On-the-run Premia, Settlement Fails, and Central Bank Access*

Fabienne Schneider†
University of Bern & Study Center Gerzensee

November 4, 2023

Abstract

The premium on "on-the-run" Treasuries is an anomaly. I explain it using a model in which primary dealers hold inventories of Treasuries. Primary dealers are more likely to hold large inventories of the most recent issues of Treasuries (i.e., on-the-run Treasuries). Because on-the-run Treasuries are easier to find, they trade at a premium. My theory is consistent with the USD 40 billion of Treasury contracts that fail to settle each day, with the median failure rate of off-the-run Treasuries being almost twice that of on-the-run Treasuries. I use the model to analyse the effects of granting access to central bank facilities to non-banks active in the Treasury market. Broad access stimulates trading and reduces the on-the-run premium, but settlement fails increase and, counterintuitively, only primary dealers benefit.

Keywords: Treasuries, on-the-run premium, settlement fails, central bank facility, non-banks
JEL Codes: G12, G19, G23

*I thank Martin Brown, Plamen Nenov, Dirk Niepelt, and especially Cyril Monnet for help, Michael Fleming for information on data, Jens Christensen, Stefania D'Amico, Refet Gürkaynak, and Antoine Martin for sharing their knowledge of the market, and Bruno Biais, Arvind Krishnamurthy, Ricardo Lagos, and the participants of the Macroeconomics PhD Seminar University of Bern (2021), Central Bankers Course on Money Markets, Liquidity, and Payment Systems (2022), Macroeconomics Workshop Hasliberg (2022), BDP Alumni Conference (2022), Brown Bag Seminar University of Bern (2022), 35th Australasian Finance & Banking Conference (2022), Swiss Society for Financial Market Research Conference (2023), and other conferences for feedback, insights, and discussion.

†E-mail: schneider.fabienne@icloud.com

1 Introduction

With an average daily trading volume of half a trillion US dollars, the US Treasury market is one of the most important and liquid markets in the US financial system, crucial to the conduct of monetary policy and a key pillar of the US economy. Despite its importance, the US Treasury market exhibits some irregularities, which I describe in detail in the next section and summarise here.[1]

First, it is well known that on-the-run Treasuries – the most recently issued Treasuries – trade at significantly lower yields, higher prices and lower repo rates than other Treasuries (known as off-the-run) with similar cash flows and maturities, giving rise to a puzzling arbitrage opportunity known as the "on-the-run" premium (see Vayanos and Weill (2008), D'Amico and Pancost (2022) and figure 2a). Second, despite trading at a premium, the volume of trades in on-the-run Treasuries is much larger than that in off-the-run Treasuries (see figure 2b). Third, on average USD 40 billion of Treasury contracts fail to settle each day (see figure 3a). Interestingly, failure rates differ by Treasury type, with on-the-run Treasuries having a median settlement failure rate almost half that of off-the-run Treasuries (see figure 3b).

These stylised facts and irregularities raise the following questions: Why is there an on-the-run premium, and why is it always on the most recently issued Treasuries? Why do the cheaper off-the-run Treasuries trade at lower volumes? How can there be settlement fails in a benchmark market such as the US Treasury market, and why do the off-the-run Treasuries fail to settle more often?

In the first part of the paper, I develop a model of the US Treasury market to answer these questions. In the second part, which I describe in more detail below, I use it to conduct policy analysis motivated by current discussions about how to restructure the market.

The US Treasury market model incorporates the key features of the market that I describe in 2 section: it is an over-the-counter (OTC) market where primary dealers are the first acquirers of Treasuries at the primary auction. To model these characteristics, I assume that there are three types of agents: sellers, buyers and primary dealers. A seller is any financial entity other than a primary dealer, such as a non-bank, that sells Treasuries short. A buyer is akin to a long-term holder of Treasuries, such as a pension fund. There are different types of Treasuries, on- and off-the-run, and to simplify the model I do not model the primary auction of Treasuries, but assume that primary dealers are endowed with the latest issue of Treasuries. Buyers have the highest valuation for Treasuries, but the market is segmented and they cannot contact primary dealers directly, only through sellers.

In a first market, sellers sell financial contracts to buyers that promise to deliver a specific type of Treasury (for short, on- or off-the-run, including maturity date). Because the seller can fail to settle, the contract is

[1] The average daily trading volume is the volume reported to TRACE between February and October 2023. The TRACE data is available here: https://www.finra.org/finra-data/browse-catalog/about-treasury/monthly-file.

secured by collateral. Next, sellers contact primary dealers to buy the desired type of Treasury in an OTC market. There, the seller is randomly matched with a primary dealer. The primary dealers always have the most recent issue of Treasuries in their inventory, as they have just been auctioned. However, depending on their trading history, they may not have enough of the desired off-the-run Treasuries. In this case, the seller fails to settle and she delivers as many Treasuries as possible to the buyer in accordance with their contracts. If necessary, the buyer seizes the collateral to cover the undelivered amount. These fails do not occur with on-the-run Treasuries because all primary dealers hold the same inventories since they were just filled up. Once all the trades have been conducted, the sellers can deposit any remaining idle balances in a central bank facility and receive an interest rate on them.[2] Sellers then go into the next sequence of trades.

In equilibrium, I show that the occurrence of settlement fails leads to a preference for the safer on-the-run Treasuries. Because they have been in the market for a shorter time, sellers are more likely to find them and they have a greater chance to settle. Therefore, on-the-run Treasuries trade at a premium and in greater volume than off-the-run Treasuries with the same cash flow and maturity, explaining the three stylised facts for the US Treasury market. I also provide empirical evidence that lower inventories imply more settlement fails in off-the-run Treasuries, and more such settlement fails imply higher on-the-run premia, as predicted by the model.

In a second part of the paper, I use the model to shed light on the current policy discussion about the need to restructure the Treasury market (see the discussion at the Jackson Hole conference by Duffie (2023)). The background to this discussion is epitomised by the US Treasury market crisis of March 2020. In the aftermath of the great financial crisis of 2007, primary dealers faced tighter regulatory constraints, forcing them to reduce their balance sheet space for Treasuries. At the same time, the US Treasury market grew strongly. Non-bank financial institutions have filled the space left by primary dealers.[3] But in March 2020, the presence of non-banks in the US Treasury market led to a rapid drying up of liquidity and a sharp decline in market depth, exacerbated by the reluctance of primary dealers to take more US Treasuries onto their balance sheets (Eren and Wooldridge (2021)). Off-the-run Treasuries were at the epicentre of the crisis (Wells (2023)). This is reflected in the on-the-run premia across all maturities, which rose sharply as shown in figure 1.[4]

[2]The facility in my model can be interpreted as a deposit facility or a reverse repo facility where I focus on the cash leg and abstract from the collateral part. First, the facility's repos are general collateral repos and the cash lender is willing to receive any securities that fall into a broad class. He does not search for a specific security (Bowman et al. (2017)). Second, even if the facility were to provide a specific security that was sought, the security would have to be returned the next day and would only facilitate the availability on a temporary basis.

[3]On the BrokerTec platform, one of the main marketplaces, non-banks, especially principal trading firms, already accounted for more than half of the trading in benchmark 5-year, 10-year and 30-year bonds in 2015. Traditional banks and dealers had a share of 30-40%.

[4]As in Christensen et al. (2017), the on-the-run yield is subtracted from the par yield of seasoned bonds. The data are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak

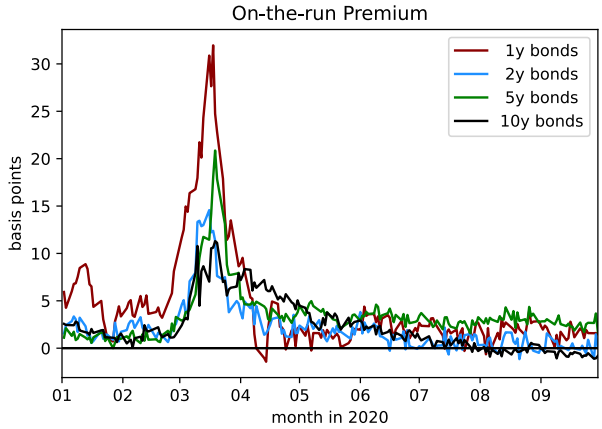


Figure 1: On-the-run premia during the "March 2020" breakdown

Among their ten recommendations for the US Treasury market, the Working Group on Treasury Market Liquidity led by Duffie, Geithner, Parkinson and Stein (2021) recommends broad access to central bank repo financing. They criticise the current facility for providing limited access to primary dealers and banks rather than a broad range of market participants.[5]

The model helps to understand the impact of broad access to central bank facilities on prices, premia, traded quantities, fails and profits. A first observation is that the facility is not a substitute for trading, but rather complements it. It generates a certain return on liquid funds between trades for those who have access. This feeds back into the overall cost of trading. If sellers gain access, an increase in the facility rate stimulates trading and prices rise. The stimulated trading implies that more Treasuries end up in the hands of buyers and less in the inventories of primary dealers. Settlement is more likely to fail. The reason is that as more Treasuries are sold early, fewer are available during the off-the-run period.

Interestingly, as I found in the data, the premium decreases as the facility rate rises. The reason is intuitive: as off-the-run Treasury prices initially rise, so does the value of the collateral. This leaves buyers better off in the event of default, and so buyers increase their demand for the contract with the off-the-run Treasuries. Therefore, the increase in the price of the off-the-run Treasuries is greater than the increase in the price of the on-the-run Treasuries. Also because of this additional demand effect, the increase in the quantity of off-the-run Treasuries traded is initially larger than the increase in the quantity of on-the-run Treasuries.[6]

et al. (2010)), https://www.federalreserve.gov/data/tips-yield-curve-and-inflation-compensation.htm, and the FRB-H15 tables, https://www.federalreserve.gov/releases/h15/.

[5]Another example is the FED's reverse repo facility. An increasing number of institutions already access this facility (Frost et al. (2015), Baklanova et al. (2015) and Marte (2021)).

[6]This policy effect is particularly relevant in the context of the recent crisis, where the market for off-the-run Treasuries

Paradoxically, in equilibrium, only the primary dealers benefit from a rise in the facility rate, and those who are granted access do not. This is because the primary dealers can now sell more Treasuries at a higher price. However, perfect competition in the contract market erodes any advantage that sellers may have. Finally, buyers of Treasuries lose, first through higher prices and second through an externality. The buyer does not take into account that if he buys more Treasuries early, fewer will be available during their off-the-run period, implying a higher default rate.[7]

Related Literature A well-known framework for the on-the-run premium is provided by Vayanos and Weill (2008). They have a setup where there are two assets with identical cash flows and agents can go long or short an asset. Since short sellers have to deliver the asset they have borrowed, they face search externalities and favour the asset that is more liquid. Liquidity is self-fulfilling in their model. As in Vayanos and Weill (2008) I also include OTC market frictions and delivery constraints in my model. However, my model is dynamic whereas theirs is static. In addition, I include a key factor to explain the on-the-run premium: the fact that one asset, the off-the-run asset, has been available in the market for a longer time. This also allows for equilibrium selection with a premium on the on-the-run asset, which is not the case in Vayanos and Weill (2008), who have two self-fulfilling equilibria with the premium on both assets.
Another theory of the on-the-run premium comes from Pasquariello and Vega (2009). In their model, the premium arises from endowment shocks. There are two frictions: information heterogeneity and imperfect competition among traders. My model is similar to theirs in the sense that I include uncertainty due to limited information. In my model, the uncertainty is about the stock of off-the-run Treasuries. In their model, it is with respect to on-the-run Treasuries, since the endowment shocks received by agents are private information.

Broadly speaking, compared to Vayanos and Weill (2008) and Pasquariello and Vega (2009), my model focuses on primary dealer inventory and settlement risk to explain the premium. All dynamics are driven solely by the fact that Treasuries are in the market for a different lengths of time since issuance. This element implies that the premium is always on the on-the-run asset. The model allows for a more general discussion of the US Treasury market and, in addition, I can also analyse the impact of access.

Empirical work attempting to explain the on-the-run premium includes, for example, Strebulaev (2002), Goldreich et al. (2005) and D'Amico and Pancost (2022). Strebulaev (2002) suggests that the premium may measure differences in tax treatment rather than liquidity premia. Goldreich et al. (2005) distinguish between current and future liquidity and suggest that expected future liquidity, not just current liquidity, determines prices and is a significant driver of the on-the-run premium. D'Amico and Pancost (2022) link the on-the-run premium to the risk of unexpected fluctuations in the collateral value of Treasuries.

froze (Eren and Wooldridge (2021)).

[7] Few of the results depend on search frictions being above a small minimum. See Appendix B.2.

A recent paper related to mine is Corradin and Maddaloni (2020). They build on Vayanos and Weill (2008) and study central bank intervention. Compared to my paper, they do not study access to facilities but central bank purchases. Specifically, they analyse how purchases by the European Central Bank affected repo specialness in the Italian government bond repo market during the euro area sovereign debt crisis. Specialness is the premium paid to procure a particular bond in the repo market, and on-the-run bonds often trade as "special". They show that purchases reduce liquidity and increase specialness in the presence of short selling. They also show that bonds in high demand and older bonds with lower turnover are more likely to fail. The probability of default increases with the specialness of the asset. In contrast, Liu and Wu (2017) show that the on-the-run premium is low when counterparty risk is high. Compared to my model, the risk in Liu and Wu (2017) and Corradin and Maddaloni (2020) refers to variations in a general risk measure or only in the specific asset. Corradin and Maddaloni (2020) focus exclusively on crisis periods, Liu and Wu (2017) state that their results are particularly pronounced in such periods.

With respect to the OTC market environment, my paper is related to Duffie et al. (2005) and Lagos and Rocheteau (2009).

The structure of the paper is as follows: Section 2 describes the US Treasury market, section 3 the environment and section 4 the value functions. The equilibrium is also defined. Section 5 proves the existence of our main equilibrium of interest. Section 6 discusses how the model explains the stylised facts and irregularities and why the premium is always on the on-the-run Treasury. The implications of broad access to central bank facilities are analysed in section 7. Section 8 presents the Treasury lifecycle. Section 9 concludes.

2 Description of the Treasury market

In this section I describe the US Treasury market, its structure and trading dynamics, and provide evidence for the stylised facts highlighted in the introduction.

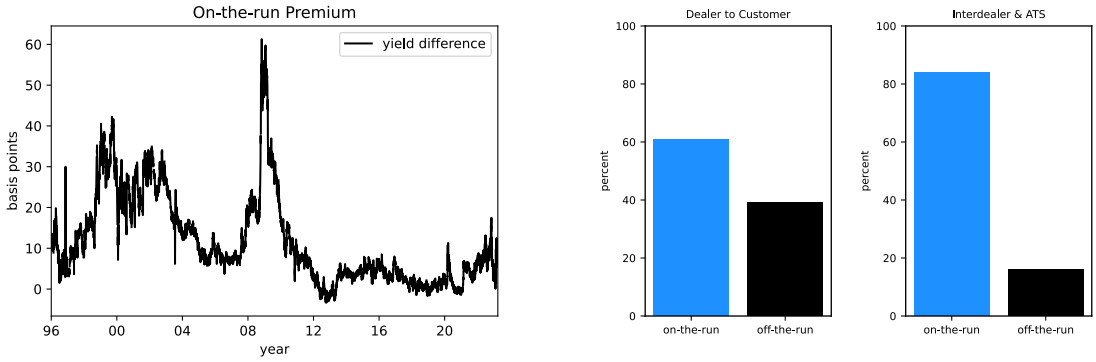
The Treasury spot market is OTC (Fleming et al. (2018)). This means that there is no all-to-all trading at a central venue and no central pricing. Depending on the security traded and the trading partners involved, the degree of friction in the OTC market varies. For example, dealer-to-dealer trading of benchmark on-the-run Treasuries on electronic platforms such as BrokerTec is less frictional than interdealer and dealer-to-customer trading of the less liquid off-the-run Treasuries intermediated on voice and more manually assisted electronic platforms (Bessembinder et al. (2020) and U.S. Department of the Treasury et al. (2015)).[8]

On-the-run Treasuries are the most recently issued Treasuries of a given maturity, and all previously issued Treasuries of the same maturity are referred to as off-the-run. As figure 2a shows for 10-year Treasuries, on-the-run Treasuries trade at significantly lower yields than off-the-run Treasuries with very similar cash

[8]The overall share of trading on all types of electronic platforms in the US Treasury market is 70 percent (Bech et al. (2016)).

flows and maturities.[9] They also have higher prices and lower repo rates. The term "very similar" refers to the fact that, in general, there are not two Treasuries in the market with exactly the same cash flow and maturity where one is on-the-run and the other is off-the-run. In fact, if you want to compare Treasuries with the same overall maturity, it is impossible to do so. In practice, therefore, one either compares Treasuries with the same overall maturity using an estimated off-the-run yield curve (see, for example, the first figure in Christensen et al. (2017)), or one abstracts from small differences in cash flows and maturity dates, or one compares on- and off-the-run Treasuries with the same maturity date that have a different overall maturity (see, for example, Christensen et al. (2020)). In figure 2a, as in Christensen et al. (2017), the on-the-run yield is subtracted from the par yield of seasoned bonds.

