The Great Recession, which began in late 2007 and continued until mid-2009, demarcates some key changes in U.S. monetary policy. In 2015, the Federal Reserve’s balance sheet is much larger than before the Great Recession. From December 2007 to October 2014, the Fed’s balance sheet (either total assets or total liabilities, which are equal) increased more than fourfold. Since late 2008, the Fed’s target interest rate, the federal funds rate, has been close to zero. This long period of a zero interest rate policy, or ZIRP, is unprecedented since the Treasury-Federal Reserve Accord of 1951 modernized the approach to central banking in the United States.

In its “Policy Normalization Principles and Plans” (Board of Governors, 2014), the Federal Open Market Committee (FOMC) proposed a program that should ultimately return the Fed’s balance sheet to a state similar to that of December 2007. This does not literally mean a return to a balance sheet of the same nominal size (in U.S. dollars) as before the financial crisis, but to a balance sheet that allows the Fed to implement monetary policy in the same way it did in December 2007, as discussed later in this article. In addition, the FOMC’s principles and plans outline a sequence of actions by which “liftoff”—departure from a ZIRP—would be achieved, followed by a reduction in the size of the Fed’s balance sheet.

The purpose of this article is first to take stock of the state of U.S. monetary policy and how that state was reached. This discussion is followed by an analysis of how the FOMC envisions normalization and the pitfalls that may arise on the road to normalization.
There are two key aspects of the Fed’s monetary policy that should be considered: (i) the size and composition of the Fed’s balance sheet and (ii) the Fed’s interest rate policy. Each of these is analyzed in turn.

**The Size and Composition of the Fed’s Balance Sheet**

Figure 1 summarizes the time series of items on the assets and liabilities sides of the Fed’s balance sheet from 2007 to the present. Clearly, there has been a remarkable increase in the size of the balance sheet since the period prior to the financial crisis. (The critical period of the financial crisis period is defined here as the fall of 2008 through early 2009). For example, total assets held by the Fed increased from $891 billion in December 2007 to $4,498 billion in December 2014, a more than fourfold increase—or, alternatively, an increase from 6.1 percent of annual gross domestic product (GDP) to 25.3 percent of GDP.

While the size of the Fed’s balance sheet has increased greatly since the Great Recession began in December 2007, such an increase is by no means unprecedented, nor is it a world record. For example, at the end of September 2014, the Swiss National Bank, the Bank of Japan,
the People’s Bank of China, and the National Bank of Denmark held assets that were, respectively, 80.1 percent, 57.0 percent, 52.3 percent, and 26.3 percent of GDP, while the Fed held assets that were only 25.3 percent of GDP. Of course, at that time, the Fed’s balance sheet was larger than those of some central banks: Bank of England (22.5 percent of GDP), the European Central Bank (20.2 percent), the Swedish Riksbank (12.1 percent), the Reserve Bank of Australia (8.9 percent), and the Bank of Canada (4.6 percent).

It is important to note that much of the increase in the size of the Fed’s balance sheet has occurred since the Great Recession ended in June 2009, when the Fed’s assets stood at $2,026 billion, or 14.1 percent of GDP. Further, the sources of the increases on the asset side of the Fed’s balance sheet during the Great Recession were quite different from the sources after the Great Recession. As shown in Figure 1, much of the increase in the Fed’s assets during the Great Recession resulted from its lending programs—both conventional and unconventional—and purchases of troubled assets and commercial paper. These interventions were responses to the global financial crisis. However, the post-recession increases in the Fed’s assets were mainly purchases of Treasury securities, mortgage-backed securities (MBS), and agency securities. Further, a key element of the Fed’s post-Great Recession policy was a lengthening in the average maturity of assets in the Fed’s portfolio. In December 2007, the Fed’s securities were entirely Treasury securities, and of these 32.1 percent were Treasury bills (short-term government debt). In December 2014, the Fed held no short-term Treasury securities; 41 percent of its security holdings were long-maturity MBS and 58.1 percent of its security holdings were long-maturity Treasury notes and bonds.

The primary liabilities on the Fed’s balance sheet before the Great Recession were currency outstanding and reserves. Currency outstanding consists of the Federal Reserve notes and coins held outside the banking system, while reserves are essentially the transaction accounts of financial institutions held with the Fed, plus a small amount of currency held in bank vaults and ATMs. In December 2007, $792 billion in currency outstanding financed most of the Fed’s asset portfolio of $891 billion. Reserves were a relatively small amount, at $4.2 billion. From December 2007 to December 2014, the stock of currency outstanding increased from $792 billion to $1,299 billion, or from 5.4 percent of annual GDP to 7.3 percent of GDP. Thus, there has been a substantial increase in the demand for currency since the beginning of the Great Recession, but that increase was far from sufficiently large to finance the expansion in the Fed’s assets. From December 2007 to December 2014, the quantity of reserves outstanding increased from 0.5 percent of total Fed liabilities to 52.9 percent. As well, reverse repurchase (repo) agreements (an interest-bearing component of Fed liabilities) accounted for 11.3 percent of Fed liabilities in December 2014.

**Interest Rate Policy**

Figure 2 shows the federal funds rate, the Fed’s policy target, from 1954 to the present. Since its meeting December 16, 2008, the FOMC’s target federal funds rate has been in the range of 0 to 0.25 percent. As Figure 2 shows, the federal funds rate was near zero at times during the 1950s and early 1960s, but the period from December 2008 to the present—more than six years of a ZIRP—is unprecedented in the United States.
A key change in the Fed’s interest rate policy was the switch to a regime that pays interest on reserves. With passage of the Financial Services Regulatory Relief Act of 2006, Congress authorized the Fed to pay interest on reserves effective October 1, 2011, but implementation was moved up to October 1, 2008, as part of the Fed’s response to the financial crisis.

