BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel—2025-19

Last updated: August 11, 2025

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

Fabienne Scheider
Financial Markets Department
Bank of Canada
fschneider@bankofcanada.ca

DOI: https://doi.org/10.34989/swp-2025-21 | ISSN 1701-9397

Acknowledgements

The views of this paper do not represent those of the Bank of Canada. All errors are my own. I wish to thank Bruno Biais, Martin Brown, Jens Christensen, Stefania D'Amico, Michael Fleming, Refet Gürkaynak, Kinda Hachem, Arvind Krishnamurthy, Ricardo Lagos, Antoine Martin, Cyril Monnet, Plamen Nenov, Andreas Uthemann and the participants of the Macroeconomics PhD Seminar of the University of Bern (2021), the Central Bankers Course on Money Markets, Liquidity, and Payment Systems (2022), the Macroeconomics Workshop Hasliberg (2022), the BDP Alumni Conference (2022), the Brown Bag Seminar of the University of Bern (2022), the 35th Australasian Finance & Banking Conference (2022), the Swiss Society for Financial Market Research Conference (2023), the seminar at the Bank of Canada (2024), the Bank of England (2024), the Sveriges Riksbank (2024), the SSES Annual Congress (2024), the Essex-UCL Workshop on Decentralised Financial Markets (2024) and other seminars and conferences for feedback, insights, and discussion.

Abstract

The premium on "on-the-run" Treasuries (i.e., the most recently issued ones) is an anomaly. I explain it using a model in which primary dealers hold inventories of Treasuries. There is less variation across primary dealers' inventories of on-the-run Treasuries compared with off-the-run Treasuries. Because there is less inventory uncertainty, on-the-run Treasuries fail to settle less frequently and trade at a premium. My theory is consistent with the USD 33 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 analyze 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.

Topics: Financial markets; Asset pricing; Market structure and pricing; Monetary policy
JEL codes: G12, G19, G23

Résumé

La prime sur les bons du Trésor faisant partie de l'émission courante (l'émission la plus récente) est une anomalie. Je l'explique à l'aide d'un modèle où les négociants principaux détiennent des stocks de bons du Trésor. Les variations observées dans ces stocks sont moindres lorsqu'il s'agit de l'émission courante, comparativement aux émissions passées. À cause de cette incertitude réduite, les bons de l'émission courante se soldent moins souvent par des échecs de règlement et se négocient avec une prime. Mon hypothèse est confortée par les échecs de règlement : chaque jour, ce sont 33 milliards de dollars américains de contrats sur des bons du Trésor qui ne sont pas réglés, mais le taux d'échec médian pour les bons d'émissions passées est presque le double de celui observé pour les bons de l'émission courante. J'utilise aussi le modèle pour analyser les effets produits lorsqu'on met les facilités d'une banque centrale à la disposition d'entités non bancaires actives sur le marché des bons du Trésor. En élargissant l'accès à ces facilités, on stimule la négociation et on réduit la prime associée à l'émission courante, mais on constate une hausse des échecs de règlement et, bien que ce soit contre-intuitif, la situation ne profite qu'aux négociants principaux.

Sujets : Marchés financiers; Évaluation des actifs; Structure de marché et établissement des prix; Politique monétaire
Codes JEL : G12, G19, G23

1 Introduction

With an average daily trading volume of around a trillion U.S. dollars, the U.S. Treasury market is one of the most important and liquid markets in the U.S. financial system, crucial to the conduct of monetary policy and a key pillar of the U.S. economy.[1] Despite its importance, the U.S. Treasury market exhibits some irregularities.

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 maturity dates, giving rise to a puzzling arbitrage opportunity known as the "on-the-run" premium (see Vayanos and Weill (2008), D'Amico and Pancost (2022), Figure 2a in the next section).[2] 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 in the next section). Third, on average, USD 33 billion of Treasury contracts fail to settle each day (see Figure 3a in the next section). Fourth, 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 in the next section).

These stylized facts and irregularities raise the following questions: How can there be a premium on certain Treasuries but not on others with almost identical cash flows? Why is the premium always on the on-the-run 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 U.S. 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 U.S. Treasury market to answer these questions. The model solves a dynamic inventory problem with a sequence of competitive forward markets and OTC spot markets with bilateral bargaining. In the second part, I use the model to conduct a policy analysis motivated by current discussions about how to restructure the market. In the third part of the paper, I test several hypotheses derived from the theory in parts 1 and 2, using U.S. Treasury market primary dealer data.

My model incorporates key features of the U.S. Treasury market: It is an over-the-counter (OTC) market where primary dealers are the first acquirers of Treasuries at the primary auction. 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.[3] Buyers have the highest valuation for Treasuries, but the market is segmented and they cannot contact primary dealers directly, only through sellers.[4]

The model is dynamic and has an infinite horizon. In a first, fully centralized competitive market,

[1]The data is provided by the Securities Industry and Financial Markets Association and can be downloaded here: https://www.sifma.org/resources/research/statistics/us-treasury-securities-statistics/. The average daily trading volume is calculated based on the year 2024.

