted rotations at each value of θ.

Sampler diagnostics. The MTM step moves θ in 1,436 of the 10,000 sampling sweeps. No candidate evaluation fails in burn-in or sampling. The weights $v(Q;\theta)/\widehat{\mathrm{E}}_z[v \mid \theta]$ stay near uniform, with an effective sample size share of 0.970 over the 10,000 draws (ARRW, Section 7.3).

Residual diagnostics. Figure D.1 plots the six equations Figure 4 leaves out. Table D.1 repeats that estimation with the four variance regimes of Table 1 in place of the single regime and standardizes each structural residual by the square root of its estimated regime variance λ_{i,s_t}. Under Gaussian errors these standardized residuals would show excess kurtosis near zero.[40] Pooled over the nine equations and the full sample they show 4.7, and the SOMA equation alone shows 20.3.

The two-step procedure of Brunnermeier et al. (2021) fits 5.46 degrees of freedom to these residuals, and the estimation uses that value. The fitted distribution matches the residuals on how often extreme residuals occur. In a sample of 1,854 observations it puts the expected number of residuals beyond six standard deviations at 0.8, and the sample holds one. Beyond five standard deviations it expects 2.0, and the sample holds three.

Table D.2 reports the estimated variance ratios $\omega_{i,s} = \lambda_{i,s}/\lambda_{i,1}$ over the retained draws.

The regime profiles differ across equations. The two-year yield variance stays above its Regime 1 level in all three later regimes. The SOMA variance rises fifteenfold in Regime 3 and falls below a

[40] Kurtosis is the expectation of the fourth power of a standardized random variable and equals three for a Gaussian. Excess kurtosis subtracts this benchmark. Positive values indicate that extreme observations occur more often than under a Gaussian. A Student-t with $\delta > 4$ degrees of freedom has excess kurtosis $6/(\delta - 4)$.

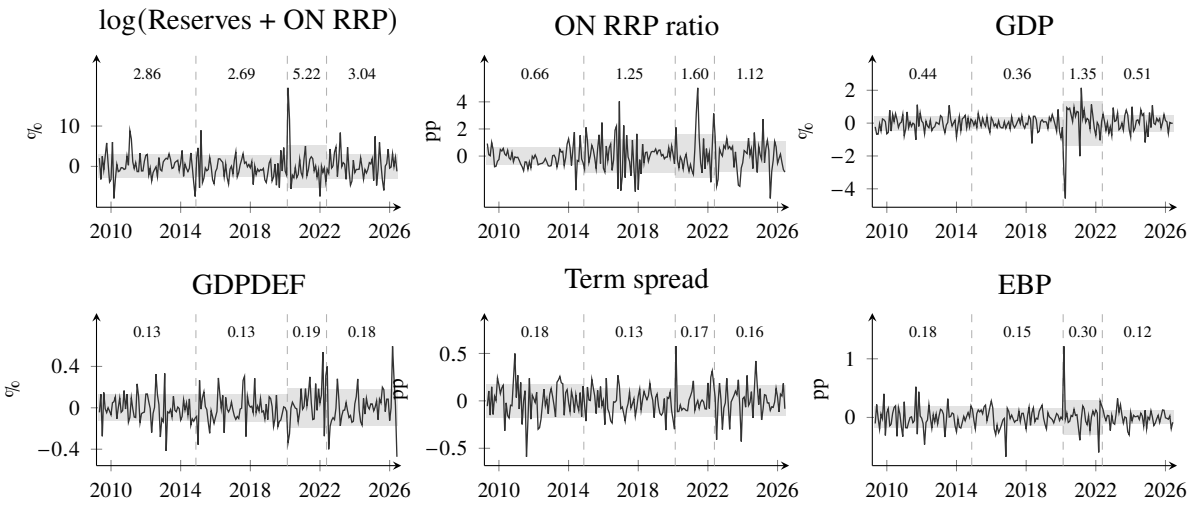


Figure D.1: Reduced-form residuals under a single variance regime, the remaining six equations.
Note. The equations Figure 4 leaves out, on the same estimation and with the same bands and subsample standard deviations. Units are percent for the log levels and percentage points for the ratio, the term spread, and the EBP.

Table D.1: Regime-standardized residuals under four variance regimes.

	Excess kurtosis	max $\lvert z \rvert$
Two-year yield	0.7	3.4
SOMA	20.3	9.1
log(Reserves + ON RRP)	0.6	3.4
ON RRP ratio	1.0	3.6
GDP	1.5	3.9
GDPDEF	0.8	3.7
Term spread	3.9	5.8
EBP	1.4	4.1
IOR − 1M bill	10.7	5.4
Pooled	4.7	9.1

Note: The table estimates equation (1) with the four variance regimes of Table 1, Gaussian errors, no identification restrictions, and the prior of Section 3.1. It reports the excess kurtosis and the largest absolute value of the structural residuals standardized by $\sqrt{\lambda_{i,s_t}}$. Entries are posterior medians. The pooled row stacks all equations over the full sample.

tenth of Regime 1 in the other two. The SOMA equation and the equation of IOR less the bill rate carry the widest bands, spanning a factor of 2.7 and 2.2 in at least one regime. The bands of the other seven equations stay inside a factor of 1.25. The variance of the equation of IOR less the bill rate rises twenty-fourfold in Regime 4.

Table D.2: Estimated variance ratios by regime, $\omega_{i,s} = \lambda_{i,s}/\lambda_{i,1}$.