Also, despite being more expensive and far fewer in number, on-the-run Treasuries trade in much larger volumes than off-the-run Treasuries, as shown in figure 2b.[10] This is true whether we look at dealer-to-customer or interdealer and automated trading system (ATS) trades.



(a) On-the-run premium on 10 year Treasury bonds

(b) Trading volumes

Figure 2: Prices and Quantities

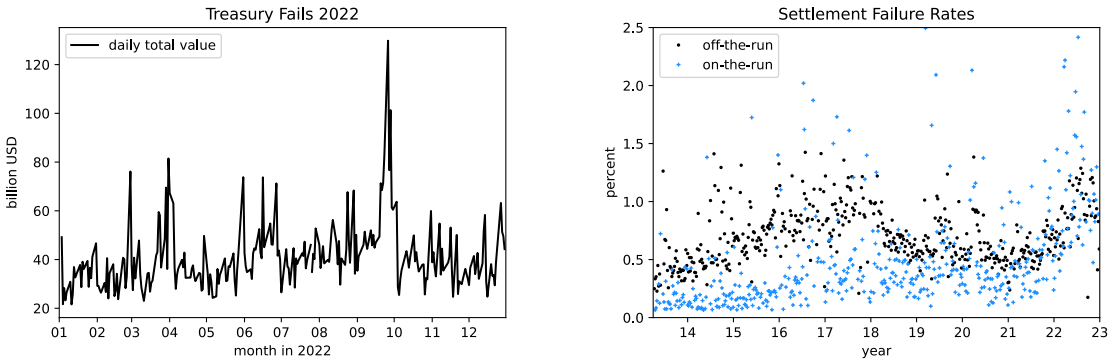
Another surprising fact is that, on average, USD 40 billion of Treasuries are not delivered on time to settle a contract each day.[11] These events are commonly referred to as "settlement fails". Figure 3a shows the

[9]The data are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al. (2010)), https://www.federalreserve.gov/data/tips-yield-curve-and-inflation-compensation.htm, and the FRB-H15 tables, https://www.federalreserve.gov/releases/h15/.

[10]Trading volumes in on-the-run and off-the-run Treasuries are the volumes reported to TRACE between February and October 2023. The TRACE data is available here:
https://www.finra.org/finra-data/browse-catalog/about-treasury/monthly-file.

[11]The data is provided by the Depository Trust and Clearing Corporation (DTCC) and can be downloaded here: https://www.dtcc.com/charts/daily-total-us-treasury-trade-fails. In times of stress, the daily value can spike. Current policy discussions consider central clearing as an effective way to significantly reduce fails in the future (Fleming and Keane (2021)). For more information on settlement fails, see Fleming and Garbade (2005).

failure rate, which is calculated by dividing the value of Treasuries that failed to be delivered on time by the value of all Treasuries traded. Interestingly, the failure rates differ depending on whether a Treasury is on- or off-the-run, and figure 3b shows that fails involving on-the-run Treasuries are less frequent than those involving off-the-run Treasuries.[12]



(a) Daily settlement fails

(b) Fails of on- and off-the-run Treasuries

Figure 3: Fails

The OTC structure implies that there are search costs that can explain settlement fails. In particular, search costs become relevant when financial contracts include delivery constraints. For example, spot market trades are often complemented by special repo trades to short-sell specific Treasuries. "Special" refers to the fact that the collateral of the repo is fixed and determined by its number, called ISIN or CUSIP, and the repo may have a rate that differs from the general collateral rate. To short-sell a particular Treasury, it is borrowed using a special repo and sold in the market today. The next day, a Treasury with the same ISIN or CUSIP is bought in the spot market, preferably at a lower price than it was sold on the previous day, and returned to the lender in the repo transaction. If such a Treasury cannot be found, a settlement fail occurs. The borrower of the Treasury in the repo transaction pays a penalty, the Treasury Market Practice Group (TMPG) fail charge.[13] Fleming and Keane (2021) write that on-the-run Treasury fails account for less than a quarter of all fails in non-crisis periods.[14] The figure in Fleming et al. (2014), presented in the appendix C, shows that

[12]The data are from the FED's primary dealer statistics. It includes outright and financing fails. The median failure rate for on-the-run Treasuries is 0.36% and the median failure rate for off-the-run Treasuries is 0.66%. The rates are not an exact measure. This is because one part of the time series used in the calculation is an average over the reporting week and the other part of the time series reports a value as of the reporting weekday. Given the high frequency, this should not matter and the observed pattern is clear. Each series is outlier adjusted, where an outlier is defined as being below the 2.5% percentile and above the 97.5% percentile. Rates up to 2.5% are shown in the figure. Few rates are higher.

[13]For more information on the TMPG fail charge, see https://www.newyorkfed.org/tmpg and Garbade et al. (2010).

[14]In addition, they also note that on-the-run Treasuries are more often involved in so-called daisy chain fails. One fail implies another as the trades are linked in a chain.

gross fails are much higher in seasoned Treasuries than in others (including on-the-run Treasuries).
Note that arbitrage to exploit the price difference between on- and off-the-run Treasuries involves short selling, but is mostly prohibited by efficient markets because repo rates for them also differ (Krishnamurthy (2002)).

3 The environment

Time is discrete and goes on forever, $t = 0, 1, ..., \infty$. The discount factor is $\beta \in (0, 1)$ and each period consists of two subperiods. There are three types of infinitely lived agents in the model: a buyer, a seller, and a primary dealer.[15] There is a continuum in each type.

There are two segmented sequential markets. The first market is called the spot market. It takes place in the first subperiod and is an OTC market. The second market is Walrasian and takes place in the second subperiod. It is called contract market. The upper part of figure 4 gives an overview over the timeline.

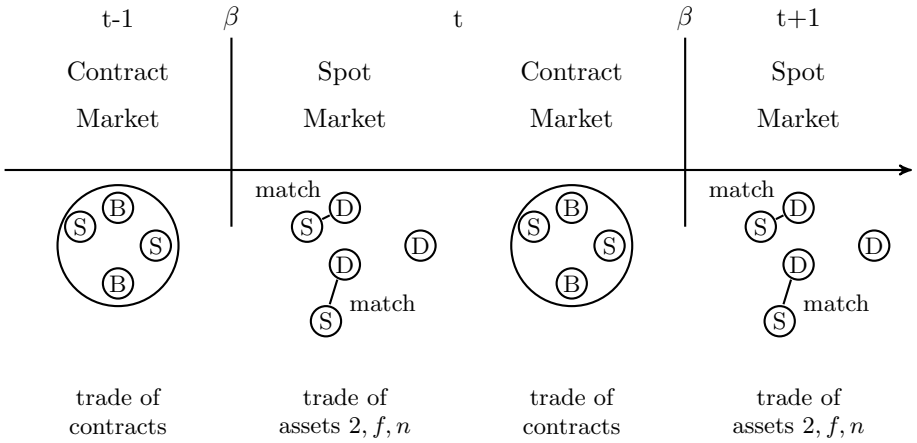


Figure 4: Timeline

There are two goods: a settlement good and real coupons. The settlement good $m \in \mathbb{R}_0^+$ is storable and divisible. Coupons δ are perishable and are given by two assets: one asset gives one coupon per each second subperiod for two consecutive periods, the other for one period. An asset is on-the-run if it belongs to the most recently issued generation of its maturity. Therefore, in each period there are two types of on-the-run and one type of off-the-run assets available for trading: the two-period assets maturing in two periods (2), the two-period assets maturing in one period (f), and the one-period assets maturing in one period (n). The letters n and f refer to their respective on-the-run (n) and off-the-run (f) state. I will refer to them

[15] A seller is any financial entity other than a primary dealer. It can for example be a non-bank. A buyer can be interpreted as a long-term holder (e.g. a pension fund).

henceforth as the on-the-run asset and the off-the-run asset. The two-period asset maturing in two periods is also on-the-run, but its state is not relevant to the analysis. The following figure gives an overview over the assets and their cash flow.

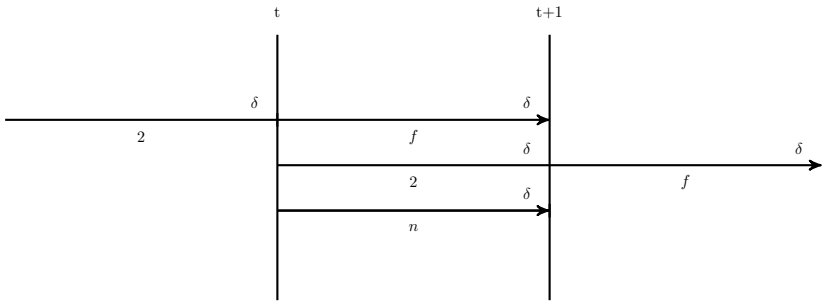


Figure 5: Assets in period t

The figure illustrates that the on-the-run (n) and off-the-run (f) assets are identical in terms of maturity date and coupon. The only difference is the issuance date. Therefore, I compare these two assets to measure the on-the-run premium.[16] At the beginning of each period, primary dealers are endowed with a stock $I^2 = \mathcal{I} \in \mathbb{R}^+$ of newly issued two-period assets and a stock $I^n = \mathcal{I} \in \mathbb{R}^+$ of newly issued one-period assets. The primary dealers' stock I^f of off-the-run assets is endogenous. I assume that buyers and sellers have knowledge of the primary dealers' inventory stock of newly issued assets. They also know the distribution of primary dealers' inventories of off-the-run assets (as they know the matching probability described below), but not the inventories of each primary dealer.

Primary dealers and buyers have linear utility from consuming the coupons in the second subperiod. Primary dealers receive utility δ from consuming the coupons, while buyers receive utility $\delta + h > \delta$.[17] Sellers do not value the coupons.[18] All agents have linear utility (disutility) from consuming (producing) the settlement good. The seller can produce it only in the second subperiod. The others can always produce it.[19] It is used to settle trades. It has properties similar to money except that it is a real asset.[20]

[16] The two assets have the same cash flow to maturity and the same maturity date as in Vayanos and Weill (2008) and Pasquariello and Vega (2009). See also, for example, Christensen et al. (2020) for how the premium can be measured. Note that if we would use the two-period on-the-run asset to calculate the premium, we would have to abstract from the second coupon rate as cash flow differences should not be the reason for the premium. The premium and all other results would remain the same.

[17] h can be interpreted as a hedging benefit from holding the asset or simply as a different valuation.

[18] It would not change any results if sellers would value the coupons as much as primary dealers.

[19] This aspect of the model ensures that there is no incentive for the seller to build up settlement goods only to deposit them in the facility. An agreement between a seller and a buyer or a primary dealer whereby the latter would produce settlement goods for the seller today so that the seller could deposit it and pay it back later with a profit is not possible because the seller cannot commit.

[20] The only difference from a nominal model is that the settlement good does not lose or gain value over time due to inflation.

In the OTC spot market, primary dealers have a match with probability $(1-\sigma) > 0$ with a seller. Sellers always have a match with a primary dealer.[21] Buyers have no access to this market. Primary dealers sell quantities of assets A^i to sellers at price p^i, where $i = \{2, f, n\}$. They face a cost of $\kappa(A^i)$ when selling assets i in the form of settlement good.[22] The function introduces a non-linearity into the model and leads to an interior solution and a determined price.[23] I assume that $\kappa(0) = 0$, $\kappa'(A^i) > 0$, $\kappa''(A^i) > 0$, and that the function is continuous.[24] Possible interpretations of the function are a nonlinear portfolio adjustment cost (see Garleanu and Pedersen (2013) and Bacchetta and van Wincoop (2021) for examples) or a regulatory cost (see, for example, Macchiavelli and Pettit (2021)).

In every other subperiod so-called contracts are traded by the seller and the buyer. The market is called contract market. The seller sells the contracts to the buyer. A contract is a list $l^i = [a^i, \omega^i, q^i]$. a^i specifies the amount of assets i promised to be delivered in the next subperiod. $\omega^i a^i$ is the collateral that the seller has to post at the moment of selling the contract. The buyer has first claim over it in the event of non-delivery.[25] q^i is the contract price in the second subperiod. $q^i a^i$ is the payment due from the buyer at settlement in the next subperiod. Figure 4 gives an overview over the markets.

In the basic model the seller has access to a central bank facility. The facility can be accessed every first subperiod for one subperiod. Sellers can deposit settlement good and receive an interest rate r on it. I assume that $\beta(1+r) < 1$. This implies that it is not worthwhile to accumulate settlement good one period in advance in order to deposit it in the facility.[26]

4 Value functions and equilibrium

4.1 Primary dealer

The primary dealer value function at the beginning of the contract market, when holding asset inventories of I_t^2, I_t^f, and I_t^n is

The dynamics would not change with a nominal model, and therefore discussing inflation would not add anything relevant.

[21] I make this assumption for simplicity. Changing it would not change the dynamics.

[22] In equilibrium they only sell assets.

[23] The main results also hold without this function. It gets relevant when I discuss access.

[24] Another possibility would be that the function depends on the sum of all assets sold. But I can show that for a positive premium this cannot be the case. Also, we need that the function is increasing and convex.

[25] The collateral in the form of settlement good is similar to the Treasury Market Practice Group (TMPG) fails charge. This fee allows a buyer of Treasuries to claim monetary compensation from the seller if the seller fails to deliver the Treasuries on time. For more information, see https://www.newyorkfed.org/tmpg and Garbade et al. (2010).

[26] We could in addition assume that primary dealers also have access to the facility, but this does not change the dynamics. That is why I am omitting it to ease notation.

$$V^D(I_t^2, I_t^f, I_t^n) = \beta(1-\sigma)\Big\{\sum_i \Big[p_t^i A_t^i - \kappa(A_t^i) + \delta(I_t^i - A_t^i)\Big] + \beta V^D(I_{t+1}^2, I_t^2 - A_t^2, I_{t+1}^n)\Big\} \\ + \beta\sigma\Big\{\sum_i \delta I_t^i + \beta V^D(I_{t+1}^2, I_t^2, I_{t+1}^n)\Big\}. \tag{1}$$

With a probability of $(1-\sigma)$, the primary dealer has a match with a seller in the spot market and can sell assets. Optimal prices and quantities are determined by the bargaining problem (3) described below. For each kind of asset $i \in \{2, f, n\}$, the primary dealer sells the optimal amount A_t^i. When selling the amount A_t^i he receives the price p_t^i for each of them and faces the cost $\kappa(A_t^i)$. In addition, he cannot consume any future coupons of these assets but only of his inventory left, $(I_t^i - A_t^i)$. With probability σ the primary dealer has no match and consumes the coupons of his inventory of assets. The inventory of off-the-run assets, I_t^f, is past dependent and therefore endogenous. Its size depends on the trading activities of the primary dealer in the former period. The primary dealer is also endowed with the newly issued assets I_t^2 and I_t^n. As they are just issued, their inventory is of size $\mathcal{I}$. The three inventory quantities are the state variables.

The amounts A_t^i sold and the price p_t^i are determined by a bargaining problem between the primary dealer and the seller. The primary dealer has full bargaining power.[27] This means that he sets the price just as high such that the seller is indifferent between delivery and non-delivery, i.e. $\left(q_{t-1}^i - p_t^i + \omega_{t-1}^i\right) a_t^i = q_{t-1}^i a_t^i$. The price therefore equals the collateral value ω_{t-1}^i:

$$p_t^i = \omega_{t-1}^i \qquad \forall\, i \text{ and } t. \tag{2}$$

The primary dealer maximizes his trade surplus. The lagrange function to his maximization problem in each period t is given by:[28]

$$\mathcal{L}(\{A_t^i, p_t^i, \lambda_t^i, \widetilde{\lambda}_t^i\}_i) = \sum_i \left(p_t^i A_t^i - \kappa(A_t^i) - \delta A_t^i - \beta\delta A_t^2\right) + \lambda_t^i \left[I_t^i - A_t^i\right] + \widetilde{\lambda}_t^i \left[a_t^i - A_t^i\right] + \beta(1-\sigma)\lambda_{t+1}^f \left[\mathcal{I} - A_t^2\right] \tag{3}$$

[27] Given that the seller already has skin in the game, due to having sold a certain number of contracts, this is a reasonable assumption.