[2]See Figure A.1 in the Appendix for a graphical representation of the on-the-run cycle.

[3]The auction is included in the extension in Section A.3.

[4]This refers to the Treasury core-periphery market structure where a client contacts an intermediary to acquire an asset from a (primary) dealer.

called contract 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 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 they deliver 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 interest on them. Sellers then go into the next sequence of trades.

I show that in equilibrium, 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, the primary dealers' inventories vary less among each other, and sellers are more likely to find them. They have a greater chance of being settled. Therefore, on-the-run Treasuries trade at a premium. Inventory uncertainty, rather than scarcity per se, is key to the emergence of the premium, although their occurrence coincides.[5] In addition to the premium, on-the-run Treasuries also trade in larger volumes than off-the-run Treasuries with the same cash flow and maturity date. This is consistent with the stylized facts for the U.S. Treasury market mentioned at the beginning.

In the 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 by Duffie (2023)). The background to this discussion is epitomized by the U.S. Treasury market crisis of March 2020. In the aftermath of the Great Financial Crisis of 2007, primary dealers faced tighter regulatory constraints, leading them to reduce their balance-sheet space for Treasuries. At the same time, the U.S. Treasury market grew strongly. Non-bank financial institutions have filled the space left by primary dealers.[6] But in March 2020, asset sales of non-banks in the U.S. 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 U.S. 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.[7]

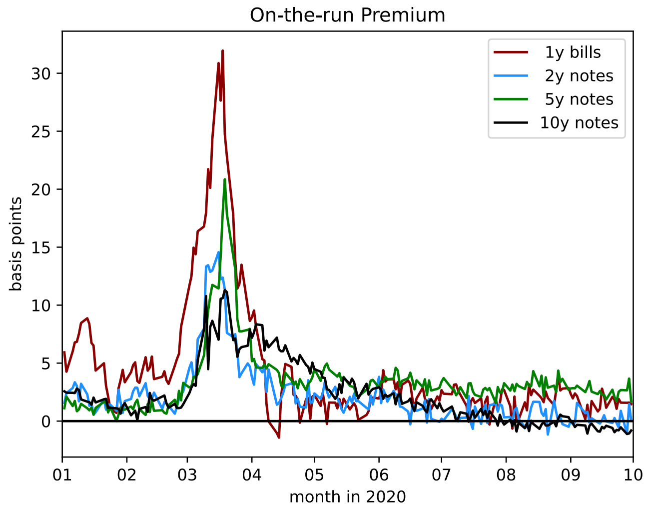
Among their ten recommendations for the U.S. Treasury market, the Working Group on Treasury Market Liquidity led by Duffie, Geithner, Parkinson and Stein, recommends broad access to central bank

[5]By scarcity, I mean the number or outstanding amount of a specific Treasury. Inventory uncertainty refers to differences in the inventory size of a specific Treasury among primary dealers. With my model, I show that if a Treasury can be found and settled with certainty, its price is not influenced by how scarce it is.

[6]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 notes and bonds in 2015. Traditional banks and dealers had a share of 30–40%.

[7]As in Christensen et al. (2017), the on-the-run yield is subtracted from the par yield of seasoned bills and notes. 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/nominal-yield-curve.htm, and the FRB-H15 Tables, https://www.federalreserve.gov/releases/h15/.

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



Notes: This figure shows the on-the-run premium at the start of the COVID-19 pandemic. Source: FED and author's calculations.

repo financing (Duffie et al. (2021)). They criticize the current repo facility for providing access only to primary dealers and banks rather than to a wide range of market participants. A facility to which access has increased substantially is the FED's reverse repo facility (Frost et al. (2015), Baklanova et al. (2015), Marte (2021)).

My model helps to understand the impact of broad access to central bank facilities on prices, premia, traded quantities, fails and profits (in a general setting in normal times). A first observation is that the facility is not a substitute for trading, but rather complements it. In this model, the facility (like a deposit or reverse repo facility) generates a certain return on liquid funds between trades for those who have access.[8] This feeds back into the overall cost of trading. If sellers, who sell the assets short, 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 fewer in the inventories of primary dealers. This is per se positive in terms of welfare as more Treasuries are held by the agent who has the highest (marginal) valuation of them. At the same time, the settlement of off-the-run Treasuries is more likely to fail. The reason being that as more Treasuries are sold early, fewer are available during the off-the-run period.

Interestingly, the on-the-run 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 a default, and 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

[8]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 security that falls into a broad class. The lender does not search for a specific security (Bowman et al. (2017)). Second, even if the facility was to provide a specific security that was sought, the security would have to be returned the next day and the facility would only provide temporary availability.

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.[9] It is worth noting that although I focus on facility access, the result can be interpreted more broadly. Any policy intervention or market change that affects settlement risk by reducing the cost of failing impacts the market and, especially, the premium in this way.

