Why They Buy: Primary Market Demand for U.S. Treasury Securities

Patrick Herb *
†

this version: September 1, 2017

Abstract

This paper provides an explanation for why investors purchase Treasury securities through the primary market, and why primary dealers compete to aid the government in the distribution process. On average, primary market prices are lower than secondary market prices upon issuance, resulting in what appears to be large excess returns. The contemporary view is that these excess returns are the result of supply shocks and represent a debt issuance cost that is borne by the Treasury. However, the bulk of these returns can be explained by volatility, so what appears to be excess is in fact compensation for risk. In addition, I show that primary market demand increases with increases in expected risk-adjusted returns. Together, these results explain why, following the 2008 financial crisis, lowering interest rates to the zero-lower-bound helped to triple primary market demand despite that debt issuance more than doubled.

JEL: G12

Keywords: Asset returns, Volatility, Time varying risk premium, Liquidity, Special repo rates, On-the-run premium, Treasury auctions, Underpricing, When-issued, Central banking, Effective-lower-bound, Sovereign debt, Supply and demand shocks, Bid-to-cover

---

*I am extremely grateful for the valuable and thoughtful comments of Michael Fleming and Kenneth Garbade at the Federal Reserve Bank of New York, Fan Dora Xia at the Bank for International Settlements, my committee members Blake LeBaron, Steve Cecchetti and George Hall, and Brandeis faculty including Dan Bergstresser, Kathryn Graddy, Catherine L. Mann, Debarshi Nandy, Carol Osler, Robert Reitano and Raphael Schoenle.

†Herb: Ph.D. Candidate, International Business School, Brandeis University, 415 South Street, Mailstop 32, Waltham, MA 02453, pmherb@brandeis.edu, people.brandeis.edu/~pmherb
I. Introduction

Since its founding, the growth and prosperity of the United States has depended on its ability to borrow, but properly incentivizing lenders has been a work in progress.\(^1\) Over at least the last century, the Treasury has tried to increase primary market demand and reduce debt issuance costs. Before 1970, most notes and bonds were issued through subscription and exchange offerings, in which undersubscribed offerings were avoided by the persuasions of famous persons\(^2\) and flat out payments to primary market buyers in the form of interest rate premiums.\(^3\) Between 1970 and 1998, most notes and bonds were issued through multi-price auctions, which tended to benefit sophisticated investors\(^4\) and were prone to collusion.\(^5\)

Despite the urgings of Milton Friedman (1959), it was not until October 1998 that all debt was issued through single-price auctions. Under the current auction regime, primary market prices tend to be lower than secondary market prices upon issuance, resulting in what appears to be large excess returns. The contemporary view is that these returns are the result of supply shocks and represent a debt issuance cost that is borne by the Treasury.\(^6\) However, the bulk of these returns can be explained by volatility, so what appears to be excess is in fact compensation for risk. In addition, I show that primary market demand increases with increases in expected risk-adjusted returns. Together, these results explain why, following the 2008 financial crisis, lowering interest rates to the zero-lower-bound helped to triple primary market demand despite debt issuance more than doubling.

\(^1\)Following the American War of Independence, Alexander Hamilton helped to establish creditable primary and secondary markets for government securities by refinancing public debt at face value despite that its value in the domestic secondary market was ten to twenty cents on the dollar (McCraw, 2012).

\(^2\)For example, Charlie Chaplin created the short film, “The Bond,” at his own expense to help sell U.S. Liberty Bonds during World War I.

\(^3\)See, for example, Cecchetti (1988), Garbade (2004).

\(^4\)Henry Kaufman, a Salomon Brothers economist argued against single-price auctions and improving efficiency as it “provides no incentives to...dealers to help in the distribution process...” (Kaufman, 1973, 170).

\(^5\)See, for example, Jordan and Jordan (1996) document that the investment firm Salomon Brothers gained control of at least 86% of 2-year notes in an attempt to corner the market and increase post auction prices.

\(^6\)See, for example, Lou, Yan, and Zhang (2013) calculate that supply shocks increased debt issuance costs by over half of a billion dollars in note issuance alone in 2007.
Underwriters make a profit when auction prices are lower than secondary market prices upon issuance. Primary dealers view this price difference as a bid-ask spread or underwriting spread. Primary dealers are essentially underwriters required to bid on a pro-rata basis in all Treasury auctions at reasonably competitive prices, taking into account when-issued prices, market volatility and other risk factors. For example, if the Treasury offers $20 billion of 2-year notes and there are 20 dealers, then each dealer must bid for a minimum of $1 billion. To offload some of the inventory risk, dealers are known to sell some of their anticipated inventory in advance of the auction. For example, suppose a dealer believes she will likely acquire $500 million from the next auction, then she might sell $250 million ahead of the auction in the when-issued or secondary market. Large short positions around auctions are thought to depress prices, which then recover sometime later, resulting in excess returns.

This paper argues that auction prices should be discounted relative to their secondary market price upon issuance, and the size of the discount depends on the volatility of the price change between auction and issuance.

The on-the-run premium, or liquidity premium, is the phenomenon that just issued on-the-run Treasury securities trade at a premium over just off-the-run securities with similar tenors and cash flows, and is thought to arise from differences in liquidity. For example, Barclay, Hendershott, and Kotz (2006) show that when Treasury securities go off-the-run, their trading volume can decline by more than 90%. Short sellers prefer liquidity and are thought to contribute to the premium. Most short selling is done through the financing market as reverse repos, in which high demand by short sellers can lower repo rates below general market rates, resulting in “specialness.” Specialness increases both short sellers borrowing costs and long investors lending fees. It is thought that long investors are willing

---

7See, for example, the Federal Reserve Bank of New York Policy on Counterparties for Market Operations.
8See, for example, Department of the Treasury et al. (1992), Simon (1994), Nyborg and Sundaresan (1996), Fleming and Rosenberg (2007).
9See, for example, Fleming and Rosenberg (2007), Lou et al. (2013).
to pay higher premiums to collect higher lending fees and short sellers prefer to borrow and sell securities that trade at a premium, as this increases profits when the bond goes off-the-run and prices converge. Banerjee and Graveline (2013) calculate that short sellers may account for 37% of the premium.

One might reasonably believe that the on-the-run premium should reduce the price difference between auction and issuance, or even buoy auction prices. However, I argue that it should exacerbate the effect. When-issued securities are not the same as issued securities: they do not accrue interest and cannot be used in the repo market, and therefore lack the utility and liquidity of issued securities. Since illiquid securities tend to have higher returns (lower prices) than liquid securities,\footnote{See, for example, Amihud and Mendelson (1986, 1991), Warga (1992), Brennan and Subrahmanyam (1996), Krishnamurthy (2002).} we should expect that when-issued securities trade at an illiquidity discount to their issued cousins. Furthermore, auction prices are based on when-issued prices, and winning auction bids are essentially when-issued securities until settlement. Therefore, we should expect that auction prices trade at an illiquidity discount.

The methods for measuring liquidity are far from perfect. However, there is ample evidence to suggest that volatility and illiquidity move together in U.S. Treasury markets.\footnote{See, for example, Engle, Fleming, Ghysels, and Nguyen (2011).} I show that volatility explains the bulk of these excess returns. It is noteworthy to point out that volatility may also be capturing illiquidity. For example, when volatility increases, expected returns increase, implying that auction prices suffer a larger illiquidity discount. If volatility and illiquidity move together, these results could also be interpreted in the following way. Using volatility as a proxy for illiquidity, I find that auction prices are more discounted when the market is less liquid, resulting in higher returns to auction buyers.

From January 2000 through June 2016, annualized returns resulting from the underwriting spread tend to be increasing in maturity and range from 0.38% for 4-week bills to 57.07% for 30-year bonds. The unconditional excess return, generated from buying at the
auction price and selling at the secondary market price upon issuance for 4-, 13-, 26-week bills and 2-year notes is 0.29, 0.77, 1.34, and 10.0 \((t = 5.86, 6.13, 6.50, 5.37)\) basis points, respectively. To evaluate the relationship between excess returns and volatility, I estimate a volatility regime switching ARMA-GARCH-M process, in which conditional volatility determines the risk premium. The volatility regimes account for the dramatic changes in risk before, during and after the 2008 financial crisis. I find that a one basis point increase in volatility increases expected returns by 0.62, 0.26, 0.27 and 0.51 \((t = 7.55, 4.66, 4.32, 3.22)\) basis points, respectively. Subtracting the risk premium from each return results in excess returns of only 0.01, 0.38, 0.71 and 0.72 \((t = 0.49, 4.35, 4.55, 0.41)\) basis points, respectively. This implies that volatility, a single factor, explains 100, 50, 47, and 100 percent of these excess returns, respectively.\(^{14}\)

To evaluate the relationship between primary market demand and risk-adjusted returns, the estimated model for returns produces forecasts of expected returns and conditional variances for each auction, using past information only. Expected returns are divided by their conditional variances to construct an expected risk-adjusted returns factor, which is then used to explain primary market demand. Estimates can be interpreted causally because expected risk-adjusted returns act like an instrument for risk-adjusted returns. Intuitively, since conditional volatility determines the conditional mean, dividing the conditional mean by its conditional variance results in a factor that feels similar to a conditional Sharpe ratio. Two empirical measures for primary market demand, auction tenders and the bid-to-cover ratio, are separately projected onto expected risk-adjusted returns space, controlling for both auction size and serial correlation. I find that a one basis point increase in expected risk-adjusted returns leads to a 5.39, 2.71, 8.76 and 122.5 \((t = 4.77, 3.59, 3.87, 2.02)\) billion dollar increase in auction tenders, respectively. Similarly, I find that a one basis point increase in expected risk-adjusted returns leads to an increase in the bid-to-cover ratio of 0.20, 0.07, 0.27 and 5.45 \((t = 6.23, 3.47, 5.18, 2.32)\), respectively. In both cases, increases in expected

---

\(^{14}\)These results are consistent with Lou et al. (2013), who find that excess returns are increasing in volatility.
risk-adjusted returns lead to increases in primary market demand.

