The Extent and Consequences of Federal Reserve Balance Sheet Shrinkage

ABSTRACT This paper discusses the process of balance sheet shrinkage that the Federal Reserve is currently undertaking. I argue that the overall balance sheet is unlikely to shrink by much and that it will remain a much larger share of nominal GDP than it was before the COVID-19 pandemic. I examine the effects of balance sheet shrinkage on asset prices, taking the perspective that these effects are mostly likely to be narrow, that is, specific to the price of the asset that the market has to absorb rather than spilling over to fixed income prices more generally. I argue that the effects of reducing the Fed’s holdings of Treasuries can be thought of as equivalent to the Treasury increasing the amount and maturity of its issuance. I estimate that this will have very small effects on term premia and bond yields. The reduction of the Fed’s holdings of mortgage-backed securities might have larger effects on the yields of these securities, especially if the Fed starts selling these securities. Any substantive macroeconomic effect of balance sheet runoff is likely to operate through mortgage rates and the housing market.

In May 2022, the Federal Open Market Committee (FOMC) announced plans for shrinking the size of the balance sheet, a plan often referred to as quantitative tightening (QT). This did not call for any outright asset sales but rather limiting reinvestment of maturing assets. The program was phased in over three months but has now reached its full extent, according to which Treasuries are reinvested only to the extent that they exceed a

Conflict of Interest Disclosure: Jonathan H. Wright is a visiting scholar at the Federal Reserve Bank of Philadelphia. Other than the aforementioned, the author did not receive financial support from any firm or person for this paper or from any firm or person with a financial or political interest in this paper. The author is not currently an officer, director, or board member of any organization with a financial or political interest in this paper.

$60 billion per month cap; for principal repayments of mortgage-backed securities (MBS), the corresponding cap is $35 billion. Most of the MBS in the system open market account (SOMA) portfolio are at low coupon rates and have low prepayment speeds given that homeowners will not choose to refinance. As such, the actual pace of shrinkage of MBS is likely to be much less than $35 billion. Ennis and Kirk (2022) projected a pace slightly above $20 billion per month over the next two years.\textsuperscript{1} Still, the total pace of Fed balance sheet shrinkage is about twice as fast as it undertook in 2017–2019. The Federal Reserve is also applying the $60 billion per month of Treasury redemptions first to coupon securities and then to bills. This will gradually lower the weighted average maturity of SOMA Treasury holdings.\textsuperscript{2}

This paper discusses the likely effects of the program of balance sheet shrinkage, starting with the likely extent of QT and followed by its asset market and macroeconomic impacts. Throughout, I am thinking of the impacts of QT as the difference in outcomes (such as Treasury yields) with the balance sheet actually chosen by the Fed relative to what would have occurred if the Fed were instead to keep the balance sheet constant as a share of nominal GDP at its peak level of 37 percentage points.

I. The Extent of QT

The implementation of monetary policy changed dramatically in the wake of the global financial crisis and the authorization that Congress gave the Fed to pay interest on reserves. In contrast to the old system of monetary policy implementation with scarce reserves, which is unfortunately still often taught in intro macroeconomics classes (Ihrig and Wolla 2020), the Fed now sets the interest rate on reserves which puts a floor on banks’ reserve demand, and then the Fed supplies an amount of reserves that ensures that equilibrium is always on the flat part of the reserve demand curve. Since only depository institutions are eligible to receive interest on reserves, it turns out that this can lead to segmentation whereby short-term interest rates are generally well below the level of interest on reserves. To counter this, the Fed has introduced a system of reverse repos which allow the Fed to effectively pay interest to other entities, such as money market

\textsuperscript{1} Their assumption, writing in spring 2022, for the terminal level of thirty-year mortgage rates was 5 percent, and by fall, rates had already soared well beyond that, but as they note, the MBS in question have such low coupons that refinancing is unattractive in any case.