The argument made by economists for paying interest on reserves appeals in part to economic efficiency. As first pointed out by Friedman (1969), when the nominal interest rate on safe assets is positive, economic agents will tend to hold too little money (in real terms) for transaction purposes. One solution to this economic inefficiency problem is for the government, or the central bank, to pay interest on all money balances—including holdings of currency and reserves, which together constitute the stock of outside money. Paying interest on currency is impractical: Currency is held by the public in widely dispersed locations, so payment of interest on this currency is extremely costly, if not infeasible, for the Fed. However, paying interest on reserves, which are account balances held by financial institutions with the Fed, is straightforward—as easy as the payment of interest by a private bank on a transaction account. If interest is not paid on reserves, there is effectively a tax on financial intermediation, because reserve requirements are still imposed in the United States on specific deposit liabilities. Thus, if reserves earn zero interest and banks are required to hold such reserves, this introduces a distortion and a loss in economic welfare.

A second rationale for interest on reserves (see, e.g., Goodfriend, 2002) is that if there is a sufficiently large stock of excess reserves outstanding overnight, then in a frictionless overnight financial market, the interest rate on reserves will determine the overnight interest rate. As an example of this mechanism, consider the following crude model of the federal funds market.
shown in Figure 3. In the figure, the nominal federal funds rate $R'$ is on the vertical axis, and the quantity of reserves $Q$ is on the horizontal axis. The demand for reserves is given by the curve $D$, which slopes downward due to a liquidity effect. Thus, to induce financial institutions to hold a higher stock of overnight reserves, the federal funds rate must be lower, as the federal funds rate is the opportunity cost of holding overnight reserves. But, the federal funds rate is bounded by the discount rate $R^d$ and the interest rate on reserves (IOER). The upper bound is determined by the fact that financial institutions with reserve accounts can borrow from the Fed at the discount rate, which is sometimes treated as a separate policy instrument by the Fed; these institutions would not borrow on the federal funds market at a rate higher than that rate. Similarly, the lower bound is determined by the ability of financial institutions to lend to the Fed at the IOER; these institutions would not wish to lend on the federal funds market at a lower rate.

Note that “IOER” typically is used to denote “interest on excess reserves,” where excess reserves are the quantity of reserves held in excess of a bank’s reserve requirement. U.S. law specifies that the interest rates on required reserves and excess reserves could be different, but for practical purposes the Fed has been setting the interest rates on both required reserves and excess reserves at 0.25 percent.

Before the financial crisis, when financial institutions in the United States were not paid interest on reserves and the quantity of reserves held in excess of reserve requirements was tiny, the federal funds market worked as captured in Figure 3, with $IOER = 0$. The position of the demand curve $D$ on any given day was determined by activity in the financial payments system. For example, large shocks can be associated with the overnight lending and borrowing activities of large financial institutions, the day of the week, the month in the year, or the day
within a particular reserve-averaging period (the period over which average reserves held for a bank must exceed the average required reserves on average deposit liabilities). Some of these large shocks are predictable, whereas others are not. Thus, on a given day in the pre-financial crisis period, the New York Fed would need to predict the position of the demand curve in Figure 3. The supply curve $S$ is vertical, determined by the intervention of the New York Fed in the overnight financial market. Typically, the New York Fed would determine the quantity of reserves outstanding overnight by varying the quantity of repo agreements (collateralized lending by the Fed to financial institutions). Standard open market operations—purchases and sales of Treasury securities by the Fed—determined the medium- and long-run quantity of reserves in the system. In terms of Figure 3, $R^T$ is the Fed’s target fed funds rate, and the goal of the Fed on a given day was to intervene so that the supply curve $S$ intersects the demand curve $D$ at the target interest rate.

In the planning stages for payment of interest on reserves in the United States, it was thought that, with a sufficiently large stock of interest-bearing reserves outstanding, the federal funds market would work as in Figure 4. That is, if the supply curve $S$ lies sufficiently far to the right, the IOER should determine the federal funds rate. Such a system of central bank intervention is called a floor system (as the floor on the overnight rate, the IOER, determines the overnight rate), while the system depicted in Figure 3 is a channel system (or corridor system) under which the overnight rate falls in a channel defined by the central bank’s lending rate (the upper bound on the channel) and IOER (the lower bound on the channel). In principle, a floor system should greatly simplify interest rate targeting for the Fed. As an intermediate step to hitting a given target for the federal funds rate, instead of having to forecast the daily demand for reserves, the New York Fed could simply ensure that a sufficiently large quantity
of reserves was outstanding each night. Then the administered IOER, set at the target federal funds rate, should determine the federal funds rate.

Economists such as Goodfriend (2002) had good reasons to think that the payment of interest on reserves in the United States would lead to the potential for a smoothly operating floor system of central bank intervention to target overnight interest rates. Indeed, some floor systems have worked well in practice, essentially as depicted in Figure 4. For example, the Bank of Canada operated a floor system from April 2009 until May 2010. The Bank of Canada usually intervenes through a channel system, whereby the target overnight interest rate is bounded by the rate at which the Bank of Canada lends and the interest rate the Bank pays on reserves, much as depicted in Figure 3. Standard operating procedure is for the Bank to set its lending rate (the “Bank Rate”) at 0.25 percent above the target overnight rate and the interest rate on reserves at 0.25 percent below the target overnight rate. But from April 2009 to May 2010, the overnight rate and the interest rate on reserves were both at 0.25 percent, with the Bank Rate at 0.50 percent. As shown in Figure 5, the Bank accomplished its goal by ensuring that about $3 billion (Canadian dollars) in reserves was outstanding each night. Note in the figure that the Bank of Canada successfully pegged the overnight rate at the interest rate on reserves (0.25 percent) during this period.