Following the 2008 financial crisis, the Federal Reserve lowered short-term interest rates to the zero-lower-bound (ZLB). At the same time, the Treasury increased debt issuance through the primary market from roughly $3.5 trillion to over $8 trillion per year. Surprisingly, primary market demand increased by even more, rising from roughly $10 trillion to $33 trillion auction tenders per year. Even more surprising, the underwriting spread significantly declined, which is the opposite of what we would expect if these excess returns are exclusively the result of supply shocks. For example, consider how 26-week bill auctions differed from 2003 to 2010: average weekly offerings increased from $16.2 billion to $27.9 billion per auction, average excess returns declined from 0.67 to 0.35 basis points, while average risk-adjusted returns and the bid-to-cover ratio increased from 0.25 to 0.58 and from 2.01 to 4.26, respectively. I hypothesize that lowering interest rates to the ZLB decreased the variance of the underwriting spread, which increased expected risk-adjusted returns and primary market demand. See section III.A. for a theoretical discussion.

These results have a number of important implications. First, the ZLB may produce a spillover effect that reduces the price impact of Treasury auctions. Increases in expected risk-adjusted returns that lead to increases in demand would likely result in more bidders who are unsatisfied with their auction allocations. Disappointed bidders may increase demand for similar securities in the secondary market, thereby offsetting dealers short positions and reducing the price impact of Treasury auctions. Second, auctions with unbounded negative interest rates may result in increased volatility, decreased expected risk-adjusted returns, and therefore suffer from lower primary market demand. Similarly, lift-off from the ZLB will likely result in increased volatility, decreased expected risk-adjusted returns, and lower primary market demand. Finally, although there is a considerable amount of research devoted to improving primary market efficiency, to my knowledge, none take into account

Lou et al. (2013) find that excess returns are increasing in auction size, but their data end before the financial crisis. However, my interpretation is consistent with Krishnamurthy (2002), who argues that increasing supply should reduce the liquidity premium.
expected risk-adjusted returns. These findings suggest the importance of including this new factor when considering optimal outcomes.

This paper is closely related to the auction underpricing literature, in particular, the finding by Nyborg, Rydqvist, and Sundaresan (2002) that underpricing increases with volatility,\footnote{See, for example, Nyborg and Sundaresan (1996), Keloharju, Nyborg, and Rydqvist (2005), Goldreich (2007).} and the finding by Lou, Yan, and Zhang (2013) that excess returns are the result of anticipated and repeated supply shocks. The contribution of my paper is to show that the underpricing of Treasury auctions relative to their secondary market price upon issuance is largely due to volatility, and that the bulk of these excess returns are in fact compensation for risk. Additionally, I show that primary market demand depends on expected risk-adjusted returns, suggesting a causal link between primary market demand for U.S. Treasury securities and the expected risk and return of acquiring these securities through the auction market. My results have important implications for the literature concerned with auction design, which seeks to improve efficiency without disrupting primary market demand.\footnote{See, for example, Goswami, Noe, and Rebello (1996), Kremer and Nyborg (2004), Nyborg and Strebulaev (2004), Goldreich (2007), Boyarchenko, Lucca, and Veldkamp (2016).} My paper is also related to the literature on the price impact of supply and demand shocks,\footnote{See, for example, Shleifer (1986), Kaul, Mehrotra, and Morek (2000), Mitchell, Pulvino, and Stafford (2004), Coval and Stafford (2007), Frazzini and Lamont (2008), Lou (2012).} and the impacts of announcements and bond supply on Treasury yields.\footnote{See, for example, Duffie (1996), Krishnamurthy (2002), Kuttner (2006), Vayanos and Vila (2009), Krishnamurthy and Vissing-Jorgensen (2012), Han, Longstaff, and Merrill (2007), Garbade and Rutherford (2007), Gagnon, Raskin, Remache, and Sack (2010), Krishnamurthy and Vissing-Jorgensen (2011), Greenwood and Vayanos (2014).} My results are also very broadly related to debt issuance that is divided between primary and secondary markets such as sovereign debt, municipal bonds, corporate debt, IPO book building, etc.

My results may challenge the finding by Goldreich, Hanke, and Nath (2005) that expected future liquidity is more important than current liquidity. If when-issued securities trade at an illiquidity discount, then current liquidity may play a more important role than was previously thought.
II. Data, Definitions and Institutions

A. Data Sample

Data are obtained from the U.S. Treasury and Bloomberg. The sample period ranges from January 2000 to June 2016. The maturities included are 4-, 13-, 26-, and 52-week bills; 2-, 3-, 5-, 7-, and 10-year notes; and 30-year bonds. Auction result variables include the bid-to-cover ratio, total offering amount, total accepted, total tendered, and both auction and issuance dates. Pricing variables for bills include the auction high rate and the last traded discount rate in the secondary market on issue day. Pricing variables for notes and bonds include the price per $100, which includes accrued interest, and the last traded price in the secondary market on issue day. Bill yields are converted to prices by the formula given in appendix B of the auction regulations Uniform Offering Circular and Amendments (31 CFR Part 256). The variable \( d_\tau \) is created as the number of trading days between auction and issuance, for an auction that closes on date \( \tau \). For example, if securities are issued the day after an auction closes, then \( d_\tau \) would equal one.

The main analysis of this paper is completed on a subset of the data, in which the main determinant of inclusion is a constant frequency of issuance over the sample period. Weekly auctions of 4-, 13-, and 26-week bills are considered from July 30, 2001 through June 28, 2016, which aligns with the introduction of 4-week bills and allows for an apples-to-apples comparison. Monthly auctions of 2-year notes are considered from January 2000 through June 2016. The constant frequency of auctions allows for the application of many time series analysis techniques that would not be possible with irregular auction frequencies.\(^\text{20}\) One observation of 26-week bills is dropped due to a missing observation in Bloomberg, which corresponds to CUSIP 912795QC8. The chosen subset of maturities is tested and found to exhibit autoregressive conditional heteroskedasticity (ARCH) effects, which allows

\(^{20}\) Additionally, one can extrapolate information from this subset and infer the main mechanism more generally to other maturities.
for effective modeling of the time varying volatility of each auction.\textsuperscript{21}

B. How Treasury Auctions Work

Since October 1998, all U.S. Treasury securities are issued through regular and predictably scheduled single-price sealed bid auctions.\textsuperscript{22} Months in advance of an auction, the Treasury makes a tentative announcement regarding future debt issuance. As the auction date approaches, the Treasury makes an official offering announcement specifying all pertinent information regarding the upcoming offering, such as the total offering amount, the auction closing time, and issuance (settlement) and expiration (maturity) dates. Following the official announcement, bidding and when-issued trading begin. Auction bids are submitted as tenders, which are price and quantity pairs.

When-issued securities are similar to futures contracts that trade between large institutions, but no money is exchanged until the securities are actually issued. Similarly, primary dealers do not need to hold funds in their account with the Federal Reserve to submit tenders, nor do they need to until settlement. For these reasons, holding a winning bid or when-issued security has no carry cost. Most tenders arrive just before the auction closes, and auction results are announced roughly two minutes later. The price is determined by the marginal bidder, as the last tender accepted determines the price paid by all winning bidders. If tenders at the stop out price exceed supply, bidders are awarded a reduced percentage of their quantity demanded. Day(s) later, the securities are issued and monies are collected from winning bidders.

For example, figure I presents a hypothetical scenario in which the Treasury sells 80 units of debt through an auction closing on date $\tau$. Tenders arrive randomly until the auction closes, and are then sorted from highest to lowest price. Starting from the highest

\textsuperscript{21}Treasury securities are typically quoted in yields or discount rates. However, since price is an inverse function of yield, we can use price interchangeably with yield. Most of the analysis in this paper is performed using returns, which can be confusing when referencing fixed income instruments. I refer only to prices and returns, in which returns are generated from changes in price.