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. In contrast, 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 internalize the fact that if they buy more Treasuries early, fewer will be available during their off-the-run period, implying a higher default rate.[10]

In the third part of the paper, I test three hypotheses derived from my theoretical results. The first two hypotheses test the main narrative of the model: First, I conjecture that lower primary dealer inventories coincide with more settlement fails in off-the-run Treasuries, and second, that more such settlement fails imply higher on-the-run premia. Both of these results are predicted by the model and confirmed in the data. Lastly, I turn to the main result in the policy part regarding the on-the-run premium. I test the following hypothesis: The on-the-run premium decreases as the FED's reverse repo rate increases. The data aligns with this hypothesis as well.

Related Literature First and foremost, my paper is related to the literature on on-the-run premia. Vayanos and Weill (2008) provide a theoretical framework for the on-the-run premium based on a set-up with two assets with identical cash flows. Agents can go long or short on 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 all models have limitations, the authors point out that a main limitation of their model is that there are two equilibria with an on-the-run premium on both of the two assets. They cannot explain why short sellers systematically concentrate on on-the-run assets and consider the lower effective supply of off-the-run assets as a possible reason. In the first part of my paper, I address this limitation and provide a fundamental reason for the preferred choice of on-the-run assets. I include a key factor to explain the on-the-run premium: the fact that the off-the-run asset has been available on the market for a longer time. This reason is directly rooted in the definition of on- versus off-the-run, as the only difference between them is age. Compared to the static model of Vayanos and Weill (2008), my dynamic model includes this time dimension while limiting the model in other dimensions. As I show, the inclusion of the time dimension allows for an equilibrium choice with the premium on the on-the-run asset. Although the effective supply of the off-the-run asset does indeed decrease over time (i.e., higher "scarcity"), as suggested by Vayanos and Weill (2008), it is the increasing inventory uncertainty over time that is key for driving the equilibrium selection, rather than scarcity per se, although the two are closely related.

[9] This policy effect is particularly relevant in the context of the recent crisis, where the market for off-the-run Treasuries froze (Eren and Wooldridge (2021)).

[10] Few of the results depend on search frictions being above a small minimum. See Appendix A.6.6.

Another theory of the on-the-run premium is that of Pasquariello and Vega (2009). They attribute the on-the-run premium to auction-driven endowment shocks, using a model with frictions from information heterogeneity and imperfect dealer competition. My model similarly incorporates uncertainty, but it concerns the stock of off-the-run Treasuries. Unlike theirs, my model assumes perfect competition and symmetric information among traders (called sellers in my model).

From a broader perspective, compared to Vayanos and Weill (2008) and Pasquariello and Vega (2009), my model concentrates on the time dimension and focuses on primary dealer inventory and intermediary settlement risk[11] to explain the premium.

The distribution of the inventories of an asset across primary dealers over time, implied by OTC frictions, is key in my model. Models of inventory management in OTC markets are provided by Cohen et al. (2024), Lagos et al. (2011) and Weill (2007). In contrast to them, my model is not about dealers optimally managing their inventories by buying and selling assets. In my model, primary dealers acquire the newly issued Treasuries (e.g., from an auction) and they "only" manage their inventory optimally by deciding when to sell it over time.

Empirical work on the on-the-run premium includes, for example, Strebulaev (2002), Goldreich et al. (2005), Graveline and McBrady (2011) and D'Amico and Pancost (2022). They all have different arguments as to why on-the-run premia arise (e.g., tax reasons, collateral value risk).[12] None of them link the premium to inventory uncertainty and settlement risk, as I do in the empirical part where I test my theoretical argument.

The second part of the paper addresses the current policy debate. Recent work on the microstructure of the Treasury market and the discussion on how to reform it includes Chaboud et al. (2025), Duffie (2020), Duffie et al. (2021), Duffie (2023), Durham and Perli (2023), Fleming and Keane (2021), He et al. (2022), Schrimpf et al. (2020) and Vissing-Jorgensen (2021). Papers specific to non-banks in the Treasury market and their access to the reverse repo facility are Doerr et al. (2023), Eren and Wooldridge (2021) and Frost et al. (2015). To the best of my knowledge, there is no paper that examines, either theoretically or empirically, the impact of access to central bank facilities on the trading of benchmark assets and their premia. The closest paper 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 analyze how purchases by the European Central Bank have affected repo specialness[13] in the Italian government bond repo market during the euro area sovereign debt crisis. They show that purchases reduce liquidity and increase specialness in the presence of short selling. They also provide evidence that older assets with lower turnover are more likely to fail, similar to my results.

The structure of the paper is as follows: Section 2 describes the U.S. Treasury market. Section 3

[11] The literature on settlement failures includes Fleming et al. (2014), Fleming and Garbade (2002), Fleming and Garbade (2004), Fleming and Garbade (2005), Garbade et al. (2010).

