Quantitative Tightening*

Vadim Elenev — Johns Hopkins Carey
Miguel Faria-e-Castro — FRB St. Louis
Daniel L. Greenwald — MIT Sloan

December 20, 2018

Preliminary and Incomplete – Not for Circulation

Abstract

We investigate how the timing with which a central bank unwinds its portfolio after asset purchases affects the dynamics of the recovering economy and the ability of the central bank to stabilize future crises. Although delayed unwinding can support house prices and borrower consumption during the recovery, it can lead to severe consequences in a future crisis if the central bank faces constraints on the size of its portfolio, doubling the drops in house prices and borrower consumption. Early unwinding provides additional room for QE, further dampening the impact of a crisis, with relatively mild short-run costs as long as conventional monetary policy is able to offset the impact of unwinding on demand. Overall, our results point to precautionary benefits of unwinding soon after the economy exits the zero lower bound.

Keywords: unconventional monetary policy, mortgages, monetary policy normalization
JEL Classification: E5, E6, G2

---

*First version: December 2018. The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System. Correspondence: miguel.fariaecastro@stls.frb.org

1 Introduction

In November 2008, facing a binding zero lower bound and concerned about spiking mortgage spreads, the Federal Reserve began a major intervention in the mortgage market, now known as Quantitative Easing (QE), purchasing what would eventually amount to \$1.78 trillion of residential mortgage backed securities. The scale of the program was massive, with the share of US residential mortgages held by the Federal Reserve, shown in Figure 1, topping out at 17.5% of the outstanding stock. This intervention was largely credited with decreasing mortgage rates and maintaining the flow of mortgage credit to homeowners. In an ongoing environment of low interest rates, asset purchase programs like QE are likely to remain an important monetary policy tool going forward.

While a growing body of research has studied the the consequences of unconventional monetary policies in which central banks purchase mortgages and other private sector assets[1], less attention has been paid to when and how a central bank should *unwind* its mortgage position following its intervention. In this paper, however, we argue that the choice of unwinding policy may have important implications for the central bank's ability to stabilize housing markets and the real economy in case of a future crisis. Our logic is based on institutional constraints: while the Federal Reserve expanded its portfolio to unprecedented levels, no central bank can take unlimited positions, with the existing QE policies perhaps already testing the limits of the Fed's balance sheet. As a result, the ability of the central bank to implement QE policies during a future trip to the zero lower bound may depend crucially on how much balance sheet room it has created through unwinding since the previous intervention.

This line of thinking provides a precautionary motive for accelerating the unwinding of the central bank's positions. At the same time, effective unwinding policy must still face two potential complications. First, just as QE purchases drove mortgage spreads down and gave a boost to the housing and mortgage markets, unwinding too quickly could have the reverse effect, with adverse consequences for these key markets at a potentially vulnerable time. Second, since mortgage market interventions like QE operate largely through borrower refinancing behavior, which is highly state dependent with respect to both interest rate incentives and equity extraction opportunities. Because policy today affects the path of interest rates and household leverage going forward, history of unwinding can influence the effectiveness of QE policies in a future crisis. In fact, since the Federal Reserve prefers to unwind by attrition as mortgages are prepaid rather than directly selling its positions to the secondary market, the *timing* of unwinding policies can have important effects on the *speed* at which the balance sheet is reduced.

To evaluate unwinding policies with respect to these considerations, we develop a general equilibrium model of the macroeconomy that allows for the key dynamics outlined above. We begin with a parsimonious specification for mortgage spreads that we demonstrate closely matches the data, and

[1] On the empirical side, notable papers are Gagnon et al. (2010), Di Maggio et al. (2016), and Stroebel and Taylor (2012). Theoretical work includes Elenev (2017), Silva (2016), Gertler and Karadi (2011), and Curdia and Woodford (2011).

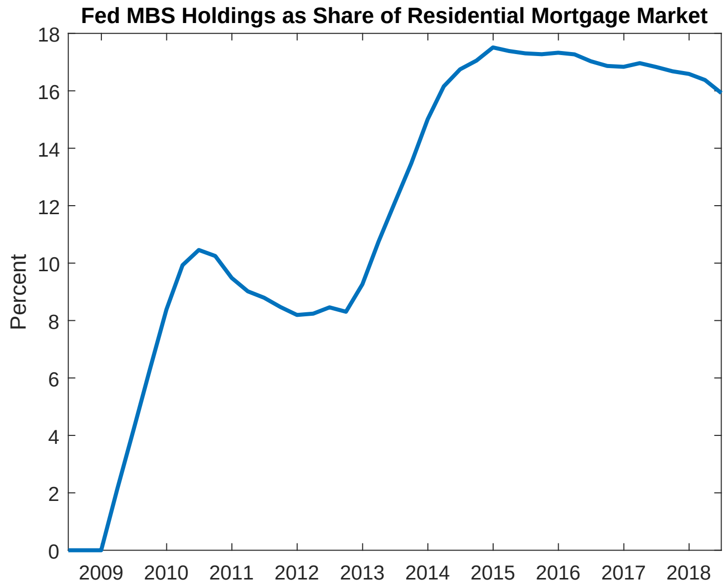

Figure 1: Federal Reserve Mortgage Holdings

implies effective QE policies during crises. We nest this mechanism in a realistic mortgage framework in which borrowers hold fixed-rate mortgages and endogenously choose when to refinance into new loans. Importantly, current borrower refinancing decisions are influenced at equilibrium by both current and past values of house prices and mortgage rates, capturing the key elements of state dependence.

Our main experiments compute the effects of alternative unwinding policies in two scenarios: a continued recovery and gradual return to steady state, and a second crisis in which mortgage spreads spike and the central bank attempts a new round of QE policies. To implement these policies in a realistic way, we respect the Federal Reserve's policy of unwinding only through attrition along a preannounced path for balances, and vary the date at which the central bank stops replenishing its portfolio as loans are prepaid. Further, since the economy is likely far from steady state, having experienced a housing crash in the near past and only recently exiting the zero lower bound, we use state space methods to estimate the initial state of the economy in 2015Q4 — the date when the FOMC announced interest rate liftoff and the beginning of conventional monetary policy normalization. This approach accommodates that the impact of unwinding at different dates can also vary due to progress in macroeconomic and mortgage market conditions as the recovery continues. Finally, motivated by concerns expressed by policymakers, we impose a cap on the size of the central bank's mortgage position that can constrain QE purchases once this limit is reached.

We find that the timing of unwinding matters. We begin with a "benchmark" path following the actual announced policy that gradually began unwinding in 2017Q3. We first contrast this policy with a "late" unwinding policy, in which unwinding begins only three years later, in 2020Q3. In the absence of a future crisis, late unwinding keeps mortgage rates lower during the recovery, supporting house prices and borrower consumption. However, in our scenario in which a second crisis arrives in 2019Q2, we find that the late unwinding path leaves QE policy severely constrained, with adverse economic consequences. Lacking central bank support, increasing spreads cause mortgage rates to rise twice as much as under the benchmark path (3.36pp vs. 1.68pp), while the fall in house prices (16% vs. 9%) and borrower consumption (8.5% vs. 4%) are similarly doubled.

Although the benchmark unwinding policy is sufficient to effectively implement a substantial amount of purchases in our 2019Q2 crisis scenario, the central bank still hits its balance sheet cap, partially constraining policy. As a result, an "early" unwinding scenario, in which balance sheet normalization counterfactually started in 2015Q4, would have led to a smaller balance sheet at the time of our second crisis, allowing for a larger QE intervention. The increased purchases under the early unwinding path have substantial benefits, cutting the falls in house prices (3.25%) and borrower consumption (1.7%) by more than half relative to the benchmark. While these benefits must be weighted against the costs of cooling mortgage markets during the recovery, we find that the costs of unwinding "too early" are relatively mild as long as the economy has exited the zero lower bound. These costs are moderate because the slowdown in demand due to early unwinding is offset by conventional monetary policy, which slows the normalization of short rates.

Overall, our results point to the potential benefits of unwinding QE policies soon after the economy exits the zero lower bound. In a recent paper, Berger et al. (2018) argue that the fact that long term rates have been persistently declining led to a favorable refinancing environment that may have boosted the effectiveness of monetary policy in the past, but may not be able to do so in the future as borrowers lock into low rates, implying weaker transmission in a future crisis. Our work indicates that while the central bank may be unable to rapidly raise the rates on existing mortgages, it *can* prepare for a future crisis, without incurring inordinate costs, by reducing the size of its balance sheet. Indeed, such policies may become increasingly important if rates continue to hover close to the zero lower bound.

Related Literature To be completed.

2 Background: Mortgage Spreads

In this section we provide evidence on mortgage originations and mortgage spreads, and estimate a parsimonious specification for issuance frictions based on gross mortgage issuance that inspires the functional form used in the model.

To estimate our measure of issuance frictions, we begin with the series of Originator Profits and Unmeasured Costs (OPUCs) constructed by Fuster et al. (2013), and displayed on the left axis of Figure 2. This series measures the gains to an originator from originating a new mortgage, securitizing it with

Fannie Mae or Freddie Mac, and selling it on the secondary market. These gains include the difference between the amount originated and the secondary market price for the secured mortgage and the present value of retained servicing fees, while subtracting off the "G-fee" paid to Fannie Mae or Freddie Mac in exchange for insuring the mortgage. For scale, a value of 2% means that originating a $100,000 mortgage would bring in revenues of $102,000 on the secondary market, after netting out the G-fee. Since our model will feature riskless mortgage lending, this is the appropriate value for the marginal revenue created by originating a mortgage.