\textsuperscript{22}See, for example, Garbade (2015).
price, tenders are accepted until all 80 units of debt are sold. The last tender accepted determines \( P^A_\tau \), which is the auction price paid by all winning bidders. Figure II depicts a general timeline of events from the official announcement to security issuance. In this case, \( P^A_\tau \) is equal to $99.2 and denotes the stop out price for the auction. After the auction price is determined, the securities are issued \( d \) trading days later, and trade in the secondary market there after. \( P^{S}_{\tau+d} \) denotes the last traded price in the secondary market on issue day.

<table>
<thead>
<tr>
<th>Tenders Submitted</th>
<th>Tenders Sorted</th>
<th>Tenders Accepted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>Quantity</td>
<td>Price</td>
</tr>
<tr>
<td>99.0</td>
<td>33</td>
<td>99.9</td>
</tr>
<tr>
<td>99.3</td>
<td>16</td>
<td>99.5</td>
</tr>
<tr>
<td>98.5</td>
<td>12</td>
<td>99.3</td>
</tr>
<tr>
<td>98.3</td>
<td>17</td>
<td>99.2</td>
</tr>
<tr>
<td>99.5</td>
<td>22</td>
<td>99.1</td>
</tr>
<tr>
<td>99.9</td>
<td>14</td>
<td>99.0</td>
</tr>
<tr>
<td>97.9</td>
<td>11</td>
<td>98.9</td>
</tr>
<tr>
<td>99.1</td>
<td>19</td>
<td>98.5</td>
</tr>
<tr>
<td>99.2</td>
<td>28</td>
<td>98.3</td>
</tr>
<tr>
<td>98.9</td>
<td>10</td>
<td>97.9</td>
</tr>
</tbody>
</table>

Figure I: Single-price sealed bid auction timeline. Suppose the Treasury sells 80 units of debt. Tenders are submitted in random order until the auction closes. Tenders are then sorted, and then accepted starting from highest to lowest price, until the total offering is sold. The last tender accepted determines \( P^A_\tau = $99.2 \), which is the price paid by all winning bidders.

C. Calculating Returns from the Underwriting Spread

Primary dealers view themselves as underwriters that guarantee the sale of all government debt and assume all distribution risk. They agree to bid on a pro-rata basis at every auction, which, from the Treasury’s perspective, is meant to reduce the risk of auction failures. As compensation for this risk, primary dealers hope to earn an underwriting spread, which is the difference between the primary market price \( P^A_\tau \), and secondary market price \( P^{S}_{\tau+d} \). One can think of this as a bid-ask spread, in which dealers buy from the primary market and sell
into the secondary market. Since there is no carry cost to holding a winning auction bid, the underwriting return is an excess return that can be expressed as the arithmetic return \( R \). Let \( t \) index a sequence of auctions that occur on date \( \tau \), then

\[
R_t = \frac{P_{\tau + d_t}^S - P_\tau^A}{P_\tau^A}.
\]

(1)

The subscript \( t \) is also added to \( d \) to denote that the number of trading days between auction and issuance varies throughout the sequence of auctions.

<table>
<thead>
<tr>
<th>Term</th>
<th>Type</th>
<th>( \bar{R} )</th>
<th>( t )-Statistic</th>
<th>Annualized %</th>
<th>Obs.</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-week</td>
<td>bills</td>
<td>0.29***</td>
<td>(5.86)</td>
<td>0.38</td>
<td>778</td>
<td>1</td>
<td>1.9</td>
<td>3</td>
</tr>
<tr>
<td>13-week</td>
<td>bills</td>
<td>0.77***</td>
<td>(6.13)</td>
<td>0.70</td>
<td>779</td>
<td>1</td>
<td>2.9</td>
<td>4</td>
</tr>
<tr>
<td>26-week</td>
<td>bills</td>
<td>1.34***</td>
<td>(6.50)</td>
<td>1.22</td>
<td>778</td>
<td>1</td>
<td>2.9</td>
<td>4</td>
</tr>
<tr>
<td>52-week</td>
<td>bills</td>
<td>2.22***</td>
<td>(4.16)</td>
<td>2.89</td>
<td>106</td>
<td>1</td>
<td>1.9</td>
<td>2</td>
</tr>
<tr>
<td>2-year</td>
<td>notes</td>
<td>10.00***</td>
<td>(5.37)</td>
<td>7.85</td>
<td>198</td>
<td>1</td>
<td>3.7</td>
<td>9</td>
</tr>
<tr>
<td>3-year</td>
<td>notes</td>
<td>10.86***</td>
<td>(3.06)</td>
<td>6.64</td>
<td>109</td>
<td>3</td>
<td>4.7</td>
<td>8</td>
</tr>
<tr>
<td>5-year</td>
<td>notes</td>
<td>21.36***</td>
<td>(4.62)</td>
<td>21.31</td>
<td>170</td>
<td>1</td>
<td>3.5</td>
<td>8</td>
</tr>
<tr>
<td>7-year</td>
<td>notes</td>
<td>28.72***</td>
<td>(5.21)</td>
<td>33.93</td>
<td>89</td>
<td>1</td>
<td>2.3</td>
<td>7</td>
</tr>
<tr>
<td>10-year</td>
<td>notes</td>
<td>20.65**</td>
<td>(2.41)</td>
<td>22.60</td>
<td>150</td>
<td>1</td>
<td>3.6</td>
<td>7</td>
</tr>
<tr>
<td>30-year</td>
<td>bonds</td>
<td>40.49***</td>
<td>(2.76)</td>
<td>57.07</td>
<td>102</td>
<td>1</td>
<td>2.7</td>
<td>6</td>
</tr>
</tbody>
</table>

Table I: Summary statistics of regressing the ex-post underwriting return, \( R_t \), on a constant, by maturity, from January 2000 to June 2016. The first two columns list the term and type of security. The third column is the average return, \( \bar{R} \), in basis points. The fourth column presents \( t \)-statistics from standard errors that are Newey-West (1987) adjusted with a max lag determined by Andrews and Monahan (1992). The fifth column presents the average annualized percentage return using a 252 trading day calendar. The sixth column reports the number of auctions in the sample. The last three columns report the min, mean, max number of trading days between auction close and security issuance.

Table I provides summary statistics for \( R_t \), the ex-post return from the underwriting spread. For example, buying at the auction price and selling at the secondary market price earns an average of 10.0 basis points for 2-year notes, with an average annualized return of 7.85% over the 198 auctions in the sample. On average, securities are issued 3.7 trading days after the auction closes; but they have been issued in as few as one, and as many as nine trading days later. The expected return is mostly increasing in the term structure, as duration risk is also increasing. The annualized return is somewhat misleading.
as this trade cannot be executed daily because no maturity has daily auctions. Also, many auction frequencies changed throughout the sample period, so comparisons between certain maturities are difficult to make. For example, 52-week bills were not offered in the sample period until June 2008.\footnote{All statistics are excluding Cash Management Bills (CMB), Treasury Inflation-Protected Securities (TIPS), and Floating Rate Notes (FRN).}

\section*{D. Debt Issuance and Asymmetric Demand at the Zero-Lower-Bound}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3}
\caption{Total annual volume of debt sold through the primary market by the Treasury for all term structures. The annual volume dramatically increased following the 2008 financial crisis.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig4}
\caption{Total annual volume of debt sold (supply) compared to total annual tenders (demand). Primary market demand surprisingly increased by more than the increase in supply at the zero-lower-bound.}
\end{figure}

Figure III plots the annual total volume of debt sold through the primary market by the Treasury. Following the 2008 financial crisis, the Treasury significantly increased the volume of debt issuance. Short term interest rates were lowered to the zero-lower-bound in December of 2008. Many investors were concerned that the market would not be able to absorb such an increase. Figure IV plots the surprising increase in primary market demand. Not only could the market absorb the large increase, but the increase in primary market demand exceeded the increase in supply. However, the response in demand was asymmetric in maturities, as the increase in demand for shorter term debt was significantly more pronounced than the roughly proportional increase in demand for longer term debt. Figures V and VI demonstrate this effect by comparing the bid-to-cover ratio of short term debt with long term debt. The

\footnote{All statistics are excluding Cash Management Bills (CMB), Treasury Inflation-Protected Securities (TIPS), and Floating Rate Notes (FRN).}
bid-to-cover ratio is the volume tendered divided by the volume offered, which provides a ratio that quantifies the relationship between supply and demand. Values greater than one imply that demand exceeds supply.

![Graph](image1)

**Figure V:** The bid-to-cover ratio of 13-week bills vs. 10-year notes. The relative increase in demand of 13-week bills is much greater than 10-year notes at the zero-lower-bound.