In contrast to the 2009-10 Canadian experience, the floor system in place in the United States since late 2008 has not been successful, if success is measured by the ability of the central bank to hit a particular overnight target rate. Despite an extremely large quantity of reserves outstanding overnight since the financial crisis, the federal funds rate has not been pegged by the IOER (Figure 6). The difference between the IOER and the federal funds rate has been variable and surprisingly large—sometimes as much as 20 basis points. This explains why the

Figure 5
Canada Overnight Rate and Bank of Canada Reserves

NOTE: CAD, Canadian dollars.
SOURCE: Canadian Socio-economic Information Management System (CANSIM).
FOMC has chosen to state its target federal funds rate as a range of 0 to 0.25 percent over this period rather than quoting a target rate.

**HOW DID WE GET HERE? THE RATIONALE FOR A ZIRP AND QUANTITATIVE EASING**

Understanding the issues involved in policy normalization requires delving into the reasons for the Fed’s unconventional steps of expanding its balance sheet and lowering its interest rate target to near zero for such a long period. As noted in the first section, much of the balance-sheet increase occurred after the Great Recession, so these unusual interventions were associated not just with the financial crisis but with its aftermath as well.

**ZIRP and Forward Guidance**

The Fed has kept its funds rate target near zero for more than six years. Why has the ZIRP lasted so long? At least part of the rationale for a ZIRP came from New Keynesian (NK) macroeconomic theory, popularized by Woodford (2003) and others. The argument is articulated in more technical terms in Werning (2011). Though the members of the FOMC have diverse views on how (or perhaps whether) to model the macroeconomy, their actions appear to be consistent with the basic NK model. Here I sidestep the issue of whether NK theory is a useful guide for monetary policy, a matter subject to some debate.

In a basic NK model, sticky prices and wages cause an inefficiency that monetary policy can correct. Figure 7 provides a simple representation of what happens in the NK model. The real interest rate $R - \pi$ is on the vertical axis; that is, the nominal interest rate $R$ on short-term government debt minus the anticipated inflation rate $\pi$. $Y$, which denotes real GDP, is on the
horizontal axis. The curve IS is the NK IS curve, which certainly looks like the familiar IS curve from traditional Keynesian analysis but is constructed differently (see, e.g., Clarida, Galí, and Gertler, 1999). Basically, the IS curve denotes a set of combinations of real interest rates and aggregate real income for which the market for government debt is in equilibrium. The supply of government debt outstanding is determined by the fiscal authority—the Treasury in the United States—and the demand for government debt is increasing in the real interest rate on that debt, increasing in current aggregate real income $Y$, and decreasing in future aggregate real income $Y_f$. These effects arise because (i) the demand for an asset increases with its rate of return and (ii) assets are held to smooth consumption over time. Thus, if current income is higher, everything else held constant, economic agents will want to smooth consumption by purchasing assets, including government debt, to allow more consumption in the future (as well as more in the present). If future income is expected to be higher, economic agents will want to sell assets to allow more consumption in the present (as well as more in the future).

In this simplified NK model, monetary policy works as follows. In Figure 7, the quantity $r^*$ denotes the “natural real rate of interest,” which is what the real interest rate would be in equilibrium if all prices and wages were flexible. To simplify things, I assume that $Y^f$ and $\pi$ are exogenous in the short run. Monetary policy sets the nominal interest rate at $R_0$ in Figure 7; I also suppose that the economy is initially in long-run equilibrium with the actual real interest rate $R_0 - \pi_0$ equal to $r^*$, the natural real rate of interest.

Suppose that the central bank lowers the nominal interest rate to $R_1$. Given the exogenous anticipated rate of inflation $\pi_0$, the real interest rate then falls to $R_1 - \pi_0$ (see Figure 7), which implies that real income must rise from $Y_0$ to $Y_1$ equilibrate to the market for government debt. Therefore, there is a short-run non-neutrality of money in that monetary policy easing—a
reduction in the nominal interest rate—increases aggregate output. In the long run, the natural real rate of interest must be the actual rate of interest, so if the nominal interest rate stays at $R_1$ forever, then the anticipated inflation rate must fall to $\pi_1$, with $R_1 - \pi_1 = r^*$. This is just the long-run Fisher effect: Monetary policy is neutral in the long run and has no long-run effect on real output or the real interest rate.

In this simplification of the basic NK framework, one way to represent the effects of the financial crisis is as a shift in the IS curve. This is not a “demand shock” in the sense of an old-style Keynesian model, but an increase in the demand for government debt. Such an effect could arise for at least two reasons. First, credit constraints could become tighter for households, causing them to value government debt more as an asset to self-insure against shocks that would otherwise cause them to borrow. Second, an effect of the financial crisis was to effectively destroy part of the stock of assets used as collateral in financial markets—principally, asset-backed securities that were no longer viewed as safe collateral. This reduction in assets would tend to increase the demand for government debt used as collateral.

In Figure 8, the IS curve shifts to the left from $IS_0$ to $IS_1$ as the result of a financial crisis shock. I also assume that the financial crisis shock is persistent, which implies that the natural real rate of interest also falls from $r^*_0$ to $r^*_1$. Initially the nominal interest rate is $R_0$ and the anticipated inflation rate is $\pi_0$ with the real interest rate equal to $R_0 - \pi_0 = r^*_0$. When the shock occurs, if the central bank keeps the nominal interest rate at $R_0$, then aggregate output will fall from $Y_0$ to $Y_1$. The goal of monetary policy in this model is essentially to remove inefficiency in the market for government debt. To remove inefficiency in these circumstances, the central bank would like to lower the nominal interest rate in the short run, with the goal of lowering the real interest rate to the new natural real rate of interest $r^*_1$. 