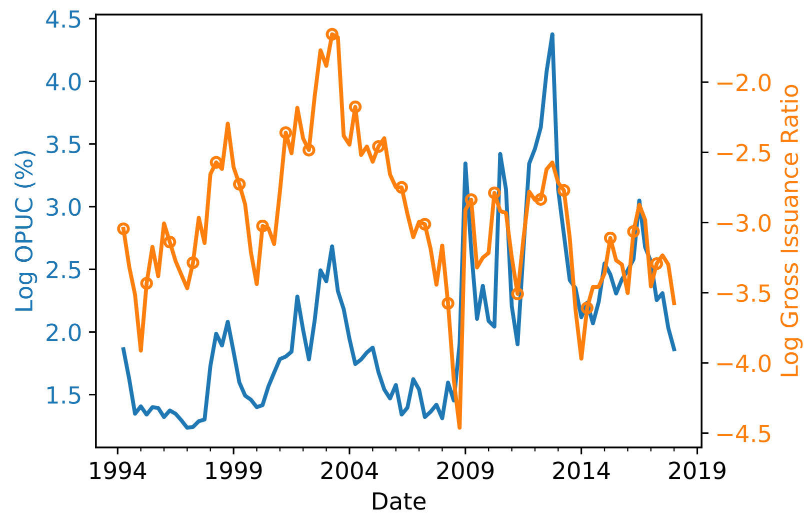

Figure 2: OPUC vs. Gross Issuance Ratio

An equilibrium OPUC value greater than zero represents marginal costs of new issuance, originator profits (i.e., due to market power), or both. Under the market power interpretation, we should see OPUCs increase when demand (and hence marginal revenue) becomes more steeply downward sloping. As can be seen from Figure 2, OPUCs increased substantially in the housing crash and have remained at an elevated level ever since, consistent with lower and more inelastic demands for mortgage backed securities following the crisis, as well as possibly an increase in originator costs.

Next, we show that much of the variation in OPUCs can be systematically explained by the volume of mortgage origination, but only after allowing for a regime shift from before to after the housing crash. Specifically, we define the "gross issuance ratio" to be the ratio of newly originated mortgages to outstanding mortgages at a given date. For the existing mortgage balance, we use Home Mortgages from the Flow of Funds.[2] Since newly originated loans either replace prepaid mortgages, increase debt

[2]Source: Federal Reserve Board of Governors.

Sample	$\beta_{s,0}$	$\beta_{s,1}$	Adj. R^2	N
Pre (to 2008 Q2)	3.183***	0.536***	0.676	58
	(0.185)	(0.065)		
Post (since 2008 Q3)	6.318***	1.159***	0.517	38
	(0.853)	(0.262)		

Table 1: Regression Results: OPUC vs. Gross Issuance Ratio

Notes: Robust standard errors are displayed in parentheses below the coefficient estimates. The notation *** indicates a p-value below 1%. The sample (at quarterly frequency) spans 1994 Q1 to 2008 Q2 for the "pre" regression, and 2008 Q3 to 2018 Q1 for the "post" regression.

balances, or both, we compute

$$GIR_t = \frac{\text{Mortgages}_t - (1 - \text{Prepayment}_t) \cdot \text{Mortgages}_{t-1}}{\text{Mortgages}_{t-1}}$$

where Prepayment_t is the share of loans in Fannie Mae MBS prepaid in quarter t (source: eMBS).

The resulting series is plotted on the right axis of Figure 2. As can be seen, this series appears to co-move strongly with the OPUC series at high frequency. However, unconditional correlation between the two series is low, largely because the relationship appears to undergo a structural break during the housing crash. Specifically, OPUCs are substantially higher post-crisis for the same level of the gross issuance ratio. To estimate this relationship, we run the regression

$$\log OPUC_t = \beta_{s,0} + \beta_{s,1} \log GIR_t + \varepsilon_t, \qquad s \in \{pre, post\} \tag{1}$$

where "pre" corresponds to the sample up to 2008 Q2, and "post" corresponds to the sample from 2008 Q3 onward. The results, shown in Table 1 show that both the intercept and the elasticity of $\log OPUC_t$ to the log gross issuance ratio essentially doubles, we find these results are consistent with a cost function for new originations of the form

$$1 + \text{Cost}_t = \exp\Big(\beta_{t,0} + \beta_{t,1} \log GIR_t\Big) = \eta_t GIR_t^{\psi_t} \tag{2}$$

for the appropriately defined η_t and ψ_t, whose values are likely to increase during times of increased mortgage market stress.

To demonstrate the importance of this shift in parameters from boom to bust, Figure 3 shows what the fitted values from the separate regressions of (1) would imply if extrapolated over the entire sample. As can be seen, the fit from each sample fits the data well during the estimation period, but displays widely counterfactual behavior in the extrapolated region. In particular, maintaining the pre-crisis fit would imply OPUC values much smaller than observed, while adjusting the parameters (increasing η_t

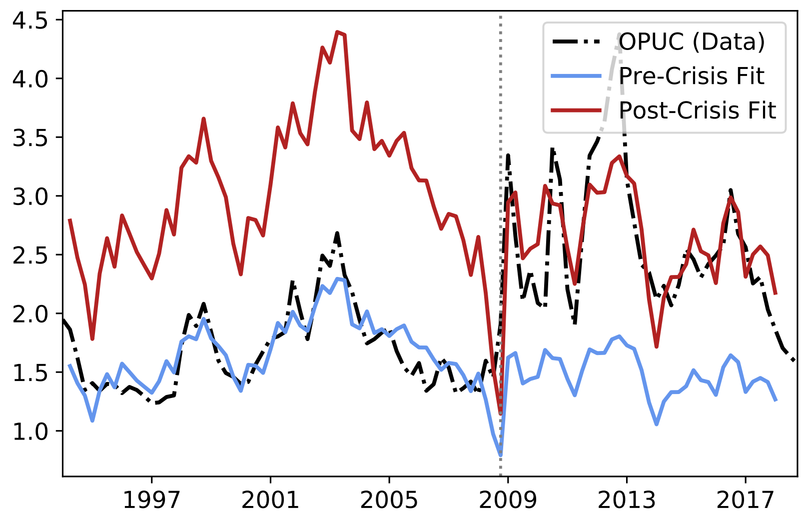

Figure 3: OPUC vs. Gross Issuance Ratio by Regime

and ψ_t) is able to reproduce the post-crisis path.

Overall, we find this evidence as supportive of an adjustment function for private sector purchases of new mortgage backed securities of the form (2), with an elasticity parameter that increases sharply during crises. Our interpretation is that the increase in OPUCs during the crisis largely represents a contraction and steepening of demand, reducing secondary market prices. This interpretation allows for a clear influence of QE policy: if the capacity of the private market to absorb new mortgage backed securities is reduced, then purchases or newly issued mortgages by the Federal Reserve can reduce gross issuance to the private sector, and therefore reduce mortgage spreads. This will be the key mechanism developed in the theoretical model below.

3 Model

In this section, we describe the model and characterize some of its mechanics.

3.1 Model Description

Time is discrete and infinite. The economy is populated by three types of agents: households, who can be borrowers or savers, firms, and a consolidated fiscal and monetary authority that conducts conventional and unconventional monetary policies. We now describe each of these agents in detail.

Households There are two types of households: borrowers and savers, indexed by $j = b, s$, respectively. Borrowers are relatively more impatient than savers $\beta_b < \beta_s$ and exist in measure χ (while savers have measure $1-\chi$). Each agent of type j chooses consumption $C_{j,t}$, housing services $H_{j,t}$, and labor $N_{j,t}$ to maximize the expected present discounted value of utility,

$$\mathbb{E}_t \sum_{k=0}^{\infty} \beta_j^k \left[\log(C_{j,k}) + \xi \log(H_{j,k}) - \eta_j \frac{N_{j,k}^{1+\varphi}}{1+\varphi} \right] \tag{3}$$

Notice that preferences are the same up to the discount factor β_j and the labor disutility parameter η_j, which is chosen such that both types of agents work the same amount of hours at the steady state.

Each type of agent has access to a different set of financial contracts, which we describe next. It is useful to define the stochastic discount factor for each type of agent as

$$\Lambda_{j,t+1} = \beta_j \frac{u^c_{j,t+1}}{u^c_{j,t}} \tag{4}$$

where $u^c_{j,t}$ is the period marginal utility of consumption for type j, i.e. $1/C_{j,t}$.

Mortgage contract The main financial contract in our model is a long-term nominal mortgage with a prepayment option. A mortgage at interest rate r is a promise to make a stream of principal payments $\nu, (1-\nu)\nu, (1-\nu)^2\nu, \ldots$ and a stream of interest payments $r, (1-\nu)r, (1-\nu)^2 r, \ldots$. Parameter ν governs the speed at which the mortgage amortizes and thus helps determine its duration.

After making the required per-period payment, borrowers can prepay the mortgage at par. When a borrower prepays, she returns the outstanding principal balance of the loan. For example, if 1 unit of a mortgage issued in period t is prepaid in period $t+2$, the borrower repays $(1-\nu^2)$ and extinguishes the promise to make any subsequent interest and principal payments.

Borrowers Borrowers receive labor income, buy houses, and borrow in the mortgage market. At the end of the period, borrower i has a nominal mortgage balance of $P_{t-1}m_{i,t}$ and an interest payment to make in the next period, whose value at current period prices is given by $P_{t-1}x_{i,t}$, where $x_{i,t}$ depends on the mortgage balance $m_{i,t}$ and the prevailing mortgage interest rate at the time the mortgage was originated. Then, at the start of the next period, the real outstanding mortgage balance is given by $\Pi_t^{-1} m_{i,t-1}$ and the real interest payment to be made is given by $\Pi_t^{-1} x_{i,t-1}$.

After making the principal payment $\Pi_t^{-1} m_{i,t-1}\nu$ and interest payment $\Pi_t^{-1} x_{i,t-1}$, each borrower draws a cost of obtaining a new mortgage $\kappa_{i,t} \sim \Gamma$. This is the cost that a borrower will have to pay if she chooses to move (optimize over housing choice) or refinance her existing mortgage. We interpret $\kappa_{i,t}$ as standing in for a variety of factors which prevent borrowers in the data from refinancing in response to minuscule rate incentives and adjusting their housing position in response to small changes in consumption. As such, we rebate the cost to borrowers so it does not destroy resources in the economy.

Re-optimizing households take out a new mortgage m_t^* at the current mortgage rate r_t^* subject to a loan-to-value constraint

$$m_{t,i}^* \leq \theta^{LTV} p_t^h h_{t,i}$$

which restricts the new mortgage balance to be no greater than a fraction of the home value.