![Graph](image2)

**Figure VI:** The bid-to-cover ratio of 26-week bills vs. 10-year notes. The relative increase in demand of 26-week bills is much greater than 10-year notes at the zero-lower-bound.

### III. A Model Relating Risk, Returns and Demand

It is well established that risk averse investors require compensation for holding risky assets. Buying Treasury securities from the primary market carries the risk of bidding too much and suffering the winners curse. Additionally, auction participants commit to a price days before settlement, which increases the uncertainty of what the security will be worth once it is issued. The incentive for primary dealers to provide underwriting services to the Treasury is not due to the benevolence of bankers, but rather for the opportunity to make a profit. A profit occurs when primary dealers or other financial intermediaries can buy securities at a discount in the primary market, in order to sell them at a premium in the secondary market. The same mechanism increases primary market demand from end-investors as well, as buying the security for less will always increase an investors return. What follows is a simple model relating demand for an asset to its risk-adjusted return.

Consider a two asset world in which investors must hold their wealth in shares of either a
risky or certain asset, in which the risky asset has normally distributed returns and the certain asset has a certain return. The risky asset is analogous to purchasing Treasury securities from the primary market, while the certain asset is analogous to auction nonparticipation. The risk in holding the risky asset is measured by the volatility of returns, and the compensation is a higher expected return, which is increasing in volatility (risk). Demand for the risky asset depends on investors’ utility functions, which determine risk preferences. Suppose the numeraire is the certain asset, which is perfectly elastically supplied with price equal to 1 and certain payoff $r$ (gross interest). Let the risky asset have price $P_A$ with a random payoff $P_S$, in which the random payoff has mean $\pi$ and variance $\omega^2$. Investors wealth $W$ can be expressed as an allocation between $X$ shares of the risky asset and $Y$ shares of the certain asset such that

$$W = P_A X + Y. \quad (2)$$

The excess return $R$ of the risky asset per dollar invested in risky shares is given by

$$R = \frac{P_S}{P_A} - r. \quad (3)$$

Since there is no carry cost to holding a winning auction bid, we can let the certain payoff equal 1, and rewrite the excess return as the arithmetic return

$$R = \frac{P_S}{P_A} - 1. \quad (4)$$

For a particular auction that closes on date $\tau$ with holding period $d$, the return takes the same form as equation 1. Define the mean and variance of $R$ such that

$$E[R] \equiv \mu = \frac{\pi}{P_A} - 1, \quad Var[R] \equiv \sigma^2 = \frac{\omega^2}{P_A^2}. \quad (5)$$

Suppose further that all investors have utility functions that exhibit constant absolute risk
aversion,\textsuperscript{24} and that investors only maximize end of period wealth so that only the first two moments matter. Investors maximize expected utility by choosing an allocation of risky and certain assets such that

\[
E[U] = \max_{X,Y} \quad 2E[P^AX + Y] - bVar[P^AX + Y]. \tag{6}
\]

Utility is maximized by choosing

\[
XP^A = \frac{\mu}{b\sigma^2}. \tag{7}
\]

The optimal solution relates the number of shares demanded \(X\) and price \(P^A\) of the risky asset to the risk-adjusted return \(\frac{\mu}{\sigma^2}\), and risk aversion parameter \(b\). In financial markets it is common for the variance of returns to change over time, a property that is commonly referred to as volatility clustering. If we allow the variance of the risky payoff \(\omega^2\) to vary from auction to auction, then the other variables will vary from auction to auction as well. Let \(t\) index a sequence of auctions. Relaxing the assumption of constant variance allows us to rewrite equation 7 as the dynamic optimal solution

\[
X_tP_t^A = \frac{\mu_t}{b\sigma_t^2}. \tag{8}
\]

A. Impact of the Zero-Lower-Bound

Under normal economic conditions, price and demand typically adjust to changes in risk-adjusted returns to ensure that the asset is fully held in equilibrium (failed auctions are avoided). Since prices are inversely related to yields, lowering short term interest rates to the zero-lower-bound (ZLB) effectively places an upper bound on Treasury security prices. In fact, by rule of the Treasury, auction tenders for negative interest rates were not allowed at the ZLB. This rule implies that prices sufficiently close to the upper bound will exhibit

\textsuperscript{24}For example, exponential utility implies constant absolute risk aversion.
inflexibility towards upward price movements. Let \( \bar{P}_t^A \) represent the upward inflexible price when the ZLB is binding. We can now rewrite the optimal solution as

\[
X_t \bar{P}_t^A = \frac{\mu_t}{b\sigma_t^2}.
\] (9)

When prices are sufficiently close to the upper bound, what is the impact on the variance of \( R_t \)? Under normal economic conditions, investors will estimate a point forecast for the expected value of \( P^S \), which will typically have a variance of possible prices with a symmetric density about the point estimate. However, when prices are sufficiently close to the binding price, investors will estimate the expected value of \( P^S \), but the density of possible prices will be asymmetric, with more mass below the point estimate than above. This change in density reduces the range of possible prices, which lowers the variance of \( R_t \). The reduction in variance decreases the denominator of the risk-adjusted returns factor \( \frac{\mu_t}{\sigma_t^2} \), thereby increasing risk-adjusted returns. Since price is sufficiently close to the binding price, it cannot adjust upwards, so the increase in risk-adjusted returns leads to an increase in demand. In fact, this is exactly what we see empirically in the data. The ZLB reduced the variance of \( R_t \), increased risk-adjusted returns \( \frac{\mu_t}{\sigma_t^2} \), and increased primary market demand \( X_t \).

IV. Model Formulation and Estimation Procedure

To test the relationship between primary market demand and risk-adjusted returns, we need an empirical measure for primary market demand and an exogenously related measure of risk-adjusted returns. Total auction tenders and the bid-to-cover ratio are good candidates for the former, but ex-post returns will likely be endogenously related to both. My identification strategy is to use expected risk-adjusted returns, which are forecasts that use past return information only. Intuitively, expected risk-adjusted returns act like an instrument for risk-adjusted returns. The challenge is to estimate a model for returns that can produce forecasts of expected returns and conditional variances, which can then be used to calculate expected
risk-adjusted returns. The model needs to be able to adapt to time-varying volatility, and by assumption, expected returns should depend on conditional volatility. This assumption reflects the assertion that excess returns are compensation for risk, in which risk is quantified by volatility. It was also argued that volatility may be an appropriate proxy for illiquidity in Treasury securities markets. Taken together, we should expect that returns are positively related to volatility, as compensation should be increasing in both risk and illiquidity. This setup suggests using a variant of the models developed by Engle (1982) and Bollerslev (1986), or so called GARCH in the Mean (GARCH-M).

Let $t$ index a sequence of auctions, let $\chi_t$ represent the empirical measure for primary market demand, and let $E_{t-1}[\frac{\mu_t}{\sigma_t^2}]$ represent expected risk-adjusted returns, which are forecasts conditional on past information only. In addition, the variable $\text{Offer}_t$ controls for auction size and the variable $\chi_{t-1}$ controls for serial correlation, since dealers are required to bid on a pro-rata basis in every auction and because auction participants’ actions are likely correlated to their actions in the previous auction. Adding these controls implies the following equation to estimate,

$$\chi_t = \beta_0 + \beta_1 E_{t-1}[\frac{\mu_t}{\sigma_t^2}] + \beta_2 \text{Offer}_t + \beta_3 \chi_{t-1} + \epsilon_t. \quad (10)$$

The parameter of interest, $\beta_1$, suggests the estimated causal relationship between primary market demand and risk-adjusted returns.

With our goal of forecasting $\frac{\mu_t}{\sigma_t^2}$ for each auction, recall that $\mu_t$ is the expected return of $R_t$, and that $\sigma_t^2$ is its conditional variance. In the simplest form, the mean equation can be expressed as the combination of an expected return and a white noise process, such that

$$R_t = \mu_t + \epsilon_t. \quad (11)$$

The expected return $\mu_t$ can be further decomposed into market frictions $\delta$ and a risk premium
\[ R_t = \delta + \gamma \sigma_t + \epsilon_t. \]  

If \( R_t \) is scaled to basis points, then a positive estimate for \( \gamma \) implies that a one basis point increase in the conditional volatility of \( R_t \) leads to a \( \gamma \) basis point increase in the risk premium, implying the same increase in expected returns. From the Treasury’s perspective, this directly quantifies the impact that volatility has on debt issuance costs. Consider what happens when the estimate for \( \delta \) is not different from zero. The expected return will equal \( \gamma \sigma_t \), which implies that \( \frac{\mu_t}{\sigma_t} \) is equal to \( \frac{\gamma \sigma_t}{\sigma_t^2} \), which is equal to \( \frac{\gamma}{\sigma_t} \). In this case, the expected risk-adjusted return is essentially an expected Sharpe (1966) ratio. If \( \beta_1 \) is positive and significant, we can interpret this result as increases in the conditional Sharpe ratio lead to increases in primary market demand; an intuitive result.