Equation	Regime 2	Regime 3	Regime 4
Two-year yield	1.97 [1.96, 1.99]	3.33 [3.29, 3.38]	6.58 [6.51, 6.62]
SOMA	0.08 [0.06, 0.12]	15.09 [14.87, 15.37]	0.08 [0.05, 0.13]
log(R + RRP)	0.85 [0.84, 0.86]	1.02 [1.00, 1.03]	0.73 [0.73, 0.74]
ON RRP ratio	4.55 [4.36, 4.70]	7.57 [7.28, 7.84]	5.19 [4.89, 5.38]
GDP	0.36 [0.35, 0.36]	1.82 [1.81, 1.83]	0.30 [0.29, 0.35]
GDPDEF	0.93 [0.93, 0.94]	1.66 [1.64, 1.68]	1.93 [1.91, 1.95]
Term spread	0.54 [0.53, 0.54]	2.45 [2.43, 2.48]	0.81 [0.80, 0.83]
EBP	0.73 [0.73, 0.73]	1.46 [1.46, 1.48]	0.29 [0.28, 0.31]
IOR − 1M bill	2.89 [2.28, 3.22]	0.33 [0.22, 0.47]	24.50 [16.95, 28.73]

Notes: Entries are posterior medians with 68% credible bands in brackets over the retained draws, weighted by the volume weights. Rows index the structural equations, labeled by the variable on which each equation is normalized. Regime 1 equals one by normalization. Regime dates follow Table 1.

E Impulse Responses to the Demand, Supply, MP, and EBP Shocks

Figure E.1 reports the impulse responses to the demand, supply, MP, and EBP shocks under the baseline identification.

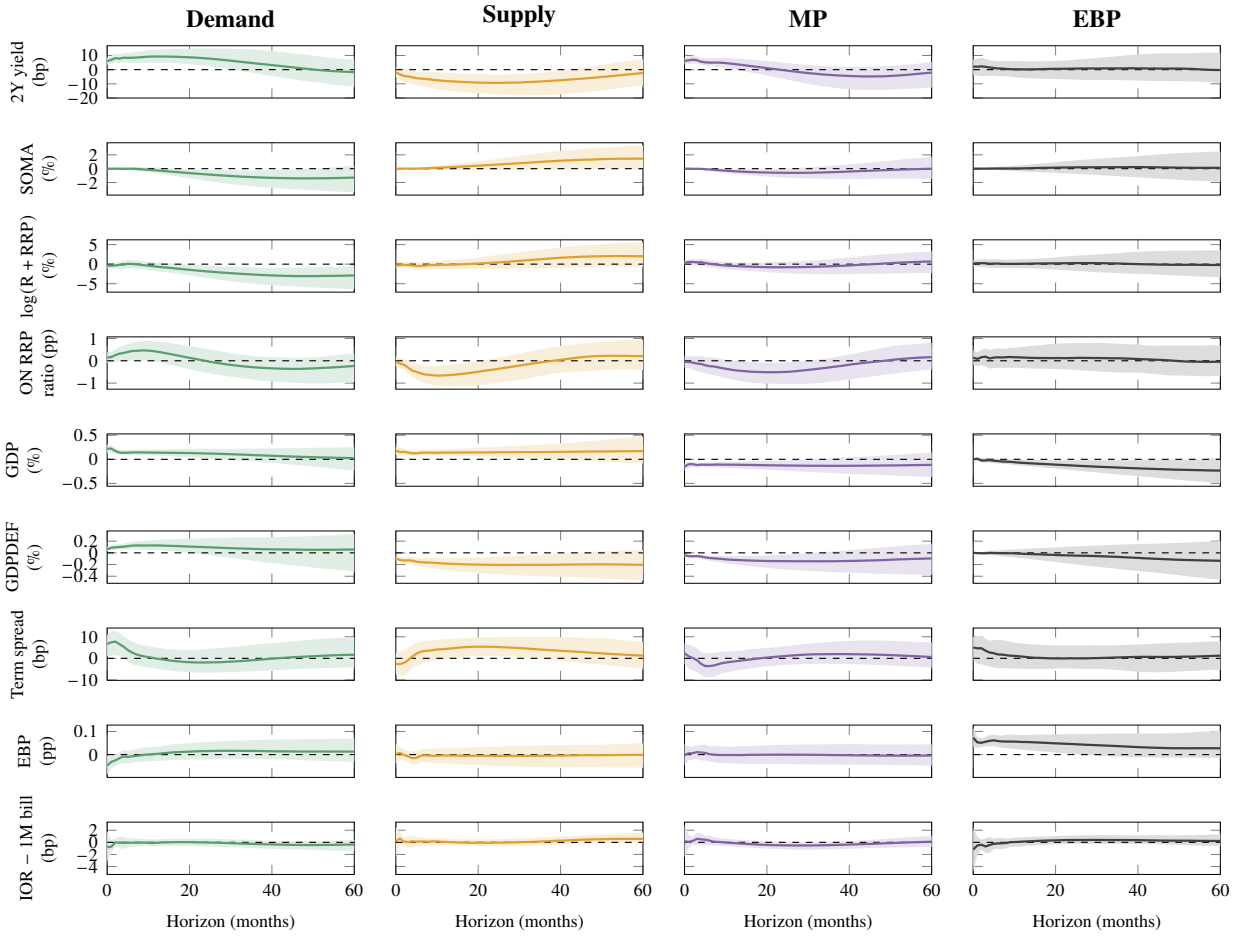


Figure E.1: Impulse responses to the demand, supply, MP, and EBP shocks.

Note. Columns report the demand, supply, MP, and EBP shocks, and rows report the nine variables with their units. I scale each shock to one standard deviation. Under Assumption 2 the SOMA impact response to all four shocks is zero, and the one-percent-SOMA normalization of the balance-sheet figures does not apply. Solid lines are posterior medians and shaded regions are 68% credible bands over the retained draws. Each row shares one vertical range across the four shocks.

F Posterior Probabilities of the Response Signs

Table F.1 reports the posterior probability that GDP and the deflator fall and the EBP rises after each balance-sheet contraction.

Table F.1: Posterior probabilities of the signs of the responses to the balance-sheet shocks.