**Figure 8**
A Financial Crisis Shock in the Simplified New Keynesian Model
However, the nominal interest rate cannot go below zero—that is, there is a zero-lower-bound problem. If the financial crisis shock is a very large one, then a situation as in Figure 8 could develop in which the nominal interest rate is lowered to zero by the central bank, so the actual real rate of interest is \(-\pi_1\). But in Figure 8 \(-\pi_0 > r^*_1\), so the economic inefficiency cannot be removed by the central bank, given the zero lower bound on the nominal interest rate. Even at the zero lower bound on the nominal interest rate, the real rate of interest is still inefficiently high.

In NK models, the central bank is not powerless at the zero lower bound, however, as it can resort to forward guidance. That is, the central bank can make promises about future monetary policy, as outlined in Werning (2011) or Woodford (2012). For example, the central bank could promise higher future inflation and thus lower the real interest rate, which is \(-\pi\), given the central bank’s zero-lower-bound policy. If the central bank is convincing enough, it may be possible to achieve an equilibrium in which the anticipated inflation rate is \(\pi_1 = -r^*_1\), so the central bank can completely remove the economic inefficiency through forward guidance and achieve a level of aggregate output \(Y_3\) in Figure 8.

Though simplified, this captures the basic NK theory behind forward guidance, which in principle rationalizes the forward guidance policies of the Fed following the financial crisis. In the FOMC statement, which follows regular FOMC meetings, the Committee at first expressed its forward guidance in terms of the future of the ZIRP, and this forward guidance changed over time. In January 2009, the Committee stated that the ZIRP would continue for “some time.” Then on March 18, this language changed to “extended period.” On August 9, 2011, this became “at least through mid-2013,” followed on January 25, 2012, by “at least through late 2014” and on September 13, 2012, by “at least through mid-2015.” In the statement on December 12, 2012, the forward guidance language became considerably longer and more specific:

[The Committee] currently anticipates that this exceptionally low range for the federal funds rate will be appropriate at least as long as the unemployment rate remains above 6½ percent, inflation between one and two years ahead is projected to be no more than a half percentage point above the Committee’s 2 percent longer-run goal, and longer-term inflation expectations continue to be well anchored. The Committee views these thresholds as consistent with its earlier date-based guidance. In determining how long to maintain a highly accommodative stance of monetary policy, the Committee will also consider other information, including additional measures of labor market conditions, indicators of inflation pressures and inflation expectations, and readings on financial developments. When the Committee decides to begin to remove policy accommodation, it will take a balanced approach consistent with its longer-run goals of maximum employment and inflation of 2 percent.

This language seems to be at least broadly consistent with what Woodford (2012) appeared to have in mind, and so in principle is in line with a clearly articulated and consistent economic theory. The forward guidance language in the quoted material expresses what the Committee cares about—the unemployment rate and the inflation rate—and gives specific numerical thresholds. However, the statement leaves some questions unanswered: for example, (i)
happens when the unemployment rate falls below 6.5 percent? and (ii) What happens if the forecast for the inflation rate is below 2 percent?

On March 19, 2014, the Committee's forward guidance language changed again, to

In determining how long to maintain the current 0 to ¼ percent target range for the federal funds rate, the Committee will assess progress—both realized and expected—toward its objectives of maximum employment and 2 percent inflation. This assessment will take into account a wide range of information, including measures of labor market conditions, indicators of inflation pressures and inflation expectations, and readings on financial developments. The Committee continues to anticipate, based on its assessment of these factors, that it likely will be appropriate to maintain the current target range for the federal funds rate for a considerable time after the asset purchase program ends, especially if projected inflation continues to run below the Committee's 2 percent longer-run goal, and provided that longer-term inflation expectations remain well anchored.

When the Committee decides to begin to remove policy accommodation, it will take a balanced approach consistent with its longer-run goals of maximum employment and inflation of 2 percent. The Committee currently anticipates that, even after employment and inflation are near mandate-consistent levels, economic conditions may, for some time, warrant keeping the target federal funds rate below levels the Committee views as normal in the longer run.

With the unemployment rate nearing 6½ percent, the Committee has updated its forward guidance. The change in the Committee's guidance does not indicate any change in the Committee's policy intentions as set forth in its recent statements.

This statement contains more information about the Committee's intentions beyond the ZIRP. For example, the language concerning “keeping the target federal funds rate below levels” indicates that the Committee expects that policy may continue to be more accommodative than usual, even after its target interest rate starts to rise.

Therefore, one explanation for the long duration of the ZIRP policy is that the extended time at the zero lower bound was a necessary part of meeting earlier commitments made by the Fed. In line with Werning (2011) and Woodford (2012), it could be argued that to mitigate the large inefficiency immediately after the Great Recession, it was important for the Fed to promise to stay with the ZIRP for a long time. If that was the idea, then one could criticize the FOMC for being vague and inconsistent in its forward guidance. Explicit commitment is important in the Werning (2011) and Woodford (2012) frameworks. With regard to explicitness, the FOMC’s forward guidance used language such as “some time,” “extended period,” and “considerable time”—phrases open to interpretation. On commitment, clearly changes in the calendar dates for the forward guidance and the failure of 6½ percent unemployment to trigger any policy action indicate, if anything, a lack of commitment. Thus, on explicitness and commitment, the Fed’s forward guidance appears to have left something to be desired.