A nonzero \( \delta \) could reflect either market frictions, inefficiencies, structural issues, or even the linearization of a nonlinear function. Generally, it represents the portion of the return that is not explained by volatility. There are several plausible reasons for a positive estimate, such as the price impact of supply and demand shocks, short-squeezes of when-issued sellers, or even barriers to entry. For example, a positive estimate could indicate a secondary market preference, in which investors prefer to purchase Treasury securities from the secondary market upon issuance to avoid the uncertainties associated with the auction process. This preference could make secondary market securities more expensive as demand shifts from the primary to the secondary market.

As Mandelbrot (1963) now famously observed, “...large changes [in price] tend to be followed by large changes - of either sign - and small changes tend to be followed by small changes...” The concept he observed is now commonly referred to as volatility clustering. One of the most common ways to model changing volatility is using a GARCH\((p,q)\) process, which consists of a constant \( \omega \), \( p \)-lagged conditional variances, and \( q \)-lagged squared innovations. However, since the number of trading days between auction close and security issuance varies
from auction to auction, we need to slightly modify the standard model. If we take price changes between auction close and security issuance to be a martingale, then we can multiply the GARCH\((p,q)\) process by the number of trading days \(d_t\). This change effectively scales the conditional variance forecast similarly to that of a random walk. The varying trading day adaption can be expressed as

\[
\sigma_t^2 = d_t [\omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2].
\]  

(13)

The financial crisis and the zero-lower-bound are associated with such extreme changes in volatility that a standard GARCH\((p,q)\) process cannot capture alone. For these periods it is more expedient to allow for a volatility regime switch. In the spirit of parameter parsimony, allowing only the constant \(\omega\) to change for each volatility regime proves effective. Additionally, the constant in the mean equation is also allowed to change for each volatility regime. To account for any serial correlation in the return series, an ARMA\((P,Q)\) process is added to the mean equation. The complete model to estimate is given by

\[
R_t = \sum_{k=1}^{K} \delta_k \mathbb{1}_{S_k}(t) + \gamma \sigma_t + \sum_{i=1}^{P} \phi_i R_{t-i} + \sum_{j=1}^{Q} \theta_j \epsilon_{t-j} + \epsilon_t,
\]

\[
\sigma_t^2 = d_t \left[ \sum_{k=1}^{K} \omega_k \mathbb{1}_{S_k}(t) + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \right],
\]

\[
\epsilon_t = \sigma_t \eta_t,
\]

\[
\mathbb{1}_{S_k}(t) = \begin{cases} 
1, & \text{if } t \in S_k, \\
0, & \text{otherwise.}
\end{cases}
\]

The innovation in the mean equation \(\epsilon_t\) can now be defined as a function of the conditional volatility and a sequence \((\eta_t)\) of mean zero independently and identically distributed random variables such that \(E[\eta_t^2] = 1\). The conditional variance is determined by a constant,
\( \omega_k \), corresponding to volatility regime \( k \), \( q \)-lagged squared innovations and \( p \)-lagged conditional variances such that \( \omega_k > 0 \) \((k = 1, 2, ..., K)\), \( \alpha_i \geq 0 \) \((i = 1, ..., q)\), \( \beta_j \geq 0 \) \((j = 1, ..., p)\), and \( \sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1 \). The general model consists of \( K \) volatility regimes estimated by \( K - 1 \) structural breaks in the variance, so that the structural breaks change the level of the variance, but do not affect the GARCH parameters. The estimated \( \omega_k \) is multiplied by the indicator function \( 1_{S_k}(t) \), which takes on the value of 1 when \( t \) is in the volatility set \( S_k \) and zero otherwise. Finally, the variance equation is multiplied by the number of trading days \( d_t \) between auction close (\( \tau \)) and settlement (\( \tau + d_t \)).

Parameter estimation is achieved by maximum likelihood, in which the likelihood function is given by

\[
L = \prod_{t=1}^{T} \left( \frac{1}{\sqrt{2\pi \sigma_t^2}} \right) \exp \left( \frac{-\epsilon_t^2}{2\sigma_t^2} \right),
\]

with the associated log-likelihood function

\[
\ln L = -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \ln \sigma_t^2 - \frac{1}{2} \sum_{t=1}^{T} \frac{\epsilon_t^2}{\sigma_t^2}.
\]

All estimations are performed using MATLAB’s constrained optimization, in which the negative log-likelihood function is minimized. The inverse of the Hessian matrix evaluated at the optimum is used as an estimate for the asymptotic covariance matrix, and standard errors are estimated as the square root of the diagonal elements of this matrix. Volatility regime switching models considered range from an ARMA(0, 0)-GARCH(0, 1)-M to an ARMA(6, 6)-GARCH(3, 3)-M. The innovations corresponding to the initial observations that are lost due to the number of lags in a given model are all set to zero. To avoid kinks and flat areas in the log-likelihood function, all models are estimated a minimum of 100 times with different random starting values for the parameters. In order for a solution to be considered valid, the chosen model and parameters must be found identically the same for at least one tenth of the estimations, and the Hessian matrix must be positive definite. The final model selected
exhibits these characteristics and is found to have the lowest Schwarz criterion (BIC).\textsuperscript{25}

To estimate the structural breaks in the variance by maturity, the underwriting return $R_t$ is regressed on a constant to obtain a vector of estimated residuals $\hat{e}_t$. The estimated residuals are then squared, and the squared residuals are then regressed on a constant using ordinary least squares. The methodology of Bai and Perron (1998, 2003) is then used to estimate global structural breaks. The number of structural breaks is determined by the modified Schwarz criterion LWZ.\textsuperscript{26} The full sample of observations is depicted as the volatility set $\{S\}$. Volatility regime $k$ contains the subset of observations $\{S_k\}$, which is determined by $K - 1$ structural breaks. For example, suppose 2-year notes have one structural break, then it has two mutually exclusive volatility sets $\{S_1, S_2\}$, which are complements that span the set $\{S\}$.

V. Empirical Results

A. Primary Market Demand and Expected Risk-Adjusted Returns

The main agenda for this paper is to test if risk-adjusted returns that result from the underwriting spread explain primary market demand for U.S. Treasury securities. First, a model is estimated for each maturity according to the methodology described in the previous section, in which the chosen models and estimated parameters are reported at the end of this section. The estimated model is then used to generate forecasts of expected returns and variances for each auction. Expected returns are divided by their variance forecasts to create the expected risk-adjusted returns factor, which acts as an instrument for risk-adjusted returns. Table II presents results for the regression of auction tenders on expected risk-adjusted returns. For each maturity, expected risk-adjusted returns are positively and significantly related to auction tenders, suggesting that the underwriting spread may in fact help the Treasury sell debt and avoid auction failures. We can interpret the estimated coefficient as the response in

\textsuperscript{25}See, for example, Schwarz (1978).
\textsuperscript{26}See, for example, Liu, Wu, and Zidek (1997).
Auction tenders to a one basis point change in expected risk-adjusted returns. For example, a one basis point increase in expected risk-adjusted returns leads to a 5.39 billion dollar increase in auction tenders for 4-week bills. This effect can be interpreted in light of the fact that the auction size ranged from 5 to 60 billion dollars over the sample period. Interestingly, the response of 2-year notes is much more sensitive to these changes, indicating that duration risk is important to auction bidders. The large adjusted $R^2$ suggests that these three factors explain most of the variation in auction tenders. Table VIII in appendix A. presents scatter plots of realized tenders versus their predicted values.

$$Tenders_t = \beta_0 + \beta_1 E_{t-1}[\mu_t] + \beta_2 Offer_t + \beta_3 Tenders_{t-1} + \epsilon_t$$

<table>
<thead>
<tr>
<th></th>
<th>4-Week Bills</th>
<th>13-Week Bills</th>
<th>26-Week Bills</th>
<th>2-Year Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>t-stat</td>
<td>Coeff.</td>
<td>t-stat</td>
</tr>
<tr>
<td>$E_{t-1}[\mu_t]$</td>
<td>5.39***</td>
<td>(4.77)</td>
<td>2.71***</td>
<td>(3.59)</td>
</tr>
<tr>
<td>Offer$_t$</td>
<td>1.53***</td>
<td>(8.22)</td>
<td>1.30***</td>
<td>(4.53)</td>
</tr>
<tr>
<td>Tenders$_{t-1}$</td>
<td>0.60***</td>
<td>(10.8)</td>
<td>0.76***</td>
<td>(15.3)</td>
</tr>
<tr>
<td>Constant</td>
<td>-12.1***</td>
<td>(-6.6)</td>
<td>-12.0***</td>
<td>(-4.1)</td>
</tr>
<tr>
<td>$R^2_{\text{adj}}$</td>
<td>0.931</td>
<td>0.963</td>
<td>0.964</td>
<td>0.930</td>
</tr>
<tr>
<td>Observations</td>
<td>777</td>
<td>777</td>
<td>776</td>
<td>197</td>
</tr>
</tbody>
</table>

Table II: Regression results of auction tenders on expected risk-adjusted returns by maturity. These results clearly indicate a positive and significant relationship between primary market demand and expected risk-adjusted returns. For example, a one basis point increase in expected risk-adjusted returns leads to a 8.76 billion dollar increase in auction tenders for 26-week bills. The large adjusted $R^2$ suggests that these three factors explain most of the variation in auction tenders. The covariance matrix is adjusted by Newey and West (1987) with a max lag determined by Andrews and Monahan (1992), so that standard errors and $t$-statistics are robust to serial correlation and heteroskedasticity.