		Horizon (months)					
Variable	Shock	6	12	24	36	48	60
Probability that GDP and GDPDEF fall and the EBP rises							
GDP	RFBS	0.67	0.85	0.92	0.94	0.95	0.95
	OFBS	0.84	0.85	0.86	0.81	0.72	0.64
GDPDEF	RFBS	0.48	0.61	0.75	0.83	0.89	0.90
	OFBS	0.61	0.73	0.66	0.55	0.49	0.48
EBP	RFBS	0.97	0.99	0.98	0.97	0.94	0.91
	OFBS	0.69	0.67	0.68	0.62	0.55	0.51

Note: Each shock is a contraction normalized to a one-percent impact decline in SOMA holdings. I evaluate the responses at the regime average of the relative variances. Entries are the shares of the retained draws, weighted by the volume weights, in which the response has the stated sign. The sign restriction on SOMA at impact fixes the sign of the normalization in every draw, and the shares do not depend on the size of the shock. A share above 0.84 means that the 68% credible band of Figures 6 and 7 excludes zero at that horizon. A share above 0.95 means that the 90% credible band excludes zero.

G The RFBS Shock with the EBP Held at Zero

Figure G.1 reports the responses of the other six variables in the counterfactual of Section 5.2.1. Figure G.2 repeats Figure 8 with the penalized offset, which keeps the EBP response near zero and keeps all 10,000 draws.

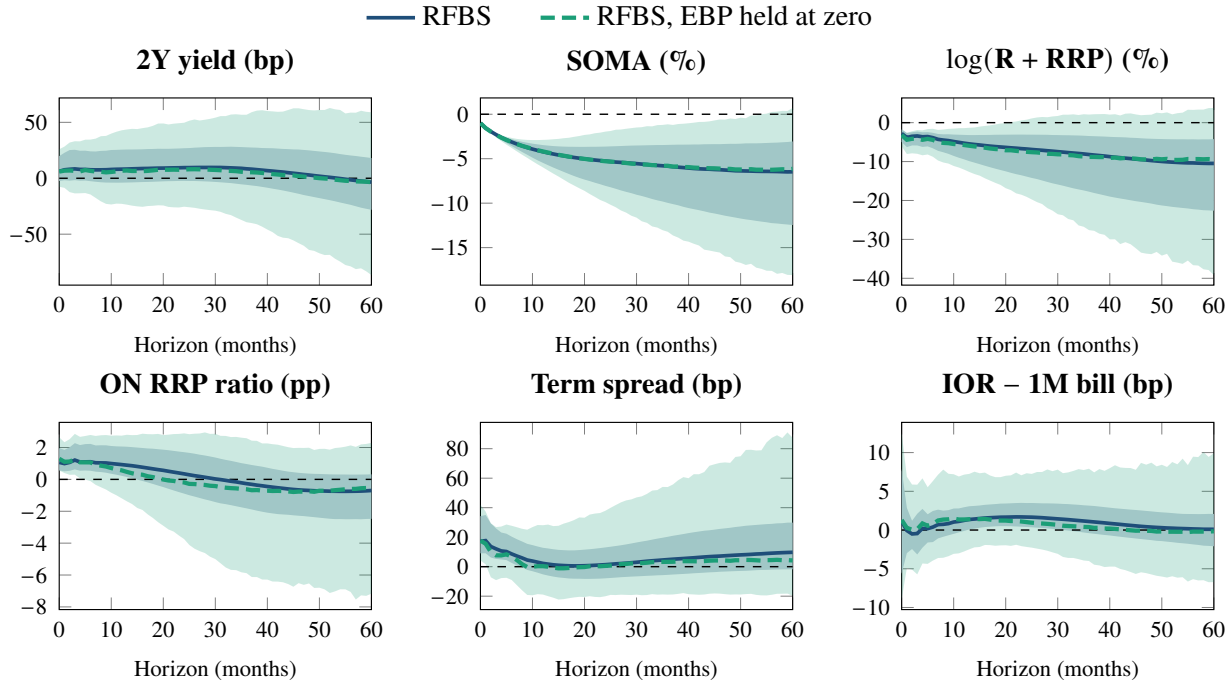


Figure G.1: Responses of the other variables to the RFBS shock with the EBP held at zero.

Note. Responses to the RFBS shock, normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid blue lines and blue regions show the RFBS shock alone. Dashed green lines and green regions show the RFBS shock combined with the EBP shocks that keep the EBP response at zero in every month. Lines are posterior medians and shaded regions are 68% credible bands over the 9,406 draws in which those shocks stay finite. Vertical axes are in basis points for the two-year yield, the term spread, and IOR less the one-month bill rate. They are in percent for SOMA and log(Reserves + ON RRP) and in percentage points for the ON RRP ratio.

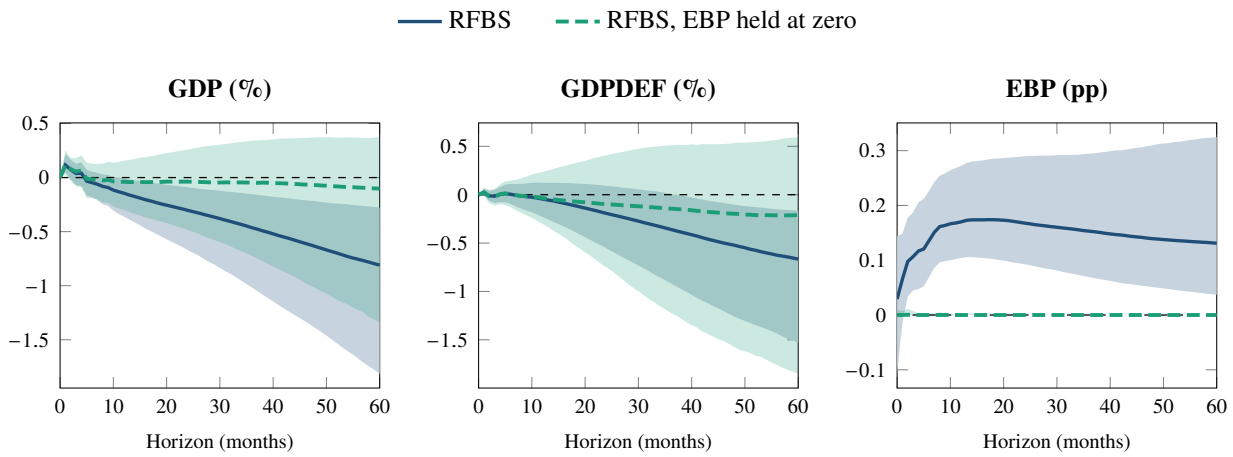


Figure G.2: Responses to the RFBS shock with the EBP held near zero by a penalized offset.