[28] If the primary dealer does not sell the two-period assets today, then the assets remain in his inventory in the next period if he has no match, or if he has a match, I assume that part is sold and the rest remains in his inventory as well. This means that his inventory constraint on the off-the-run assets is not binding if he still has the full inventory available in the next period. I will show later why I assume that this holds in equilibrium and that I can always find equilibria where it does. The surplus can be written as $p_t^i A_t^i - \kappa(A_t^i) - \delta A_t^i - \beta\left[\sigma\delta A_t^2 + (1-\sigma)\left(p_{t+1}^f A_{t+1}^f - \kappa(A_{t+1}^f) + \delta\left(A_t^2 - A_{t+1}^f\right)\right)\right]$.

The trade surplus is given by the income generated, $p_t^i A_t^i$, minus the costs $\kappa(A_t^i)$ and the opportunity costs in terms of coupons, $-\delta A_t^i - \beta\delta A_t^2$. Note that for the asset maturing in two periods, the primary dealer takes into account that if he sells the asset today, he not only forgoes the coupon today but also tomorrow. For each type of asset, the primary dealer faces an inventory constraint, $A_t^i \leq I_t^i$. He cannot sell more than what he has. In addition, he cannot sell more than what the seller is willing to buy of each type of asset given by a_t^i. Therefore the following constraint needs to hold: $A_t^i \leq a_t^i$. The primary dealer makes a take-it-or-leave-it offer to the seller subject to his delivery constraints. He sets the price as such that the seller is just as well off with the purchase as without it. In both cases the seller receives the contract price q_{t-1}^i. If he delivers he has to buy the asset at price p_t^i but he can keep his collateral ω_{t-1}^i.

The first order conditions are:

$$\begin{aligned} p_t^2 &= \kappa'(A_t^2) + \delta + \beta\delta + \lambda_t^2 + \widetilde{\lambda}_t^2 + \beta(1-\sigma)\lambda_{t+1}^f \\ p_t^f &= \kappa'(A_t^f) + \delta + \lambda_t^f + \widetilde{\lambda}_t^f \\ p_t^n &= \kappa'(A_t^n) + \delta + \lambda_t^n + \widetilde{\lambda}_t^n. \end{aligned} \tag{4}$$

The prices equal the marginal costs faced when selling the assets and takes into account potentially binding constraints. λ_t^i is the lagrange multiplier of the inventory constraint $A_t^i \leq I_t^i$. $\widetilde{\lambda}_t^i$ is the lagrange multiplier of the demand constraint $A_t^i \leq a_t^i$.

The complementary slackness conditions are:

$$\begin{aligned} \lambda_t^i(I_t^i - A_t^i) &= 0 \qquad \forall\, i \text{ and } t \\ \widetilde{\lambda}_t^i(a_t^i - A_t^i) &= 0 \qquad \forall\, i \text{ and } t. \end{aligned} \tag{5}$$

The inventories of the newly issued assets equal $\mathcal{I}$, i.e. $I_t^2 = I_t^n = \mathcal{I}$. The inventory of off-the-run assets I_t^f can take two values. The primary dealers who had a match the period beforehand have an inventory of $I_t^{f,h} \equiv I_t^f = \mathcal{I}$. The one who did not have a match have an inventory of $I_t^{f,l} \equiv I_t^f = \mathcal{I} - A_{t-1}^2$. The letter h stands for high and the letter l for low, corresponding to the higher and lower inventories, respectively. Given the probability of having no match is σ, by the law of large numbers a share σ of primary dealers has an inventory of $I_t^{f,h}$ and a share $(1-\sigma)$ has an inventory of $I_t^{f,l}$. I denote the lagrange multiplier of the inventory constraint of the high inventory group $\lambda_t^{f,h}$ and that of the low inventory group $\lambda_t^{f,l}$. I denote

the sold off-the-run assets of the high inventory group $A_t^{f,h}$ and those of the other group $A_t^{f,l}$. The other lagrange multipliers are the same for both groups. In the following I assume that $\lambda_t^n = \lambda_t^2 = \lambda_t^{f,h} = 0$. The equilibrium where this holds is our main equilibrium of interest. In section A I discuss other equilibria. In these other equilibria, the main dynamics are the same. They are extreme cases of the main equilibrium.

4.2 Buyer

The buyer and seller anticipate which of the primary dealers' inventory constraints are non-binding and which are potentially binding when trading. They do this because they know the distribution over inventories. As mentioned above, I assume that $\lambda_t^n = \lambda_t^2 = \lambda_t^{f,h} = 0$. This means that primary dealers are unconstrained in selling assets if they still have the full amount of the assets in their inventory. Given this assumption, we have to distinguish two cases: $\lambda_t^{f,l} > 0$ and $\lambda_t^{f,l} = 0$. A fraction $(1-\sigma)$ of primary dealers has already sold part of its stock of assets in the previous period and if $\lambda_t^{f,l} > 0$ they face a demand for assets today that will exhaust the remaining stock. If $\lambda_t^{f,l} = 0$ the demand is lower than the remaining stock.

The buyer's value function at the beginning of the contract market is

$$\begin{aligned} V^b(a_{t-1}^2) = \max_{\{a_t^i\}_i} \sum_i & -\beta q_{t-1}^i a_t^i + \beta(\delta+h)\left(\sum_i a_t^i + a_{t-1}^2\right) \\ & + \beta(1-\sigma)\left[\omega_{t-1}^f - (\delta+h)\right](a_t^f - I_t^{f,l})\mathbb{I}_{\lambda_{t+1}^{f,l}>0} + \beta V^b(a_t^2). \end{aligned} \tag{6}$$

The buyer's state variable is a_{t-1}^2. These are the assets he bought last period and did not mature yet. For each asset, the buyer chooses how many he wants to buy from the seller. For each asset he wants to buy, he must build up the payment $q_{t-1}^i a_t^i$ in the form of settlement goods the next period at settlement. The assets are delivered in the next period at settlement, and the buyer receives utility $(\delta+h)$ from each asset he holds. If we are in the case where $\lambda_t^{f,l} > 0$, then with a probability $(1-\sigma)$ his seller encounters a primary dealer who is constrained in his inventory of off-the-run assets and only $I_t^{f,l}$ instead of a_t^f assets are delivered. For the amount of assets for which there is a settlement failure $(a_t^f - I_t^{f,l})$, he receives the collateral $\omega_{t-1}^f a_t^f$.

The first order conditions are

$$\begin{aligned} q_{t-1}^2 &\geq (1+\beta)(\delta+h) \\ q_{t-1}^n &\geq (\delta+h) \\ q_{t-1}^f &\geq (\delta+h) + (1-\sigma)\left[\omega_{t-1}^f - (\delta+h)\right]\mathbb{I}_{\lambda_{t+1}^{f,l}>0}. \end{aligned} \tag{7}$$

The prices are greater or equal the discounted marginal utilities. The price of the two-period on-the-run asset is twice the price of the one-period on-the-run asset before adjusting for discounting. If we are in the case where settlement fails can occur, i.e. $\lambda_t^{f,l} > 0$, then the price of the off-the-run asset also reflects the risk of a settlement failure.

4.3 Seller

The seller's value function at the beginning of the contract market is

$$\begin{aligned} V^s = \max_{\{a_t^i\}_i} & \sum_i -\omega_{t-1}^i a_t^i + \beta(1+r)\sum_i (q_{t-1}^i - p_t^i + \omega_{t-1}^i)a_t^i \\ & + \beta(1+r)(1-\sigma)(p_t^f - \omega_{t-1}^f)(a_t^f - I_t^{f,l})\mathbb{I}_{\lambda_{t+1}^{f,l}>0} + \beta V^s. \end{aligned} \tag{8}$$

The seller chooses the optimal number of contracts to sell to the buyer. That is, he chooses for each asset the amount he is willing to deliver in the next period. He has to build up collateral $\omega_{t-1}^i a_t^i$ in the form of settlement goods to support a contract. ω_t^i is taken as given.

In the next period, the seller goes to the spot market and buys the assets. For each asset he can deliver, he receives the price q_{t-1}^i from the buyer, he pays the spot market price p_t^i to the primary dealer, and he can keep his accumulated collateral (all in the form of settlement goods).[29] Any remaining funds after the trade, he can deposit in the central bank facility and receive an interest rate r on them.[30]

If we are in the case where $\lambda_t^{f,l} > 0$, then with probability $(1-\sigma)$ he is matched with a primary dealer who is constrained and can only deliver $I_t^{f,l}$ instead of a_t^f. For this amount of non-deliverable assets $(a_t^f - I_t^{f,l})$, the seller's collateral is seized and given to the buyer. The seller does not buy this amount on the spot market and therefore does not have to pay the spot market price.

The first order conditions are

$$\begin{aligned} \omega_{t-1}^2 &\geq \beta(1+r)(q_{t-1}^2 - p_t^2 + \omega_{t-1}^2) \\ \omega_{t-1}^f &\geq \beta(1+r)(q_{t-1}^f - p_t^f + \omega_{t-1}^f) + \beta(1+r)(1-\sigma)(p_t^f - \omega_{t-1}^f)\mathbb{I}_{\lambda_{t+1}^{f,l}>0} \\ \omega_{t-1}^n &\geq \beta(1+r)(q_{t-1}^n - p_t^n + \omega_{t-1}^n). \end{aligned} \tag{9}$$

[29] The seller delivers an asset if $p_t^i \leq \omega_{t-1}^i$, which is the case in equilibrium for all i and t. This means that the value of the collateral seized in case of non-delivery must be as high as the value of the assets he buys on the spot market. Since this constraint is always satisfied (see section 4.1), it is not added to the maximization problem.

[30] In equilibrium, the funds are not negative.

The collateral value that has to be built up today is greater or equal to the contract price he receives tomorrow and the value of the collateral he can keep minus the spot price he has to pay to acquire the asset. If $\lambda_t^{f,l} > 0$ non-delivery occurs with probability $(1-\sigma)$ and in this case he does not have to pay the spot price but he cannot keep the collateral.

4.4 Equilibrium and premium definition

Before I define the equilibrium I make two assumptions.

If the buyer and the seller are indifferent to buying more or less assets (after accounting for the probability of a settlement failure), I assume that they are willing to buy the maximum amount of assets that is profitable and that the primary dealer can sell. This implies that in each equilibrium

$$\begin{aligned} a_t^2 &= A_t^2 \\ a_t^n &= A_t^n \\ a_t^f &= A_t^{f,h}. \end{aligned} \tag{10}$$

Note that in any equilibrium, as soon as an inventory constraint starts to bind, so does the corresponding constraint on the contracts. Also, if one is slack, the other is slack. The only exception is $\widetilde{\lambda}_t^{f,l}$, which can be zero even if $\lambda_t^{f,l} > 0$, but not vice versa. Therefore, as we concentrate on equilibria where $\lambda_t^n = \lambda_t^2 = \lambda_t^{f,h} = 0$, then also $\widetilde{\lambda}_t^n = \widetilde{\lambda}_t^2 = \widetilde{\lambda}_t^{f,h} = 0$.

I assume in the following that due to perfect competition and market regulation the contract price and the collateral values adjust in equilibrium such that the first order conditions of the buyer and the seller hold with equality. This means that there is a non-zero finite amount of contracts sold in all assets, $a_t^i \in (0, \infty)\ \forall i$.

Definition 4.1 (Equilibrium)**.** An equilibrium consists of the contract prices of all assets ($q_{t-1}^i\ \forall i$), the spot prices of all assets ($p_t^i\ \forall i$), the amount of assets contracted by sellers and buyers ($a_t^i\ \forall i$) and sold by primary dealers ($A_t^2, A_t^n, A_t^{f,h}$, and $A_t^{f,l}$), the collateral values ($w_{t-1}^i\ \forall i$), the shadow values (λ_t^2, $\widetilde{\lambda}_t^2$, λ_t^n, $\widetilde{\lambda}_t^n$, $\lambda_t^{h,l}$, $\widetilde{\lambda}_t^{h,l}$, $\lambda_t^{f,l}$, $\widetilde{\lambda}_t^{f,l}$), and satisfies with equality the first-order conditions of the primary dealer (4), the buyer (7), and the seller (9), the delivery constraints (2), and the market clearing conditions (10) with respect to all assets in all periods.

Next, I define the on-the-run premium.

Definition 4.2 (On-the-run premium)**.** The on-the-run premium is defined as $\Delta_t \equiv p_t^n - p_t^f$.

As mentioned above, the on-the-run and off-the-run assets have the same cash flow to maturity and mature

on the same day. The only difference is their issuance date. To measure the on-the-run premium we compare these two assets. A positive (negative) premium implies that the yield to maturity of the on-the-run asset is lower (higher) than that of the off-the-run asset.

5 Existence

This section focuses on the case where $\lambda_t^2 = \lambda_t^n = \lambda_t^{f,h} = 0$. This means that primary dealers are unconstrained if they still have the full stock of assets available, i.e. $\mathcal{I} > max(A_t^2, A_t^n, A_t^{f,h})$.

I argue that in an equilibrium where there is a non-zero premium, it must be the case that $\lambda_t^{f,l} > 0$ and $\widetilde{\lambda}_t^{f,l} = 0$.[31]

Therefore in this equilibrium $\mathcal{I} \in (max(A_t^2, A_t^n, A_t^{f,h}), (A_{t-1}^2 + A_t^{f,h}))$.[32] This means that primary dealers who sold some of their inventory in the previous period and can sell again today, are constrained. All other primary dealers are not constrained.

Proposition 5.1. In an equilibrium where $\mathcal{I} > max(A_t^2, A_t^n, A_t^{f,h})$, a necessary condition for the on-the-run premium to be non-zero is $\mathcal{I} < (A_{t-1}^2 + A_t^{f,h})$.

Suppose it is not the case that $\lambda_t^{f,l} > 0$. Then $\lambda_t^{f,l} = \widetilde{\lambda}_t^{f,l} = 0$.[33] From the first-order conditions of the seller and the buyer and $\omega_t^f = p_t^f$, it follows that $p_t^n = p_t^f = \beta^2(1+r)(\delta+h)$. Since both assets are found with certainty, there is no difference in their prices. Therefore, there is no premium, $\Delta = 0$. Without settlement fails, both assets are priced according to their marginal utility to the buyer, regardless of how abundant or scarce they are in the market. The marginal utility is the same for both assets, on- and off-the-run. I use a linear utility function, but this is true for any utility function where the assets are perfect substitutes, i.e., only the sum of the two assets matters. Given the possession of an asset, there is no reason why they should not be substitutes, given that they have the same coupons and are held to maturity.

Therefore, our equilibrium candidate is the equilibrium where $\lambda_t^{f,l} > 0$. From the equilibrium conditions it follows that

[31] The case where $\lambda_{t+1}^{f,l} > 0$, $\widetilde{\lambda}_{t+1}^{f,l} > 0$, and $\lambda_{t+1}^{f,h} = 0$ is not possible.

[32] Note that we could also add the knife-edge case where $\mathcal{I} = max(A^2, A^n, A^{f,h})$ and $\lambda_t^2 = \lambda_t^n = \lambda_t^{f,h} = 0$. To make the notation easier, we omit it.

[33] The case $\lambda_t^{f,l} = 0$ and $\widetilde{\lambda}_t^{f,l} > 0$ is not possible.

$$p_t^n = \beta(1+r)(\delta+h)$$
$$p_t^f = \frac{\sigma}{1-(1-\sigma)\beta(1+r)}\beta(1+r)(\delta+h)$$
$$p_t^2 = (1+\beta)\beta(1+r)(\delta+h).$$

This implies a positive premium, i.e. $\Delta = \left[1 - \frac{\sigma}{1-(1-\sigma)\beta(1+r)}\right]\beta(1+r)(\delta+h) > 0$. Intution for the prices and the premium is given in section 6. Next, I state a necessary and sufficient condition for the existence of this equilibrium.

Proposition 5.2. A necessary and sufficient condition for the existence of the equilibrium with $\mathcal{I} \in (max(A_t^2, A_t^n,\ A_t^{f,h}), (A_{t-1}^2 + A_t^{f,h}))$ is $\sigma > \frac{1-\beta(1+r)}{\beta(1+r)}\frac{\delta}{h}$.