[12] 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. Graveline and McBrady (2011) link the premium to the demand for hedging interest rate risk. D'Amico and Pancost (2022) link the on-the-run premium to the risk of unexpected fluctuations in the collateral value of Treasuries.

[13] Specialness is the premium paid to obtain a particular security in the repo market. On-the-run securities are often traded as "special". Important early work on this topic was done by Duffie (1996) and Krishnamurthy (2002).

contains the theoretical part of the paper. Subsection 3.1 describes the environment and Subsection 3.2 the value functions. The equilibrium is also defined. Subsection 3.3 proves the existence of my main equilibrium of interest. Subsection 3.4 discusses how the model explains the stylized facts and why the premium is always on the on-the-run Treasury. The implications of broad access to central bank facilities are also analyzed. Section 4 contains the empirical part of the paper. Subsection 4.1 derives three hypotheses from the theoretical results, which are tested in the next two subsections. Subsection 4.2 empirically analyzes the relationship between the on-the-premium, settlement fails and primary dealer inventories. Subsection 4.3 empirically examines the dependence of the on-the-run premium on the reverse repo facility rate. Section 5 concludes.

2 Description of the Treasury market

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

The Treasury spot market, particularly the dealer-to-client segment, is mostly OTC (Fleming et al. (2018), Chaboud et al. (2025)). This means that there is no all-to-all trading at a central venue and no central pricing.[14] 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), U.S. Department of the Treasury et al. (2015)).[15]

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 flows and maturity dates.[16] They also have higher prices and lower repo rates.[17]

Also, despite being more expensive and far lower in outstanding amount, on-the-run Treasuries trade in much larger volumes than off-the-run Treasuries, as shown in Figure 2b.[18] This is true whether I look at dealer-to-customer or interdealer and automated trading system (ATS) trades.[19]

Another surprising fact is that, on average, USD 33 billion of Treasuries are not delivered on time

[14]Chaboud et al. (2025) discuss the advantages and disadvantages of introducing all-to-all trading in the U.S. Treasury market.

[15]The overall share of trading on all types of electronic platforms in the U.S. Treasury market is 70% (Bech et al. (2016)).

[16]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/nominal-yield-curve.htm, and the FRB-H15 Tables, https://www.federalreserve.gov/releases/h15/.

[17]The term "very similar" refers to the fact that, in general, no two Treasuries in the market have exactly the same cash flow and maturity date 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 compares Treasuries with the same overall maturity using an estimated off-the-run yield curve (see, e.g., the first figure in Christensen et al. (2017)), abstracts from small differences in cash flows and maturity dates, or compares on- and off-the-run Treasuries with the same maturity date that have a different overall maturity (see, e.g., 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 bills and notes.

[18]Trading volumes in on-the-run and off-the-run Treasuries are the volumes reported to TRACE between February 2023 (first available data) and December 2024. The TRACE data is available here:
https://www.finra.org/finra-data/browse-catalog/about-treasury/monthly-file.

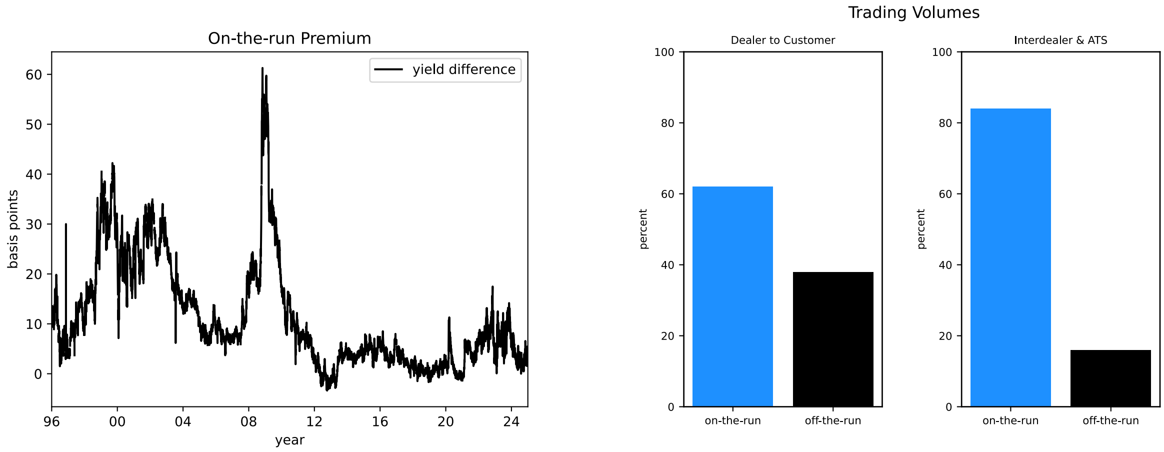
[19]See Figure 1 in Barclay et al. (2006) for a graphical representation of how the trading volume collapses on the day the Treasury goes off-the-run.

to settle a contract each day.[20] These events are commonly referred to as “settlement fails”. Figure 3a shows the 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.[21]

Figure 2: Yields and volumes

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

(b) Trading volumes



Notes: The left figure shows the on-the-run premium for 10-year Treasuries. The right figure shows the trading volumes for on- and off-the-run Treasuries separately in the Dealer-to-Customer and Interdealer & ATS segments. Source: a) FED and authors calculations, b) TRACE and authors calculations.