Note. As Figure 8, with the offsetting EBP shocks chosen by the penalized offset of Section 5.2.1 and the bands computed over all 10,000 draws. Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

H Robustness of the Responses to the Balance-Sheet Shocks

Figures H.1 and H.2 plot the responses of GDP, the deflator, and the EBP to both balance-sheet shocks with the TGA and with the bill supply shock of Section 5.3.1. Figures H.3 to H.7 plot the responses to both balance-sheet shocks against the baseline, one figure for each specification of Section 5.3.2. They follow the format of Figure 9.

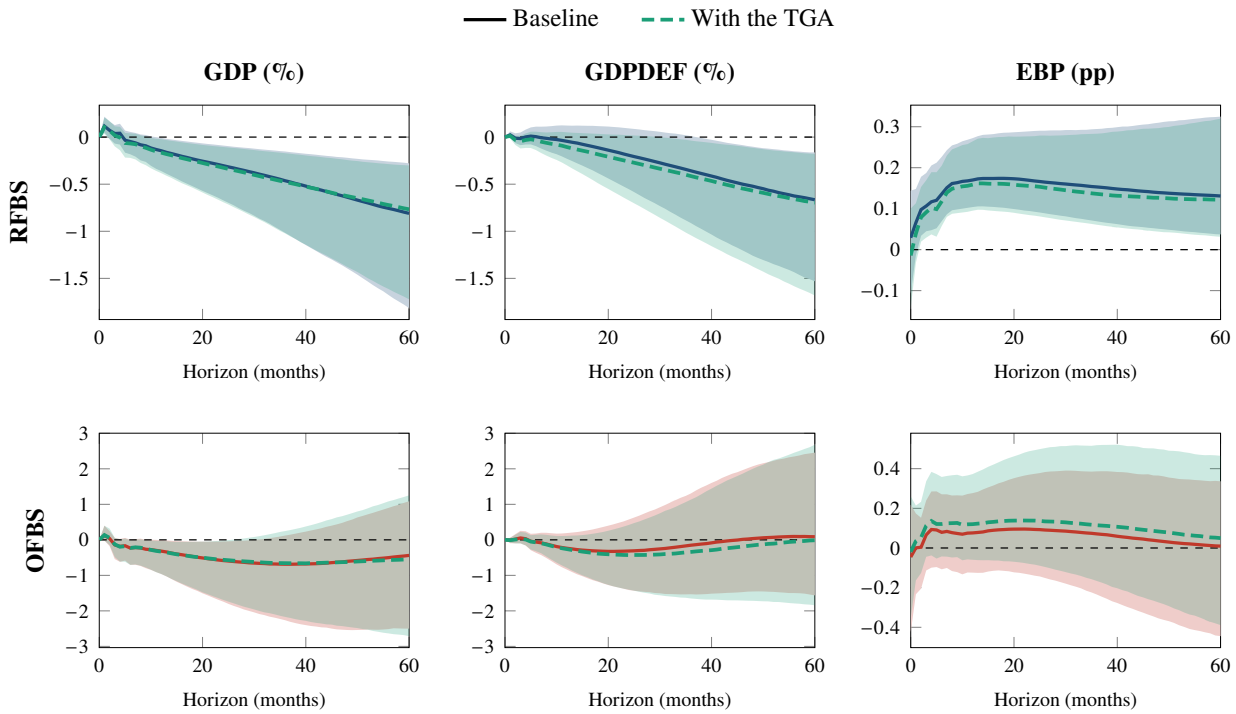


Figure H.1: Responses to the two balance-sheet shocks with the TGA.

Note. Rows report the RFBS and OFBS shocks, each normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid lines and shaded regions in the color of each shock are the baseline posterior medians and 68% credible bands of Figures 6 and 7. Dashed green lines and green regions are the posterior medians and 68% credible bands of the baseline system with the log of the TGA as a tenth variable. They come from one chain of 10,000 draws. Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