\textsuperscript{2} At present, the weighted average maturity of the SOMA Treasury holdings is 8.3 years, whereas that of marketable Treasuries outstanding is 6.2 years.
mutual funds and government-sponsored enterprises. Thus, in effect, the Fed supplies ample reserves and puts two floors on interest rates, one via interest on reserves and the other via interest on reverse repos.\(^3\)

Figure 1 shows the evolution over time of the liabilities of the Federal Reserve System, all scaled by nominal GDP. The size of the Fed’s balance sheet soared after the financial crisis and the subsequent shrinkage was quite limited.\(^4\) During the COVID-19 pandemic, the balance sheet expanded again to a peak of 37 percent of GDP and has now begun to shrink.

3. See Dawsey, English, and Sack (forthcoming) for a clear exposition of the Fed’s new implementation system.
4. I am defining the size of the Fed balance sheet as the line in the H.4.1 release labeled “Total factors supplying reserve funds,” which is a bit bigger than “Securities held outright” because the Fed has assets other than securities.
As the Fed shrinks the balance sheet, it has made it clear that it intends to keep this new system of monetary policy implementation but with as small a balance sheet as possible. In May 2022, the Federal Reserve Bank of New York laid out projections for the process of balance sheet shrinkage (Federal Reserve Bank of New York 2022). They projected that the balance sheet would shrink to 22 percent of nominal GDP, consisting of bank reserve balances of 8 percent of GDP, negligible reverse repos, and 14 percent of nominal GDP in currency and other liabilities (so-called autonomous factors). At the current pace of balance sheet shrinkage, this would continue until 2025 with a Fed balance sheet of about $5.9 trillion before growth would resume.5

I am very skeptical that the Fed will ultimately shrink the balance sheet by anything like that much. Bank reserve balances of 8 percent of GDP would get very close to the point where, as in September 2019, banks ended up being on the steep part of their reserve demand curve, causing sharp spikes in the federal funds rate. Copeland, Duffie, and Yang (2021) find that both before and after the spike in September 2019, there were strains in intraday payments that can be tied directly to a shortage of reserves—the sharp rise in rates was an extreme manifestation of a broader shortage of reserves.6 Furthermore, there is reason to believe that the kink in bank reserve demand is now at a higher level relative to nominal GDP than it was in September 2019. For one thing, since the end of 2019, whereas nominal GDP has risen 16 percent, the total assets of commercial banks in the United States have climbed 28 percent, and that seems a more natural way of scaling reserve demand (Afonso and others 2022). Afonso and others (2022) estimate a nonlinear reserve demand curve using a time-varying vector autoregression and find that the curve has shifted upward and to the right over time. Having bank reserves at around 10 percent of nominal GDP seems a more likely steady state and is close to, but a bit below, the estimates of Afonso and others (2022).7

In September 2022, the Fed was offering an overnight reverse repo facility at a fixed offering rate that is 10 basis points below the rate of interest on

5. The projections assume a period of tapering at the end of QT, but I abstract from this for simplicity.
6. The Fed has put in place a standing repo facility as a backstop to prevent spikes in rates, but it is not being used much and may well end up subject to the same problem of stigma that has arisen with the discount rate.
7. Afonso and others (2022) scale reserve demand by bank assets and estimate that the reserve demand curve now becomes steep with reserves below 13–14 percent of bank assets, which corresponds to about 11–12 percent of nominal GDP.
reserves, and the usage of this facility was around 10 percent of nominal GDP, as shown in figure 1. The Fed can, of course, make this facility less attractive by widening the interest spread relative to interest on reserves or by restricting access. But that then weakens its ability to control short-term interest rates and runs the risk that the interest paid to commercial banks on reserves will be substantially above the overall level of short-term interest rates. Moreover, once the Fed shrinks the balance sheet to the point that it does not have the capacity to offer a large overnight reverse repo facility, there is no ready way of going back to restart it again. For these reasons, I think that the Fed will keep room on its balance sheet for a reverse repo facility of at least 6 percent of nominal GDP, which would still involve shrinking it substantially from its current size.