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 for 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.[22] Fleming and Keane (2021) write that on-the-run Treasury fails account for less than a quarter of all fails in non-crisis periods.[23] Fleming

[20] The data are 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).

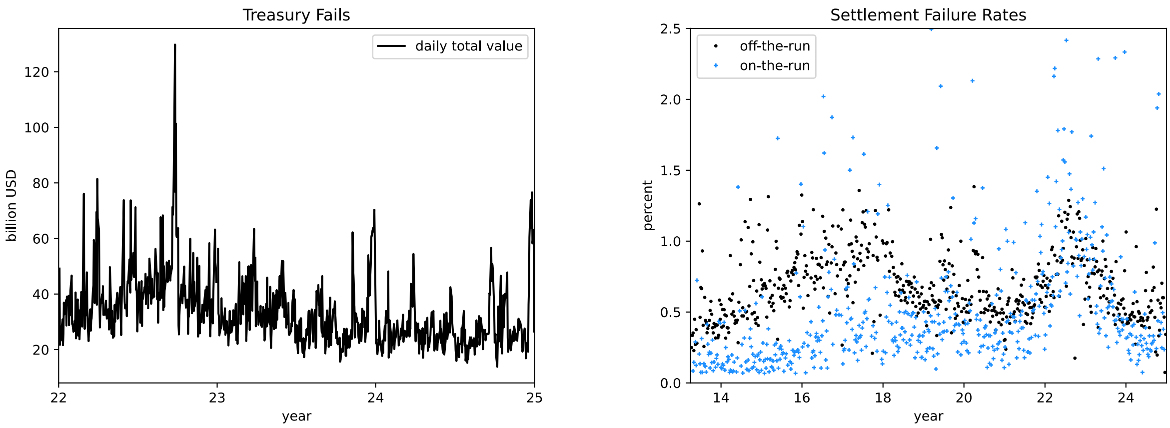
[21] The data are from the FED’s Primary Dealer Statistics. They include outright and financing fails. The median failure rate for on-the-run Treasuries is 0.38% and the median failure rate for off-the-run Treasuries is 0.64%. The rates are not an exact measure. This is mainly because one part of the 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.

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

[23] 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.

Figure 3: Fails

(a) Daily settlement fails (b) Fails of on- and off-the-run Treasuries



Notes: The left figure shows the value of Treasuries that were not delivered on time to fulfill a contract. The figure on the right shows the percentage of Treasuries that failed to settle separately for on- and off-the-run Treasuries. Source: a) DTCC, b) FED and authors calculations.

et al. (2014) show that gross fails are much higher in seasoned Treasuries (issued more than 180 days ago) 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 absent as relatively expensive Treasuries also have lower repo rates (Krishnamurthy (2002)).

3 Theory

In this section, I develop a model of the Treasury market to explain how differences in yields, trading volumes and settlement fails among different generations of Treasuries are interconnected. I use the model to explain the on-the-run phenomenon. Based on the model, I also provide an analysis of broadening access to central bank facilities, speaking to the current policy debate. In addition, I lay the foundation for three hypotheses that I test in the empirical section that follows.

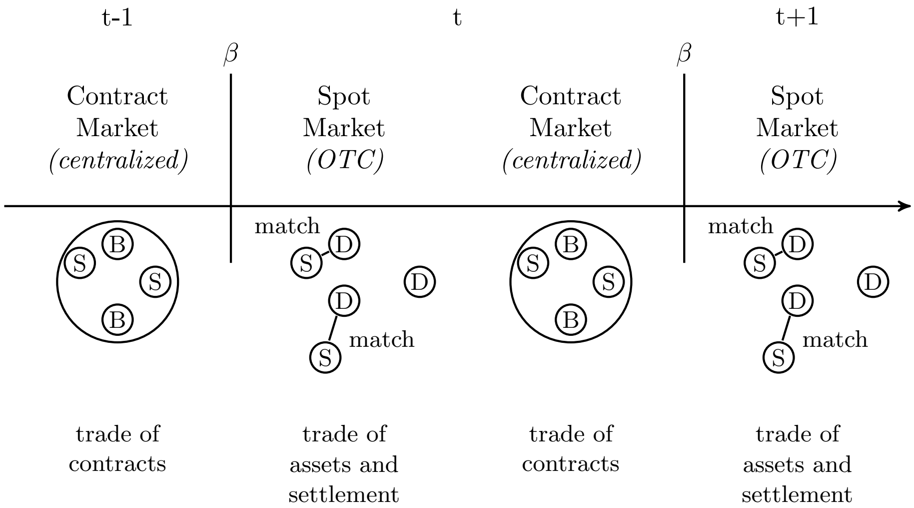
3.1 The environment

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.[24] 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. Figure 4 gives an overview of the timeline.