The proof can be found in the appendix section B.1. For positive demand in off-the-run assets, the probability of a settlement fail, $(1-\sigma)$, cannot be too high. The function $\kappa(\cdot)$ can be interpreted as nonlinear portfolio adjustment costs or as regulatory costs.
Since the two conditions of the proposition are satisfied, I show in the proof that I can always find an $\mathcal{I}$ where our equilibrium of interest exists, i.e. $\mathcal{I} \in (max(A_t^2, A_t^n,\ A_t^{f,h}), (A_{t-1}^2 + A_t^{f,h}))$. I call this equilibrium from now on "premium equilibrium".

Definition 5.1 (Premium equilibrium)**.** The premium equilibrium is the equilibrium where $\lambda_t^{f,l} > 0$ and all other lagrange multipliers are zero.

I will restrict any further analysis on the premium equilibrium. In the appendix A I discuss other equilibria. The dynamics and intuition are the same as in the premium equilibrium.

6 Premium Equilibrium

6.1 Dynamics

To discuss the relevant dynamics in the premium equilibrium, I illustrate the life cycle of two two-period assets issued in period t in the figure 6 below.
The example shows the main dynamics of the model. In period $t-1$ the buyer buys a contract from the seller. In period t the assets are issued and on-the-run. Every primary dealer receives one asset in his inventory. I restrict here the inventory to one asset for illustrative purposes. On the spot market the seller is matched with one of the two primary dealers, buys the asset and delivers it to the buyer against a payment

in the form of settlement good (not illustrated).[34] The other primary dealer was not matched with a seller and keeps his asset in his inventory. After the spot market, the contract market takes place in the second subperiod. The buyer buys again one contract. In period $t+1$ the assets are off-the-run. In my model the assets are always off-the-run after one period because new assets are issued (not illustrated). The seller is matched on the market with the primary dealer who was able to sell his asset already the period beforehand. A settlement fail occurs. Nevertheless it is optimal for the buyer to initially buy one contract. He takes the probability of a settlement fail into account when taking his decision.

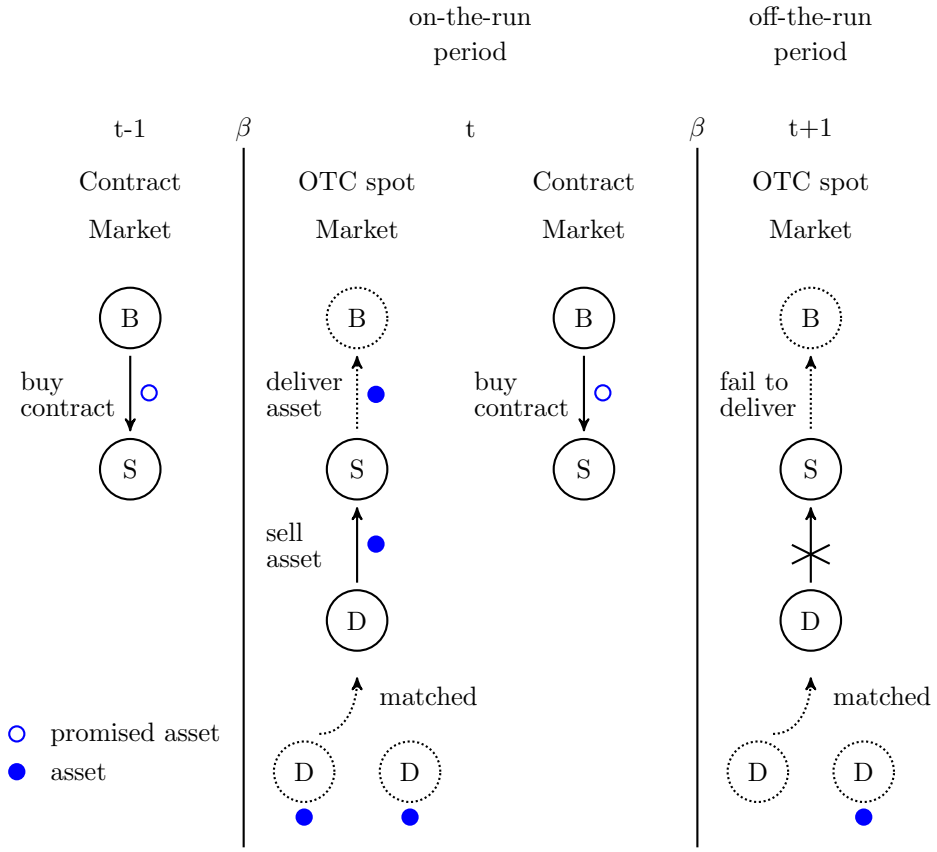


Figure 6: Dynamics

The figure shows only one kind of asset, the two-period asset. In the model there is also the one-period on-the-run asset. It's easy to see that settlement fails occur for off-the-run assets, but not for the on-the-run assets. Inventories do not differ for on-the-run assets because they are less long in the market. There is a premium for on-the-run assets because they do not fail to settle compared to off-the-run assets where the probability of failure is priced in. It is therefore the only characteristic in which the two assets differ that gives in combination with OTC market frictions and delivery constraints rise to the premium: time since

[34] The asset stays in the portfolio of the buyer until it matures in $t+1$.

issuance.[35]

Using the simple example above again, another important observation can be made. The amount of assets available in the market becomes scarcer over time. In the off-the-run period, some of the assets are already locked up in the buyer's portfolio, while in the on-the-run period, the full amount of issued assets is available. This observation is consistent with what has been suggested as a possible reason for the emergence of a positive premium: scarcity due to the lock-up in buy-and-hold portfolios. However, contrary to popular belief, the results of our model suggest that it is not scarcity per se that leads to the premium, but uncertainty about the inventory of the primary dealer with whom the seller is matched. Without uncertainty, there would be no premium even if the asset is scarcer. To understand this fact, suppose an asset is scarce but all primary dealer inventories are equal. Since the buyer knows the distribution of assets, he knows this. Settlement fails never occur as the buyer never orders more assets than are maximally available in any of the inventories. Given there are no settlement fails, there is no premium.

6.2 Equilibrium prices, quantities, and premium

In the premium equilibrium, the following equations for prices and quantities hold:

$$
\begin{aligned}
p_t^n &= \beta(1+r)(\delta+h) \\
p_t^f &= \frac{\sigma}{1-(1-\sigma)\beta(1+r)}\beta(1+r)\beta(\delta+h) \\
p_t^2 &= (1+\beta)\beta(1+r)(\delta+h)
\end{aligned}
$$

and

$$
\begin{aligned}
A_t^n &= \kappa'^{-1}\left[\beta(1+r)(\delta+h)-\delta\right] \\
A_t^{f,h} &= \kappa'^{-1}\left[\frac{\sigma}{1-(1-\sigma)(1+r)\beta}\beta(1+r)\beta(\delta+h)-\delta\right] \\
A_t^{f,l} &= \mathcal{I} - A_t^2 \\
\kappa'(A_t^2) &= W + \beta(1-\sigma)\kappa'(\mathcal{I}-A_t^2)
\end{aligned}
$$

[35] Without settlement fails, both assets would be priced only according to the marginal utility that the coupons give to the buyer. This is true not only for any linear utility function like the one used here, but also for any non-additively separable non-linear function. Once a buyer has obtained the assets, there is no reason why the on- and off-the-run assets should not be substitutes, given that they have the same coupons and are held to maturity.

where $W \equiv \left[(1+\beta) - \beta(1-\sigma)\frac{\sigma}{1-(1-\sigma)\beta(1+r)}\right]\beta(1+r)(\delta+h) - (1+\beta\sigma)\delta$.

Spot prices are determined by the marginal utility of the buyer. The on-the-run price of the two-period asset is $(1+\beta)$ times the on-the-run price of the one-period asset, because the buyer receives twice the utility from it. All of my results would also hold if I would compare the two-period off-the-run asset to the two-period on-the-run asset and abstract from the cash flow in the second period to make them equal in terms of cash flow.

The quantity of the two-period asset maturing in two periods and the off-the-run asset traded by constrained dealers depends on the issue size $\mathcal{I}$. The reason for the first quantity is that when these assets are sold, it is taken into account that less can be sold tomorrow due to the binding inventory constraint. This binding inventory constraint is then also the reason why the amount of off-the-run assets traded by constrained primary dealers depends on $\mathcal{I}$.

Proposition 6.1. The on-the-run premium is given by: $\Delta = \left[1 - \frac{\sigma}{1-(1-\sigma)\beta(1+r)}\right]\beta(1+r)(\delta+h) > 0$. If $\sigma \to 1$, then $\Delta \to 0$.

The on-the-run premium depends on the asset valuation of the buyer $(\delta + h)$, the probability to find the off-the-run assets σ, and $\beta(1+r)$ which is the discount factor cost to the seller for binding collateral. If assets would be found with certainty in the second period, i.e. $\sigma \to 1$, then the premium vanishes.[36]

In equilibrium not only $p_t^n > p_t^f$ but also $A_t^n > \sigma A_t^{f,h} + (1-\sigma)A_t^{f,l}$, i.e. on-the-run assets not only have a higher price but are also traded in larger quantities. Both equilibrium results can be explained by the fact that off-the-run assets are less attractive because they fail to settle more often. It is contrary to common intuition that a scarcer asset has the lower price, but this observation is consistent with what is observed in the market (U.S. Department of the Treasury (2022)).

To summarize, our equilibrium is consistent with the three stylized facts and irregularities described in the introduction. First, on-the-run assets are more expensive than off-the-run assets. Second, they trade in larger volumes. And third, settlement fails occur and on-the-run assets fail to settle less often.

Our discussion leads us to state our first summary of results:

Summary of results I *The on-the-run premium is due to differences in inventories of off-the-run assets as they are longer in the market. The reason is as follows: Some of the off-the-run assets are locked up in buy-and-hold portfolios because they have already been sold during their on-the-run period. Since not all primary dealers faced the same demand during the on-the-run period due to the OTC market structure, there*

[36] The premium also vanishes if $\sigma = 0$ but then we are not anymore in the premium equilibrium and the formula above does not hold.

are differences in their inventories at the start of the off-the-run period. This implies uncertainty about the amount of assets available in an upcoming match with them in the OTC market. This leads to a higher frequency of settlement fails for off-the-run assets, as contracts promising their delivery cannot always be fulfilled. There is a preference for on-the-run assets because their settlement is not risky. Compared to off-the-run assets they are safe in this aspect. This implies that they are traded in larger quantities.

6.3 Dependence of premium, fails, and inventories in the data

Next, I look at the net outright positions in Treasury bonds to analyse if my theory is consistent with the emprical evidence.[37] We can see from figure 7 that the position has been mostly positive and rising over the past years. Net outright positions are a good indicator of the primary dealer inventory available to traders.

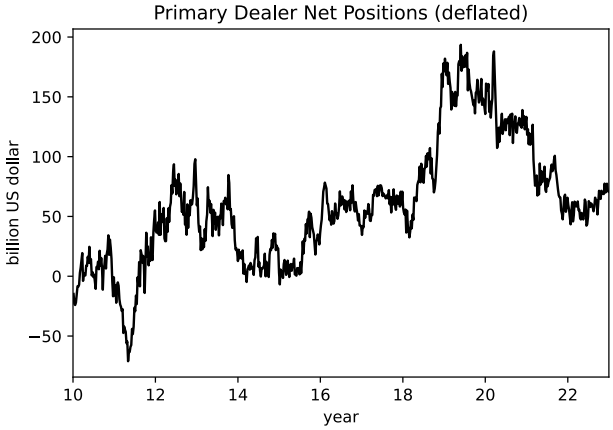


Figure 7: The primary dealer net outright positions

For illustrative purposes, I show the scatter plot between the 10 year on-the-run premium and the net outright positions of primary dealers in Treasury bonds in figure 8.[38] We observe that there is a negative correlation between the 10 year on-the-run premium and the net outright positions of the primary dealers with tight 95% confidence bands. This is consistent with my theory, as I predict that frequent settlement fails, implying a high premium, occur when primary dealer inventories are low.

[37]The data can be downloaded from the FED's Primary Dealer Statistics, https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics. The frequency is weekly. The data are deflated using the "Consumer Price Index for All Urban Consumers: All Items in U.S. City Average", which can be downloaded from FRED. I set the index to 1 when the series starts in 2010.

[38]The data sources are the same as in figures 2a and 7. The net position of primary dealers is deflated as in figure 7. The time horizon is 2010-2022 and the frequency weekly.

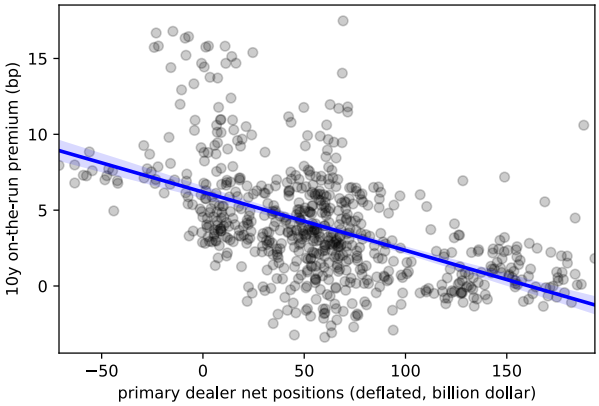


Figure 8: The correlation between the premium and the net positions

To systematically test my theory, I regress on-the-run premia data on the failure rate of off-the-run assets in a first step (see table 1), and the failure rate of off-the-run assets on the net outright positions of primary dealers in a second step (see table 2). If my theory is correct, I should observe a positive coefficient in the first regression and a negative one in the second. In the regressions, I control for the logarithm of the VIX, the 10 year - 2 year yield spread, and the general collateral financing repo rate.

For the first regression (see table 1), I use data on the 1, 2, 3, 5, 7, 10, 20, and 30 year on-the-run premia. Each maturity and day in the sample is a separate observation. The failure rate of off-the-run Treasuries is not available for different maturities. The independent variables are therefore the same for each on-the-run premium maturity.

The coefficient of interest is the failure rate. As expected, the coefficient is positive and is significant at the 1% level. If the failure rate increases by 1 percentage point, the premium increases by 2.4 basis points, which is about half a standard deviation.
The interpretation of the other variables we control for is as follows: A one percent increase in the VIX increases the premium by 1.2 basis points. A one percentage point increase in the 10 year - 2 year yield spread increases the premium by 0.3 basis points. Finally, a one percentage point increase in the general collateral financing repo rate has little impact on the premium. Looking at all other control variables next to the failure rate, only the impact of the VIX is significant.

Next, I regress the failure rate of off-the-run Treasuries on the deflated net outright positions in Treasury bonds of different maturity baskets and the other variables I already controlled for in the first regression (see table 2). Specifically, I have the following baskets for net outright positions in Treasury bonds: below or equal 3 years, above 3 years to 6 years, above 6 years to 11 years, above 11 years. Each basket and day

	On-the-run premium
Failure rate off-the-run Treasuries	2.4206***
	(0.563)
lnVIX	1.1706**
	(0.570)
10 year - 2 year yield spread	0.2605
	(0.364)
General collateral financing repo rate	0.0128
	(0.314)
Constant	-2.5491
	(2.054)
No. Observations:	3760
R-squared:	0.025
Adj. R-squared:	0.024

I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. ** indicates significance at the 5% level. The sample period is April 2013-2022 and the frequency is weekly.
The data source on the premia (in basis points) is the same as in figure 2a. The failure rate data (in percent) is the same as in figure 3a. The VIX data are taken from FRED. The 10 year and 2 year yields (par yields of seasoned bonds in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al. (2010)), https://www.federalreserve.gov/data/tips-yield-curve-and-inflation-compensation.htm. The general collateral financing repo rate (in percent) is provided by the Depository Trust and Clearing Corporation (DTCC) and can be downloaded here: https://www.dtcc.com/charts/dtcc-gcf-repo-index.

Table 1: Premium on off-the-run failure rate regression

is a separate observation.

As expected, the coefficient on deflated primary dealers' net positions is negative and significant at the 1% level.[39] As net positions are denominated in dollars, the coefficient is best interpreted by first multiplying it by the standard deviation of net outright positions. It follows that an increase in net positions by one standard deviation reduces the failure rate by 6 basis points, equivalent to a quarter of a standard deviation.

[39]The result still holds if the first difference in the primary dealers' net positions is taken.

	Failure rate off-the-run Treasuries
Primary dealer net positions	-0.0031***
	(0.000)
lnVIX	-0.0769***
	(0.033)
10 year - 2 year yield spread	-0.1559***
	(0.019)
General collateral financing repo rate	-0.0256*
	(0.014)
Constant	1.1508***
	(0.111)
No. Observations:	1880
R-squared:	0.132
Adj. R-squared:	0.130

I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. ** indicates significance at the 5% level. * indicates significance at the 10% level. The sample period is April 2013-2022 and the frequency is weekly.