Let $\mathbb{1}_{i,t}$ take the value of 1 if borrower i refinances and 0 otherwise. Then, borrower i ends the period with a real mortgage balance given by

$$m_{i,t} = \mathbb{1}_{i,t} m_t^* + (1 - \mathbb{1}_{i,t})(1 - \nu)\Pi_t^{-1} m_{t-1,i}$$

and the real value of next period's interest payment given by

$$x_{i,t} = \mathbb{1}_{i,t} m_t^* r_t^* + (1 - \mathbb{1}_{i,t})(1 - \nu)\Pi_t^{-1} x_{t-1,i}$$

Between themselves, borrowers can trade the full menu of state-contingent contracts and thus make collective consumption, labor, and asset choices. The collective borrower family chooses which borrowers re-optimize their housing and mortgage choices, and which continue with their existing allocations, as a function of the transaction cost $\kappa_{i,t}$ and the aggregate state variables. We guess and verify that the optimal refinancing policy follows a cutoff rule i.e.

$$\mathbb{1}_{i,t} = \begin{cases} 1, & \kappa_{i,t} \leq \kappa_t^* \\ 0, & \text{otherwise} \end{cases}$$

Let $\rho_t = \Gamma(\kappa_t^*)$ indicate the share of borrowers who refinance i.e. who obtain a new loan at time t. Then aggregate mortgage balance and interest payments evolve according to

$$m_t^B \equiv \int m_{i,t} di = \rho_t m_t^{*,B} + (1 - \rho_t)(1 - \nu)\Pi_t^{-1} m_{t-1}^B$$
$$x_t^B \equiv \int x_{i,t} di = \rho_t m_t^{*,B} r_t^{*,B} + (1 - \rho_t)(1 - \nu)\Pi_t^{-1} x_{t-1}^B$$

Each borrower enters the period owning $h_{t-1,i}$ of housing and consumes its housing services. Each house costs p_t^h, creates one unit of housing services, which enters the agents' utility function, and requires a maintenance payment of δp_t^h to prevent its full depreciation. Refinancing borrowers can buy or sell housing, while other borrowers continue with their previous allocation. The aggregate borrower housing stock evolves according to

$$h_t^B \equiv \int h_{i,t} di = \rho_t h_t^{*,B} + (1 - \rho_t) h_{t-1}^B$$

Borrowers pay a lump-sum tax T_t to the government and choose how many hours N_t^B to work at wage w_t and how much to consume C_t^B. To summarize, the borrower family optimally chooses the threshold refinancing cost κ_t^*, new mortgages $m_t^{*,B}$ and houses $h_t^{*,B}$, and consumption C_t^B and labor N_t^B subject to the budget constraint

$$
\begin{aligned}
C_t^B + \underbrace{\rho_t p_t^h (h_t^{*,B} - h_{t-1}^B)}_{\text{New houses}} = \\
\underbrace{w_t N_t^B}_{\text{Labor income}} - T_t - \underbrace{\delta p_t^h h_t^{-1}}_{\text{Maintenance}} - \underbrace{\Pi_t^{-1}(x_{t-1}^B + \nu m_{t-1}^B)}_{\text{Mtge payments}} + \underbrace{\rho_t \left[m_t^{*,B} - \Pi_t^{-1}(1-\nu) m_{t-1}^B \right]}_{\text{New borrowing}} - \underbrace{\left(\Psi_t^B(\kappa_t^*) - \Psi_t^B(\cdot) \right)}_{\text{Rebated refi cost}}
\end{aligned}
$$

the evolutions of mortgage balances m_t^B, total interest payments x_t^B, houses h_t^B, and the aggregate borrowing

$$
m_t^{*,B} \leq \theta^{LTV} p_t^h h_t^B
$$

Note that this constraint only restricts new borrowings $m_t^{*,B}$. Thus, a drop in house prices may limit the borrower's ability to take out new loans but will not compel a margin-call-like prepayment.

Savers Savers lend to borrowers in the mortgage market and to the government in the Treasury market. Like borrowers, they trade a full set of state-contingent securities between themselves, and thus their problem can also be expressed in aggregate terms. They begin the period with a real portfolio of principal-only (PO) mortgage strips $\Pi_t^{-1} m_{t-1}^S$, interest-only (IO) mortgage strips $\Pi_t^{-1} x_{t-1}^S$, and Treasuries $\Pi_t^{-1} b_{t-1}^S$ and receive payments on these securities.

A PO strip is a claim to all principal payments of a mortgage, while an IO strip is a claim to its interest payments. Thus, one PO strip pays ν in scheduled payments, $\rho_t(1-\nu)$ in unscheduled prepayments, and has an ex-payment value of $q_t^m(1-\nu)(1-\rho_t)$, where q_t^m is the ex-payment price and $(1-\nu)(1-\rho_t)$ is the remaining balance after amortization and prepayment. We define the total value of a PO strip at the start of a period to be

$$
Z_t^m = \nu + (1-\nu)\rho_t + (1-\nu)(1-\rho_t) q_t^m
$$

An IO strip pays 1 in scheduled payments and has an ex-payment value of $q_t^a(1-\nu)(1-\rho_t)$, where q_t^a is the ex-payment price and $(1-\nu)(1-\rho_t)$ is the remaining balance after amortization and prepayment. Note that prepayments reduce the balances of POs and IOs equally, but because prepayments are made at par, they extinguish a fraction of future interest payments $(1-\nu)\rho_t$. We define the total value of an

IO strip at the start of a period to be

$$Z_t^a = 1 + (1-\nu)(1-\rho_t)q_t^a$$

A treasury is a short-term nominal bond issued by the government that has a riskless payoff of 1 at the start of the period.

The new asset positions of savers consist of legacy POs and IOs purchased on the secondary market as well as new issuances of treasuries and new securitizations of mortgages tranched into POs and IOs. A newly originated mortgage ℓ_t^* is split into ℓ_t^* unit of a PO bond and $\ell_t^* r_t^*$ units of an IO bond. The origination and securitization of a new mortgage is subject to a cost $\Psi_t^S(\ell^{*,I})$ that is increasing and convex in new originations retained by the savers $\ell_t^{*,I}$. This share is equal to total originations ℓ_t^* as , which we parameterize as

$$\Psi_t^S(\ell) = \eta_{m,t} \frac{\ell}{1+\psi^m} \left(\frac{\ell}{\ell_{ss}^{*,I}} \right)^{1+\psi^m}$$

where $\ell_{ss}^{*,I}$ is the steady-state level of originations and $\eta_{m,t}$ is a time-varying level shock to the cost of the deviations of originations from steady state, which follows the following AR(1) process in logs

$$\log \eta_{m,t} = (1-\rho_\eta)\log \eta_{m,ss} + \rho_\eta \log \eta_{m,t-1} + \sigma_\eta \epsilon_t^\eta, \epsilon_t^\eta \sim iid\mathcal{N}(0,1)$$

We do not explicitly model financial intermediaries and frictions between them and savers. Here, savers directly operate the mortgage lending technology. Instead, we interpret the cost Ψ_t^S as standing in for a wide variety of intermediation frictions, which make it costly for mortgage lenders to originate and securitize new mortgages. Positive shocks to $\eta_{m,t}$ increase the marginal cost of originating a new mortgage and thus represent periods of greater intermediation frictions. Whatever the underlying cause of these frictions, we do not want to impose a direct effect on the real economy, and thus we choose to rebate $\Psi_t^S(\ell_t^{*,S})$ to the savers.

Along with labor income and dividends from owning the productive sector (firms), savers use their wealth to finance consumption, housing maintenance, taxes, and a portfolio of POs, IOs, and treasuries. The budget constraint of the savers is

$$C_t^S + \underbrace{q_t^m(m_t^S - \ell_t^*) + q_t^a(x_t^S - \ell_t^* r_t^*)}_{\text{Legacy PO/IO purchases}} + \underbrace{l_t^*}_{\substack{\text{new}\\\text{mortgages}}} + \underbrace{q_t b_t^S}_{\substack{\text{new}\\\text{treasuries}}} =$$

$$= \underbrace{w_t N_t^s}_{\substack{\text{labor}\\\text{income}}} - T_t + \underbrace{\Pi_t^{-1}(Z_t^m m_{t-1}^S + Z_t^a x_{t-1}^S + b_{t-1}^S)}_{\text{Financial wealth}} + \underbrace{\Theta_t}_{\text{Firm dividends}} - \underbrace{\left(\Psi_t^S(\ell_t^*) - \Psi_t^S(\cdot)\right)}_{\text{Rebated orig cost}}$$

Like borrowers, savers own and receive a service flow from housing, but they do not optimize the quantity of housing, holding instead a fixed stock $\bar{h}_t^S$.

While we consign most of the derivations to the Appendix, we note here that the first-order condition for new originations implies that

$$q_t^m + q_t^a r_t^* = 1 + \Psi_t^{S,\prime}(\ell_t^{*,S})$$

In other words, the rate that borrowers have to pay on new mortgages is implied by the market prices of PO and IO strips as well as the time-varying cost of origination.

Firms We model firms in the standard New Keynesian way. Competitive retail firms combine a measure one continuum of intermediate goods into a final good using a CES aggregation technology

$$Y_t = \left(\int_0^1 Y_t(i)^{1-\frac{1}{\varepsilon}} di \right)^{\frac{\varepsilon}{\varepsilon-1}}$$

Each intermediate good is produced by a single wholesale firm. The symmetry of the final good technology implies identical choices by each wholesale firm, so we drop the i subscript in describing their problem.

Wholesale firms hire labor N_t to produce $Y_t = A_t N_t$ goods subject to the productivity A_t, which follows an AR(1) process

$$\log A_t = \rho_A \log A_{t-1} + \sigma_A \epsilon_t^A$$

Their earn profits equal to revenues, less wage expenses and the Rotemberg cost of adjusting prices

$$\Theta_t = Y_t - w_t N_t - \mathcal{R}(\Pi_t) + \mathcal{R}(\cdot)$$
$$\mathcal{R}(\Pi_t) = \frac{\zeta}{2} \left(\frac{\Pi_t}{\bar{\Pi}} - 1 \right)^2 Y_t$$

which they pay out to their shareholders – savers – as dividends. Like we do with costs faced by borrowers and savers, we rebate the Rotemberg price adjustment costs to firms so that they do not enter the aggregate resource constraint.