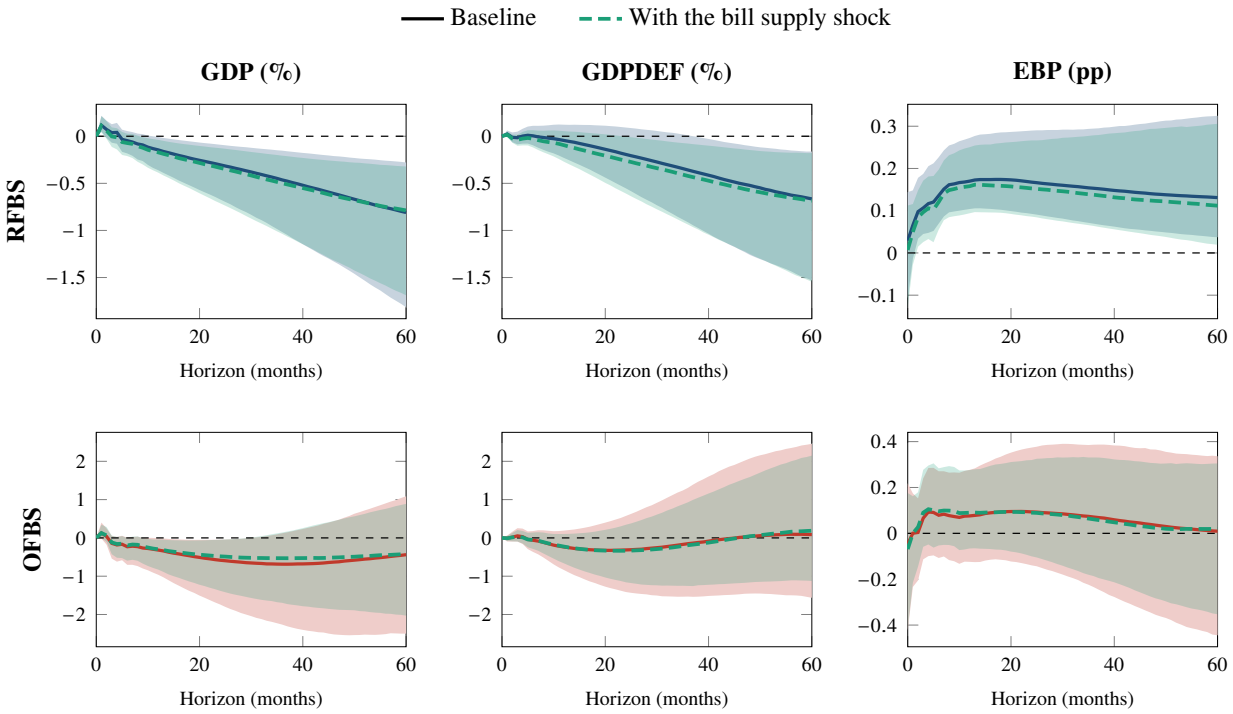


Figure H.2: Responses to the two balance-sheet shocks with the bill supply shock.

Note. Rows report the RFBS and OFBS shocks, each normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid lines and shaded regions in the color of each shock are the baseline posterior medians and 68% credible bands of Figures 6 and 7. Dashed green lines and green regions are the posterior medians and 68% credible bands of the ten-variable system with the TGA and the Treasury bill supply shock. They come from one chain of 10,000 draws. Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

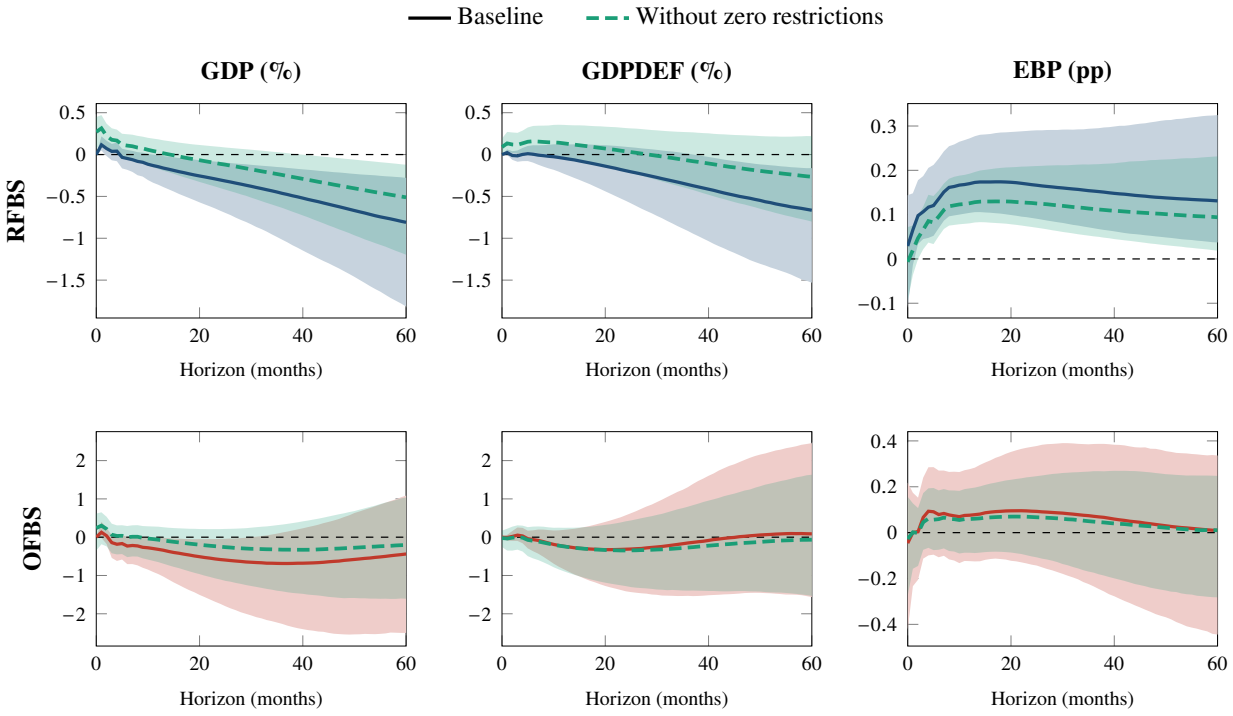


Figure H.3: Responses to the two balance-sheet shocks without zero restrictions.

Note. Rows report the RFBS and OFBS shocks, each normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid lines and shaded regions in the color of each shock are the baseline posterior medians and 68% credible bands of Figures 6 and 7. Dashed green lines and green regions are the posterior medians and 68% credible bands of the alternative. It is the baseline system without the zero restrictions of Assumptions 1, 2, and 3. Without zero restrictions every draw carries the same weight. Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