The data source for the net outright positions (in billion US dollars) is the same as in figure 8. The data are also deflated in the same way (with the index set to 1 when the series starts in April 2013). The failure rate data (in percent) is the same as in figure 3a. The VIX data are taken from FRED. The 10 year and 2 year yields (par yields of seasoned bonds in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al. (2010)), https://www.federalreserve.gov/data/tips-yield-curve-and-inflation-compensation.htm. The general collateral financing repo rate (in percent) is provided by the Depository Trust and Clearing Corporation (DTCC) and can be downloaded here: https://www.dtcc.com/charts/dtcc-gcf-repo-index.

Table 2: Failure rate on net positions regression

The interpretation of the other variables we control for is as follows: A one percent increase in the VIX reduces the failure rate by 8 basis points. A one percentage point increase in the 10 year - 2 year yield spread reduces the failure rate by 16 basis points. Finally, a one percentage point increase in the general collateral financing repo rate reduces the failure rate by 3 basis points. All variables are significant.

In a last step, we find evidence in the data consistent with our theory that off-the-run inventories are much

more affected by search and matching frictions than on-the run inventories.

On-the-run	Maturity	2y	3y	5y	7y	10y	30y
	Volatility	4.1	4.0	3.5	2.7	4.1	2.3
All	Maturity	≤3y	(3y,6y]	(6y,7y]	(7y,11y]	≥11y	
	Volatility	21.2	11.7	7.6	7.0	12.3	

Table 3: Primary dealer net positions volatilities

I analyse the volatility of deflated primary dealer net positions in Treasury bonds (see table 3).[40] In addition to the primary dealers' net position in all Treasuries (on- and off-the-run), I have separately the position in on-the-run Treasuries. We can see from the table that the volatility of the latter is much lower. This is consistent with our theory if we assume that there is a slight variation in σ over time, which is most likely the case in reality. The effect of the search and matching frictions on off-the-run inventories can then be observed. They are volatile, while on-the-run inventories are much more stable.

7 Central bank facility access

What would be the impact on the Treasury market of providing broad access to central bank facilities? How do prices, premia, traded quantities, fails, and profits change? This section provides answers to these questions.

In my model, the central bank facility is comparable to a deposit facility or a reverse repo facility, where I abstract from the collateral provided. The collateral part would not make the existing dynamics disappear. First, the repos from the facility are general collateral repos. For such repos the cash lender is willing to accept any collateral that falls into a broad class, and is not looking for any particular collateral (Bowman et al. (2017)). Second, rehypothecation is not allowed with the collateral provided by the FED's reverse repo facility. Third, even if the facility were to provide a specific Treasury sought, the Treasury would have to be returned the next day and would only temporarily ease availability.

[40] The data can be downloaded from the FED's Primary Dealer Statistics, https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics. The data are deflated using the "Consumer Price Index for All Urban Consumers: All Items in U.S. City Average", which can be downloaded from FRED. I set the index to 1 when the on-the-run net positions time series start in April 2013. I use data from April 2013 to the end of 2022. The frequency is weekly. To calculate the volatilities, I use the average of the data over the full time horizon.

7.1 Homogenous sellers

I first analyse the situation where all sellers are equal. They gain or already have access to the central bank facility. The facility can be accessed each first subperiod for one subperiod. Sellers deposit settlement good and receive an interest rate r on it. The seller represents any type of financial institution (e.g. a hedge fund or a non-primary dealer). The goal of the analysis is to find, in a general setting, the effect of the reverse repo or deposit facility rate on trading when any type of financial firm other than a primary dealer is given access to the facility.

I assume an unanticipated increase in the facility rate r at the beginning of the contract market. Next to the situation where all sellers already had access and the facility rate increases, the increase can also represent the situation where all sellers gain access to the facility and due to the facility they now face a higher interest rate than beforehand (with no interest or a lower market rate).

An increase in r initially increases the profitability of the trade for the seller. The facility is always available, and excess liquidity can be deposited until the seller enters the next trade. The facility is not a substitute for trading, but a complement to it. Increased profitability leads to a positive supply shock in the contract market. This in turn implies a positive demand shock in the spot market. In equilibrium, spot prices rise, i.e. $\frac{dp_t^i}{dr} > 0 \, \forall i$. The off-the-run price reacts more strongly than its on-the-run equivalent, i.e. $\frac{dp_t^f}{dr} > \frac{dp_t^n}{dr}$. This implies that the premium decreases, $\frac{d\Delta}{dr} < 0$. Compared to the on-the-run asset, we need to consider the impact on the case of non-delivery for the off-the-run asset. The case of non-delivery is less unprofitable than before because the collateral values must increase due to the increase in prices. This reduces the spread between the real valuation of holding the asset and the collateral. Therefore, the first demand shock from the seller triggers an increase in demand from the buyer. This implies that the price of the off-the-run asset reacts more strongly.

The quantity of on-the-run and off-the-run assets traded by unconstrained dealers increases, i.e. $\frac{dA_t^n}{dr} > 0$ and $\frac{dA_t^{f,h}}{dr} > 0$. The latter is stimulated more, i.e. $\frac{dA_t^{f,h}}{dr} > \frac{dA_t^n}{dr}$. The same reasoning applies as for the effect on the on-the-run premium. The amount of off-the-run assets traded, $\sigma A_t^{f,h} + (1-\sigma) A_t^{f,l}$, increases in the first period after the rate hike because $A_t^{f,l} = \mathcal{I} - A_{t-1}^2$ is determined from the period before and does not change. This makes the policy particularly interesting in the context of the Treasury market crisis during the pandemic, where the market for off-the-run assets froze (Eren and Wooldridge (2021)). However, the settlement failure rate of off-the-run, assets also initially increases as more assets are traded. It is defined as the value of the assets involved in a failure divided by the total amount of assets promised to be delivered, $f_{t+1}^f \equiv \frac{(1-\sigma) p_{t+1}^f (a_{t+1}^f - I_{t+1}^{f,l})}{p_{t+1}^f a_{t+1}^f}$. When search frictions are above a small minimum value[41] A_t^2 increases, i.e. $\frac{dA_t^2}{dr} > 0$. Thus, in the second period after the rate hike, $A_{t+1}^{f,l}$ decreases given the increase in A_t^2. Whether or not the total amount of off-the-run assets traded increases is ambiguous. The settlement failure rate

[41] See Appendix B.2.

increases even more as A_t^2 increases.
Given the increase in A_t^2, A_t^n, and $A_t^{f,h}$ from the first period after the rate hike onward, more total assets end up in the portfolio of the buyer, the agent with the highest marginal asset valuation.

I summarize the above results as follows:

Summary of results II *If sellers gain access, then an increase in the facility rate r increases all spot prices and decreases the on-the-run premium. The quantities of on-the-run and off-the-run assets traded by unconstrained dealers increase. Trade in the off-the-run asset is (initially) more stimulated than trade in the on-the-run asset, and the settlement failure rate of the off-the-run asset (initially) increases. If search frictions are above a small minimum value, then overall more assets are offloaded from the inventories and end up in the portfolio of the holder with the highest marginal asset valuation.*

Putting all of the above results in a broader context, I can conclude that any kind of policy that lowers the costs of trade and intermediation can trigger the effects described. Access is one possibility. See appendix section B.2 for details on the derivatives discussed in this paragraph.

The next result is about who benefits from an increase in the facility rate r. The profits are denoted by π^i.

In equilibrium, $\frac{d\pi^s}{dr} = 0$, $\frac{d\pi^D}{dr} > 0$, and $\frac{d\pi^b}{dr} < 0$ (if search frictions are above a small minimum value).[42] See appendix section B.3 for how the derivatives are derived.
Competition among sellers in the contract market erodes any positive profits for them to zero. Initially increased profitability is offset by higher spot prices in equilibrium.
Primary dealers benefit from the policy by providing the assets that are in higher demand. They sell more assets, they sell them earlier, and they sell them at a higher price. In equilibrium, primary dealers sell assets until their marginal nonlinear cost equals the price. Therefore, each asset sold yields a small positive marginal surplus until the last asset is sold, where the marginal surplus equals zero. Since prices are higher in the new equilibrium, breakeven is reached at a higher quantity of assets. Primary dealer profits increase.
Buyers' utility falls. To gain intuition, I first explain how buyers' utility or profits materialize in equilibrium. Since the price of the off-the-run asset reflects the utility of the last unit bought, it incorporates the probability that a settlement fail occurs. But the first part of the bought assets is found with certainty, the part I_t^f. The value the buyer places on these assets is therefore higher. Nevertheless, in equilibrium he pays the same price for all the assets and therefore he has a small profit.
An increase in the facility rate r increases the off-the-run price and decreases the quantity of off-the-run assets

[42] For a small range of small σ, increasing r leads to an increase in p_t^2 but a decrease in Q_t^2. See Appendix B.2. The increase in p_t^2 is then solely due to the increase in p_{t+1}^f, the opportunity cost of selling an asset early. As A_t^2 decreases, the effects are no longer straightforward.

found with certainty (as the quantity of on-the-run assets traded with a two-period maturity increases). Both effects lead to a lower profit. In summary, buyers' profits decrease as the spread between the value of an uncertain unit and a certain unit decreases and there are fewer certain units. The decrease in profits is an externality problem. The buyer does not consider how the purchase of two-period assets affects the availability and price of the same assets in the next period. This is in contrast to the primary dealer, who manages his inventory and takes into account that a two-period asset sold today cannot be sold in the next period.

We solved our baseline model under the assumption that the primary dealer has full bargaining power. We can relax this assumption and prove that even with positive bargaining power of the seller, in equilibrium the seller does not make a profit in contrast to the primary dealer. When maximizing the joint surplus of both agents, the first-order conditions of the spot market problem change to:

$$
\begin{aligned}
p_t^2 &= \kappa'(A_t^2) + \delta + \beta\delta + (\omega_{t-1}^2 - p_t^2) + \lambda_t^2 + \widetilde{\lambda}_t^2 + \beta(1-\sigma)\lambda_{t+1}^f \\
p_t^f &= \kappa'(A_t^f) + \delta + (\omega_{t-1}^f - p_t^f) + \lambda_t^f + \widetilde{\lambda}_t^f \\
p_t^n &= \kappa'(A_t^n) + \delta + (\omega_{t-1}^n - p_t^n) + \lambda_t^n + \widetilde{\lambda}_t^n.
\end{aligned}
$$

The surplus is divided according to their bargaining power: $\left(\omega_{t-1}^i - p_t^i\right) A_t^i = \frac{(1-\theta)}{\theta} S_t^i$ where S_t^i is the surplus of the primary dealer when selling assets i. The first-order conditions of the seller and the buyer when making their decision about the optimal number of contracts to sell and buy do not change. This is crucial because we see from the seller's first-order conditions and his value function that he makes no profit in equilibrium. Even if in equilibrium $\omega_{t-1}^2 - p_t^2 > 0$ (due to the new pricing scheme) and the seller receives part of the positive total surplus of the OTC trade, the collateral value ω_{t-1}^i in the Walrasian contract market adjusts in such a way that he has no total profits. Otherwise, the seller would supply an infinite number of contracts or no contracts at all. Therefore, even with positive bargaining power, the seller never profits from an increase in the facility rate r if he has access.

The above discussion leads to the following summary of results:

Summary of results III *If sellers gain access, they paradoxically do not benefit from an increase in the facility rate r. In equilibrium (if search frictions are above a small minimum value) only primary dealers profit from it, even with positive bargaining power of sellers, and buyers' utility even decreases.*

7.2 Heterogenous sellers

In this second subsection I extend my model. I assume that there is a measure ξ of sellers which have access to the facility and a measure $(1-\xi)$ which does not have access. Sellers which have access are denoted by a and the ones which don't by na. The facility rate is given by r and the market rate by r^m. We assume that $r > r^m$. Agents with no access face the market rate, while the others will use the facility.

I first analyse the prices and quantities of the newly issued assets. Analogously to section 6, the contract and spot prices are given by

$$\begin{aligned} q_t^n &= (\delta + h) \\ p_t^{n,na} &= \beta(1+r^m)(\delta+h) \\ p_t^{n,a} &= \beta(1+r)(\delta+h) \end{aligned}$$

and

$$\begin{aligned} q_t^2 &= (1+\beta)(\delta+h) \\ p_t^{2,na} &= (1+\beta)\beta(1+r^m)(\delta+h) \\ p_t^{2,a} &= (1+\beta)\beta(1+r)(\delta+h). \end{aligned}$$

The traded quantities are determined by the following equations:

$$\begin{aligned} A_t^{n,na} &= \kappa'^{-1}\left[\beta(1+r^m)(\delta+h) - \delta\right] \\ A_t^{n,a} &= \kappa'^{-1}\left[\beta(1+r)(\delta+h) - \delta\right] \end{aligned}$$

and

$$\begin{aligned} \kappa'(A_t^{2,na}) &= W^{na} + \beta(1-\sigma)\kappa'(\mathcal{I} - A_t^{2,na}) \\ \kappa'(A_t^{2,a}) &= W^{a} + \beta(1-\sigma)\kappa'(\mathcal{I} - A_t^{2,a}) \end{aligned}$$

where
$W^{na} \equiv p_t^{2,na} - \beta(1-\sigma)\left[\xi p_t^{f,a} + (1-\xi)p_t^{f,na}\right] - (1+\beta\sigma)\delta$ and

$$W^a \equiv p_t^{2,a} - \beta(1-\sigma)\left[\xi p_t^{f,a} + (1-\xi)p_t^{f,na}\right] - (1+\beta\sigma)\delta.$$

Sellers without access sell lower quantities and have lower spot prices of newly issued assets, i.e. $A_t^{2,na} < A_t^{2,a}$, $A_t^{n,na} < A_t^{n,a}$, $p_t^{2,a} < p_t^{2,na}$, and $p_t^{n,na} < p_t^{n,a}$.

The off-the-run prices and quantities are less straightforward. We now do not only need to distinguish if a primary dealer has met a seller the previous period or not but also if the match was with a seller with or without access. Depending on the type, more or less assets were sold and therefore the inventory differs. A mass σ of agents has a high inventory of $\mathcal{I}$ because they did not face any demand the period beforehand. A mass $(1-\sigma)\xi$ of primary dealers has sold assets to a seller with access in the previous period and they have an inventory of $(\mathcal{I} - A_{t-1}^{2,a})$. Lastly, a mass $(1-\sigma)(1-\xi)$ of sellers have an inventory of $(\mathcal{I} - A_{t-1}^{2,na})$ as they met a seller without access the previous period.

Let us denote by $\mathbb{P}^a$ the probability of a seller with access finding the off-the-run assets. Analogously, we define $\mathbb{P}^{na}$ as the probability for the sellers without access.

Analogously to section 6, the off-the-run prices are given by

$$\begin{aligned} p_t^{f,a} &= \frac{\mathbb{P}^a}{1-(1-\mathbb{P}^a)\beta(1+r)}\beta(1+r)(\delta+h) \\ p_t^{f,na} &= \frac{\mathbb{P}^{na}}{1-(1-\mathbb{P}^{na})\beta(1+r^m)}\beta(1+r^m)(\delta+h). \end{aligned}$$

As in section 5 I assume that the primary dealers with a full inventory are unconstrained. Therefore

$$\begin{aligned} A_t^{f,a,h} &= \kappa'^{-1}\left[\frac{\mathbb{P}^a}{1-(1-\mathbb{P}^a)\beta(1+r)\beta}\beta(1+r)(\delta+h) - \delta\right] \\ A_t^{f,na,h} &= \kappa'^{-1}\left[\frac{\mathbb{P}^{na}}{1-(1-\mathbb{P}^{na})\beta(1+r^m)\beta}\beta(1+r^m)(\delta+h) - \delta\right]. \end{aligned}$$

We define the on-the-run premium by $\Delta \equiv [\xi p_t^{n,a} + (1-\xi)p_t^{n,na}] - [\xi p_t^{f,a} + (1-\xi)p_t^{f,na}]$.

For a positive premium either $\mathbb{P}^a$ or $\mathbb{P}^{na}$ or both must be below 1. If they are below one, they either equal $\sigma\xi$ or $\sigma[\xi + (1-\xi)] = \sigma$ given that $(\mathcal{I} - A_{t-1}^{2,a}) > (\mathcal{I} - A_{t-1}^{2,na})$.