As monopolists, wholesale firms choose prices taking as given retail firm demand. They maximize the stream of current and future dividends discounted at the savers' stochastic discount factor

$$\max \mathrm{E}_t \left[\sum_{k=0}^{\infty} \Lambda_{t,t+k}^S \Theta_{t+k} \right]$$

Optimal price-setting behavior implies the following Phillips curve

$$\varepsilon\left(\frac{\varepsilon-1}{\varepsilon}-\frac{w_t}{A_t}\right)=\zeta \mathrm{E}_t\Lambda_{s,t+1}\frac{Y_{t+1}}{Y_t}\frac{\Pi_{t+1}}{\Pi}\left(\frac{\Pi_{t+1}}{\Pi}-1\right)-\zeta\frac{\Pi_t}{\Pi}\left(\frac{\Pi_t}{\Pi}-1\right)$$

Government The government is a consolidated fiscal and monetary entity. Its central bank sets the nominal interest rate on treasuries subject to a Taylor-type rule constrained by a zero lower bound, and subject to a monetary policy shock:

$$\frac{1}{q_t}=\max\left\{0,\left[\frac{1}{q_{t-1}}\right]^{\rho_i}\left[\frac{1}{\bar{q}}\left(\frac{\Pi_t}{\bar{\Pi}}\right)^{\phi_\pi}\left(\frac{Y_t}{\bar{Y}}\right)^{\phi_y}\right]^{1-\rho_i} mp_t\right\} \tag{5}$$

The central bank can also conduct unconventional monetary policy through large-scale asset purchases of mortgages in the primary market. In other words, the central bank can choose to buy a fraction f_t^{QE} of POs and IOs associated with newly originated mortgages ℓ_t^* and finance these purchases by issuing reserves.[3].

Finally, the government conducts fiscal policy. Every period it finances government consumption G through lump-sum taxes T_t and short-term debt B_t^G. The government sets taxes to increase in the amount of government debt outstanding:

$$T_t=\bar{T}\left(\frac{B_t^G}{\bar{B}_t^G}\right)^{\phi_T}, \phi_T>0$$

Collectively, fiscal policy and large-scale asset purchases imply the following government budget constraint

$$T_t+\Pi_t^{-1}(Z_t^m m_{t-1}^G+Z_t^a x_{t-1}^G)+q_t B_t^G=G+\Pi_t^{-1}B_{t-1}^G+q_t^m m_t^G+q_t^a x_t^G$$

and the evolution of the central bank balance sheet

$$\begin{aligned} m_t^G &= f_t^{QE}\ell_t^*+(1-\nu)(1-\rho_t)\Pi_t^{-1}m_{t-1}^G \\ x_t^G &= f_t^{QE}\ell_t^*+(1-\nu)(1-\rho_t)\Pi_t^{-1}x_{t-1}^G \end{aligned}$$

In the quantitative exercise section of the paper, we consider a variety of policies $\{f_t^{QE}\}_t$.

Equilibrium In equilibrium, the following markets clear

- Housing: $\chi h_t^B+(1-\chi)\bar{h}_t^S=1$

[3]In our model, reserves and treasuries are both one-period securities issued by the government and held by the savers, so we use these terms interchangeably

- New originations: $\chi\rho_t m_t^* = \ell_t^* = \ell_t^{*,S} + f_t^{QE}\ell_t^*$
- POs: $(1-\chi)m_t^S + m_t^G = \chi m_t$
- IOs: $(1-\chi)x_t^S + x_t^G = \chi x_t$
- Treasuries/reserves: $(1-\chi)b_t^S = B_t^G$
- Labor: $\chi N_t^B + (1-\chi)N_t^s = N_t$
- Resource constraint (by Walras Law): $\chi C_t^B + (1-\chi)C_t^S + \delta p_t^h + G = Y_t$

3.2 Model Mechanics

In this section, we briefly describe some of the forces at work in the model.

Refinancing Incentives The first-order condition for the refinancing cost threshold is given by

$$\kappa_{C,t}^* = \Omega_t^x(\bar{r}_t - r_t^*) + \left[1 - \frac{\Pi_t^{-1}(1-\nu)m_{t-1}}{m_t^*}\right](1 - \Omega_t^m - \bar{r}_r\Omega_t^x)$$

A higher κ_t^* denotes a larger incentive to refinance, and the total fraction of mortgages that are refinanced is given by $\rho_t = F_\kappa(\kappa_t^*)$. The expression highlights the two motives for refinancing: the first part of the expression is the rate incentive, while the second part is the cash-out incentive.

The rate incentive is directly proportional to the interest rate gap $\bar{r}_t - r_t^*$, the difference between the average rate on outstanding mortgage debt $\bar{r}_t$ and the interest rate on new mortgages r_t^*. A larger interest rate gap means that the household is currently paying a high interest rate on its mortgage balance relative to the deal that could be obtained at current market conditions; this increases the incentive to refinance.

The cash-out incentive is proportional to $1 - \frac{\Pi_t^{-1}(1-\nu)m_{t-1}}{m_t^*}$, a measure of how much equity the borrower can extract by refinancing. The borrowing constraint makes m_t^*, the balance that can be obtained on a new loan, proportional to current house prices p_t^h. Thus, high house prices relative to the outstanding mortgage balance indicate that the household can generate additional current income by refinancing the mortgage.

In summary, the borrower household will have an incentive to refinance when (i) current mortgage rates are low relative to those in the outstanding balance; and (ii) current house prices are high relative to outstanding mortgage balances. Since the LTV constraint tends to be binding for the borrower household, its marginal propensity to consume will be relatively high. This means that an increase in refinancing is associated with an increase in borrower consumption and, due to aggregate demand effects, a rise in aggregate activity and inflation.

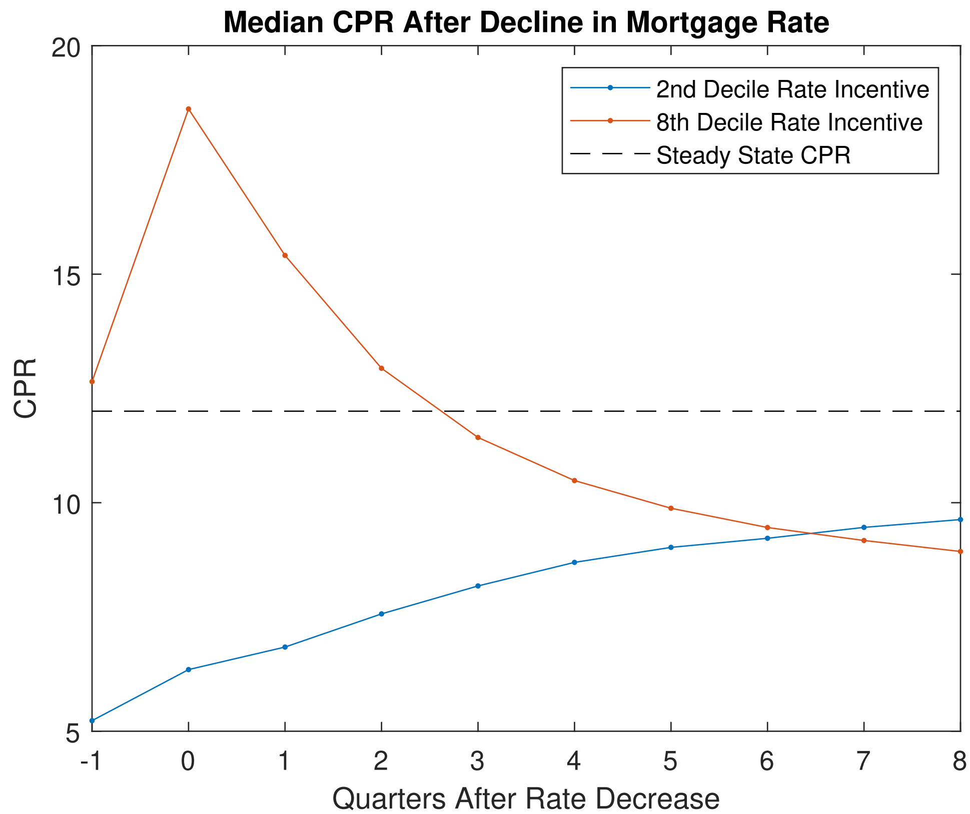

Figure 4: Response of endogenous variables to a monetary policy shock that lowers the FFR.

State- and History-Dependent Effects of Conventional Monetary Policy The Fed directly sets the short-rate $R_t = 1/q_t - 1$, which in turn affects the mortgage rate r_t^* via the saver's first-order conditions. More specifically, saver optimality implies

$$r_t^* = \frac{\Psi_t + q_t^m}{q_t^a}$$

Via the conventional New Keynesian channel of monetary policy transmission, the Fed affects the intertemporal consumption profile of the saver, which in turn affects the prices of long-term securities q_t^a, q_t^m, and is transmitted to mortgage rates r_t^*. When the Fed raises rates, for example, $q_t \downarrow$ which implies $\mathbb{E}_t \frac{\Lambda_{s,t+1}}{\Pi_{t+1}} \downarrow$ and $q_t^m, q_t^a \downarrow$. Under the assumption that $q_t^a >>> \Psi_t + q_t^m$ (which is locally true if r_t^* is a number close to zero), this raises mortgage rates $r_t^* \uparrow$. Conversely, mortgage rates tend to fall when the Fed lowers the policy rate.

Importantly, through its effect on the mortgage rate and refinancing incentives, the effects of conventional monetary policy will depend on the history of past interest rates. Figure 4 illustrates these effects. Using the event study approach for a long simulation of the model, the figure plots the share of borrowers that refinance (conditional prepayment rate, or CPR) just before and for 8 quarters after a decline in r_t^*. When the refinancing incentive $\bar{r} - r_t^*$ is high – in the 8th decile of all simulated refinancing incentives – a decline in r_t^* indeed leads to substantially more prepayments. But when the refinancing incentive is low e.g. because $\bar{r}$ is low, a drop in r_t^* has no meaningful effect on prepayments. This also means that the increase in borrower consumption and potentially on aggregate activity will be lower. Long periods of low interest rates can therefore reduce the future effectiveness of conventional monetary policy through this refinancing channel, as in Eichenbaum et al. (2018).

Unconventional Monetary Policy Unconventional monetary policy affects directly origination costs for the lender. Recall that the FOC for originations is

$$q_t^m + q_t^a r_t^* = \eta_{m,t} \left[\frac{\ell_t^{*,I}}{\ell_{ss}^{*,I}} \right]^{\psi^m}$$

By purchasing assets, the Fed is essentially lowering $\ell_t^{*,I}$ for a given level of total originations ℓ_t^*. Active purchases of new mortgages can, via this channel, stabilize mortgage rates r_t^* and asset prices q_t^a, q_t^m. The real effects of QE are transmitted through its effect on mortgage rates: by stabilizing r_t^*, QE raises the incentives to refinance and raises borrower income, which in turn translates into higher borrower consumption and aggregate activity.