Lopez-Salido and Vissing-Jorgensen (2022) estimate the relationship between the spread of the effective federal funds rate less interest on reserves and the size of reserves plus overnight reverse repos. They conclude that the sum of reserves and overnight reverse repos can be reduced to 15.3 percent of nominal GDP, while avoiding daily spikes. This is almost identical to the 16 percent that I am assuming. With autonomous factors at 14 percent of nominal GDP, as in the SOMA projections, all this implies a steady-state level of the Fed balance sheet of about 30 percent of nominal GDP. At the current pace of balance sheet shrinkage and with the Survey of Professional Forecasters projections for nominal GDP, this will be attained in the middle of 2023 at a level a bit below $8 trillion. And while I can certainly see circumstances in which the balance sheet shrinkage would proceed further, there are also circumstances in which it could end earlier still. Measures of liquidity in the Treasury market are rather poor at the moment. Figure 2 shows the average absolute fitting

8. In the reverse repo facility, the Fed transfers a security to a counterparty, that counterparty deposits cash with a bank, and that bank’s reserves with the Fed are debited. The transaction is unwound the next day. Thus the reverse repo extinguishes reserves. If reverse repo facilities are largely ended and the balance sheet is shrunk so that it is equal to the sum of autonomous factors and the minimum level of reserves demanded by banks on the flat part of the demand curve, then subsequently restarting reverse repos in size in a crisis will create reserve scarcity.


10. One of the autonomous factors that has grown substantially is the Treasury General Account (TGA), which is both high and volatile. Swings in the TGA mechanically cause shifts in reserves. Before the financial crisis, the Treasury would instead hold its cash at private banks. If the Fed could push the TGA back to private banks, then they could keep balance sheet shrinkage going a bit longer.
error on Treasury securities—a measure of liquidity proposed by Hu, Pan, and Wang (2013)—and this is at a high level at present. Further deterioration in Treasury market liquidity might lead the Fed to end balance sheet shrinkage early. Demand for the overnight reverse repo facility may remain elevated, and the Fed may be reluctant to do much to restrict its usage. An economic downturn that comes sooner than expected might also lead the process to stop early.

When the Fed does stop shrinking the overall size of its balance sheet, it intends to continue shrinking holdings of MBS with a view to eventually reverting to something close to a Treasuries-only balance sheet, and I assume that the Fed will do this. But that would mean that its holdings of Treasuries would have to expand somewhat faster than nominal GDP to keep the overall balance sheet constant as a share of GDP. All in all, I expect
that the decline in the Fed’s holdings of Treasuries will be both small (relative to nominal GDP) and transitory. Figure 3 illustrates a prospective path of Treasuries and MBS as shares of nominal GDP computed on the assumption that Treasuries fall by $60 billion per month until the total balance sheet hits 30 percent of nominal GDP and that MBS fall by $22 billion per month for the remainder of the projection period. The decline in Treasuries as a share of GDP gets completely reversed as MBS continue to run off.

The Fed has left open the possibility of outright sales of MBS, and it will have to do this if its goal of holding primarily Treasuries on the Fed balance
sheet is to be anything more than an abstract aspiration. But of course if there are outright sales of MBS, then for a given balance sheet size, the Fed has to hold more Treasuries.

There is clearly something of a ratchet effect going on here—the expansion of the balance sheet creates more demand for reserves and other Fed liabilities, meaning that the expansion of the balance sheet that we saw in the global financial crisis never got reversed, and I am arguing that the same is likely to happen again as a result of the response to the pandemic. Nelson (2019) and Smith and Valcarel (2021) show that interest rate responses to declining reserves are steeper than to increasing reserves. Acharya and others (2022) argue that quantitative easing (QE) induces banks to expand their short-term liabilities, and this in turn boosts their demand for reserves—in this model, once the balance sheet expands, it becomes hard to shrink it back. A ratchet effect is something of a downside to QE that probably will and should be taken into account before its use in the future.