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

Figure 4: Timeline



Notes: The figure gives an overview of the timeline. In each period, a spot market with matching frictions is followed by a fully centralized market. In the spot market, assets are traded and settlement takes place. In the centralized market, contracts are traded.

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. Assets are storable and divisible. 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 on-the-run and off-the-run states, respectively. I will refer to them 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.[25] Figure 5 gives an overview of the assets and their cash flow.

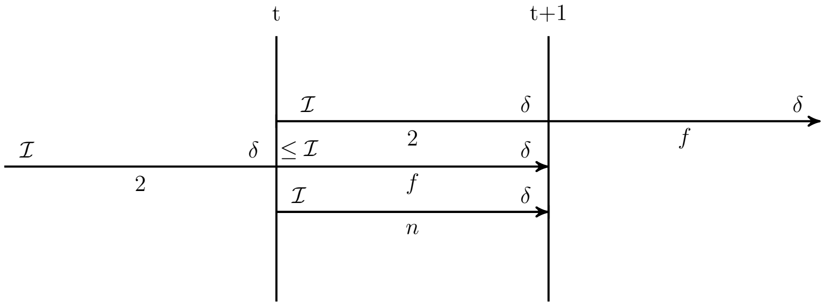
The figure illustrates that the on-the-run (n) and off-the-run (f) assets are identical in terms of maturity date and coupon. The only difference is the issuance date. Therefore, I compare these two assets to measure the on-the-run premium.[26] At the beginning of each period, primary dealers are endowed with a stock of newly issued two-period assets, $I^2 = \mathcal{I} \in \mathbb{R}^+$, and a stock of newly issued one-period assets, $I^n = \mathcal{I} \in \mathbb{R}^+$. 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.

All agents have linear utility δ from consuming the coupons in the second subperiod. Buyers addi-

[25] In only a very few contexts are both types of on-the-run assets meant when using the term "on-the-run".

[26] 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 I would use the two-period on-the-run asset to calculate the premium, I 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.

Figure 5: Assets in period t



Notes: The figure gives an overview of the available assets in period t. The new one-period and two-period assets (denoted by n and 2, respectively) and the two-period assets issued one period earlier (denoted by f) are available.

tionally receive utility g each period when holding an asset.[27] Sellers do not value the coupons.[28] 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.[29] It is used to settle trades. It has properties similar to money except that it is a real asset.[30]

In the OTC spot market, primary dealers and sellers trade assets. Primary dealers have a match with probability $(1 - \sigma) > 0$ with a seller. Only matched primary dealers can trade. Sellers always have a match with a primary dealer.[31] 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 an adjustment cost of $\kappa(A^i)$ when selling assets i in terms of settlement good.[32] The function introduces a nonlinearity into the model and leads to an interior solution and a determined price.[33] I assume that $\kappa(0) = 0$, $\kappa'(A^i) > 0$, $\kappa''(A^i) > 0$, and that the function is continuous.[34] Possible interpretations of the function are a nonlinear portfolio adjustment cost (see Gârleanu and Pedersen (2013), Bacchetta and van Wincoop (2021) for examples) or a regulatory cost (see, e.g., 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. Sellers cannot commit. $\omega^i a^i$ is the collateral (in settlement good) that the seller has to post at the moment of selling the contract. The buyer has first claim over the collateral in the event of non-delivery.[35] q^i is the contract price in

[27] g can be interpreted as a hedging benefit from holding the asset or simply as a different valuation.
[28] It would not change any results if sellers would value the coupons as much as primary dealers.
[29] 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 buyer or primary dealer would produce settlement goods for the seller so that the seller could deposit it and pay it back later with a profit is not possible because the seller cannot commit.
[30] The only difference from a nominal model is that the settlement good does not lose or gain value over time due to inflation. The dynamics would not change with a nominal model and, therefore, discussing inflation would not add anything relevant.
[31] I make this assumption for the sake of simplicity. Changing it would not change the dynamics.
[32] In equilibrium they only sell assets.
[33] The main results also hold without this function. It becomes relevant when I discuss access.
[34] Another possibility would be that the function depends on the sum of all assets sold. But I show that for a positive premium this cannot be the case. Also, the function needs to be increasing and be convex.
[35] The collateral in the form of settlement good is similar to the 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,

the second subperiod. $q^i a^i$ is the payment (in settlement good) due from the buyer at settlement in the next subperiod. Figure 4 gives an overview of 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_t on it. I assume that $\beta(1+r_t) < 1$. This implies that it is not worthwhile to accumulate settlement good one period in advance in order to deposit it in the facility.[36]

3.2 Value functions and equilibrium

3.2.1 Primary dealer

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

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

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 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, the dealer receives the price p_t^i and is faced with the cost $\kappa(A_t^i)$. In addition, the dealer cannot consume any future coupons of these assets but only of what is left of the inventory, $(I_t^i - A_t^i)$. With probability σ, the primary dealer has no match and consumes the coupons of their inventory of assets. The inventory of off-the-run assets, I_t^f, depends on whether the primary dealer had a match in the previous period and, if so, how much was traded. The inventory is therefore endogenous. 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. I assume that the primary dealer has full bargaining power and the seller therefore does not make a profit.[37] This means that the primary dealer sets the price just high enough 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:

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

The primary dealer maximizes their trade surplus. The Lagrange function to the dealer's maximization problem in each period t is given by[38]

see https://www.newyorkfed.org/tmpg and Garbade et al. (2010).