4 Quantitative Exercise

We now use the model described in the previous section to perform a quantitative analysis of monetary policy normalization in the United States. In particular, we focus on the following questions:

1. Given the commitment to an interest rate rule, what are the effects of different balance sheet normalization plans?

2. Do these different plans matter should the US economy experience a new financial crisis?

3. Do these different plans matter if the Fed wants to engage in new asset purchases?

We proceed as follows: first, we calibrate the model to match features of the US economy during the period of interest. Then, we feed model with time series data to estimate values for the endogenous and exogenous states of the US economy in 2015Q4, the period when the Fed announces the first interest rate increase (i.e., the first period of policy normalization). Starting the economy from those estimated states, we consider nonlinear transitions back to the economy's steady state. Different policy normalization plans are treated as perfectly anticipated paths for policies, consistent with the Federal Reserve's current communication policy. New crises, and new asset purchase programs depolyed in response to these crises are treated as unanticipated shocks that hit the economy along its transition path. We now describe each of these steps in more detail, along with our findings.

4.1 Calibration

The calibration of the model is relatively standard and summarized in Table 2.

Parameter	Description	Value	Target
	Demographics and Preferences		
χ	Fraction of borrowers	0.45	Avg share w/ neg fixed income pos, SCF 93-16
β_s	Discount factor savers	0.9959	Avg level of federal funds rate 2000-2018
β_b	Discount factor borrowers	0.9829	Value of housing to income of 8.89
φ	Frisch elasticity	1	Standard
ξ	Housing preference parameter	0.25	Davis and Ortalo-Magné (2011)
η_b	Borrower labor disutility	14.13	$N_t^b = 0.33$
η_s	Saver labor disutility	8.28	$N_t^s = 0.33$
	Production		
ε	Micro elasticity of substitution across varieties	6	20% markup in SS
ζ	Rotemberg Menu Cost	98.37	Prices adjust once every five quarters
	Government		
$\bar{G}$	SS Govt. Spending	$0.2 \times Y$	20% for the US
$\bar{B}^g$	SS Govt. Debt	$0.14 \times Y$	Avg. maturity of 20 months, 70% of GDP (Faria-e-Castro, 2018)
$\bar{\Pi}$	Trend Inflation	$1.02^{0.25}$	2% for the US
ϕ_π	Taylor rule: Inflation	1.5	Standard
ϕ_y	Taylor rule: Output	0.5/4	Standard
ρ_i	Taylor rule: Smoothing	0.8	Standard
ϕ_τ	Fiscal Rule	0.01	Faria-e-Castro (2018)
	Housing and Mortgages		
θ^{LTV}	Maximum LTV at origination	0.80	Max LTV for GSE conforming loans
ν	Contractual duration of mortgages	0.005	Standard
δ	Maintenance cost of housing	0.0065	2.5% annual, standard
$\bar{H}$	Total stock of housing	1	Normalization
s_κ	SD of prepayment shock	0.152	Greenwald (2018)
μ_κ	Mean of prepayment cost shock	0.2902	$\rho_{ss} = 0.0376$
$\eta_{m,ss}$	Mean financial friction	1.0969	Annual. mortgage spread of 2%
ϕ_m	Elasticity of Ψ to originations	2.5	
	Shock Parameters		
ρ_a	Persistence of TFP	0.90	Standard
σ_a	SD of TFP Innovations	0.01	Standard
ρ_i	Persistence of nominal rate	0.80	Standard
ρ_r	Persistence of MP Shock	0.80	Standard
σ_r	SD of MP Shock Innovations	0.005	Standard
ρ_{QE}	Persistence of QE	0.75	Estimated
σ_{QE}	SD of QE Innovations	1	Normalization
ρ_η	Persistence of financial shock	0.75	
σ_η	SD of financial shock Innovations	1	Normalization

Table 2: Summary of the calibration.

4.2 Estimating the State of the US Economy in 2015Q4

To analyze the different normalization scenarios, we need to estimate the constellation of endogenous states that characterized the US economy at the time when normalization began in 2015Q4. As the previous section highlighted, some of these states can be quite important for the analysis of both conventional and unconventional policies, such as the average rate on outstanding mortgages $\bar{r}_{t-1}$, or the total stock of mortgages owned by the Fed m^G_{t-1}.

We estimate values for these endogenous states by using standard state space methods that combine a linearized version of the model and the Kalman Filter (An and Schorfheide, 2007). We linearize the calibrated model and use the Kalman Filter to extract sequences for the four exogenous shocks $\{\varepsilon_t^a, \varepsilon_t^r, \varepsilon_t^m, \varepsilon_t^{QE}\}_{t=0}^T$, given four observable time series that we match from the data.

Choice of Observables Since the focus of our analysis is monetary policy normalization, we need to ensure that the exercise adequately captures the stance as of 2015Q4. For this reason, we choose as observables the short-term nominal interest rate R_t and the fraction of mortgages held by the Fed in its balance sheet $\frac{m_t^G}{m_t}$. Since our model features four shocks, we can match two more observables. We choose observables that we believe are informative for the estimation of the paths of the TFP and Financial shocks: aggregate consumption C_t, and the growth rate of outstanding mortgages $g_{m,t} = \frac{m_t}{m_{t-1}} - 1$.

Data All of our data is taken from FRED. For the nominal interest rate, we use the 3-Month Treasury Bill rate (TB3MS). For the fraction of mortgages held by the Fed, we divide the amount of mortgage-backed securities held outright by Federal Reserve Banks (WSHOMCB) by the stock of mortgages in the US (ASHMA). For consumption, we take real personal consumption expenditures (PCECC96) and divide them by population (B230RC0Q173SBEA). We then detrend the log of this series using the procedure suggested by Hamilton (2017) for quarterly series. That is, we run the following regression

$$\log \tilde{c}_{t+8} = \alpha + \sum_{i=0}^{4} \beta_i \log \tilde{c}_{t-i} + \epsilon_t^c$$

and obtain detrended consumption as the estimated residuals $\{\hat{\epsilon}_t^c\}_t$. Finally, our real mortgage growth series is obtained by computing the quarterly log difference of the ASHMA series and subtracting the quarterly growth rate of the implicit price deflator for PCE (DPCERD3Q086SBEA). Figure 5 plots these time series.

Estimation Results Figure 6 plots the resulting estimates for the exogenous states of the model. The Great Recession is captured as a combination of negative TFP realizations, contractionary monetary policy shocks (proxying for the absence of a zero lower bound in the linearized model), and a large negative financial shock. The QE_t variable captures, by construction, surges in Fed purchases of mortgage securities during QE1 and QE3.

By construction, observables are matched exactly as we do not have measurement error in the model. Figure 7 plots additional endogenous variables that are not directly targeted. During the boom period, mortgage rates fall and house prices boom. These two forces conspire to generate a refinancing boom, which is also reflected in a boom for borrower consumption. The crisis then corresponds to a large collapse in house prices, borrower consumption, and GDP. Mortgage rates rise, and refinancing also

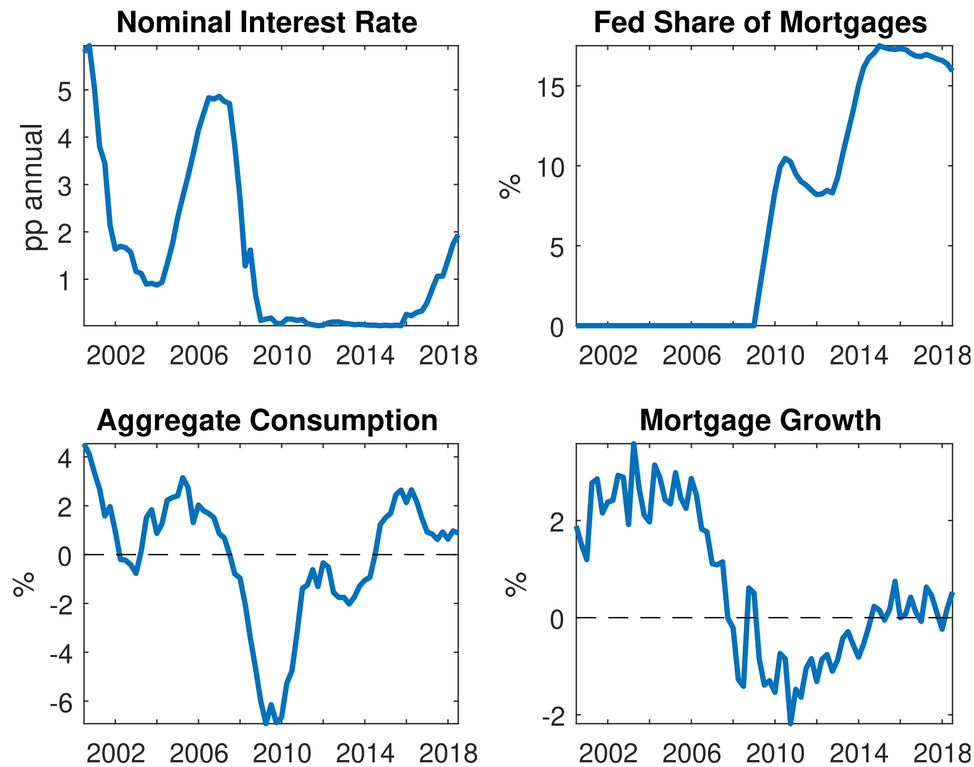

Figure 5: Time series used to match model observables.

falls. Most variables remain below their steady state levels until 2015, when the data implies that a new expansion takes place, but with much more moderate house price and refinancing behavior than in the pre-crisis boom period.

4.3 Monetary Policy Normalization Scenarios

We use the resulting estimates for endogenous and exogenous state variables in 2015Q4 as starting points for the analysis of monetary policy normalization. We choose this date as it includes the December 2015 meeting at which the FOMC announced that interest rate liftoff would begin. Consistent with the Fed's communication policy, we model policy normalization as consisting of a completely predictable (or anticipated) set of policy rules:

1. Interest rate normalization follows the pre-crisis Taylor Rule as specified in 5, subject to the zero lower bound.

2. Balance sheet normalization follows the instructions directed by the FOMC after its September 2017 meeting to the Open Market Trading Desk of the New York Fed:[4]

    (a) From 2015Q4 to 2017Q3, we assume that the Fed follows a maintenance regime, purchasing

[4]See https://www.newyorkfed.org/markets/opolicy/operating_policy_170920 for a description of these plans.