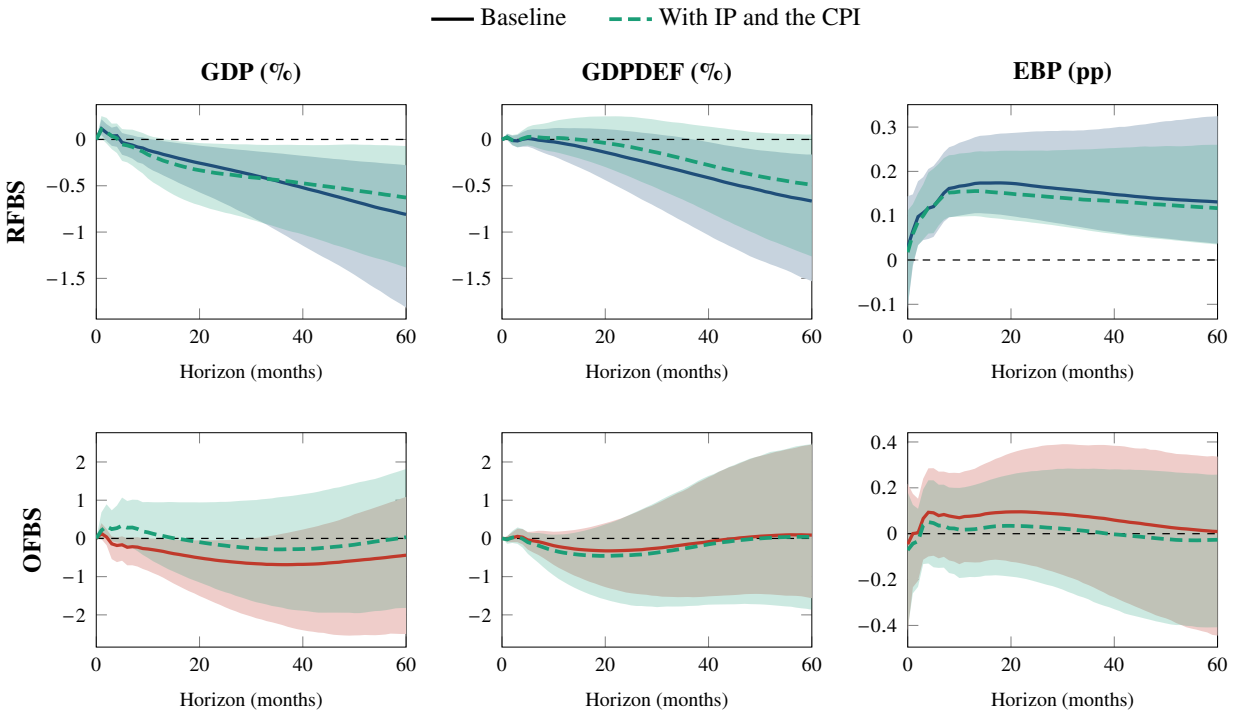


Figure H.4: Responses to the two balance-sheet shocks with industrial production and the CPI.

Note. Rows report the RFBS and OFBS shocks, each normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid lines and shaded regions in the color of each shock are the baseline posterior medians and 68% credible bands of Figures 6 and 7. Dashed green lines and green regions are the posterior medians and 68% credible bands of the alternative. It is the baseline system with log industrial production (IP) and the log CPI in place of log GDP and the log GDP deflator. The first two panels of each row plot GDP and GDPDEF for the baseline and IP and the CPI for the alternative. Vertical axes are in percent for GDP, GDPDEF, IP, and the CPI and in percentage points for the EBP.

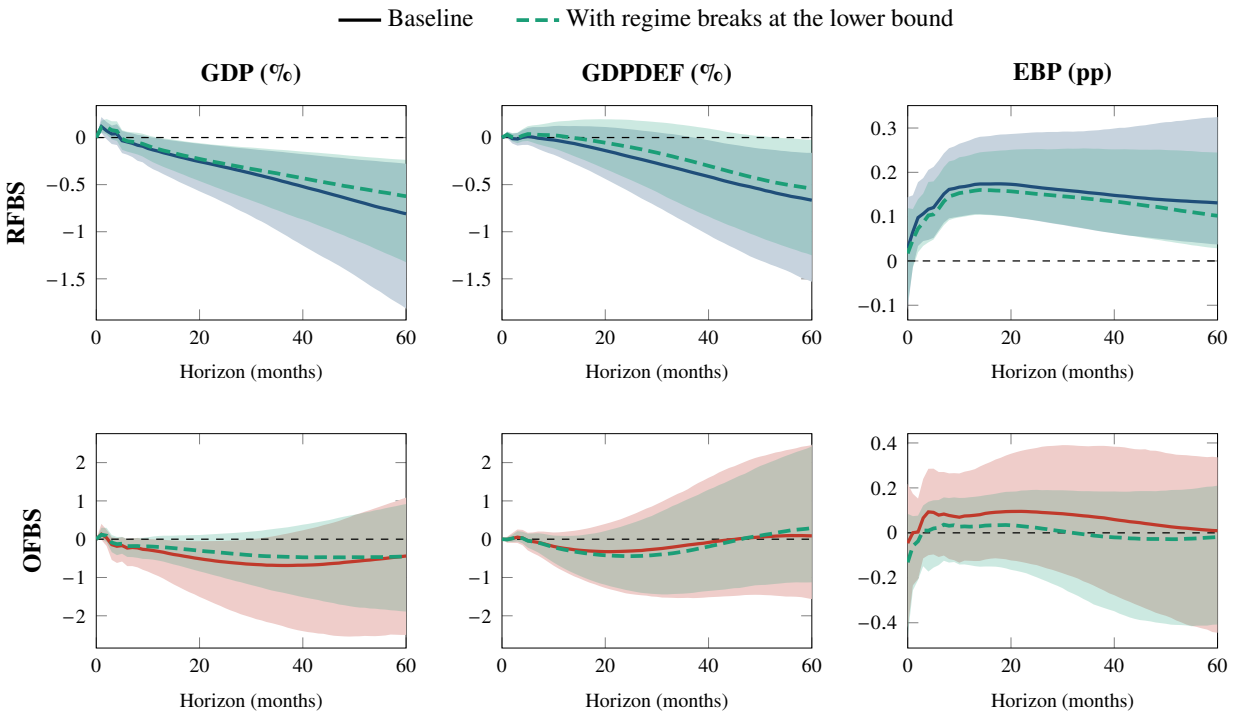


Figure H.5: Responses to the two balance-sheet shocks with regime breaks at the lower bound.
Note. Rows report the RFBS and OFBS shocks, each normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid lines and shaded regions in the color of each shock are the baseline posterior medians and 68% credible bands of Figures 6 and 7. Dashed green lines and green regions are the posterior medians and 68% credible bands of the alternative. It is the baseline system with the regime breaks at 2015:M11, 2020:M3, and 2022:M2 in place of the balance-sheet episodes. Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