An alternative, and perhaps more popular measure for primary market demand is the bid-to-cover ratio, which, recall, is the volume of auction tenders divided by the offering amount (auction size). Table III presents results for the regression of the bid-to-cover ratio on expected risk-adjusted returns. The results show a positive and significant relationship for each maturity, suggesting again that primary market demand depends on expected risk-adjusted returns that result from the underwriting spread. The estimated coefficient represents the response of the bid-to-cover ratio to a one basis point change in expected risk-
adjusted returns. For example, a one basis point increase in expected risk-adjusted returns leads to an increase in the bid-to-cover ratio of 0.2 for 4-week bills. Again, the response of 2-year notes is much more sensitive to changes in expected risk-adjusted returns, suggesting the importance of duration risk to primary market buyers. The large adjusted $R^2$ suggests that these three factors explain most of the variation in the bid-to-cover ratio. Table IX in appendix A. presents scatter plots of realized bid-to-cover ratios versus their predicted values.

\[ \text{Bid-to-cover}_t = \beta_0 + \beta_1 E_{t-1} \left( \frac{\mu_t}{\sigma^2_t} \right) + \beta_2 \text{Offer}_t + \beta_3 \text{Bid-to-cover}_{t-1} + \epsilon_t \]

<table>
<thead>
<tr>
<th>Maturity</th>
<th>Coeff.</th>
<th>$t$-stat</th>
<th>Coeff.</th>
<th>$t$-stat</th>
<th>Coeff.</th>
<th>$t$-stat</th>
<th>Coeff.</th>
<th>$t$-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-Week Bills</td>
<td>0.20***</td>
<td>(6.23)</td>
<td>0.07***</td>
<td>(3.47)</td>
<td>0.27***</td>
<td>(5.18)</td>
<td>5.45***</td>
<td>(2.32)</td>
</tr>
<tr>
<td>13-Week Bills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26-Week Bills</td>
<td>-0.01***</td>
<td>(-2.70)</td>
<td>0.015***</td>
<td>(6.12)</td>
<td>0.011***</td>
<td>(5.09)</td>
<td>0.002</td>
<td>(0.46)</td>
</tr>
<tr>
<td>2-Year Notes</td>
<td>0.77***</td>
<td>(24.1)</td>
<td>0.85***</td>
<td>(42.0)</td>
<td>0.86***</td>
<td>(48.4)</td>
<td>0.71***</td>
<td>(15.2)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.54***</td>
<td>(7.14)</td>
<td>0.08**</td>
<td>(2.49)</td>
<td>0.12***</td>
<td>(3.80)</td>
<td>0.55***</td>
<td>(4.59)</td>
</tr>
</tbody>
</table>

Table III: Regression results of the bid-to-cover ratio on expected risk-adjusted returns by maturity. These results clearly indicate a positive and significant relationship between primary market demand and expected risk-adjusted returns. For example, a one basis point increase in expected risk-adjusted returns leads to a 0.27 increase in the bid-to-cover ratio for 26-week bills. The large $R^2_{adj}$ suggests that these three factors explain most of the variation in the bid-to-cover ratio. The covariance matrix is adjusted by Newey and West (1987) with a max lag determined by Andrews and Monahan (1992), so that standard errors and $t$-statistics are robust to serial correlation and heteroskedasticity.

B. Volatility and Time-Varying Risk Premia

Using return volatility as a measure of risk, we should expect that risk averse investors earn higher returns when investing becomes more risky. This rationale implies that when underwriting returns become more volatile, primary market buyers should be compensated with a larger risk premium. We have already seen that primary market buyers earn what appears to be large excess returns, and that primary market demand depends on expected risk-adjusted returns. This subsection demonstrates that underwriting returns move with, and are explained by, volatility.
Figure VII demonstrates how the volatility of returns, and therefore the risk premium, move together and evolve through time. Each maturity has both upper and lower plots. Upper plots are the underwriting return $R_t$ in basis points, and lower plots are the estimated risk premium $\gamma \sigma_t$. Recall that $\sigma_t$ is a forecast of the volatility of $R_t$ for each auction using previous return information only. One can easily see that the volatility of $R_t$ is time varying, and that the mean return is larger when volatility is greater. Black stems denote nonnegative returns, red stems denote negative returns, and dashed lines denote estimated volatility regime switch dates. The risk premium plot is shaded to differentiate between estimated volatility regimes $S_k(t)$, in which the actual zero-lower-bound (ZLB) is marked with a dashed line.

Bills and notes appear to have responded differently to the financial crisis: bills are estimated to have three volatility regimes, while 2-year notes only have two. The increase in volatility of 2-year notes around the financial crisis did not justify a volatility regime switch. However, both bills and notes were affected by the ZLB, which significantly reduced both return volatility and the risk premium. As a result, the ZLB somewhat serendipitously reduced both debt issuance costs and debt servicing costs, all while increasing primary market demand by increasing expected risk-adjusted returns to auction buyers.

Since volatility is related to the risk premium, a natural question is how much of the expected return is compensation for risk? Or, how much of the expected return is explained by volatility? One way to answer this question is by comparing unconditional excess returns to returns with the risk premium removed. Table IV compares these for each maturity. The first row reports the estimates of $\gamma$, which determine the risk premium $\gamma \sigma_t$, and are the marginal change in returns to changes in volatility. The estimate is positive and significant for all maturities, implying that increases in volatility lead to increases in expected returns. Take 4-week bills for example, if the volatility of $R_t$ increases by one basis point, then the risk premium, and therefore expected return, will increase by 0.62 basis points. The second
Figure VII: Time series plots of the underwriting return $R_t$ and estimated risk premium $\gamma \sigma_t$, in basis points by maturity. In the upper plots, black stems denote nonnegative returns, red stems denote negative returns, and dashed lines denote estimated volatility regime switch dates. In the lower plots, the risk premium is calculated from forecasts of $\sigma_t$, in which the shading distinguishes volatility regimes $S_k(t)$, and ZLB is the actual zero-lower-bound date. Visually, the risk premium moves with the volatility of $R_t$. Note: a large negative outlier for 13-week bills exceeds the displayed range of $R_t$.

Row presents the unconditional expected return of $R_t$. For example, 2-year notes earn an average of 10.0 basis points from buying at the primary market price and selling at the secondary market price upon issuance. The third row subtracts the forecasted risk premium $\gamma \sigma_t$ from each return and then reports the remaining unconditional return. This effectively removes the portion of returns that are explained by volatility. It turns out that for 4-week bills and 2-year notes, removing the risk premium leaves an expected excess return that is not different than zero. In other words, volatility explains all of what is considered excess.
For 13- and 26-week bills, volatility explains about half of the excess returns.

<table>
<thead>
<tr>
<th></th>
<th>4-Week Bills</th>
<th>13-Week Bills</th>
<th>26-Week Bills</th>
<th>2-Year Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coeff.</td>
<td>0.62</td>
<td>0.26</td>
<td>0.27</td>
<td>0.51</td>
</tr>
<tr>
<td>t-stat</td>
<td>(7.55)</td>
<td>(4.66)</td>
<td>(4.32)</td>
<td>(3.22)</td>
</tr>
<tr>
<td>( \gamma )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( E[R_t] )</td>
<td>0.29***</td>
<td>0.77***</td>
<td>1.34***</td>
<td>10.0***</td>
</tr>
<tr>
<td>t-stat</td>
<td>(5.86)</td>
<td>(6.13)</td>
<td>(6.50)</td>
<td>(5.37)</td>
</tr>
<tr>
<td>( E[R_t - \gamma \sigma_t] )</td>
<td>0.01</td>
<td>0.38***</td>
<td>0.71***</td>
<td>0.72</td>
</tr>
<tr>
<td>t-stat</td>
<td>(0.49)</td>
<td>(4.35)</td>
<td>(4.55)</td>
<td>(0.41)</td>
</tr>
</tbody>
</table>

Table IV: Volatility and returns. The first row reports estimates of \( \gamma \), which determines the risk premium \( \gamma \sigma_t \), and is the marginal impact of volatility on returns. A one basis point change in volatility leads to a \( \gamma \) basis point change in expected returns. All estimates of \( \gamma \) are positive and significant, indicating that expected returns are increasing in risk. The second row presents unconditional expected returns. The third row subtracts the risk premium from each return and then reports the remaining unconditional return, essentially removing the portion of returns that is explained by volatility.