In every equilibrium it must be that if $\mathbb{P}^a > \mathbb{P}^{na}$ then $A_t^{f,a,h} < A_t^{f,na,h}$ and if $\mathbb{P}^{na} > \mathbb{P}^a$ then $A_t^{f,na,h} < A_t^{f,a,h}$. A higher probability to find the assets must go hand in hand with a lower quantity searched for. From our equilibrium equations it follows that the case $\mathbb{P}^a > \mathbb{P}^{na}$ is not possible given that $r > r^m$. Therefore in equilibrium it must be that $\mathbb{P}^{na} \geq \mathbb{P}^a$ which implies that $A_t^{f,a} > A_t^{f,na}$. Given that $A_t^{f,a} > A_t^{f,na}$ it follows

from our equilibrium equations that $p_t^{f,a} > p_t^{f,na}$.

As a next step I analyse the impact of giving all sellers access. The interest rate for sellers gaining access increases from r^m to r. Applying the analogous reasoning as in section 7.1, we immediately observe that the spot prices and quantities of the on-the-run assets traded by sellers who gain access increase. In the first period, this also holds for the off-the-run prices and quantities sold by unconstrained dealers to sellers who gain access. The reaction of the prices and quantities of the off-the-run assets is stronger. The on-the-run premium decreases.

Compared to section 7.1, we must take into account how the propabilities to find the assets change in the second period. Otherwise the reasoning is analogous. We assume that settlement fails still occur such that there is still a positive premium. The propabilities to find the assets are given by $\mathbb{P}^a = \mathbb{P}^{na} = \sigma$. Probabilities are therefore weakly higher than beforehand. We can apply the same reasoning as in section 7.1. If the probabilities are strictly higher, then they only strengthen the effects. Therefore the quantities $A_t^{f,na,h}$ and $A_t^{f,a,h}$ increase if the rate for the non-access sellers increases from r^m to r. The same holds for the prices $p_t^{f,na}$ and $p_t^{f,a}$. If probabilities are strictly higher, the premium decreases even more in the second period after the shock.

Given that off-the-run prices on average increase, it follows that $A_t^{2,a}$ decreases and therefore $dA_t^{f,a,l}$ increases. Compared to section 7.1, if $A_t^{2,na}$ increases or decreases not only depends on the probability to find the assets but also on how these probabilities are affected by the increase in r^m. The overall effect is ambiguous. Therefore it is also ambiguous if the amount of off-the-run assets traded by constrained dealers with sellers who gain access, $A_t^{f,na,l}$, increases or decreases.

Summary of results IV *If sellers who did not had access, gain access, their spot prices increase and the on-the-run premium decreases. The quantities of on-the-run and off-the-run assets traded by unconstrained dealers with them increase. Trade in the off-the-run asset is initially more stimulated than trade in the on-the-run asset.*

7.3 Dependence of premium and facility rate in the data

For illustrative purposes, I show the scatter plot between the 10-year on-the-run premium and the reverse repo rate in figure 9.[43] We observe that there is a negative correlation between the 10-year on-the-run premium and the reverse repo rate with tight 95% confidence bands. This is consistent with my theory as I

[43]The data source for the premium is the same as in figure 2a. The reverse repo rate data is provided by the New York FED and can be downloaded here: https://www.newyorkfed.org/markets/omo_transaction_data#rrp. I take daily averages. The time horizon is 23 September 2013-2021 Q2 and the frequency is weekly.

predict a high premium when the reverse repo rate is low.

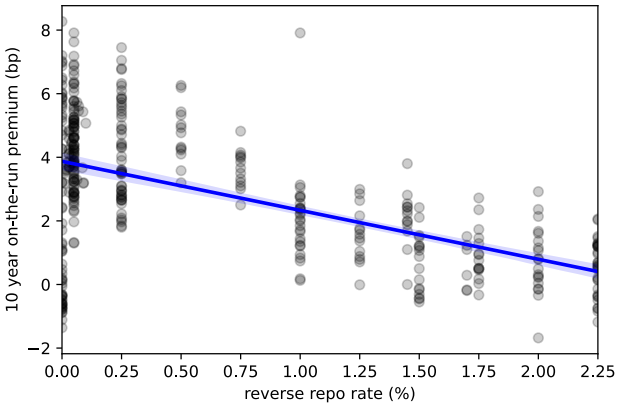


Figure 9: The correlation between the premium and the reverse repo rate

To systematically test my theory, I regress on-the-run premia data on the reverse repo rate (see table 4). If my theory is correct, I should observe a negative coefficient. In the regressions, I control for the deflated reverse repo trade amounts, the logarithm of the VIX, the 10 year - 2 year yield spread, and the general collateral financing repo rate.

I use data on the 1, 2, 3, 5, 7, 10, 20, and 30 year on-the-run premia. Each maturity and each day in the sample is a separate observation. The first observations I use are from 23 September 2013, as this was the first day the facility was available on a large scale.[44]

As expected, the coefficient on the reverse repo rate is negative and significant at the 1% level. A one percentage point increase in the reverse repo rate reduces the on-the-run premium by 2.2 basis points.[45] The interpretation of the other variables we control for is as follows: Since the deflated trade amounts are denominated in dollars, the coefficient is best interpreted by first multiplying it by the standard deviation of the trade amount. It follows that a one standard deviation increase in trade amounts, increases the premium by 0.2 basis points, or 0.08 standard deviations. A one percent increase in the VIX increases the premium by 0.8 basis points. A one percentage point increase in the 10 year - 2 year yield spread increases the premium by 0.2 basis points. Finally, a one percentage point increase in the general collateral financing repo rate increases the premium by 1.63 basis points. Next to the reverse repo rate, the trade amounts, the general

[44]See https://www.newyorkfed.org/markets/opolicy/operating_policy_130920.html for more information.

[45]The data on each reverse repo transaction, including the rate, the amount traded and other details, are only publicly available after two years. However, a reverse repo rate series can be downloaded without any time lag. When we use these data (without controlling for trade size) and extend the time horizon to 31 August 2023, the coefficient on the rate is still negative but insignificant.

	On-the-run premium
Reverse repo rate	-2.2225***
	(0.446)
Reverse repo trade amounts	0.1498***
	(0.048)
lnVIX	0.8326***
	(0.321)
10 year - 2 year yield spread	0.2183
	(0.161)
General collateral financing repo rate	1.6331***
	(0.406)
Constant	-0.1528
	(1.051)
No. Observations:	13076
R-squared:	0.051
Adj. R-squared:	0.051

I use Newey-West standard errors with 6 lags. *** indicates significance at the 1% level. The sample period is 23 September 2013-2021 Q2 and the frequency is daily.

The data source for the premium (in basis points) is the same as in figure 2a. The reverse repo rate and trade amount data are provided by the New York FED and can be downloaded here: https://www.newyorkfed.org/markets/omo_transaction_data#rrp. I take daily averages. The rate is in percent, the trade amount in billion US dollars. I deflate the trade amounts using the "Consumer Price Index for All Urban Consumers: All Items in U.S. City Average", which can be downloaded from FRED. I set the index to 1 when the trade amount series starts on the 23 September 2013. The VIX data are taken from FRED. The 10 year and 2 year yields (par yields of seasoned bonds in percent) are taken from the FED yield curve, which is an updated version of the original Gürkaynak-Sack-Wright curve (Gürkaynak et al. (2010)), https://www.federalreserve.gov/data/tips-yield-curve-and-inflation-compensation.htm. The general collateral financing repo rate (in percent) is provided by the Depository Trust and Clearing Corporation (DTCC) and can be downloaded here: https://www.dtcc.com/charts/dtcc-gcf-repo-index.

Table 4: Premium on reverse repo rate regression

collateral financing repo rate, and the VIX are significant.

8 Life cycle model

This section presents an extension of the basic model. It shows that the model can be used as a tool to describe the life cycle of a Treasury. In addition we can explain two more stylized facts:

When-issued Treasuries trade at a premium compared to previously issued Treasuries.

The primary market price is lower than the secondary market price.

The observations are illustrated in the appendix C.

The life cycle of a Treasury can be divided into three periods as illustrated in figure 10: the when-issued period, the on-the-run period, and the off-the-run period. The auction takes place between the announcement and the issuance of the assets. Each period and the auction are characterized by a different price. The chart below illustrates the life cycle.

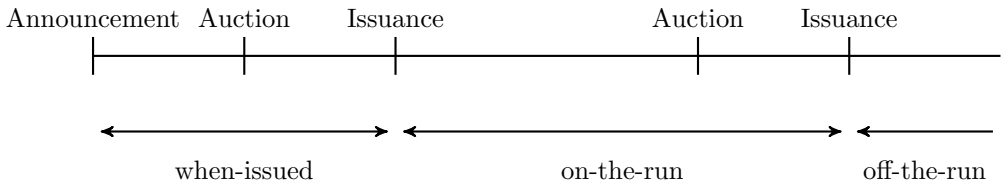


Figure 10: Life cycle

The when-issued market takes place after the auction of the security is announced but before it is issued. The when-issued market is important for price discovery (Durham and Perli (2022)). One possible interpretation of the prices q_t^2 and q_t^n is as the when-issued price of the respective assets.

Thus, the model covers all three periods with its prices. The fourth important price during the life cycle, the auction price, is added to the model below. Since I know the on- and off-the-run prices, I can derive the auction price.

I will first discuss the auction price of the one-period asset. A primary dealer taking part in the auction chooses to bid with a bundle consisting of price and quantity which I denote by $\left\{p_t^{1,A}, A_t^{1,A}\right\}$. The bundle specifies the quantity the primary dealer wants to buy and the price he is willing to pay. For each possible price $p_t^{1,A}$ that could be chosen, the optimal quantity is given by the solution to the following maximization problem:

$$\max_{A_t^{1,A}} -p_t^{1,A}A_t^{1,A} + (1-\sigma)\left[p_t^n A_t^n - \kappa(A_t^n) + \delta(A_t^{1,A} - A_t^n)\right] + \sigma\delta A_t^{1,A}$$

s.t.

$$A_t^{1,A} \geq A_t^n.$$

If the primary dealer buys an asset at the auction, he must pay the price $p_t^{1,A}$ today. With probability $(1-\sigma)$ he can enter the market in the same subperiod and sell the quantity A_t^n of the asset. With probability σ he will hold the asset until maturity and consume the coupon.

We know that in the premium equilibrium the constraint is not binding. Therefore, $p_t^{1,A} = \delta$. The price must equal the marginal utility of the asset. Otherwise there is infinite or no demand. Bidding will therefore drive the price to this value. The supply side is given by the Treasury, which auctions the amount $I_t^n = \mathcal{I}$. Therefore, in equilibrium, $A_t^{1,A} = I_t^n = \mathcal{I}$ and $p_t^{1,A} = \delta$.

The maximization problem for the two-period asset is analogous.

$$\begin{aligned}\max_{p_t^{2,A}, A_t^{2,A}} & -p_t^{2,A}A_t^{2,A} + (1-\sigma)\Big\{p_t^2 A_t^2 - \kappa(A_t^2) + \delta(A_t^{2,A} - A_t^2) \\ & + \beta(1-\sigma)\left[p_{t+1}^f A_{t+1}^{f,l} - \kappa(A_{t+1}^{f,l}) + \delta(A_t^{2,A} - A_t^2 - A_{t+1}^{f,l})\right] + \beta\sigma\delta(A_t^{2,A} - A_t^2)\Big\} \\ & + \sigma\left\{\delta A_t^{2,A} + \beta(1-\sigma)\left[p_{t+1}^f A_{t+1}^{f,h} - \kappa(A_{t+1}^{f,h}) + \delta(A_t^{2,A} - A_{t+1}^{f,h})\right] + \beta\sigma\delta A_t^{2,A}\right\}\end{aligned}$$

s.t.

$$A_t^{2,A} \geq A_t^2$$

$$A_t^{2,A} \geq A_{t+1}^{f,h}$$

$$A_t^{2,A} - A_t^2 \geq A_{t+1}^{f,l}$$

We know that in the premium equilibrium, the last constraint is binding. The others are not.

The Treasury auctions the stock $I_t^2 = \mathcal{I}$. Therefore in equilibrium $A_t^{2,A} = I_t^2 = \mathcal{I}$ and $p_t^{2,A} = \delta + (1-\sigma)\beta\left\{(1-\sigma)[p_{t+1}^f - \kappa'(A^{f,l})] + \sigma\delta\right\} + \sigma\beta\delta$.

After adding the auction price, we can now summarize and compare all the prices. In the case of the one-period asset, these are

$$
\begin{aligned}
p_t^{1,A} &= \delta \\
q_t^n &= (\delta + h) \\
p_t^n &= \beta(1+r)(\delta + h).
\end{aligned}
$$

In the case of the two-period asset, these are

$$
\begin{aligned}
p_t^{2,A} &= \delta + (1-\sigma)\beta \left\{ (1-\sigma)[p_{t+1}^f - \kappa'(A^{f,l})] + \sigma\delta \right\} + \sigma\beta\delta \\
q_t^2 &= (1+\beta)(\delta + h) \\
p_t^2 &= \beta(1+r)(1+\beta)(\delta + h) \\
p_t^f &= \frac{\sigma}{1-(1-\sigma)\beta(1+r)} \beta(1+r)(\delta + h).
\end{aligned}
$$

In addition to the empirical regularity that the on-the-run price is above the off-the-run price, the model explains two other regularities about prices as pointed out in the beginning of this section. First, a detailed analysis by Durham and Perli (2022) showed that when-issued prices carry a premium compared to already issued Treasuries. In their comparison, they include both on- and off-the-run Treasuries. Second, the secondary market price is known to be higher than the auction price (Goldreich (2007), Spindt and Stolz (1992), and Fleming et al. (2022)). See appendix section C for an illustration.

I will explain the first of the two stylized facts first. It follows from the equations that the spot price (in the case of no fail) is the discounted when-issued price. This can be explained from the seller's perspective. When selling a when-issued contract, the seller receives the when-issued price tomorrow and can deposit it. This income has to cover the spot price he has to pay tomorrow, which in equilibrium is equal to the collateral value he has to build up today. Therefore the spot price in case of no fail is the discounted when-issued price. Compared to off-the-run prices, the when-issued price is even higher, because the off-the-run price takes into account the probability of a settlement fail.

To explain the second stylized fact, the most natural comparison is to compare the auction to the on-the-run price in our model. The auction price is given by the marginal utility of the last unit purchased. The last unit remains in the inventory of the primary dealer and he consumes the coupon δ, except it is sold as an off-the-run asset of a constrained primary dealer. The empirical results of Fleming et al. (2022) suggest that some of the Treasuries remain in the portfolios of the primary dealers until maturity, while the other part is sold. This is consistent with our observation and is reflected in the auction price.
On the other hand, the on-the-run price is determined by the value of the asset to participants in the

secondary market. In our case, the buyer is the ultimate owner of the asset, so his marginal utility determines the price. In equilibrium, the spot price is above the coupon rate δ because the buyer has a higher marginal valuation of the asset. The on-the-run prices are therefore higher than the auction prices. Since the when-issued price is higher than the on-the-run price, it is also higher than the auction price.

Our discussion leads us to state our last summary of results:

Summary of results V *The on-the-run price is driven by the valuation of the buyer of the asset in the secondary market. The auction price also contains the valuation of the asset if it stays in the inventory of the primary dealer until maturity. The when issued price is higher than the spot price due to the cost of collateralization.*

9 Conclusion

I developed a model of the on-the-run phenomenon. I provide a novel explanation using inventory which I also test empirically. The model was then used to discuss broad access to central bank facilities. The analysis is motivated by the increase in intermediation and participation by non-bank financial institutions (Eren and Wooldridge (2021)). This raises the question of why only a limited number of participants currently have access to central bank facilities.[46] I analysed the implications of providing broad access to a reverse repo or deposit facility. Finally, I added to the literature by extending the model to cover the full life cycle of Treasuries. It includes the auction price, the when-issued price, and the on- and off-the-run prices (all four relevant prices), and can accommodate stylized facts about the relationship between them.

To explain the premium, I endogenized settlement fails and found, consistent with the data, that they are more frequent for off-the-run Treasuries. The probability of a settlement fail is higher for off-the-run Treasuries because they are longer in the market and more dispersed. This leads to a preference for and premium on on-the-run Treasuries. This also explains why the premium is always on the on-the-run Treasury, an aspect that Vayanos and Weill (2008) could not explain. I also test my theory in the data. I show that the premium is higher in times when the probability of failure of off-the-run Treasuries is higher. The probability of off-the-run Treasuries is higher, when the inventories of primary dealers are lower. Future research should analyse whether the fundamental dynamics described imply a self-fulfilling dynamic that adds to the size of the premium.