**The Size of the Fed's Balance Sheet: Quantitative Easing**

The post-Great Recession increase in the size of the Fed's balance sheet was driven by asset purchases—quantitative easing (QE). There were four different components to QE after the Great Recession: (i) purchases of long-maturity Treasury debt, (ii) purchases of MBS and
agency securities, (iii) increases in the average maturity of the Fed’s security holdings through exchanges of short-maturity Treasury debt for long-maturity Treasury debt, and (iv) reinvestment—that is, purchases of long-maturity Treasury debt and MBS as those securities mature to hold constant the size of the balance sheet.

When the Great Recession ended in June 2009, the Fed was in the midst of purchasing a substantial quantity of MBS and agency securities, and these purchases continued well into 2010. This program was part of what is sometimes calledQE1. Another element of QE—reinvestment of maturing securities—began in August 2010 and continues today. From November 2010 until June 2011, the Fed purchased a stock of about $600 billion in long-maturity Treasury securities in an asset purchase program sometimes called QE2. Then from September 2011 until December 2012, the Fed engaged in “Operation Twist,” which involved swaps by the Fed of short-maturity Treasury securities for long-maturity Treasury securities.1 Finally, purchases of MBS resumed in September 2012: From January 2013 to October 2014, the Fed purchased a large quantity of MBS and long-maturity Treasury securities in what is sometimes called QE3.

What is the rationale for QE? In principle, QE is another unconventional monetary policy instrument—that allows the central bank to ease monetary conditions when a ZIRP is in place and the short-term nominal interest rate cannot go any lower. But how does it work, exactly? To address this question one needs to think about the outstanding debt of the consolidated government, which is defined as including the Treasury and the Fed. Thus, the consolidated government debt for the United States includes currency and reserves (Fed liabilities) plus debt issued by the U.S. Treasury and held by the public (not including any Treasury debt held by the Fed).

When the nominal interest rate is zero, there is a liquidity trap in the sense that reserves and short-term government debt are roughly identical assets. So, if the central bank swaps reserves for short-term government debt, this should not matter at all—the central bank’s intervention will be neutral, affecting no prices or quantities of consequence in the economy. But QE is not a swap of reserves for short-term government debt. For example, the most straightforward QE intervention by the central bank is a Fed purchase of long-maturity government debt with reserves.

Even if the short-term nominal interest rate is zero, long bond yields can be positive, as they have been in the United States in the post-Great Recession period. This reflects the fact that long-maturity Treasury securities and reserves are different in ways that short-maturity Treasuries and reserves are not at the zero lower bound. This might mean that QE will not be neutral at the zero lower bound—possibly, purchases of long-maturity Treasuries by the Fed will lower long-term bond yields.

To the extent that QE is justified by any existing economic theory, policymakers typically appeal to segmented markets financial theory (see, e.g., Bernanke, 2012). Some segmented markets theories of the term structure of interest rates fall under the umbrella of “preferred habitat theory” (see Vayanos and Vila, 2009). Preferred habitat theory posits that different financial market participants have different preferences for assets of particular maturities. For example, life insurance companies have long-maturity liabilities and these institutions might prefer long-maturity assets to manage maturity risk. Preferred habitat preferences tend to seg-
ment the market for assets by maturity, according to the theory. Alternatively, policymakers sometimes articulate a segmented asset markets theory of QE in terms of “portfolio balance” models (see, e.g., Tobin, 1969). The idea here is a close cousin of preferred habitat theory, as portfolio balance models specify demand and supply functions for assets as functions of the rates of return and wealth. Then, the critical parameters in such a theory are those determining whether, and to what degree, assets are substitutes or complements.

Whether economists are appealing to preferred habitat or portfolio balance models to rationalize QE, the mechanism by which QE is imagined to work is similar. According to the story, if the Fed purchases long-maturity assets, arbitrage will not be complete, and this will tend to increase the prices of long-maturity assets and reduce their yields. Then, lower long-term bond yields might act to increase investment, consumption, and output through a conventional Keynesian mechanism. Policymakers also envision an effect of QE on inflation, though this mechanism is perhaps less clear.

What is notable is an absence of modern economic theory that provides a role for QE. Indeed, some monetary and financial theory shows how particular central bank asset purchases can have no effect at all. For example, Wallace (1981) provides a general theorem for the neutrality of open market operations. Wallace’s theorem is akin to the Modigliani-Miller theorem in corporate finance, which gives conditions under which the financial structure of a firm can be irrelevant. In Wallace’s theorem, under certain conditions the financial structure of the central bank is irrelevant for economic activity, asset prices, and goods prices.

There may be strong reasons to believe that theorems such as Wallace’s are not relevant for thinking about reality, as the conditions required for the theorems to hold are far too stringent. However, the case can be made that monetary policy matters—and can potentially increase economic welfare—because the central bank has some special powers that private financial intermediaries do not. For example, the Fed has a monopoly on issuing currency, and many of the nation’s daytime payments involve the transfer of Fed liabilities—reserves—among financial institutions. But QE may not be an activity for which the Fed has a special advantage. For example, turning long-maturity Treasury securities into overnight assets (reserves) is very close to the private sector activity of turning Treasury securities into overnight repos. If the Fed engages in QE, why would there not be an equal and offsetting effect in the private sector, yielding no net change in any aggregate economic variables?

In some widely used macroeconomic models QE is irrelevant. For example, in basic NK models (as Woodford, 2012, notes) there are complete markets and Ricardian equivalence. Financial arbitrage is perfect, and the quantity and composition of government debt are irrelevant. Further, while some macroeconomists (e.g., Gertler and Karadi, 2013) have studied the effects of QE, they do so without explicitly integrating a central bank and monetary policy into their models. Thus, there is no broader theory integrating mainstream NK theory—which motivates forward guidance—with a theory of QE.