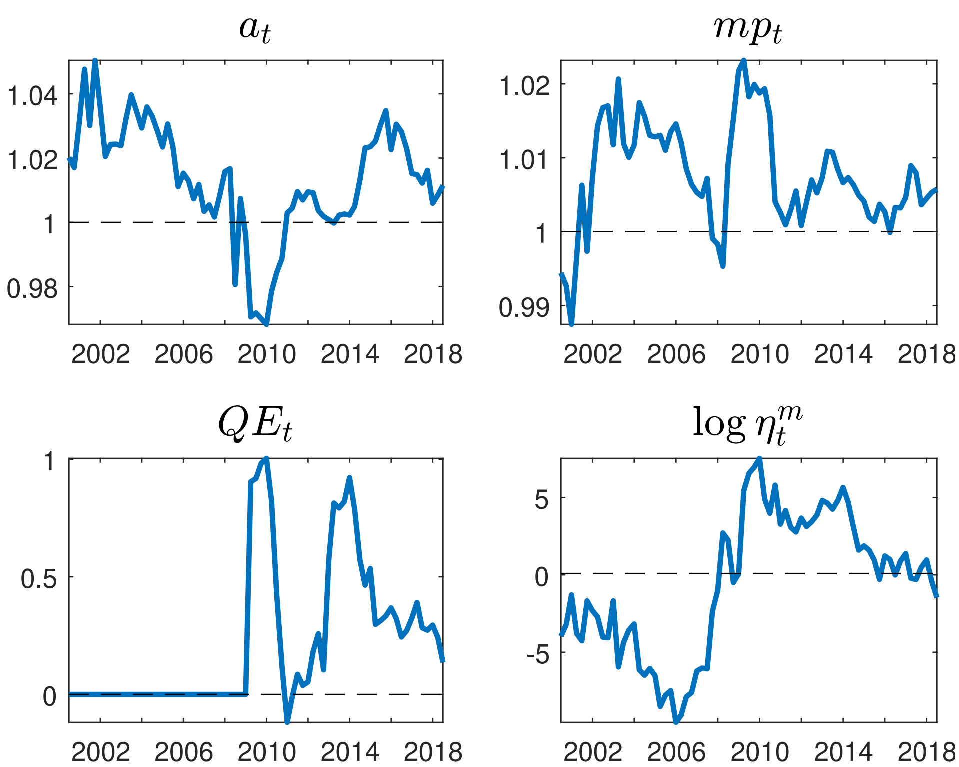

Figure 6: Smoothed series for the exogenous states.

whatever amount of newly issued mortgages in order to maintain $m_t^G = m_{max}^G$, where m_{max}^G is the size of MBS holdings as of 2015Q4.

(b) From 2017Q3 onwards, these reinvestments are subject to caps that become gradually larger, i.e. the amount reinvested by the Fed becomes gradually smaller, leading to a gradual shrinking of the balance sheet.

This describes our benchmark scenario, which is a description of what the Fed is actually doing. We also consider two other scenarios, one in which the reinvestment caps are imposed in 2015Q4, at the time of rate liftoffs (*early unwinding*, EU), and one in which normalization is delayed by three years, and the caps are only imposed in 2020Q3 (*late unwinding*, LU). In all of these alternative scenarios, conventional policy follows the same Taylor Rule as in the benchmark scenario. This means that nominal rates can differ across scenarios, to the extent that they respond to endogenous macroeconomic variables that could take different paths. What is important is that all of these policies are completely predictable in any of the scenarios we consider.

Figure 8 plots the transition paths for several endogenous variables in the model. These transitions are plotted starting in 2015Q4, through 2025Q4, and assume that no further shocks — anticipated or not — hit the US economy. The first panel shows the path of the Fed's balance sheet: the blue line is the benchmark scenario, where it starts shrinking in 2017Q3. The red line is early unwinding, and the orange line is late unwinding. The second panel shows the path of Fed *purchases*: they drop abruptly for

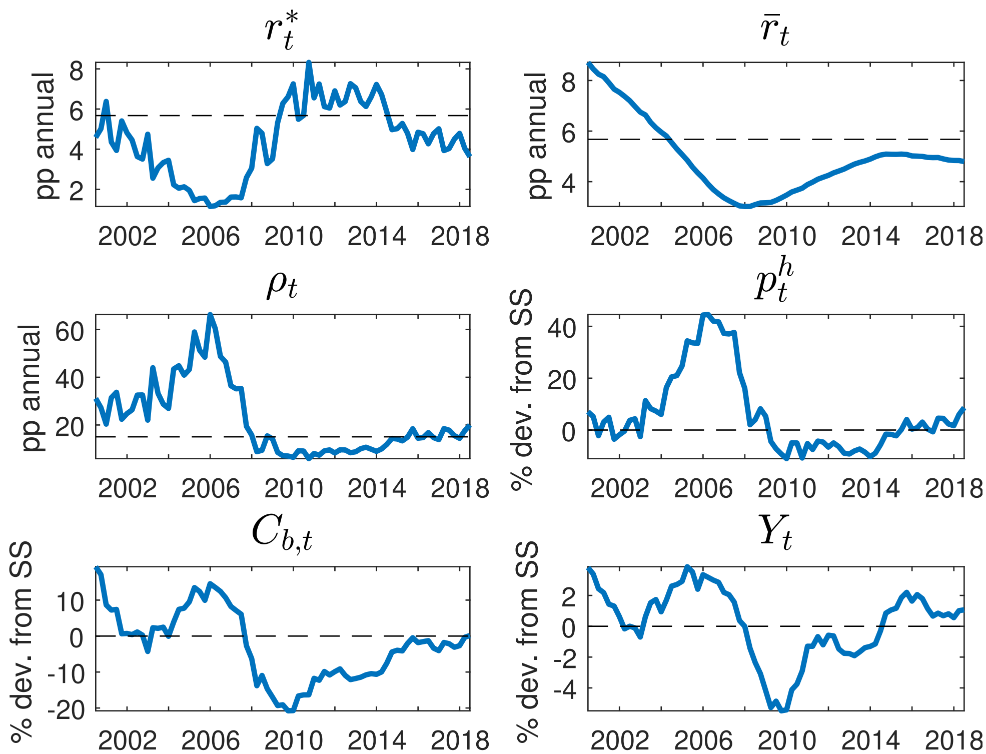

Figure 7: Smoothed series for selected endogenous variables.

the case of early unwinding, but remain relatively high up to the unwinding date in the other cases. The third panel displays an important consequence of these purchases: the longer the Fed keeps reinvesting, the more financial frictions are mitigated. This maps into a prolonged fall in mortgage rates r_t^* and, consequently, in average rates on outstanding mortgages $\bar{r}_t$. This interest rate incentive, along with booming house prices lead to an increase in the share of borrowers that refinance.

In summary, the timing of unwinding has an important effect on finance frictions, which then pass through to mortgage rates, refinancing rates, and house prices. The longer the Fed waits to unwind its balance sheet, the larger is the housing and refinancing booms that it generates.

4.4 New Crisis and Normalization

We now consider the effects of these three normalization scenarios in the event of a new financial crisis hitting the US economy in the second quarter of 2019. We assume that in 2019Q2 there is a large realization of ε_t^η that is unexpected (and persistent, via the parameter ρ_η).

The transitions are shown in Figure 9. Since the crisis is unexpected, the transition paths prior to it are identical to the ones in the no-shock normalization scenarios that we discussed above. The effects of the crisis are similar across normalization regimes. Due to the house price, borrower consumption, and refinancing boom generated in LU, the relative drop in these variables is larger. One important thing to notice, however, is the recovery from this shock is faster under LU (and there is no discernible difference

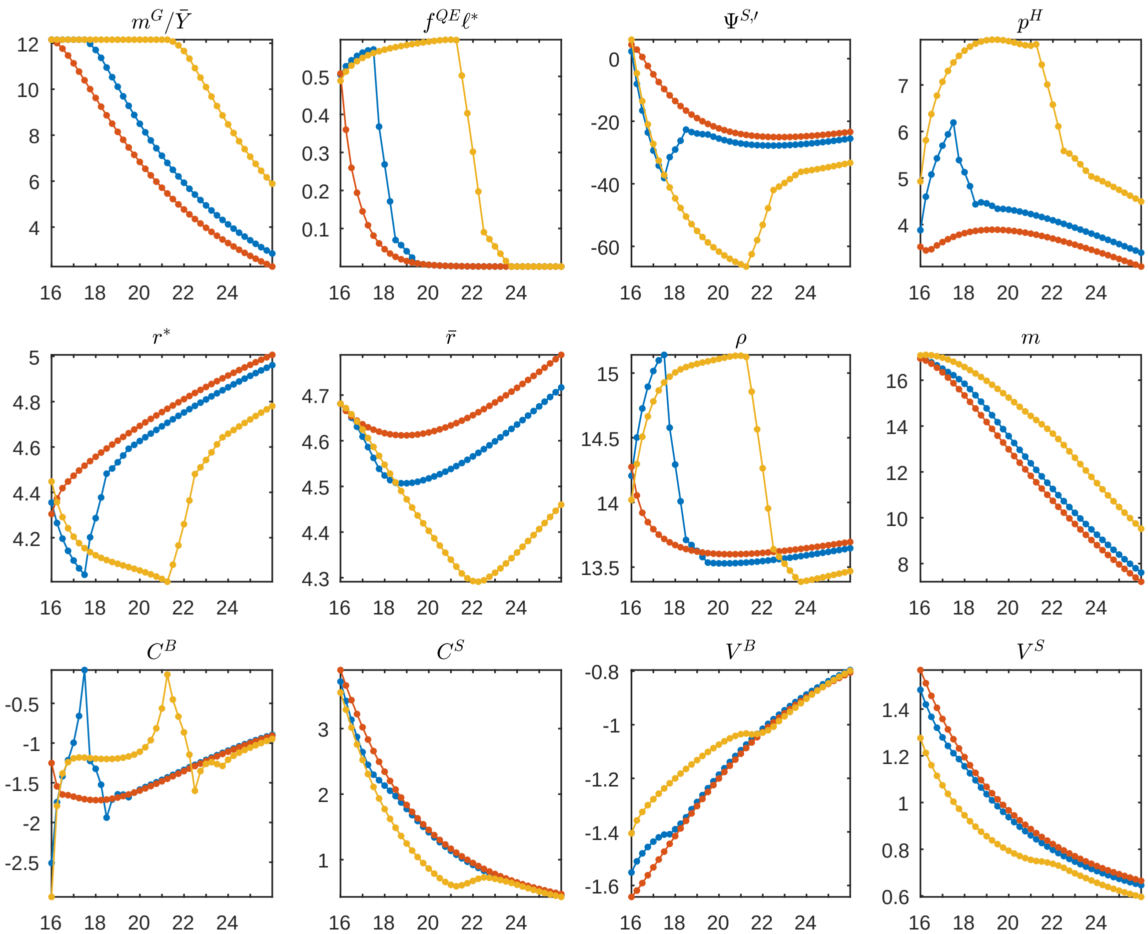

Figure 8: Monetary policy normalization scenarios. These figures show the transition dynamics from 2016Q1 to 2025Q4. In each panel, the blue line shows the baseline balance sheet normalization scenario, consistent with how the Federal Reserve has been reducing the balance sheet in the data starting in 2017Q4. The red line shows a scenario in which balance sheet normalization begins 7 quarters earlier. The orange line plots a scenario in which balance sheet reductions do not begin until 2020Q3. Balance sheet paths are plotted in the top-left panel, while the purchases they imply are plotted directly to the right. Interest rates r^* and $\bar{r}$ and the refinancing share ρ are annualized and in percent. Value functions V^B and V^S are deviations from steady state. All other variables are in percent deviations from steady state.