The second part focused on the impact of the facility rate when broad access is provided. The result is that

[46] See for example the FAQ on the repo and the reverse repo facility of the New York FED for eligibility criteria, https://www.newyorkfed.org/markets/repo-agreement-ops-faq and https://www.newyorkfed.org/markets/rrp_faq.

an increase in the facility rate leads to a reduction in the cost of trading. The facility complements a trade in the sense that it provides a certain return on liquid funds between trades for those who have access to it. I show that access stimulates trading and prices rise. An increase in the facility rate decreases the premium, which I also confirm in the data. Interestingly, if the facility rate increases, fails increase as more off-the-run Treasuries are traded in contracts promising their delivery but fewer are available. Also, giving access to all institutions is not an act of increasing fairness. Only primary dealers benefit from an increase in the facility rate, not those who gained access. An interesting way to extend the model would be to take into account more specific characteristics of non-bank financial institutions. This could be, for example, leverage in the case of a hedge fund. In addition, given the current discussions on how to restructure the Treasury market (see e.g. Duffie et al. (2021)), it would be interesting to explore the effects of broad access to a repo facility to complement my analysis.

References

Bacchetta, P. and E. van Wincoop (2021). Puzzling exchange rate dynamics and delayed portfolio adjustment. *Journal of International Economics 131*.

Baklanova, V., A. Copeland, and R. McCaughrin (2015). Reference guide to U.S. repo and securities lending markets. *Federal Reserve Bank of New York Staff Reports* (740).

Barclay, M. J., T. Hendershott, and K. Kotz (2006). Automation versus intermediation: Evidence from Treasuries going off the run. *Journal of Finance 61*(5), 2395–2414.

Bech, M., A. Illes, U. Lewrick, and A. Schrimpf (2016). Hanging up the phone – electronic trading in fixed income markets and its implications. *BIS Quarterly Review*.

Bessembinder, H., C. Spatt, and K. Venkataraman (2020). A survey of the microstructure of fixed-income markets. *Journal of Financial and Quantitative Analysis 55*(1), 1–45.

Bowman, D., J. Louria, M. McCormick, and M.-F. Styczynski (2017). The cleared bilateral repo market and proposed repo benchmark rates. *FEDS Notes*.

Chambers, D. and E. Dimson (2009). IPO underpricing over the very long run. *The Journal of Finance 64*(3), 1407–1443.

Christensen, J. H. E., J. A. Lopez, and P. Shultz (2017). Do all new Treasuries trade at a premium? *FRBSF Economic Letter* (3).

Christensen, J. H. E., J. A. Lopez, and P. J. Shultz (2020). Is there an on-the-run premium in TIPS? *The Quarterly Journal of Finance 10*(2).

Corradin, S. and A. Maddaloni (2020). The importance of being special: Repo markets during the crisis. *Journal of Financial Economics 137*(2), 392–429.

D'Amico, S. and N. A. Pancost (2022). Special repo rates and the cross-section of bond prices: The role of the special collateral risk premium. *Review of Finance 26*(1), 117–162.

Duffie, D. (2023). Resilience redux in the US Treasury market. Jackson Hole Symposium.

Duffie, D., N. Garleanu, and L. H. Pedersen (2005). Over-the-counter markets. *Econometrica 73*(6), 1815–1847.

Duffie, D., T. F. Geithner, P. Parkinson, and J. Stein (2021). U.S. Treasury markets: steps toward increased resilience. https://group30.org/publications/detail/4950.

Durham, J. B. and R. Perli (2022). *The Research Handbook of Financial Markets*, Chapter "The Treasury and the When-Issued Market". Forthcoming from Edward Elgar.

Eren, E. and P. Wooldridge (2021). Non-bank financial institutions and the functioning of government bond markets. *BIS Papers* (119).

Fleming, M. and F. Keane (2021). The netting efficiencies of marketwide central clearing. *Federal Reserve Bank of New York Staff Reports* (964).

Fleming, M., G. Nguyen, and J. Rosenberg (2022). How do Treasury dealers manage their positions? *Federal Reserve Bank of New York Staff Reports* (299).

Fleming, M. J. and K. D. Garbade (2005). Explaining settlement fails. *Current Issues in Economics and Finance 11*(9).

Fleming, M. J., F. M. Keane, A. Martin, and M. McMorrow (2014). What explains the June spike in Treasury settlement fails? Liberty Street Economics.

Fleming, M. J., B. Mizrach, and G. Nguyen (2018). The microstructure of a U.S. Treasury ECN: The BrokerTec platform. *Journal of Financial Markets 40*, 2–22.

Frost, J., L. Logan, A. Martin, P. McCabe, F. Natalucci, and J. Remache (2015). Overnight RRP operations as a monetary policy tool: Some design considerations. *Finance and Economics Discussion Series* (2015-010).

Garbade, K. D., F. M. Keane, L. Logan, A. Stokes, and J. Wolgemuth (2010). The introduction of the TMPG Fails Charge for U.S. Treasury securities. *FRBNY Economic Policy Review*.

Garleanu, N. and L. H. Pedersen (2013). Dynamic trading with predictable returns and transaction costs. *The Journal of Finance 68*(6), 2309–2340.

Goldreich, D. (2007). Underpricing in discriminatory and uniform-price Treasury auctions. *Journal of Financial and Quantiative Analysis 42*(2), 443–466.

Goldreich, D., B. Hanke, and P. Nath (2005). The price of future liquidity: time-varying liquidity in the U.S. Treasury market. *Review of Finance 9*(1), 1–32.

Gürkaynak, R. S., B. Sack, and J. H. Wright (2010). The tips yield curve and inflation compensation. *American Economic Journal: Macroeconomics 2*(1), 70–92.

Krishnamurthy, A. (2002). The bond/old-bond spread. *Journal of Financial Economics 66*(2-3), 463–506.

Lagos, R. and G. Rocheteau (2009). Liquidity in asset markets with search frictions. *Econometrica 77*(2), 403–426.

Liu, S. and C. Wu (2017). Repo counterparty risk and on-/off-the-run Treasury spreads. *Review of Asset Pricing Studies 7*(1), 81–143.

Macchiavelli, M. and L. Pettit (2021). Liquidity regulation and financial intermediaries. *Journal of Financial and Quantitative Analysis 56*(6), 2237 – 2271.

Marte, J. (2021). NY Fed eases eligibility requirements for reverse repo facility. Reuters article.

Pasquariello, P. and C. Vega (2009). The on-the-run liquidity phenomenon. *Journal of Financial Economics 92*(1), 1–24.

Spindt, P. A. and R. W. Stolz (1992). Are US Treasury bills underpriced in the primary market? *Journal of Banking and Finance 16*(5), 891–908.

Strebulaev, I. (2002). Liquidity and asset pricing: Evidence from the U.S. Treasury securities market. Unpublished working paper.

U.S. Department of the Treasury (2022). Additional public transparency in Treasury markets.

U.S. Department of the Treasury, Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, U.S. Securities and Exchange Commission, and U.S. Commodity Futures Trading Commission (2015). Joint staff report: The U.S. Treasury market on October 15, 2014.

Vayanos, D. and P.-O. Weill (2008). A search-based theory of the on-the-run phenomenon. *The Journal of Finance 63*(3), 1361–1398.

Wells, M. (2023). Averting a Treasury market crisis. *Econ Focus*, 14–17.

A Other equilibria

Until now I focused on the case where $\lambda_t^2 = \lambda_t^n = \lambda_t^{f,h} = 0$. This means that primary dealers are unconstrained if they still have the full inventory of assets i. I showed that we can always find issuance sizes $\mathcal{I}$ where such an equilibrium exists.
For the sake of completeness we can also adjust and think of equilibria where λ_t^2, λ_t^n, and $\lambda_t^{f,h}$ are non-zero. In this case, the quantity of the corresponding assets sold is $\mathcal{I}$. The prices in each of these equilibria are the same as in the premium equilibrium 6 as long as $\lambda_t^{f,l} > 0$, which we need for a positive premium. Since the dynamics and intuition are exactly the same in these equilibria as well, I don't discuss these equilibria in the following, as no additional insights are gained. The only results that would change are the effects of central bank access on quantities. In these other equilibria, there are no quantity effects on assets for which constraints are binding, and only the lagrange multipliers would change in magnitude.

B Proofs

B.1 Proposition 5.2

Proposition 5.2 states that a necessary and sufficient condition for existence of the equilibrium with $\mathcal{I} \in (max(A_t^2, A_t^n,\, A_t^{f,h}), (A_{t-1}^2 + A_t^{f,h}))$ is $\sigma > \frac{1-\beta(1+r)}{\beta(1+r)}\frac{\delta}{h}$.

First, I show that for an equilibrium with trade in all assets to exist, we need $\sigma > \frac{1-\beta(1+r)}{\beta(1+r)}\frac{\delta}{h}$.

In equilibrium $p_t^f = \kappa'(A_t^{f,h}) + \delta = \frac{\sigma}{1-(1-\sigma)\beta(1+r)}\beta(1+r)(\delta+h)$. For $\kappa'(A_t^{f,h}) > 0$ and $A_t^{f,h} > 0$ we must have that $\frac{\sigma}{1-(1-\sigma)\beta(1+r)}\beta(1+r)(\delta+h) - \delta > 0$, respectively $\sigma > \frac{1-\beta(1+r)}{\beta(1+r)}\frac{\delta}{h}$.

Second, I show that $A_t^2 > A_t^n > A_t^{f,h}$.

The first order conditions of the primary dealer with respect to the assets maturing in one period are $p_t^n = \kappa'(A_t^n) + \delta$ and $p_t^f = \kappa'(A_t^f) + \delta$. As $p_t^n > p_t^f$ and $\kappa(A_t^i)$ being a strictly convex function it follows that $A_t^n > A_t^f$. In addition, I can show that $A_t^2 > A_t^n$. The first order condition of the primary dealer with respect to the new two-period asset is:

$$\begin{aligned}
\kappa'(A_t^2) - \beta(1-\sigma)\kappa'(\mathcal{I} - A_t^2) &= p_t^2 - (1+\beta)\delta - \beta(1-\sigma)\kappa'(A_{t+1}^{f,h}) \\
\kappa'(A_t^2) - \beta(1-\sigma)\kappa'(\mathcal{I} - A_t^2) &= (1+\beta)p_t^n - (1+\beta)\delta - \beta(1-\sigma)\kappa'(A_{t+1}^{f,h}) \\
\kappa'(A_t^2) - \beta(1-\sigma)\kappa'(\mathcal{I} - A_t^2) &= (1+\beta)\kappa'(A_t^n) - \beta(1-\sigma)\kappa'(A_{t+1}^{f,h})
\end{aligned}$$

Further rearrangement yields $\kappa'(A_t^2) - \kappa'(A_t^n) = \beta\kappa'(\mathcal{I} - A_t^2) + \beta[\kappa'(A_t^n) - \kappa'(A_{t+1}^{f,h})] + \beta\sigma[\kappa'(A_{t+1}^{f,h}) - \kappa'(\mathcal{I} - A_t^2)] > 0$. Therefore $A_t^2 > A_t^n$.

Lastly I show that I can always find an $\mathcal{I}$, where the equilibrium exists.

In equilibrium $(A^2_{t-1} + A^{f,h}_t) > \mathcal{I} > max(A^2_t, A^n_t, A^{f,h}_t)$. I know that $A^2_t > A^n_t > A^{f,h}_t$. Therefore it follows that in equilibrium $(A^2_{t-1} + A^{f,h}_t) > \mathcal{I} > A^2_t$.

In addition we know that in equilibrium

$$\begin{aligned} A^n_t &= \kappa'^{-1}\left[\beta(1+r)(\delta+h)-\delta\right] \\ A^{f,h}_t &= \kappa'^{-1}\left[\frac{\sigma}{1-(1-\sigma)\beta(1+r)}\beta(1+r)(\delta+h)-\delta\right] \\ \kappa'(A^2_{t-1}) &= W + \beta(1-\sigma)\kappa'(\mathcal{I}-A^2_{t-1}) \end{aligned}$$

where $W \equiv \left[(1+\beta)-\beta(1-\sigma)\frac{\sigma}{1-(1-\sigma)\beta(1+r)}\right]\beta(1+r)(\delta+h)-(1+\beta\sigma)\delta$.

The LHS of the third equation, $f_1(x) = \kappa'(x)$, is a strictly increasing function with $x \in [0, \mathcal{I}]$ and $min(f_1(x)) = 0$ and $max(f_1(x)) = \kappa'(\mathcal{I})$. The RHS of the third equation, $f_2(x) = W + \beta(1-\sigma)\kappa'(\mathcal{I}-x)$, is a strictly decreasing function with $x \in [0, \mathcal{I}]$ and $min(f_2(x)) = W$ and $max(f_2(x)) = W + \beta(1-\sigma)\kappa'(\mathcal{I})$.

Therefore for the equilibrium to exist it must be that $\kappa'(\mathcal{I}) > W$ respectively $\mathcal{I} > \kappa'^{-1}(W)$. By continuity it then follows that $\mathcal{I} > A^2_{t-1}$.

As $\mathcal{I} - A^2_{t-1}(\mathcal{I})$ is a continuous function (as $\kappa'(A^i_t)$ and $\kappa'^{-1}(A^i_t)$ are both continuous) with minimal value 0 for $\mathcal{I} = \kappa'^{-1}(W)$ I can always find an $\mathcal{I}$ where $\mathcal{I} - A^2_{t-1}(\mathcal{I}) < A^{f,h}_t$ for any $A^{f,h}_t > 0$.

B.2 Result II

Result II states that if all sellers gain access, an increase in the facility rate r increases all spot prices and decreases the on-the-run premium. The quantities of on-the-run and off-the-run assets traded by unconstrained dealers increase. Trade in the off-the-run asset is (initially) more stimulated than trade in the on-the-run asset, and the settlement failure rate of the off-the-run asset (initially) increases. If search frictions are above a small minimum value, then overall more assets are offloaded from the inventories and end up in the portfolio of the holder with the highest marginal asset valuation.

I assume an unanticipated increase in the interest rate that occurs at the beginning of the contract market. The stock of available off-the-run assets in the next spot market is given and can reach a new steady state only after a period.

First, I show that an increase in the facility rate raises all prices. In equilibrium

$$
\begin{aligned}
p_t^n &= \beta(1+r)(\delta+h) \\
p_t^2 &= \beta(1+r)(1+\beta)(\delta+h) \\
p_t^f &= \frac{\sigma}{1-(1-\sigma)\beta(1+r)}\beta(1+r)(\delta+h).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\frac{dp_t^n}{dr} &= \beta(\delta+h) > 0 \\
\frac{dp_t^2}{dr} &= \beta(1+\beta)(\delta+h) > 0 \\
\frac{dp_t^f}{dr} &= \left[\frac{\sigma}{1-(1-\sigma)\beta(1+r)} + \frac{\sigma(1-\sigma)\beta(1+r)}{\left[1-(1-\sigma)\beta(1+r)\right]^2}\right]\beta(\delta+h) > 0.
\end{aligned}
$$

Second I show that off-the-run prices react stronger to an increase in the facility rate compared to their on-the-run equivalents. In equilibrium $\frac{dp_t^n}{dr} < \frac{dp_t^f}{dr}$ if $\frac{\sigma}{1-(1-\sigma)\beta(1+r)} + \frac{\sigma(1-\sigma)\beta(1+r)}{[1-(1-\sigma)\beta(1+r)]^2} > 1$ which is the case if $\sigma > \frac{[1-\beta(1+r)]^2}{[\beta(1+r)]^2} \approx 0$.[47]

Third I show that the on-the-run premium decreases given an increase in r. The premium is given by $\Delta = p_t^n - p_t^f$. The derivative is $\frac{d\Delta}{dr} = \frac{dp_t^n}{dr} - \frac{dp_t^f}{dr} < 0$.

Fourth I discuss the impact on the quantities.