In previous work (Williamson, 2014a), I develop an alternative theory of QE. In the theory, collateral is important in the banking sector for supporting financial intermediation activity, and collateral can be in short supply because the fiscal authority will not expand the quantity of government debt, even though that might be appropriate. Binding collateral constraints then tend to lower the real interest rate on government debt. At the zero lower bound on the
nominal interest rate, this tends to make the inflation rate high, given the Fisher effect. In the model, QE increases the effective stock of eligible collateral, as QE is a swap of short-maturity assets for long-maturity assets, and short-maturity assets are better collateral. Therefore, QE relaxes collateral constraints, decreases nominal long bond yields, increases the short-term real interest rate, and decreases the inflation rate. The latter two effects (the real interest rate increase and the inflation rate decrease) are certainly not what central bankers typically imagine will happen when they resort to QE interventions, though these effects are consistent in the model with an increase in economic welfare.

In general, one can conclude there should be little confidence in our knowledge of the quantitative effects of QE, let alone the qualitative effects. The existing theories are not well developed and have not been confronted by the empirical evidence in ways that shed light on how and to what extent QE works. There is some evidence that QE has important announcement effects on asset prices (see Neely, 2015, and Williamson, 2014b), though such evidence is consistent with a signaling role for QE. That is, QE may matter only to the extent that it acts as forward guidance—signaling the future path of market interest rates. There is little solid evidence about the effects of QE on economic quantities of interest—for example, employment, unemployment, GDP, and inflation.

POLICY NORMALIZATION

The FOMC announced its “Policy Normalization Principles and Plans” on September 17, 2014 (Board of Governors, 2014). There are two key aspects to the Fed’s normalization plans. First, the FOMC will raise its overnight interest rate target when it deems there is no longer as great a need for monetary accommodation. Second, the FOMC wishes to ultimately reduce the Fed’s balance sheet to a size commensurate with operating a channel system—that is, to a range such that the quantity of reserves outstanding will be on the order of pre-financial crisis levels. However, this balance-sheet reduction will occur slowly, with no plans by the FOMC to sell assets, though that option has not been completely ruled out. Primarily, the Fed’s assets will decline as Treasury securities and MBS mature. Treasury securities mature in a predictable manner, but the rate at which MBS decline is uncertain, as this depends in part on refinancing and default on the private mortgages backing the MBS.

QE ended in October 2014, so the next steps by the Fed toward normalizing monetary policy are (i) taking actions to increase short-term nominal interest rates—deciding on “liftoff” from the zero lower bound—and (ii) reducing the size of the balance sheet. The FOMC has made clear that liftoff will occur before any steps are taken toward balance-sheet reduction, so liftoff will happen in the context of a large stock of excess reserves in the financial system. This timing of policy actions (liftoff before balance-sheet reduction) is perhaps puzzling. Indeed, since the Fed first reduced its interest rate target and then resorted to balance-sheet-expanding asset purchases, one might think these actions would occur in reverse as the Fed exits its accommodative stance, with balance-sheet reduction occurring before liftoff. In any case, no rationale appears to have been provided by the FOMC for the sequence of actions laid out in its normalization plans.
Interest Rate Control and the Overnight Reverse Repurchase Agreement Facility

As indicated earlier, the U.S. financial system presents some unique problems for central bank interest rate control under a floor system. While the IOER should effectively determine the federal funds rate, it has already been shown that this has not happened (see Figure 6). That is, the margin between the IOER and the federal funds rate is significant—and variable. Why is this? Under the Financial Services Regulatory Relief Act of 2006, the Fed was not authorized to pay interest on the reserve balances of government-sponsored enterprises (GSEs), including Fannie Mae, Freddie Mac, and the Federal Home Loan Banks (FHLBs). In principle, this should not cause a problem for interest rate control, as the GSEs could earn overnight interest by lending on the federal funds market. That is, a financial institution with a reserve account earning interest on reserves at 0.25 percent could borrow on the federal funds market overnight from a GSE and hold the funds overnight as reserves. In the absence of risk or costs of lending and borrowing, and under perfect competition, the interest rate on such a federal funds loan should be 0.25 percent. But since the federal funds rate has been substantially below IOER, there must be some friction or frictions that inhibit arbitrage in the federal funds market.

At least two factors potentially explain the margin between the IOER and the federal funds rate (see Martin et al., 2013). The first is balance-sheet costs for the financial institutions that can earn interest on reserves. These balance-sheet costs currently arise for two reasons. First, under the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010, the rules for deposit insurance premiums paid by banks to the Federal Deposit Insurance Corporation changed in February 2011. Formerly, a bank’s deposit insurance premium depended only on domestic deposits, but now the premium depends instead on total assets. This means that if a bank borrows on the federal funds market overnight and invests the proceeds in its reserve account with the Fed, its total deposit insurance premium will be higher. This regulatory change was instituted as a means to correct incentive problems for larger banks, which depend more on nondeposit funding than do small banks. But the change has the side effect of increasing a bank’s balance-sheet costs when it borrows on the federal funds market and holds the proceeds as reserves.

Second, given how capital requirements are imposed on banks, if a bank holds more reserves over a long period, its capital requirement will increase. That is, some of the increase in reserves must be financed by an increase in capital and cannot all be financed by borrowing on the federal funds market. To the extent that capital is a more costly source of funding than borrowing on the federal funds market, capital requirements then add another balance-sheet cost—over and above the cost of additional deposit insurance premiums—to the cost of holding interest-bearing reserves for a financial institution that cannot receive interest on reserves.

Finally, in addition to balance-sheet costs, it is possible that imperfect competition could explain part of the margin between IOER and the federal funds rate. For example, because GSEs have few counterparties in the federal funds market, it may be the case that these counterparties extract some rent in federal funds transactions from the GSEs. However, even if the GSEs have few counterparties, it is not clear why the threat of entry from other counterparties would not enforce competitive pricing. Thus, balance-sheet costs seem the more likely source...
of the interest rate margin, though we do not have a good explanation for why the interest rate margin in Figure 6 is so variable.