The Fed could, of course, decide to make fundamental changes in its monetary policy implementation framework. US regulators, including the Fed but also the Office of the Comptroller of the Currency and the Federal Deposit Insurance Commission, have encouraged banks to hold reserves. They view reserves as the most liquid asset on a bank’s balance sheet, preferable even to Treasury securities, which might be hard to sell quickly in a crisis without triggering a destabilizing fire sale (Bush and others 2019). A big balance sheet can support financial stability by expanding the supply of safe short-term assets and diminishing the incentive of the private sector to create assets that purport to be both safe and liquid but are in fact neither (Greenwood, Hanson, and Stein 2016). And yet, the objective of regulators to some extent works at cross-purposes with monetary policy that wants to achieve a smaller balance sheet. Materially shrinking the balance sheet would involve substantial changes and agreement among regulators. Nelson (2019, 2020) argues for going back to something like the scarce reserves regime that existed before the global financial crisis. He highlights costs to a big balance sheet, including the fact that tightening monetary policy with a big balance sheet can lead to the Fed’s income being insufficient to cover expenses and a deferred asset being created, an outcome that appears imminent at the time of writing. This clearly has a political and optical cost to the Federal Reserve and is conceivably even a threat

11. McAndrews and Kroeger (2016) also argue that a bigger balance sheet promotes resilience in the payments system.
to its long-run independence (English and Kohn 2022). On the other hand, the current monetary policy implementation framework operates well, in the sense of keeping all short-term rates, including the federal funds rate and repo rates, close to the target with little day-to-day volatility (Duffie and Krishnamurthy 2016; Logan 2019). The decline in the Fed’s income is a consequence of long-duration assets being on the balance sheet, which was a choice to stimulate the economy during the pandemic, not a necessary implication of a large balance sheet. The Fed is (wisely, I think) reluctant to make big changes to a monetary policy framework that works, and so I expect them to continue to stick to an ample reserves approach while making mild to moderate efforts to shrink the balance sheet size.

II. The Effects of QT

There is now a large body of literature on understanding the effects of QE. This is made possible in part by the fact that many of the QE announcements came as a surprise and so the event study methodology can identify the impact effects quite precisely. Economists debate whether QE operates through broad channels—affecting the expected path of policy and term premia on all fixed-income assets—or narrow channels, with the price impact limited to the specific security being purchased. While it surely has elements of both, there is now a lot of evidence that narrow or local supply channels are a very large part of the mechanism of QE.¹²

Identifying the effects of QT is more challenging. Because central banks had a long time to prepare for QT, the announcements did not come as big surprises and so the event study methodology is not as powerful, although some authors have looked at the effects of QT announcements (D’Amico and Seida 2020; Smith and Valcarel 2021). And while it is tempting to think of QT as the inverse of QE, there are many reasons why they are quite different. QE happens in part during a time of financial instability; QT occurs during mostly stable financial markets. QE might reinforce forward guidance and affect the expected future path of policy; QT is very explicitly disconnected from the future path of policy. QE might have been understood to signal that more purchases would be undertaken if needed to drive bond prices higher (Haddad, Moriera, and Muir 2022); no such signal

could possibly be construed from QT. And in the past QE happened at the zero lower bound whereas QT is now occurring away from it, which might make the impact of asset purchases on prices larger (Gagnon and Jeanne 2020). Related to this, D’Amico and Seida (2020) find that balance sheet policy announcements have larger effects when there is more interest rate uncertainty.

To the extent that the effects of QE are thought to operate mainly through narrow channels, it might be reasonable to suppose that the same is true of QT. But otherwise it is hard to draw much inference about the likely impacts of QT from the experience with QE. The Treasury component of QT essentially increases the supply of Treasuries that the arbitrageurs have to absorb—in the framework of Vayanos and Vila (2021)—and also increases the maturity of that supply. As such, we might get some guidance on the likely effects of QT by studying the effects of the supply and composition of Treasury debt during the period before the zero lower bound, as considered by Greenwood and Vayanos (2014). This considers changes in the amount of Treasuries that the market has to absorb without any of the other special features associated with QE.