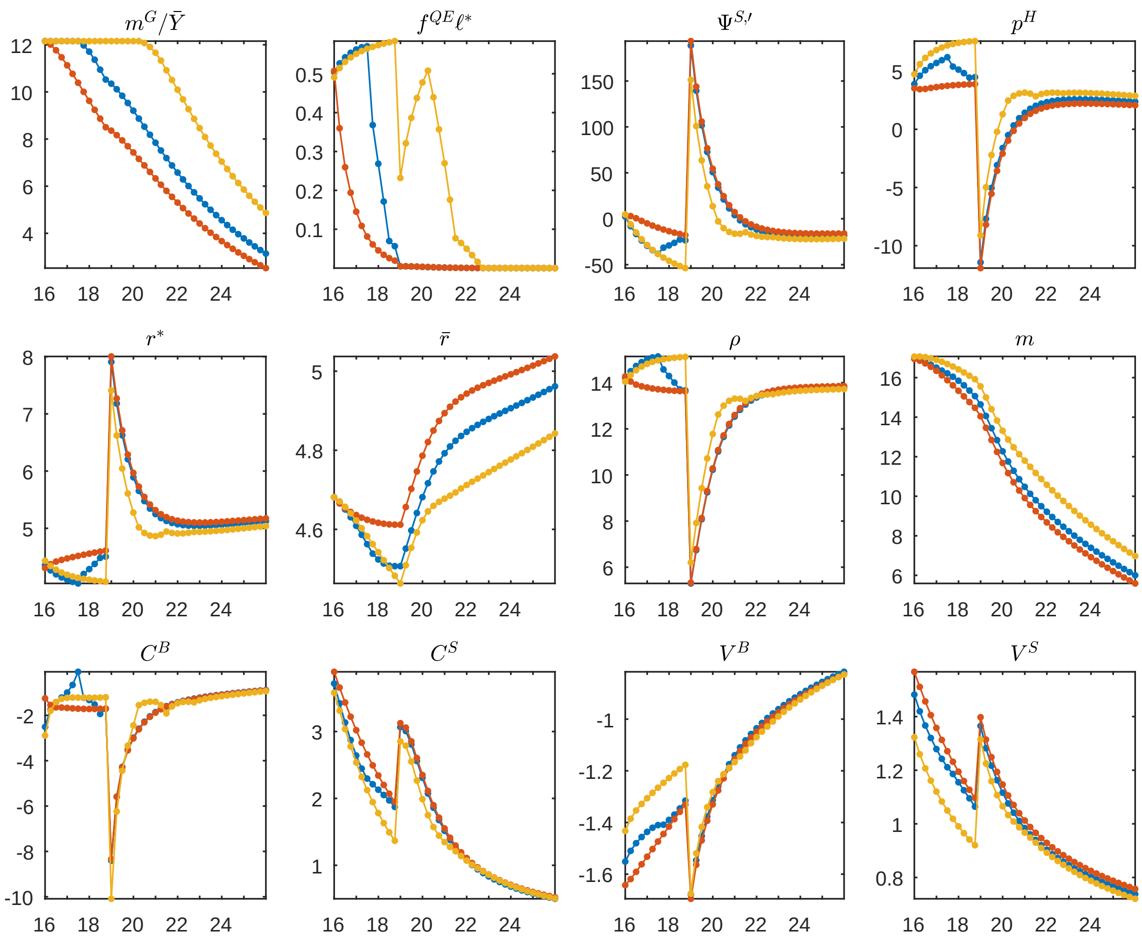

Figure 9: Monetary policy normalization scenarios with unexpected crisis in 2019Q2. These figures show the transition dynamics from 2016Q1 to 2025Q4. In each panel, the blue line shows the baseline balance sheet normalization scenario, consistent with how the Federal Reserve has been reducing the balance sheet in the data starting in 2017Q4. The red line shows a scenario in which balance sheet normalization begins 7 quarters earlier. The orange line plots a scenario in which balance sheet reductions do not begin until 2020Q3. Balance sheet paths are plotted in the top-left panel, while the purchases they imply are plotted directly to the right. Interest rates r^* and $\bar{r}$ and the refinancing share ρ are annualized and in percent. Value functions V^B and V^S are deviations from steady state. All other variables are in percent deviations from steady state.

between the benchmark and EU scenarios). The reason is that, by reinvesting to maintain the size of its balance sheet, the Fed is still doing "passive QE" at the time of the crisis: since refinancing drops, both the share and absolute amount of new mortgages that are purchased by the Fed are larger under the LU scenario. This cuts intermediation costs and allows mortgage rates to return faster to their transition path, thus allowing for a speedier recovery.

4.5 New Crisis, QE4, and Political Constraints

We now consider the possibility that the Fed might respond to an unexpected crisis in 2019Q2 by, also unexpectedly, conducting another round of purchases of mortgage-backed securities. That is, we augment the previous crisis scenario with another unexpected shock to ε_t^{QE}.

Additionally, we assume that the Fed faces a political economy constraint in the form of a maximum size of its balance sheet: it is not allowed to exceed m_{max}^G, the maximum level of MBS holdings that it held prior to normalization. While not explicitly bound by it in the course of the three first rounds of QE, FOMC members were well aware of potential political issues that could be raised by its asset purchase programs.[5] We assume that this constraint is known from 2015Q4 onwards (therefore not affecting the beginning of the transition).

The results are shown in 10. The effects of the political constraint are visible in the first two panels: in the LU scenario, the hands of the Fed are tied, and it cannot do more beyond reinvesting to maintain the size of its balance sheet. In the benchmark scenario, some QE can be done, and the stabilizing effects of this small dosage are visible in the following panels: this reduces financing frictions and mortgage spreads, preventing house prices, borrower consumption and refinancing from falling by as much. The EU scenario allows the Fed, however, to go much further with its QE program due to its smaller balance sheet by the time the crisis shock hits. In this case, mortgage rates are contained and so are the drops in consumption and house prices.

5 Conclusion

In this paper, we developed a macroeconomic model with a rich mortgage sector and a role for realistic quantitative easing policy. Our results indicate that the timing of unwinding matters, most importantly that late unwinding can leave the central bank constrained in a crisis, worsening economic outcomes. In contrast, early unwinding provides additional room for unconventional monetary policy, improving crisis outcomes, without introducing major disruptions during the recovery.

[5] See for example the transcript of the April 26-27, 2011 meeting, available at FRASER: https://fraser.stlouisfed.org/title/677/item/23299/content/pdf/FOMC20110427meeting. Stanley Fisher argues for reducing Fed exposures to long-term assets to prevent capital losses, which could lead to political issues. In that same meeting, then Governor Elizabeth Duke claims that "[she] can't quarrel with the political risk of maintaining a large balance sheet(...)".

References

An, S. and F. Schorfheide (2007): "Bayesian Analysis of DSGE Models," *Econometric Reviews*, 26, 113–172.

Berger, D. W., K. Milbradt, F. Tourre, and J. Vavra (2018): "Mortgage Prepayment and Path-Dependent Effects of Monetary Policy," Tech. rep., National Bureau of Economic Research.

Curdia, V. and M. Woodford (2011): "The central-bank balance sheet as an instrument of monetary policy," *Journal of Monetary Economics*, 58, 54–79.

Davis, M. A. and F. Ortalo-Magné (2011): "Household expenditures, wages, rents," *Review of Economic Dynamics*, 14, 248–261.

Di Maggio, M., A. Kermani, and C. Palmer (2016): "Unconventional Monetary Policy and the Allocation of Credit," *Columbia Business School Research Paper*.

Eichenbaum, M., S. Rebelo, and A. Wong (2018): "State Dependent Effects of Monetary Policy: the Refinancing Channel," Working Paper 25152, National Bureau of Economic Research.

Elenev, V. (2017): "Mortgage credit, aggregate demand, and unconventional monetary policy," .

Faria-e-Castro, M. (2018): "Fiscal Multipliers and Financial Crises," Working Papers 2018-23, Federal Reserve Bank of St. Louis.

Fuster, A., L. Goodman, D. Lucca, L. Madar, L. Molloy, and P. Willen (2013): "The Rising Gap between Primary and Secondary Mortgage Rates," *FRBNY Economic Policy Review*.

Gagnon, J., M. Raskin, J. Remache, and B. P. Sack (2010): "Large-scale asset purchases by the Federal Reserve: did they work?" *FRB of New York Staff Report*.

Gertler, M. and P. Karadi (2011): "A model of unconventional monetary policy," *Journal of monetary Economics*, 58, 17–34.

Greenwald, D. L. (2018): "The Mortgage Credit Channel of Macroeconomic Transmission," Unpublished, MIT Sloan.

Hamilton, J. D. (2017): "Why You Should Never Use the Hodrick-Prescott Filter," NBER Working Papers 23429, National Bureau of Economic Research, Inc.

Silva, D. H. (2016): "The risk channel of unconventional monetary policy," Tech. rep., Working Paper, MIT.

Stroebel, J. and J. B. Taylor (2012): "Estimated Impact of the Federal Reserve Mortgage-Backed Securities Purchase Program," *International Journal of Central Banking*, 8, 1–42.