As explained in Afonso, Entz, and LeSueur (2013a,b), because most financial institutions that can hold reserve accounts in the United States are flush with reserves, the majority of lending and borrowing on the federal funds market consists of lending by FHLBs to the branches of foreign banks in the United States. Thus, activity on the federal funds market is much different than it was before the financial crisis—it is now mostly an attempt by GSEs to earn interest on overnight balances.

Given our imperfect understanding of the determinants of the margin between the IOER and the federal funds rate, it is uncertain what will happen to that margin as the IOER rises. And given that the FOMC wishes to retain its focus on the federal funds rate as the relevant target overnight rate, it could be problematic if the IOER/federal funds rate margin increases or decreases as IOER rises. A potential solution to this problem involves use of the overnight reverse repurchase agreement (ON RRP) facility at the Fed. A reverse repo is a loan to the Fed with Fed assets posted as collateral. An ON RRP is thus another interest-bearing Fed liability—effectively reserves by another name. The ON RRP facility at the Fed has an expanded list of counterparties, including GSEs and money market mutual funds (MMMFs). First, the GSEs can lend directly to the Fed by way of the ON RRP facility as a means of earning overnight interest. Second, institutions such as the MMMFs, which cannot hold reserves, now have access to interest-bearing Fed liabilities.

As shown in Figure 9, the ON RRP facility was in use prior to the financial crisis, but it was not used as a policy tool. Daily intervention to target the federal funds rate typically consisted of varying the quantity of repos (collateralized loans by the Fed) rather than ON RRPs.
In the figure, the increased recent activity at the ON RRP facility is due to experimentation by the New York Fed in an attempt to understand the mechanics of the market before liftoff occurs. The ON RRP facility should, in principle, set a floor under the federal funds rate. If the Fed fixes an ON RRP interest rate and then borrows whatever quantity is forthcoming from its counterparties at that rate (a procedure called “fixed rate, full allotment”), then no market participant should wish to lend on the federal funds market at a rate below the ON RRP rate. But two key questions then arise: (i) What should the ON RRP rate be relative to the IOER? (ii) What interest rate should the Fed treat as its overnight target? These two questions are closely related.

Clearly, an ON RRP rate greater than the IOER makes no sense, as the Fed should not want to pay a higher interest rate on its overnight liabilities than the rate it pays to financial institutions with interest-bearing reserve accounts. But suppose that the Fed sets the ON RRP rate equal to the IOER. Since MMMFs do not have the balance-sheet costs of commercial banks (no deposit insurance and no capital requirements), MMMFs should be able to attract depositors away from commercial banks and cause a migration of Fed liabilities from reserves to ON RRPs held by MMMFs. Ultimately, reserves would shrink to a small amount—on the order of what was outstanding before the financial crisis, and ON RRPs would account for most of the Fed's outstanding interest-bearing liabilities. Relative to a regime without the ON RRP facility, this would (i) reduce the Fed's interest costs of a given quantity of outstanding interest-bearing liabilities and (ii) peg the overnight rate to the IOER. However, federal funds market activity would dwindle to a very small amount and, given the riskiness of unsecured federal funds lending, the federal funds rate would be greater than the IOER. Under these circumstances, the Fed could focus on the ON RRP rate or the IOER (the same thing in this case) as the relevant policy rate and ignore the federal funds rate.

Alternatively, the FOMC could set the ON RRP rate below the IOER. The wider the margin between the IOER and the ON RRP rate, the smaller would be the quantity of ON RRPs outstanding relative to reserves and the larger would be the quantity of federal funds market activity (because of lending by GSEs to banks earning interest on reserves). If under these circumstances the Fed were to focus on the federal funds rate as the FOMC's policy rate, then this procedure should bound the federal funds rate between the ON RRP rate and the IOER. Where the federal funds rate would reside in that corridor is unpredictable, in part because future changes in regulation could affect the balance-sheet costs of financial institutions borrowing on the federal funds market.

There are other potential solutions to the problem of setting a floor under the federal funds rate. For example, Congress could authorize the Fed to pay interest on the reserve balances of the GSEs. Another possibility is that banks could be permitted to hold segregated accounts with the Fed, which are not subject to balance-sheet costs and would therefore allow more efficient arbitrage and a federal funds rate closer to the IOER. However, neither of these alternatives appears to be under consideration.

**Reducing the Size of the Balance Sheet**

In “Policy Normalization Principles and Plans” (Board of Governors, 2014), the FOMC indicates that it plans to reduce the size of the balance sheet slowly, primarily by not reinvest-
ing as securities mature. There are no plans for outright sales of assets in the Fed’s portfolio. Further, the reinvestment program will not end until after liftoff occurs.

After reinvestment ends, how long will it take to shrink the balance sheet to a size commensurate with a low level of reserves—in line with the level outstanding prior to the financial crisis? The assets in the Fed’s portfolio are financed primarily by currency outstanding, reserves, and now ON RRPs as well. So, as government debt and MBS mature and assets fall, this must be met by a reduction in the quantity of one of the primary liabilities—currency, reserves, or ON RRPs. One can think of the quantity of currency outstanding as being driven by demand, so as Fed assets mature, reserves and ON RRPs must fall. Note also that reserves and ON RRPs will fall as the demand for currency rises, everything else held constant. Therefore, the speed of normalization for the balance sheet depends mainly on the rate at which the Fed’s assets mature and the rate of increase in currency demand over time.