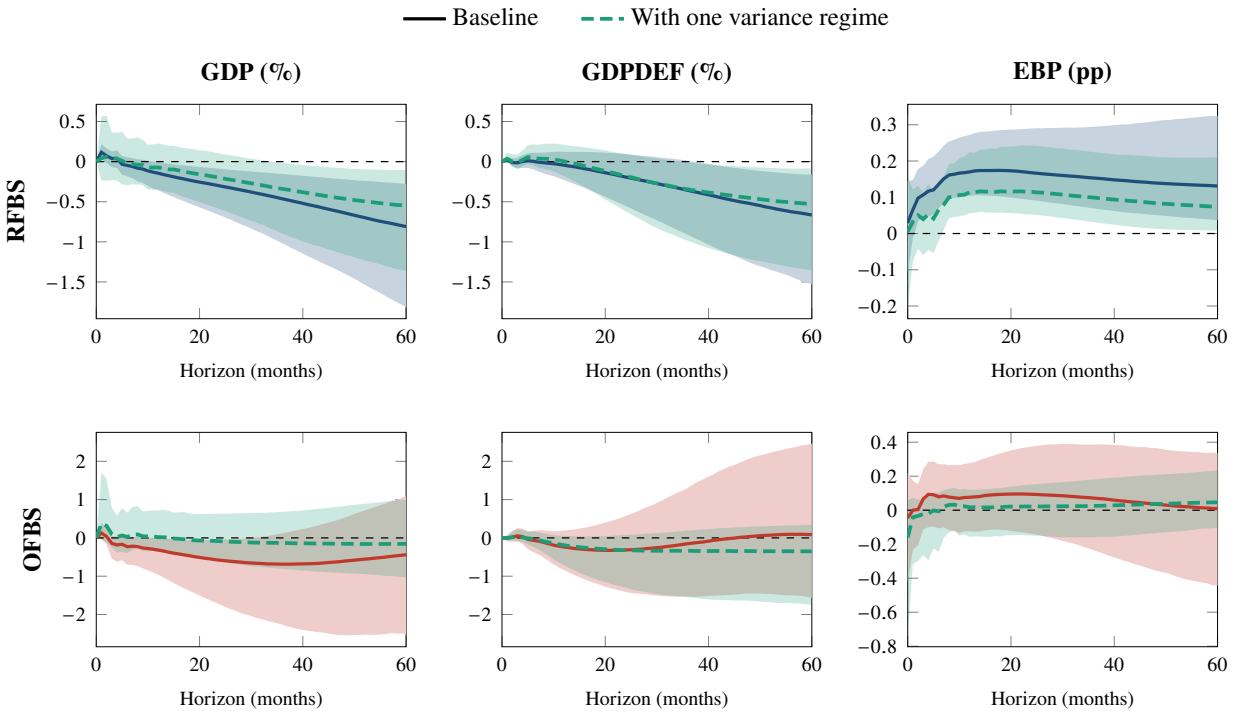


Figure H.6: Responses to the two balance-sheet shocks with one variance regime.

Note. Rows report the RFBS and OFBS shocks, each normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid lines and shaded regions in the color of each shock are the baseline posterior medians and 68% credible bands of Figures 6 and 7. Dashed green lines and green regions are the posterior medians and 68% credible bands of the alternative. It is the baseline system with one variance regime (homoskedasticity) in place of four variance regimes (heteroskedasticity). Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

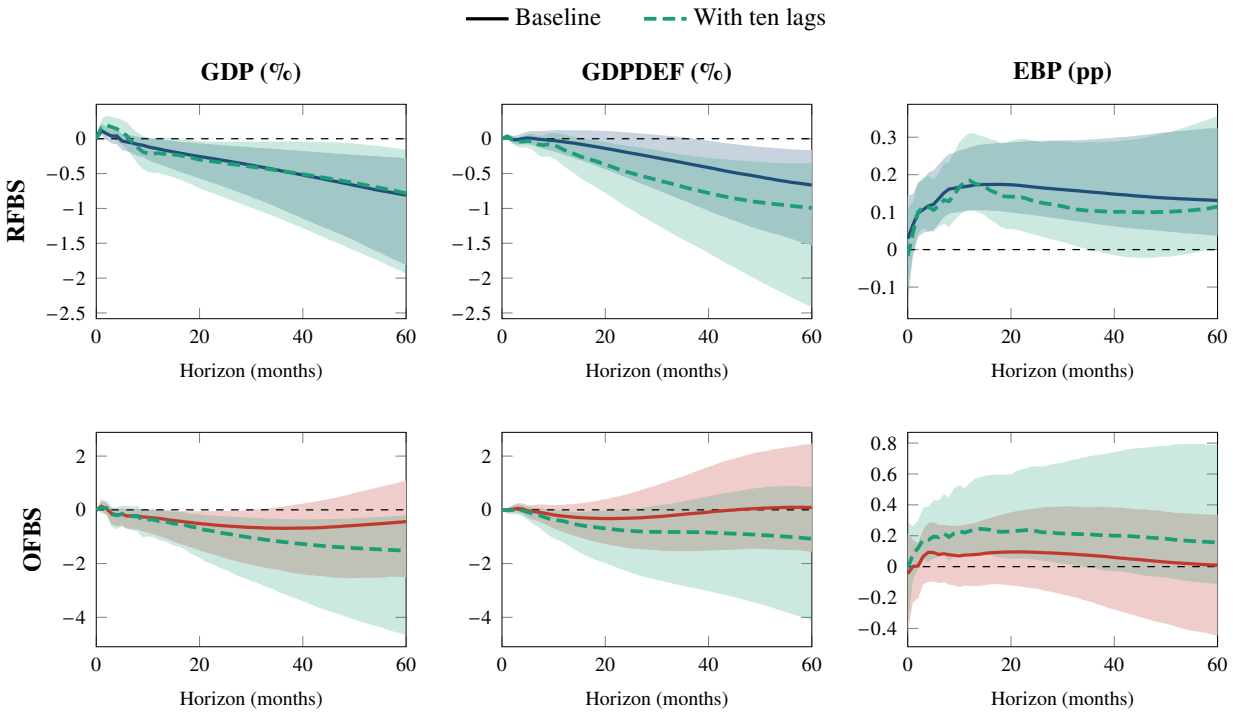


Figure H.7: Responses to the two balance-sheet shocks with ten lags.