C. Model Estimation by Maturity

This subsection reports empirically estimated return processes by maturity. Each maturity has its own estimated mean and variance equations, and volatility regimes. Table V reports all of the estimated parameters and regime dates for the best fit model. It is important to note that neither moving average terms \( \theta_j \) nor GARCH terms \( \beta_j \) were found to be significant for any maturity. Abbreviated mean and variance equations, which exclude insignificant variables, are included in the table to aid the readers understanding. The mean and variance parameters are scaled to basis points so that an estimated coefficient value of 1.5 is equal to 1.5 basis points. Appendix B. reports the same results in equation form for each maturity.

The number of estimated volatility regimes differ between bills and notes. All bills have an estimated three volatility regimes corresponding to the volatility set \( \{S_1(t), S_2(t), S_3(t)\} \), while 2-year notes have only two regimes corresponding to the volatility set \( \{S_1(t), S_2(t)\} \). The estimated break dates can be considered in light of actual events.\(^{27}\) Volatility regimes of Treasury bills can be characterized by the period of moderate volatility associated with economic expansion, high volatility during the financial crisis, and low volatility at the zero-lower-bound (ZLB). Interestingly, the financial crisis began to affect Treasury bill auctions

\(^{27}\)See, for example, https://www.stlouisfed.org/financial-crisis/full-timeline for a timeline of events around the financial crisis presented by the Federal Reserve Bank of St. Louis.
in the summer of 2007, which is roughly a year before the stock market crash. The volatility of 2-year notes was not significantly different between the periods of economic expansion and financial crisis. However, some events are relevant to all maturities. In particular, the period of low volatility associated with the ZLB has at least two significant dates. On October 29th,
2008, the FOMC lowered the federal funds rate to 1.0 percent; and on December 16th, 2008, the FOMC voted to establish a target range for the effective federal funds rate of 0 to 0.25 percent. The estimated volatility regime switch dates of all Treasury bills occurred between these two events, while the less frequently issued 2-year notes experienced a volatility regime switch just after the federal funds target range was established.

There are several similarities and patterns that emerge from the estimated mean equations. First, the constants for 4-week bills and 2-year notes are not different between volatility regimes, nor are they significantly different from zero. This is consistent with the finding in the previous subsection that volatility explains all of the excess returns for these maturities. However, 13- and 26-week bills have a positive and significant constant for each volatility regime: both have small and positive constants that are the same for the first and third regimes, but have a large and positive constant during the financial crisis. For a given regime, the constants are larger for 26- than for 13-week bills. The positive and significant constant could be the result of some market frictions, inefficiencies, structural issues, or even a secondary market preference. Second, the estimated parameter $\gamma$ is positive and significant for each maturity, indicating that expected returns are increasing in volatility (risk). We can interpret this parameter as the marginal change in expected returns to changes in volatility, which can be used to compute the risk premium $\gamma \sigma_t$. Third, only 13-week bills had a negative and significant autoregressive term, revealing a negative return correlation from auction to auction. Finally, no moving average terms were significant for any maturity.

For each maturity, the variance equations consist of constants that differ by volatility regime and lagged squared innovations (ARCH effects). First, it is interesting to note that the constants are increasing in maturity, indicating that the volatility of underwriting returns is increasing in the term structure. This result helps to explain why average underwriting returns tend to be increasing in the term structure as well: if expected returns are compensation for risk, then maturities with higher volatility should have higher expected returns.
Second, the ZLB resulted in a profound reduction in volatility. For example, the volatility of 13-week bills was 243 times lower at the ZLB than during the financial crisis. Third, 4-week bills have only one significant ARCH term, while 13- and 26-week bills have two. Finally, there are no GARCH terms that are significant for any maturity. These results indicate that the variance is driven by a constant that varies over volatility regimes, some shocks to the variance, and the scaling factor $d_t$. In fact, for 2-year notes, the ARCH term is not significant, indicating that changes in the variance of underwriting returns, and therefore expected returns, are in large part due to the variation in the number of trading days between auction and issuance. Furthermore, this suggests that a large portion of the risk premium could be reduced by simply reducing the time to settlement.

VI. Conclusion

Evidence of what appears to be large excess returns could be cause for concern. Primary dealers are obviously making a profit from their role as underwriter and distributor. But is the Treasury paying too much, not enough, or is the payment just right? The main finding in this paper is that the bulk of these returns can be explained by volatility, so what appears to be excess is in fact compensation for risk. Perhaps that is a justifiable payment. However, the results in this paper suggest that the Treasury can reduce debt issuance costs by reducing the risk to auction participants. In particular, shortening the number of days between auction close and security issuance will decrease the volatility of the underwriting spread, which will reduce the risk premium and therefore costs borne by the Treasury.

While reducing debt issuance costs is important, changing the current auction structure could result in unintended consequences. In particular, the results in this paper show that risk-adjusted returns to auction buyers are a critical component of primary market demand. If it were possible to eliminate the cost to zero, would primary dealers still want to guarantee the sale of all government debt? Although reducing debt issuance costs to zero is extremely unlikely, consider what would happen if the Treasury required immediate settlement following
an auction. Auction participants would have to hold precautionary capital in case all of their
bids were accepted. This requirement would induce a carry cost, which would likely be offset
by lower auction prices. Therefore, it is likely that some cost to the Treasury will always
exist, but when the cost is compensation for risk, perhaps it is justifiable. Lowering the
cost by reducing risk to primary market buyers seems a tangible goal. Any proposed future
change to the primary market design should consider its impact on risk-adjusted returns, so
that the dual goal of increasing primary market demand and decreasing debt issuance costs
may be achieved.
References


———. 2016. The Early Years of the Primary Dealer System. *FRB of New York Staff Report* (777).


Online Appendix

A. Scatter Plots: Actual vs. Predicted Demand

Figure VIII presents scatter plots of actual auction tenders versus their predicted values by maturity. Predicted values are generated from the regression of auction tenders on expected risk-adjusted returns, controlling for auction size and serial correlation. The clear linear relationship suggests that larger expected risk-adjusted returns predict greater primary market demand for U.S. Treasury securities. Figure IX presents scatter plots of actual bid-to-cover ratios versus their predicted values by maturity. Predicted values are generated from the regression of the bid-to-cover ratio on expected risk-adjusted returns, controlling for auction size and serial correlation. Similar to auction tenders, there is a clear linear relationship, suggesting that larger expected risk-adjusted returns predict greater primary market demand.

Figure VIII: Scatter plots of realized auction tenders (vertical axis) versus their predicted values (horizontal axis) in billions of dollars by maturity. Predicted values are generated from the regression of auction tenders on expected risk-adjusted returns, controlling for auction size and serial correlation.

Figure IX: Scatter plots of realized bid-to-cover ratios (vertical axis) versus their predicted values (horizontal axis) by maturity. Predicted values are generated from the regression of the bid-to-cover ratio on expected risk-adjusted returns, controlling for auction size and serial correlation.
B. Estimated ARMA-GARCH-M Models in Equation Form

B.1 4-Week Bills

The estimated return process for 4-week bills is given by the mean equation 16 and variance equation 17, and is a volatility regime switching ARMA(0,0)-GARCH(0,1)-M process. The mean equation consists of a constant that is not different from zero, a risk premium, and noise. The risk premium is positively related to volatility, and can be interpreted as a one basis point increase in volatility leads to a 0.62 basis point increase in expected return. This implies that primary market buyers are compensated for risk, and that the compensation increases when security valuation becomes less certain.

\[ R_t = -0.02 + 0.62\sigma_t + \epsilon_t \]
\[ \sigma_t^2 = d_t(0.07[t \in S_1] + 1.54[t \in S_2] + 0.01[t \in S_3] + 0.21\epsilon_{t-1}^2) \]

The variance equation consists of a constant that varies by volatility regime, one lagged squared innovation, and the number of trading days between auction close and security issuance. Table VI presents the estimated volatility regimes start and end dates. There are an estimated three volatility regimes \( S_1, S_2, S_3 \); corresponding to the periods of moderate volatility experienced during the economic expansion, of high volatility during the financial crisis, and of low volatility while short term interest rates were at the ZLB, respectively. The variance at the ZLB is 154 times lower than during the crisis period, and 7 times lower than during the expansionary period. The effect of the financial crisis on 4-week bill auctions began around the June 19\textsuperscript{th}, 2007 auction, and lasted through the November 4\textsuperscript{th}, 2008 auction.