In an exercise very similar to that undertaken by Greenwood and Vayanos (2014), I regress the ten-year Treasury term premium as estimated by Adrian, Crump, and Moench (2013) onto the maturity-weighted debt-to-GDP ratio of Greenwood and Vayanos (2014). Maturity-weighted debt-to-GDP takes the debt-GDP ratio for each maturity, in decimal form, multiplies it by the maturity in years, and sums them up over all maturities. There is a potential endogeneity problem here in that a larger term premium might motivate debt managers to issue shorter maturity debt. I follow Greenwood and Vayanos (2014) and others by instrumenting the maturity-weighted debt-to-GDP ratio by the unweighted debt-to-GDP ratio, which is purely a function of past fiscal decisions. Table 1 shows the estimates using both ordinary least squares and instrumental variables, which are very similar. Both imply that a one unit increase in maturity-weighted debt-to-GDP increases the ten-year term premium by about 0.34 percentage points. Greenwood and Vayanos (2014) instead regressed the ten-year yield on maturity-weighted debt-to-GDP and controlled for the one-year yield, and they also obtained very similar results. For Treasury securities, we can look up the SOMA holdings and work out both what the maturity-weighted

The debt-to-GDP ratio is at the start of QT and what it will be at any future date, assuming that QT continues at the current pace and making assumptions about reinvestment decisions. From August 2022 to July 2023 (a plausible end date for QT as discussed above), the maturity-weighted debt-to-GDP ratio held by the SOMA portfolio in Treasuries will decline by 0.25. These Treasury securities have to be absorbed by the market, which would drive the ten-year term premium up but only by about 10 basis points. Of course, the actual scope of QT will become clearer over time: if QT carries on for a longer or shorter period than I assume, then the term premium impact would scale up or down.

II.A. Comparison with Other Estimates

Belton and others (2018), writing in the context of QE and surveying the QE literature, propose a rule of thumb: that adding 1 percent of nominal

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MWGDP_t$</td>
<td>$0.34^*$</td>
<td>$0.32^*$</td>
</tr>
<tr>
<td></td>
<td>$(0.13)$</td>
<td>$(0.14)$</td>
</tr>
<tr>
<td>First-stage $F$</td>
<td>869.00</td>
<td></td>
</tr>
</tbody>
</table>

Source: Author’s calculations.

Note: This table reports the results of regressions of the ten-year term premium of Adrian, Crump, and Moench (2013) onto the maturity-weighted debt-to-GDP ratio as calculated by Greenwood and Vayanos (2014) using monthly data from June 1961 to December 2007. The term premium is measured in percentage points. Maturity-weighted debt-to-GDP takes the debt-to-GDP ratio for each maturity, in decimal form, multiplies it by the maturity in years, and sums the results over all maturities. Newey-West standard errors are included in parentheses. Following Lazarus and others (2018), the lag truncation parameter is set to $1.3T^{1/2}$ (rounded to the nearest integer) where $T$ is the sample size.

$^*$ $p < .05$, using the nonstandard fixed-b critical values of Kiefer and Vogelsang (2005).