[36] I could, in addition, assume that primary dealers also have access to the facility, but this does not change the dynamics. Therefore, I omit it to ease notation.

[37] Where relevant, I relax this assumption later.

[38] If the primary dealer does not sell the two-period assets today, then the assets remain in their inventory in the next period if there is no match. If there is a match, I assume that a part is sold and the rest remains in the

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

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 they sell the asset today, they not only forgoe the coupon today but also tomorrow. For each type of asset, the primary dealer faces an inventory constraint, $A_t^i \leq I_t^i$. The dealer cannot sell more than what they have. In addition, the dealer 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 the seller's delivery constraints. The dealer 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 they deliver, they have to buy the asset at price p_t^i but they can keep their collateral ω_{t-1}^i.

The first-order conditions are

$$\begin{aligned} p_t^2 &= \kappa'(A_t^2) + \delta + \beta\delta + \lambda_t^2 + \widetilde{\lambda}_t^2 + \beta(1-\sigma)\lambda_{t+1}^f \\ p_t^f &= \kappa'(A_t^f) + \delta + \lambda_t^f + \widetilde{\lambda}_t^f \\ p_t^n &= \kappa'(A_t^n) + \delta + \lambda_t^n + \widetilde{\lambda}_t^n. \end{aligned} \tag{3}$$

The prices equal the marginal costs faced when selling the assets and take 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

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

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 did not have a match in the prior period have an inventory of $I_t^{f,h} \equiv I_t^f = \mathcal{I}$. The primary dealers who had a match have an inventory of $I_t^{f,l} \equiv I_t^f = \mathcal{I} - A_{t-1}^2$. For the sake of simplicity, trading between primary dealers is not considered in this model. I could also model trade between primary dealers. As long as there are frictions in the interdealer market that lead to primary dealers not fully balancing their inventories, the results are valid.[39] h stands for high and l for low, corresponding to higher and lower inventories, respectively. The probability of having no match with a seller is σ. Therefore, by the law of large numbers, a share σ of the primary dealers has

dealer's inventory as well. This means that the inventory constraint on the off-the-run assets is not binding if the dealer 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 $p_t^i A_t^i - \kappa(A_t^i) - \delta A_t^i - \beta\left[\sigma\delta A_t^2 + (1-\sigma)\left(p_{t+1}^f A_{t+1}^f - \kappa(A_{t+1}^f) + \delta\left(A_t^2 - A_{t+1}^f\right)\right)\right]$.

[39]See, e.g., Eisfeldt et al. (2023) for an analysis of price dispersion between dealers and intermediation capacity.

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 as $\lambda_t^{f,h}$ and that of the low inventory group as $\lambda_t^{f,l}$. I denote the sold off-the-run assets of the high inventory group as $A_t^{f,h}$ and those of the other group as $A_t^{f,l}$. 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 the main equilibrium of interest. In Section A.2 I discuss other equilibria. In these other equilibria, the main dynamics are the same. They are extreme cases of the main equilibrium.

3.2.2 The 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, I have to distinguish two cases: $\lambda_t^{f,l} > 0$ and $\lambda_t^{f,l} = 0$. A fraction $(1-\sigma)$ of primary dealers have already sold a part of their 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 less than the remaining stock.

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

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

The buyer's state variable is a_{t-1}^2. These are the assets they bought last period and which have not yet matured. For each asset, the buyer chooses how many they want to buy from the seller. For each asset they want to buy, they must build up the payment $q_{t-1}^i a_t^i$ in the form of settlement goods. The assets are delivered in the next period at settlement, and the buyer receives utility $(\delta+g)$ from each asset they hold. If $\lambda_t^{f,l} > 0$, then with the probability $(1-\sigma)$, the seller encounters a primary dealer who is constrained in their 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 $(a_t^f - I_t^{f,l})$, the buyer receives the collateral $\omega_{t-1}^f a_t^f$.

The first-order conditions are

$$\begin{aligned} q_{t-1}^2 &\geq (1+\beta)(\delta+g) \\ q_{t-1}^n &\geq (\delta+g) \\ q_{t-1}^f &\geq (\delta+g) + (1-\sigma)\left[\omega_{t-1}^f - (\delta+g)\right]\mathbb{I}_{\lambda_{t+1}^{f,l}>0}. \end{aligned} \tag{4}$$

The prices are greater than or equal to 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 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.