<table>
<thead>
<tr>
<th></th>
<th>( S )</th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Auction Date</td>
<td>7/30/01</td>
<td>7/30/01</td>
<td>6/19/07</td>
<td>11/12/08</td>
</tr>
<tr>
<td>Last Auction Date</td>
<td>6/28/16</td>
<td>6/12/07</td>
<td>11/4/08</td>
<td>6/28/16</td>
</tr>
<tr>
<td>Observations (( t ))</td>
<td>1-778</td>
<td>1-306</td>
<td>307-379</td>
<td>380-778</td>
</tr>
</tbody>
</table>

Table VI: Estimated volatility sets for 4-week bills. The entire sample of \( R_t \) belongs to the set \( S \). There are an estimated two structural breaks in the variance, implying three mutually exclusive subsets \( S_1, S_2, S_3 \) that span the set \( S \). The first row lists the first auction date included in a set, the second row lists the last auction date included in a set, and the third row lists the auction sequence values of \( t \) included in a set.
B.2 13-Week Bills

The estimated return process for 13-week bills is given by the mean equation 18 and variance equation 19, and is a volatility regime switching ARMA(1,0)-GARCH(0,2)-M process. The mean equation consists of a constant, a risk premium, a moving average, and noise. The constant is the same for the first and third volatility regimes, but is much larger during the financial crisis. A nonzero constant indicates that there may be some market friction that is orthogonal to volatility. The risk premium is positively related to volatility, and can be interpreted as a one basis point increase in volatility leads to a 0.26 basis point increase in expected return. This implies that primary market buyers are compensated for risk, and that the compensation increases when security valuation becomes less certain. The moving average is picking up some level of serial correlation in the return series.

\[
R_t = 0.14[t \in S_1 \cup S_3] + 3.21[t \in S_2] + 0.26\sigma_t - 0.15R_{t-1} + \epsilon_t
\]  
(18)

\[
\sigma_t^2 = d_t(0.15[t \in S_1] + 7.29[t \in S_2] + 0.03[t \in S_3] + 0.35\epsilon_{t-1}^2 + 0.15\epsilon_{t-2}^2)
\]  
(19)

The variance equation consists of a constant that varies by volatility regime, two lagged squared innovations, and the number of trading days between auction close and security issuance. Table VII presents the estimated volatility regimes start and end dates. There are an estimated three volatility regimes \(S_1, S_2, S_3\); corresponding to the periods of moderate volatility experienced during the economic expansion, of high volatility during the financial crisis, and of low volatility while short term interest rates were at the ZLB, respectively. The variance at the ZLB is 243 times lower than during the crisis period, and 5 times lower than during the expansionary period. The effect of the financial crisis on 13-week bill auctions began around the August 13th, 2007 auction, and lasted through the October 27th, 2008 auction.

<table>
<thead>
<tr>
<th></th>
<th>(S)</th>
<th>(S_1)</th>
<th>(S_2)</th>
<th>(S_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Auction Date</td>
<td>7/30/01</td>
<td>7/30/01</td>
<td>8/13/07</td>
<td>11/3/08</td>
</tr>
<tr>
<td>Last Auction Date</td>
<td>6/27/16</td>
<td>8/6/07</td>
<td>10/27/08</td>
<td>6/27/16</td>
</tr>
<tr>
<td>Observations ((t))</td>
<td>1-779</td>
<td>1-315</td>
<td>316-379</td>
<td>380-779</td>
</tr>
</tbody>
</table>

Table VII: Estimated volatility sets for 13-week bills. The entire sample of \(R_t\) belongs to the set \(S\). There are an estimated two structural breaks in the variance, implying three mutually exclusive subsets \(S_1, S_2, S_3\) that span the set \(S\). The first row lists the first auction date included in a set, the second row lists the last auction date included in a set, and the third row lists the auction sequence values of \(t\) included in a set.
B.3 26-Week Bills

The estimated return process for 26-week bills is given by the mean equation 20 and variance equation 21, and is a volatility regime switching ARMA(0,0)-GARCH(0,2)-M process. The mean equation consists of a constant, a risk premium, and noise. The constant is the same for the first and third volatility regimes, but is much larger during the financial crisis. A nonzero constant indicates that there may be some market friction that is orthogonal to volatility. The risk premium is positively related to volatility, and can be interpreted as a one basis point increase in volatility leads to a 0.27 basis point increase in expected return. This implies that primary market buyers are compensated for risk, and that the compensation increases when security valuation becomes less certain.

\[
R_t = 0.21[t \in S_1 \cup S_3] + 4.64[t \in S_2] + 0.27\sigma_t + \epsilon_t \tag{20}
\]

\[
\sigma_t^2 = d_t(0.65[t \in S_1] + 17.83[t \in S_2] + 0.12[t \in S_3] + 0.09\epsilon_{t-1}^2 + 0.24\epsilon_{t-2}^2) \tag{21}
\]

The variance equation consists of a constant that varies by volatility regime, two lagged squared innovations, and the number of trading days between auction close and security issuance. Table VIII presents the estimated volatility regimes start and end dates. There are an estimated three volatility regimes \(S_1, S_2, S_3\); corresponding to the periods of moderate volatility experienced during the economic expansion, of high volatility during the financial crisis, and of low volatility while short term interest rates were at the ZLB, respectively. The variance at the ZLB is 148 times lower than during the crisis period, and 5 times lower than during the expansionary period. The effect of the financial crisis on 26-week bill auctions began around the August 13\(^{th}\), 2007 auction, and lasted through the November 17\(^{th}\), 2008 auction.

<table>
<thead>
<tr>
<th></th>
<th>(S)</th>
<th>(S_1)</th>
<th>(S_2)</th>
<th>(S_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Auction Date</td>
<td>7/30/01</td>
<td>7/30/01</td>
<td>8/13/07</td>
<td>11/24/08</td>
</tr>
<tr>
<td>Last Auction Date</td>
<td>6/27/16</td>
<td>8/6/07</td>
<td>11/17/08</td>
<td>6/27/16</td>
</tr>
<tr>
<td>Observations ((t))</td>
<td>1-778</td>
<td>1-314</td>
<td>315-381</td>
<td>382-778</td>
</tr>
</tbody>
</table>

Table VIII: Estimated volatility sets for 26-week bills. The entire sample of \(R_t\) belongs to the set \(S\). There are an estimated two structural breaks in the variance, implying three mutually exclusive subsets \(S_1, S_2, S_3\) that span the set \(S\). The first row lists the first auction date included in a set, the second row lists the last auction date included in a set, and the third row lists the auction sequence values of \(t\) included in a set.
B.4 2-Year Notes

The estimated return process for 2-year notes is given by the mean equation 22 and variance equation 23, and is a volatility regime switching ARMA(0,0)-GARCH(0,1)-M process. The mean equation consists of a constant that is not different from zero, a risk premium, and noise. The risk premium is positively related to volatility, and can be interpreted as a one basis point increase in volatility leads to a 0.51 basis point increase in expected return. This implies that primary market buyers are compensated for risk, and that the compensation increases when security valuation becomes less certain.

\[
R_t = 0.64 + 0.51\sigma_t + \epsilon_t
\]

\[
\sigma^2_t = d_t(195.72[t \in S_1] + 21.71[t \in S_2] + 0.02\epsilon^2_{t-1})
\]

The variance equation consists of a constant that varies by volatility regime, one lagged squared innovation, and the number of trading days between auction close and security issuance. Table VI presents the estimated volatility regimes start and end dates. There are an estimated two volatility regimes \(S_1, S_2\); corresponding to the periods of higher volatility before the ZLB, and lower volatility at the ZLB. The variance at the ZLB is 9 times lower than the previous regime. Unlike Treasury bills, 2-year notes did not experience a structural break in the volatility during the financial crisis. Also, the coefficient on the lagged squared innovation is not significantly different than zero, which indicates there is not significant updating of the variance within a volatility regime. Within a regime, the volatility is roughly constant, however, the variation in the risk premium is mostly due to the variation in \(h_t\), indicating the importance of this time delay. The effect of the ZLB on 2-year note auctions is right in line with the introduction of the actual ZLB.

<table>
<thead>
<tr>
<th>Set</th>
<th>S</th>
<th>S_1</th>
<th>S_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Auction Date</td>
<td>1/26/00</td>
<td>1/26/00</td>
<td>1/27/09</td>
</tr>
<tr>
<td>Last Auction Date</td>
<td>6/20/16</td>
<td>12/22/08</td>
<td>6/20/16</td>
</tr>
<tr>
<td>Observations (t)</td>
<td>1-198</td>
<td>1-108</td>
<td>109-198</td>
</tr>
</tbody>
</table>

Table IX: Estimated volatility sets for 2-year notes. The entire sample of \(R_t\) belongs to the set \(S\). There is an estimated one structural break in the variance, implying two mutually exclusive subsets \(S_1, S_2\) that span the set \(S\). The first row lists the first auction date included in a set, the second row lists the last auction date included in a set, and the third row lists the auction sequence values of \(t\) included in a set.