14. The calculation is laborious but not complicated. For each month, I take the par value of maturing coupon securities from the Federal Reserve Bank of New York, “System Open Market Account Holdings of Domestic Securities,” https://www.newyorkfed.org/markets/soma-holdings. If this amount is less than $60 billion, I assume that the difference is subtracted from the holdings of bills. If it is greater than $60 billion, I assume that the amount of bills is kept fixed and the excess of maturing coupons over the cap are reinvested in coupon securities with an assumed maturity of one hundred months (the approximate weighted average maturity of newly issued nominal notes and bonds). I then compute the exact maturity-weighted value of the portfolio, treating all bills as having a maturity of six months, on August 31, 2022, and July 31, 2023, and scale these by nominal GDP for 2022:Q3 and 2023:Q3 respectively. The nominal GDP numbers are the projections from the August 2022 Survey of Professional Forecasters. This calculation results in the maturity-weighted SOMA Treasury debt-to-GDP ratio declining from 1.73 to 1.48.
GDP to the supply of ten-year equivalent Treasuries raises the term premium by about 6 basis points. As the duration of a ten-year Treasury is about 8.5 years, this rule would say that adding 8.5 percentage points to the maturity-weighted debt-GDP ratio would increase the term premium by 6 basis points. As the increase in the maturity-weighted debt-GDP ratio to be absorbed by the market is 25 percentage points, this corresponds to an 18 basis point increase in the term premium. That’s a little larger than my estimate above, but since this rule was calibrated to QE, it may well have larger effects than QT. Crawley and others (2022) use the FRB/US model to estimate the impact of QT on ten-year term premia and get an estimate of around 50 basis points, but this is partly because they are following the New York Fed’s SOMA projections, which are for a much bigger extent of QT. If one takes the term premium in Crawley and others (2022) as of mid-2023, when I expect QT will end, it is an increase of about 20 basis points—also in the same ballpark. Wei (2022) estimates the effect of a $2.2 trillion runoff of Treasuries, which is also much bigger than I am assuming, and finds an effect of only 6 basis points.

I don’t attempt to estimate the macroeconomic impacts of this increase in the term premium, but it is a small and temporary shock to the slope of the yield curve, and generally small and temporary shocks have small macro effects.

II.B. MBS

A different question is the impact of the redemptions of the Fed’s MBS portfolio. Unless a new crisis in the housing market develops, the Fed is intent on reducing these holdings permanently and so there may be larger effects here, assuming as I do that the reduction is permanent. Taking the view of asset purchases having predominantly narrow or local effects, any potential impact would be mainly in MBS yields and consequently in mortgage rates and the housing market.

As noted by many authors (for example, Krishnamurthy and Vissing-Jorgensen 2011), the spread of MBS rates over Treasuries, once the MBS yields have been adjusted for their embedded prepayment option, is a natural place to look for the impacts of asset purchases. The current-coupon options-adjusted spread is shown in figure 4. The effects of the QE operations that involved MBS purchases can clearly be seen in this figure: during both the third round of QE and the pandemic, this spread turned negative. In contrast, the limited MBS redemptions in 2017–2019 did not show any effect on this spread. As noted earlier, prepayment rates are
very slow at the moment because refinancing is uneconomic. Nonetheless, MBS spreads have risen a bit since the QT program reached its full tilt in September. We have very little to go on, but either MBS redemptions are going to go on for a long time or outright sales will begin, and in either case it seems quite plausible that MBS spreads will widen further. At least, this is the place where we should look for material asset price implications of QT. And the effect of MBS balance sheet shrinkage (relative to the counterfactual of holding the Fed’s MBS holdings fixed as a share of nominal GDP) is a permanent one. The macroeconomic effect of the MBS spread widening could be important by slowing the housing market, which is already being cooled substantially by the effect of tighter-than-expected conventional monetary policy.

Source: Bloomberg.

Note: The figure shows the spread between the current coupon Fannie Mae MBS—adjusted for the embedded option—and the corresponding Treasury security.
III. Summary and Conclusions

I expect the Fed to resume balance sheet growth with a much higher level of the balance sheet, scaled by nominal GDP, than before the COVID-19 pandemic. The evidence that I have pointed to mostly suggests very small effects on asset prices. If there are to be substantive impacts, I would look to rising MBS spreads adding to the cooling of the housing market.

ACKNOWLEDGMENTS I am grateful to Bill English for helpful comments on an earlier draft. All errors are my sole responsibility.
References