A Model Appendix

A.1 Full List of Equilibrium Conditions

Borrower household,

$$
\begin{aligned}
\chi C_t^b &= \chi w_t N_t^b - \chi T_t + \rho_t \left[(m_t^* - \Pi_t^{-1}(1-\nu)m_{t-1}) \right] - \delta p_t^h \chi \bar{H} - \Pi_t^{-1}[x_{t-1} + \nu m_{t-1}] \\
w_t &= \chi_b (N_t^b)^{\varphi} C_t^b \\
m_t^* &\leq \theta^{LTV} p_t^h \chi \bar{H} \perp \mu_t \geq 0 \\
m_t &= \rho_t m_t^* + (1-\rho_t)(1-\nu)\Pi_t^{-1} m_{t-1} \\
x_t &= \rho_t r_t^* m_t^* + (1-\rho_t)(1-\nu)\Pi_t^{-1} x_{t-1} \\
p_t^h(1-\mu_t \theta^{LTV}) &= \Omega_t^h \\
1 - \mu_t &= \Omega_t^m + r_t^* \Omega_t^x \\
\kappa_t^* &= \Omega_t^x(\bar{r}_t - r_t^*) + \left[\frac{m_t^* - \Pi_t^{-1}(1-\nu)m_{t-1}}{m_t^*} \right] (1 - \Omega_t^m - \bar{r}_t \Omega_t^x) \\
\rho_t &= \frac{1}{4} \frac{1}{1 + \exp\left(-\frac{\kappa_t^* - \mu_\kappa}{s_\kappa} \right)} \\
\Omega_t^h &= \mathbb{E}_t \Lambda_{t,t+1}^B \left\{ \xi \frac{C_{t+1}^b}{\bar{H}} - \delta p_{t+1}^h + \rho_{t+1} p_{t+1}^h + (1-\rho_{t+1})\Omega_{t+1}^h \right\} \\
\Omega_t^m &= \mathbb{E}_t \frac{\Lambda_{t,t+1}^B}{\Pi_{t+1}} \left\{ \nu + (1-\nu)\rho_{t+1} + (1-\nu)(1-\rho_{t+1})\Omega_{t+1}^m \right\} \\
\Omega_t^x &= \mathbb{E}_t \frac{\Lambda_{t,t+1}^B}{\Pi_{t+1}} \left\{ 1 + (1-\nu)(1-\rho_{t+1})\Omega_{t+1}^x \right\} \\
\Lambda_{b,t+1} &= \beta_b \frac{C_t^b}{C_{t+1}^b} \\
\bar{r}_t &= \frac{x_{t-1}}{m_{t-1}}
\end{aligned}
$$

Saver households,

$$
\begin{aligned}
q_t &= \mathbb{E}_t \frac{\Lambda_{s,t+1}}{\Pi_t} \\
q_t^m &= \mathbb{E}_t \frac{\Lambda_{s,t+1}}{\Pi_{t+1}} Z_{t+1}^m \\
q_t^a &= \mathbb{E}_t \frac{\Lambda_{s,t+1}}{\Pi_{t+1}} Z_{t+1}^a \\
\Lambda_{s,t+1} &= \beta_s \frac{C_t^s}{C_{t+1}^s} \\
w_t &= \eta_s C_t^s (N_t^s)^\varphi \\
\Psi_t &= r_t^* q_t^a + q_t^m \\
1 + \Psi_t &= \eta_{m,t} \left(\frac{(1 - f_t^{QE}) \rho_t m_t^*}{\rho_{ss} m_{ss}^*} \right)^{\phi_m}
\end{aligned}
$$

Government,

$$
\begin{aligned}
T_t + \Pi_t^{-1}(Z_t^m m_{t-1}^G + Z_t^a x_{t-1}^G) + q_t B_t^G = G + \Pi_t^{-1} B_{t-1}^G + q_t^m m_t^G + q_t^a x_t^G \\
m_t^G =& f_t^{QE} \ell_t^* + (1-\nu)(1-\rho_t)\Pi_t^{-1} m_{t-1}^G \\
x_t^G =& f_t^{QE} \ell_t^* + (1-\nu)(1-\rho_t)\Pi_t^{-1} x_{t-1}^G \\
\frac{1}{q_t} &= \max \left\{ 0, \left[\frac{1}{q_{t-1}} \right]^{\rho_i} \left[\frac{1}{\bar{q}} \left(\frac{\Pi_t}{\bar{\Pi}} \right)^{\phi_\pi} \left(\frac{Y_t}{\bar{Y}} \right)^{\phi_y} \right]^{1-\rho_i} mp_t \right\} \\
f_t^{QE} &= \frac{\exp(QE_t) - 1}{1 + \exp(QE_t)} \\
T_t &= T_{ss} \left[\frac{B_t^G}{B_{ss}^G} \right]^{\phi_T}
\end{aligned}
$$

Aggregate conditions,

$$
\begin{aligned}
Y_t &= A_t N_t \\
N_t &= \chi N_t^b + (1-\chi) N_t^s \\
\varepsilon \left(\frac{\varepsilon - 1}{\varepsilon} - \frac{w_t}{A_t} \right) &= \zeta \mathbb{E}_t \Lambda_{s,t+1} \frac{Y_{t+1}}{Y_t} \frac{\Pi_{t+1}}{\Pi} \left(\frac{\Pi_{t+1}}{\Pi} - 1 \right) - \zeta \frac{\Pi_t}{\Pi} \left(\frac{\Pi_t}{\Pi} - 1 \right) \\
Z_t^m &= \nu + (1-\nu)\rho_t + (1-\nu)(1-\rho_t) q_t^m \\
Z_t^a &= 1 + (1-\nu)(1-\rho_t) q_t^a \\
Y_t &= \chi C_t^b + (1-\chi) C_t^s + G + \delta p_t^h \bar{H}
\end{aligned}
$$

Exogenous shocks,

$$\begin{aligned}
\log A_t &= \rho_a \log A_{t-1} + \sigma_a \epsilon_t^a \\
\log mp_t &= \rho_r \log mp_{t-1} + \sigma_r \epsilon_t^r \\
QE_t &= \rho_{QE} QE_{t-1} + \sigma_{QE} \epsilon_t^{QE} \\
\log \eta_{m,t} &= (1-\rho_\eta) \log \eta_{m,ss} + \rho_\eta \log \eta_{m,t-1} + \sigma_\eta \epsilon_t^\eta
\end{aligned}$$

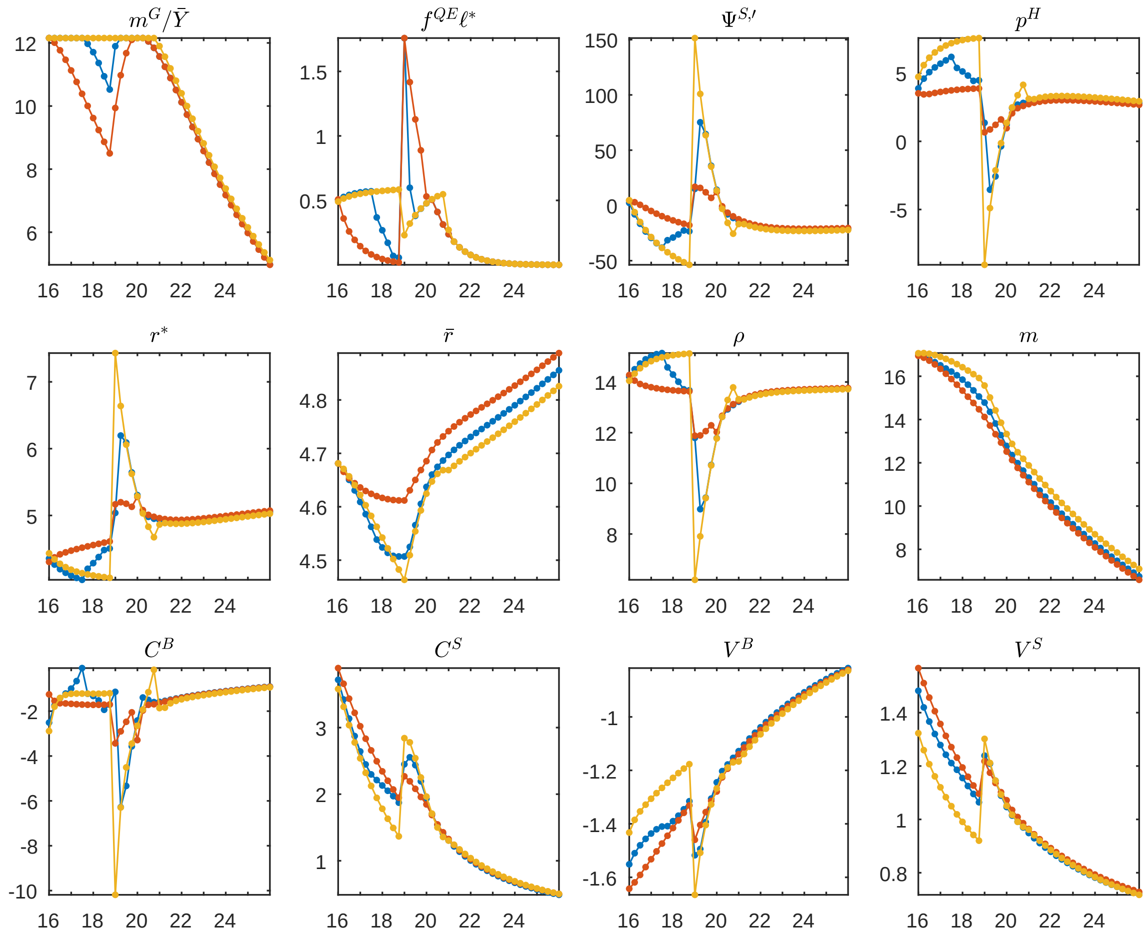

Figure 10: Monetary policy normalization scenarios with unexpected crisis in 2019Q2, QE4, and a political constraint on the Fed's balance sheet. These figures show the transition dynamics from 2016Q1 to 2025Q4. In each panel, the blue line shows the baseline balance sheet normalization scenario, consistent with how the Federal Reserve has been reducing the balance sheet in the data starting in 2017Q4. The red line shows a scenario in which balance sheet normalization begins 7 quarters earlier. The orange line plots a scenario in which balance sheet reductions do not begin until 2020Q3. Balance sheet paths are plotted in the top-left panel, while the purchases they imply are plotted directly to the right. Interest rates r^* and $\bar{r}$ and the refinancing share ρ are annualized and in percent. Value functions V^B and V^S are deviations from steady state. All other variables are in percent deviations from steady state.