Carpenter et al. (2013) provide projections that estimate 2020 or so as when the Fed balance sheet normalizes—that is, the date at which reserves are at a level similar to what was outstanding before the financial crisis. Their projections were made in early 2013 under the assumptions that QE3 would end in June 2014, that liftoff would occur in the third quarter of 2015, and that the reinvestment program would end prior to liftoff. If liftoff occurs as they projected, then the facts that QE3 continued until October 2014 and that the Board of Governors (2014) normalization plan includes a plan to stop reinvestment after liftoff occurs imply that the date of balance-sheet normalization would be later than 2020. This is because the process of balance-sheet normalization is expected to start later and with a larger balance sheet than what Carpenter et al. (2013) assumed.

The projections of Carpenter et al. (2013) also assume there are no future reinvestment or QE programs on the path to normalization. There is no guarantee of that. Much depends on how soon the Fed achieves liftoff, how rapidly market interest rates rise after liftoff, and the size and timing of future shocks to the economy. It is conceivable that liftoff could be postponed or proceed very slowly, increasing the chance of a large negative macroeconomic shock to the U.S. economy before short-term market interest rates reach normal levels. Thus, the Fed could possibly find itself following a ZIRP once again, with the real economy performing poorly. Then, more QE could be a possibility.

However, suppose liftoff occurs sooner rather than later and short-term interest rates increase at a moderate pace, with no large unanticipated negative macroeconomic shocks. What then are the implications of a large balance sheet, other than the differences in the Fed’s operating procedure for targeting interest rates as already discussed? The Fed’s balance sheet is currently structured with a level of maturity mismatch that is far different from before the financial crisis, as already noted. The Fed currently borrows short—primarily overnight—and lends long. At the end of 2012, the average maturity of the Fed’s asset portfolio was calculated as 10.4 years (see Bukhari et al., 2013) and has increased since then. If the Fed were a private financial institution, this mismatch would put it in a precarious position. Maturity mismatch is risky for private financial intermediaries because, if short-term interest rates rise, the financial institution’s profit margin can disappear and default can ensue.
But the Fed is not a private financial institution. Essentially, it cannot default on its liabilities, except for ON RRP (which are identical to the repo agreements traded by private repo market participants). Currency and reserves are not promises to deliver anything in the future, so there are no promises concerning these liabilities that the Fed could fail to meet.

What about the Fed’s profits? Typically, the Fed earns interest on the securities in its portfolio; it pays the interest on its liabilities (except for currency, which of course does not bear interest); it pays its costs, including wages to Fed employees; and it remits the remainder to the U.S. Treasury. So, if short-term interest rates rise, then Fed profits will be lower, and it is feasible for short-term interest rates to rise sufficiently that remittances to the Treasury fall to zero.

But is this scenario likely to happen over the normalization period? The projections of Carpenter et al. (2013) suggest not. According to their estimates, remittances to the Treasury by the Fed will fall, at worst, to pre-financial crisis levels over the normalization period. Even if remittances did fall to zero, in what sense would that matter? A period with zero remittances to the Treasury by the Fed is one in which the Fed has effectively increased the interest payments on the consolidated government debt. That is, by purchasing long-maturity government debt with short-term liabilities and suffering a loss, the Fed has created a situation in which the consolidated government debt has shorter maturity than it otherwise would (if the Fed had not made the asset purchases) and the consolidated government is making larger interest payments than if the long-term Treasury debt had not been purchased by the Fed. But QE has also reduced interest payments on the consolidated government debt up to the present, and that is projected to continue for some period. Indeed, the Fed has been making record profits since the onset of the financial crisis in 2008. Further, Carpenter et al. (2013) argue that the Fed’s QE programs will ultimately produce a larger cumulative flow of remittances to the Treasury than if QE had not happened.

Perhaps one should not be so sanguine, though. Goodfriend (2014) argues that the expansion of the Fed’s balance sheet is a kind of “carry trade,” whereby the Fed lends long and borrows short. Goodfriend’s point is that it would have been preferable for the Fed to (i) retain interest earnings while it was making large profits on this carry trade and (ii) then later draw down the accumulated surplus on the back end of the carry trade when short-term interest rates rise. By not doing so, Goodfriend argues that the Fed has compromised its monetary policy goals.

**CONCLUSION**

The Fed is currently faced with an unprecedented problem: how to normalize policy in the face of a large balance-sheet expansion and a very long period with short-term nominal interest rates close to zero. The FOMC has laid out a set of plans (Board of Governors, 2014) that specify a gradualist program for normalization. Balance-sheet reduction is expected to occur through the maturing of long-term Treasury debt and MBS in the Fed’s portfolio after the Fed’s reinvestment program ends. The end of reinvestment will occur sometime after liftoff occurs—the date at which the Fed ends its zero interest rate policy and raises short-term market interest rates.
NOTES

1 The first Operation Twist occurred in 1961. The Fed wanted to “twist,” or flatten, the yield curve on U.S. Treasury securities, thereby increasing short-term interest rates while reducing long-term bond yields. This objective is slightly different from its objective in 2011-12, when the Fed wanted to flatten the yield curve but keep short-term interest rates pegged essentially at zero.

2 This pattern of borrowing and lending on the federal funds market occurs because (i) FHLBs cannot receive interest on reserves and (ii) the branches of foreign banks in the United States have interest-bearing reserve accounts and low balance-sheet costs.

REFERENCES


Gertler, Mark and Karadi, Peter. “QE 1 vs. 2 vs. 3…: A Framework for Analyzing Large-Scale Asset Purchases as a Monetary Policy Tool.” International Journal of Central Banking, January 2013, 9(Suppl. 1), pp. 5-53; http://www.iacb.org/journal/iacb13q0a1.pdf.