In equilibrium

$$
\begin{aligned}
\kappa'(A_t^n) &= p_t^n - \delta \\
\kappa'(A_t^2) - \beta(1-\sigma)\kappa'(I_t^2 - A_t^2) &= p_t^2 - \beta(1-\sigma)p_{t+1}^f - (1+\beta\sigma)\delta \\
\kappa'(A_t^{f,h}) &= p_t^f - \delta.
\end{aligned}
$$

An increase in spot prices implies an increase in the following traded quantities: A_t^n and $A_t^{f,h}$. Therefore $\frac{dA_t^n}{dr} > 0$, $\frac{dA_t^{f,h}}{dr} > 0$. Furthermore, $\frac{dA_t^2}{dr} > 0$ iff $\frac{dp_t^2}{dr} > \frac{dp_{t+1}^f}{dr}$. Also, $\frac{dA_t^2}{dr} > 0$ if search frictions, measured by σ, are above a small minimum value.[48]

[47] Let us define $k \equiv (1-\sigma)\beta(1+r)$. It follows: $\frac{\sigma}{1-k} + \frac{\sigma k}{(1-k)^2} > 1$, respectively $\sigma > (1-k)^2$. Further rearranging yields $0 > 1 - 2\beta(1+r) + (1-\sigma)\left[\beta(1+r)\right]^2$, respectively $\sigma > \frac{[1-\beta(1+r)]^2}{[\beta(1+r)]^2} \approx 0$.

[48] From the equations it follows that $\frac{dp_t^2}{dr} > \frac{dp_{t+1}^f}{dr}$ iff $(1+\beta) > \beta(1-\sigma)\left[\frac{\sigma}{1-(1-\sigma)\beta(1+r)} + \frac{\sigma(1-\sigma)\beta(1+r)}{[1-(1-\sigma)\beta(1+r)]^2}\right]$, respectively, $(1-\beta(1+r))^2 > \frac{\sigma}{(1-\sigma)}\left[(1-\sigma)\left(\frac{\beta}{(1+\beta)} + (\beta(1+r))^2\right) - 1\right]$, where $(1-\beta(1+r))^2 \approx 0$. Only for a limited range of small values of σ it can be that $\frac{dp_t^2}{dr} > \frac{dp_{t+1}^f}{dr}$ is not true. For example if $\beta = 0.96$ and $r = 0.01$, then only for $\sigma \in (0.003, 0.25)$ it follows that $\frac{dp_t^2}{dr} < \frac{dp_{t+1}^f}{dr}$.

The quantity of off-the-run assets traded by constrained dealers, $A_t^{f,l} = \mathcal{I} - A_{t-1}^2$, stays in the first spot market after the interest rate hike the same as the inventory is given from the period beforehand but then decreases (increases) if A_t^2 increased (decreased).

Lastly I define an off-the-run settlement failure rate as the value of off-the-run assets involved in a fail divided by the overall amount of off-the-run assets promised to be delivered:

$$f_{t+1}^f \equiv \frac{(1-\sigma)p_{t+1}^f(a_{t+1}^f - I_{t+1}^f)}{p_{t+1}^f a_{t+1}^f} = (1-\sigma)\left(1 - \frac{\mathcal{I} - a_t^2}{a_{t+1}^f}\right).$$

The rate increases initially after the rate hike. It also increases in the longterm if $\frac{\partial A_t^2}{\partial r} = \frac{\partial a_t^2}{\partial r} > 0$.

B.3 Result III

Result III states that if sellers gain access, they paradoxically do not benefit from an increase in the facility rate r. In equilibrium (if search frictions are above a small minimum value) only primary dealers profit from it, even with positive bargaining power of sellers, and buyers' utility even decreases.

Profits (starting in the contract market, averages[49]) are given by

$$\begin{aligned}
\pi_t^b &= \sum_i \left[\beta(\delta+h) - \beta q_t^i\right] a_{t+1}^i + \beta^2(\delta+h)a_{t+1}^2 - \beta(1-\sigma)[(\delta+h) - \omega_t^f](a_{t+1}^f - I_{t+1}^f) + \beta\pi_{t+1}^b \\
\pi_t^s &= \sum_i \left\{-\omega_t^i + \beta\left[q_t^i - p_{t+1}^i + \omega_t^i\right](1+r)a_{t+1}^i\right\} - \beta(1-\sigma)[\omega_t^f - p_{t+1}^f](1+r)(a_{t+1}^f - I_{t+1}^n) + \beta\pi_t^s \\
\pi_t^D &= \beta\{(1-\sigma)^2[p_{t+1}^f(I_t^2 - A_t^2) - \kappa(I_t^2 - A_t^2) - \delta(I_t^2 - A_t^2)] + (1-\sigma)\sigma[p_{t+1}^f A_{t+1}^{f,h} - \kappa(A_{t+1}^{f,h}) - \delta A_{t+1}^{f,h}] \\
&\quad + (1-\sigma)\sum_{j\in\{n,2\}}\left[p_{t+1}^j A_{t+1}^j - \kappa(A_{t+1}^j) - \delta A_{t+1}^j\right] + \delta(I_{t+1}^n + I_{t+1}^2 + I_t^2 - (1-\sigma)A_t^2)\} + \beta\pi_{t+1}^D.
\end{aligned}$$

I assume that the unanticipated increase in the interest rate occurs at the beginning of the contract market. The stock of available off-the-run assets in the next spot market is given and can reach a new steady state only after a period. New steady state values have no time index. I can simplify the buyer's and seller's profit using the first-order conditions in equilibrium. This yields

$$\begin{aligned}
\pi_t^b &= \beta(1-\sigma)[(\delta+h) - p_{t+1}^f]I_{t+1}^f + \frac{\beta^2}{1-\beta}(1-\sigma)\left[(\delta+h) - p^f\right]I^f \\
\pi_t^s &= 0.
\end{aligned}$$

[49] One group of primary dealers has an inventory of $\mathcal{I}$ off-the-run assets and another has one of $\mathcal{I} - A_t^2$. I take the profit of both groups and average according to their share in the population.

The derivatives are

$$
\begin{aligned}
\frac{d\pi_t^b}{dr} &= -\beta(1-\sigma)\frac{dp_{t+1}^f}{dr}(\mathcal{I}-a_t^2) - \frac{\beta^2}{1-\beta}(1-\sigma)\left\{\frac{dp^f}{dr}(\mathcal{I}-a^2) + \left[(\delta+h)-p^f\right]\frac{da^2}{dr}\right\} \\
\frac{d\pi_t^d}{dr} &= 0 \\
\frac{d\pi_t^D}{dr} &= \beta\Bigg\{(1-\sigma)^2\frac{dp_{t+1}^f}{dr}(\mathcal{I}-Q_t^2) + (1-\sigma)\sigma\frac{d[\kappa'(A_{t+1}^{f,h})A_{t+1}^{f,h} - \kappa(A_{t+1}^{f,h})]}{dr} \\
&\quad + (1-\sigma)\sum_{j\in\{n,2\}}\frac{d[\kappa'(A_{t+1}^{j})A_{t+1}^{j} - \kappa(A_{t+1}^{j})]}{dr}\Bigg\} \\
&\quad + \frac{\beta^2}{1-\beta}\Bigg\{(1-\sigma)^2\frac{d[\kappa'(\mathcal{I}-A^2)(\mathcal{I}-A^2) - \kappa(\mathcal{I}-A^2) - \beta(1-\sigma)\kappa'(\mathcal{I}-A_t^2)\mathcal{I}]}{dr} \\
&\quad + (1-\sigma)\sigma\frac{d[\kappa'(A^{f,h})A^{f,h} - \kappa(A^{f,h})]}{dr} + (1-\sigma)\sum_{j\in\{n,2\}}\frac{d[\kappa'(A^{j})A^{j} - \kappa(A^{j})]}{dr}\Bigg\}.
\end{aligned}
$$

We can simplify the last equation. This yields

$$
\begin{aligned}
\frac{d\pi_t^D}{dr} =& \beta(1-\sigma)^2\frac{dp_{t+1}^f}{dr}(\mathcal{I}-A_t^2) + \frac{\beta}{1-\beta}\Bigg\{\beta(1-\sigma)^2\kappa''(\mathcal{I}-A^2)A^2\frac{\partial A^2}{\partial r} + (1-\sigma)\sigma\kappa''(A^{f,h})A^{f,h}\frac{\partial A^{f,h}}{\partial r} \\
&+ (1-\sigma)\sum_{j\in\{n,2\}}\kappa''(A^j)A^j\frac{\partial A^j}{\partial r}\Bigg\}.
\end{aligned}
$$

Regarding the profits of the buyer, note that $(\delta+h)-p^f > 0$. If $\frac{\partial A^2}{\partial r} = \frac{\partial a^2}{\partial r} > 0$ (which is the case if σ is above a small minimal value, see B.2), then $\frac{d\pi_t^b}{dr} < 0$, $\frac{d\pi_t^s}{dr} = 0$, and $\frac{d\pi_t^D}{dr} > 0$.

C Graphs

The following graphs show the stylized facts and irregularities that are discussed in the paper.

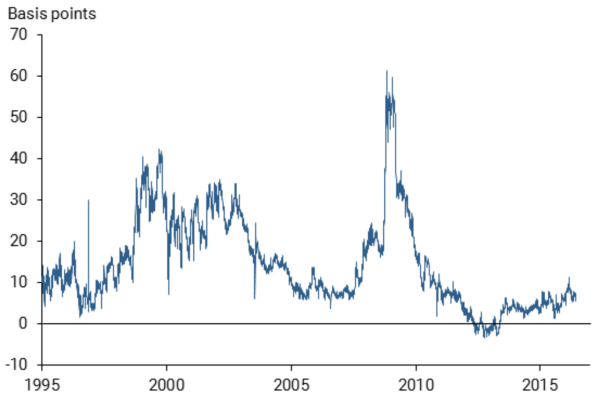


Figure 11: On-the-run premium on 10 year bonds

Figure 11 is taken from Christensen et al. (2017) and shows the on-the-run premium on 10 year Treasury bonds. The on-the-run yield is subtracted from the par yield of seasoned bonds.

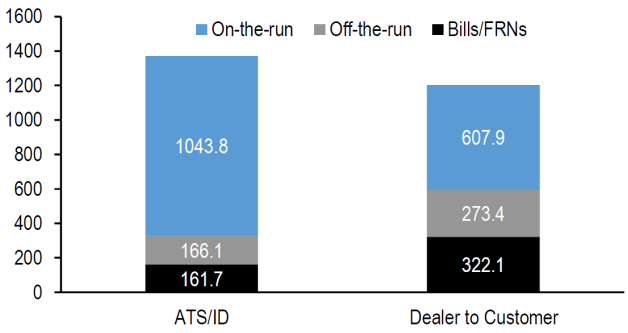


Figure 12: Treasury trading volumes by counterparty in billion USD

Figure 12 shows Treasury trading volumes by counterparty. ATS stands for automated trading system and ID for interdealer (U.S. Department of the Treasury (2022)). We can see that the trading volume in on-the-run Treasuries is much higher than in off-the-run Treasuries, regardless of the counterparty. This is also shown in the next chart.

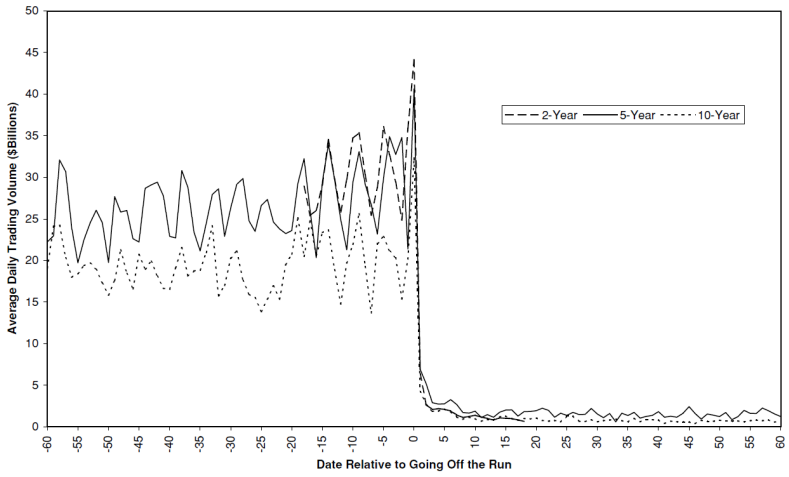


Figure 13: Trading volume for on-the-run and off-the-run Treasuries

Figure 13 shows the average trading volume for on- and off-the-run 2-year, 5-year, and 10-year Treasury notes relative to the date they went off-the-run (Barclay et al. (2006)).

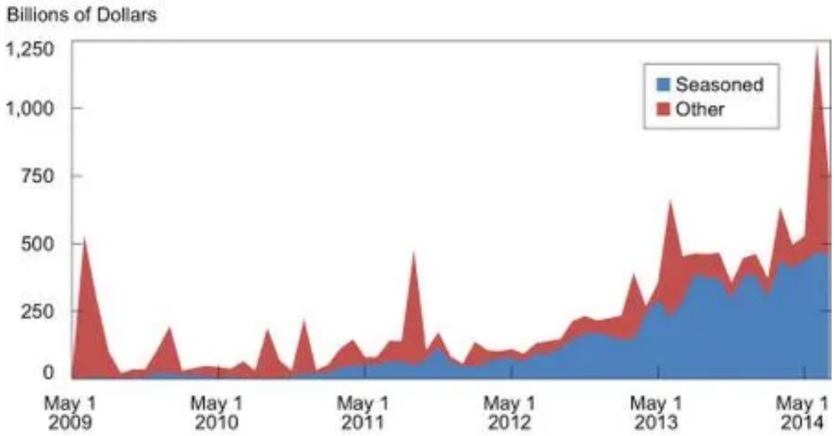


Figure 14: Fails in seasoned and other Treasuries

Figure 14, provided by Fleming et al. (2014), plots the cumulative gross fails by month in seasoned Treasuries and all other Treasuries (including on-the-run Treasuries). Seasoned Treasuries are Treasuries issued more than 180 days ago. Back in 2014, Fleming et al. (2014) pointed out that there has been a steady increase in the number of fails of seasoned Treasuries over the past few years.

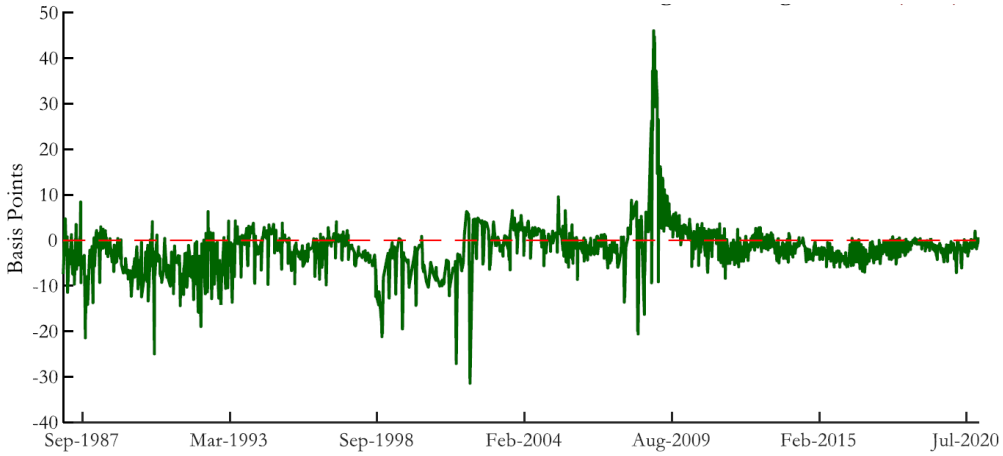


Figure 15: When-issued premium

The conditional when-issued premium, shown in figure 15, was measured by Durham and Perli (2022). The premium is computed by regressing the difference between the actual and the fitted yield on control variables and a dummy variable indicating whether the Treasury is when-issued.

Time of Day	No. of Auctions	Underpricing (relative to when-issued transactions)	Standard Error	Underpricing (relative to when-issued bid quote)	Standard Error
12:50 PM–1:00 PM	178	0.32***	0.12	0.20**	0.12
12:30 PM–1:00 PM	178	0.40***	0.12	0.29***	0.12
1:00 PM–1:30 PM	176	0.32***	0.10	0.15*	0.10

Figure 16: Underpricing

Fleming et al. (2022) document in their paper that Treasury dealers appear to be compensated for taking inventory risks at the auction by price increases in subsequent weeks. This is evidence that auction prices are lower than secondary market prices after the auction. A direct comparison of primary and secondary market prices at the time of the auction has been done by Goldreich (2007) and Spindt and Stolz (1992), among others. Both show that the primary market price is lower than the secondary market price of the same security.[50] The table above from Goldreich (2007) compares the auction yields with the when-issued yields in the minutes before and after the auction. As Goldreich (2007) explains, for example, an underpricing of 0.32 basis points (first row, third column) is equivalent to 1.3 cents per 100$.

[50] Underpricing is also an observed phenomenon in Initial Public Offerings (IPO's). See for example Chambers and Dimson (2009).