Note. Rows report the RFBS and OFBS shocks, each normalized to a one-percent impact decline in SOMA holdings and shown in the contractionary direction. Solid lines and shaded regions in the color of each shock are the baseline posterior medians and 68% credible bands of Figures 6 and 7. Dashed green lines and green regions are the posterior medians and 68% credible bands of the alternative. It is the baseline system with ten lags in place of six. The alternative pools three chains of 10,000 draws each. Vertical axes are in percent for GDP and GDPDEF and in percentage points for the EBP.

I Variance Decomposition at Three Years

Table I.1 reports the shares of Figure 10 at 36 months with their 68% intervals.

Table I.1: Forecast-error variance decomposition.

	Share of forecast-error variance at 36 months, percent					
	Demand	Supply	MP	RFBS	OFBS	EBP
GDP	7.0 [3.1, 16.9]	10.0 [3.2, 23.2]	7.0 [2.1, 18.8]	4.9 [1.1, 13.6]	3.6 [0.7, 11.5]	4.3 [0.7, 15.6]
GDPDEF	8.7 [2.3, 26.2]	20.3 [5.5, 42.0]	7.1 [1.9, 22.5]	3.4 [0.7, 11.6]	3.7 [0.7, 13.1]	4.0 [0.8, 15.7]
EBP	5.0 [1.6, 15.0]	4.0 [1.1, 13.5]	5.0 [1.4, 15.7]	19.7 [7.3, 37.0]	4.3 [1.2, 12.6]	14.4 [4.6, 32.9]

Notes: Shares of the forecast-error variance in percent at a horizon of 36 months, under the average transmission, with the relative variances at their average of one across regimes. The first row of each pair is the posterior median and the second is the 68% credible interval, both over the retained draws. The six labeled shocks appear here, and the three unlabeled shocks account for the remainder.

J Historical Decompositions for All Shocks

Figure J.1 reports the decomposition of Figure 11 for the six variables that figure does not show.

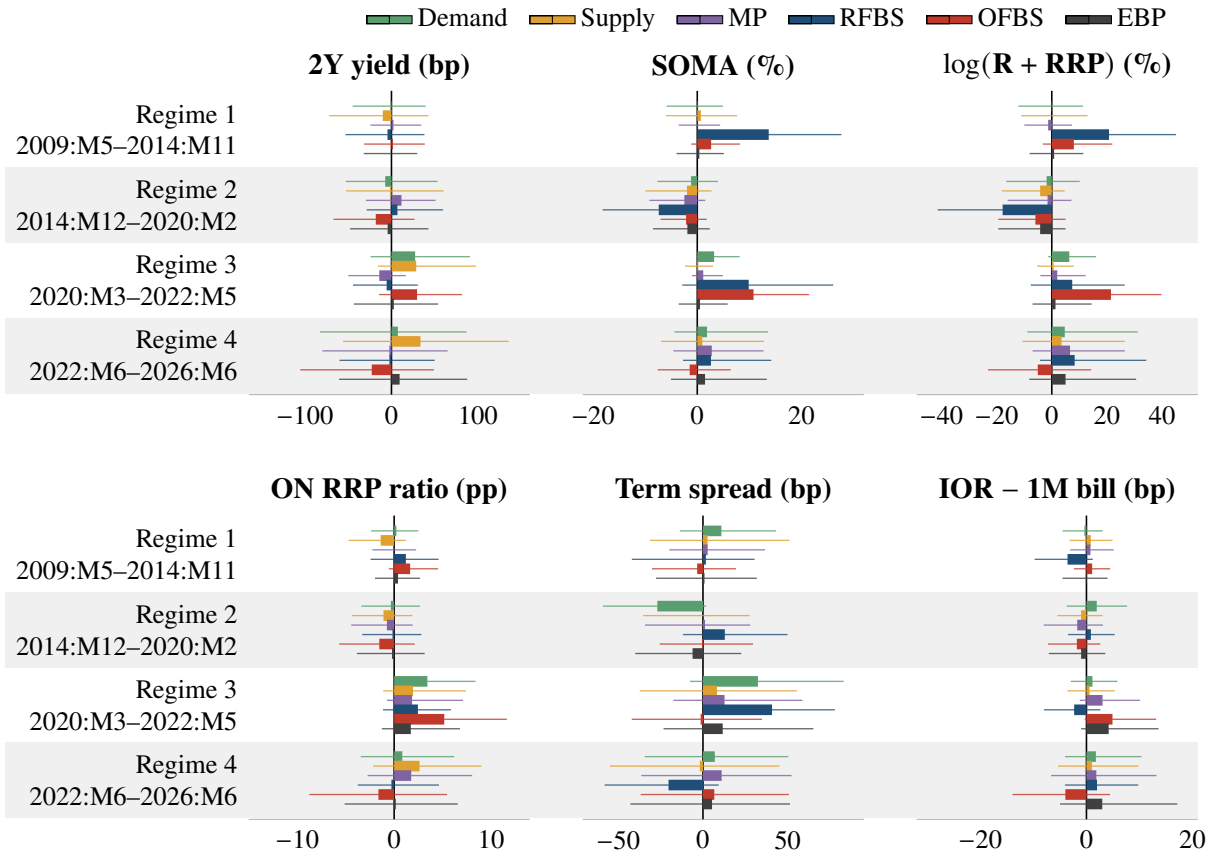


Figure J.1: Historical decomposition by volatility regime, the other six variables.

Note. Contribution of each labeled shock to the accumulated forecast error of the variable over the window. Units are basis points for the two-year yield, the term spread, and IOR less the one-month bill rate. They are percent for SOMA holdings and log(Reserves + ON RRP) and percentage points for the ON RRP ratio. Bars are posterior medians and whiskers are 68% credible intervals, both over the retained draws. The windows are those of Figure 11. The three unlabeled shocks account for the remainder. In the size panel the RFBS contribution excludes zero in Regime 2, and the OFBS contribution excludes zero in Regime 3. In the ratio panel the demand and RFBS contributions in Regime 3 exclude zero.