3.2.3 Seller

The seller's value function at the beginning of the contract market is[40]

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

The seller chooses the optimal number of contracts to sell to the buyer, i.e., for each asset, they choose the amount they are willing to deliver in the next period. The seller 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 buys the assets on the spot market. For each asset the seller can deliver, they receive the price q_{t-1}^i from the buyer, they pay the spot market price p_t^i to the primary dealer, and they can keep their accumulated collateral (all in the form of settlement goods).[41] If there are any remaining funds after the trade, the seller can deposit them in the central bank facility and receive the interest rate r on them.[42] If $\lambda_t^{f,l} > 0$, then with probability $(1-\sigma)$ the seller 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

$$
\begin{aligned}
&\omega_{t-1}^2 \geq \beta(1+r_t)(q_{t-1}^2 - p_t^2 + \omega_{t-1}^2) \\
&\omega_{t-1}^f \geq \beta(1+r_t)(q_{t-1}^f - p_t^f + \omega_{t-1}^f) + \beta(1+r_t)(1-\sigma)(p_t^f - \omega_{t-1}^f)\mathbb{I}_{\lambda_{t+1}^{f,l}>0} \\
&\omega_{t-1}^n \geq \beta(1+r_t)(q_{t-1}^n - p_t^n + \omega_{t-1}^n).
\end{aligned} \tag{5}
$$

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

3.2.4 Equilibrium and premium definition

Next, I define the equilibrium.

Definition 1 (Equilibrium). *An equilibrium consists of*

a) the contract and spot prices of all assets (q_{t-1}^i $\forall i$ and p_t^i $\forall i$),

[40]The seller never buys assets for themself. The reason is that there are negative gains from trade because the primary dealer and the seller value the asset the same but if they were to trade, they would face the adjustment cost.

[41]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 the seller buys on the spot market. Since this constraint is always satisfied (see Section 3.2.1), it is not added to the maximization problem.

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

b) the assets contracted (a_t^i $\forall i$) and sold ($A_t^2, A_t^n, A_t^{f,h}$, and $A_t^{f,l}$),

c) the collateral values (w_{t-1}^i $\forall i$)

and the primary dealer, the buyer, and the seller behave optimally given contract prices q_{t-1}^i and collateral values w_{t-1}^i ((3), (4), (5)), and the delivery constraints (1) are satisfied.

Given the equilibrium definition, I make two additional assumptions. First, 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 contract the maximum amount of assets that is profitable for the primary dealers to sell (if they are not constrained). This implies that in each equilibrium

$$\begin{aligned} a_t^2 &= A_t^2 \\ a_t^n &= A_t^n \\ a_t^f &= A_t^{f,h}. \end{aligned} \tag{6}$$

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 I concentrate on equilibria where $\lambda_t^n = \lambda_t^2 = \lambda_t^{f,h} = 0$, then also $\widetilde{\lambda}_t^n = \widetilde{\lambda}_t^2 = \widetilde{\lambda}_t^{f,h} = 0$.

Second, 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, $a_t^i \in (0, \infty)$ $\forall i$.

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

Definition 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, I 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.

3.3 Existence

In this section, $\lambda_t^2 = \lambda_t^n = \lambda_t^{f,h} = 0$. This means that primary dealers are unconstrained if they still have the full stock of assets available, i.e., $\mathcal{I} > max(A_t^2, A_t^n, A_t^{f,h})$. As argued in the previous section, 3.2.4 $\widetilde{\lambda}_t^n = \widetilde{\lambda}_t^2 = \widetilde{\lambda}_t^{f,h} = 0$ holds as well.

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 $\widetilde{\lambda}_t^{f,l} = 0$. Therefore, in this equilibrium $\mathcal{I} \in (max(A_t^2, A_t^n, A_t^{f,h}),\ A_{t-1}^2 + A_t^{f,h})$.[43] 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.

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

Proposition 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 $\mathcal{I} < A_{t-1}^2 + A_t^{f,h}$.*

Proof. See Appendix A.6.1. □

I need constrained primary dealers for an equilibrium with a positive premium because this gives rise to settlement fails. The fails imply the premium (see Section 3.4). Without settlement fails, both assets are priced the same, since they are perfect substitutes in this case.[44] Therefore, the equilibrium candidate is the equilibrium where $\lambda_t^{f,l} > 0$.

Proposition 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 $\sigma > \frac{1-\beta(1+r_t)}{\beta(1+r_t)} \frac{\delta}{g}$, which implies positive trade in all assets. I can always find issuance sizes $\mathcal{I}$ where this equilibrium exists.*

Proof. See Appendix A.6.2. □

For positive demand in off-the-run assets, the probability of a settlement fail, $(1 - \sigma)$, cannot be too high. Therefore the condition. Given the condition that the proposition is satisfied, I show in the proof that I can always find an $\mathcal{I}$ where the 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 "premium equilibrium".

Definition 3 (Premium equilibrium). *The premium equilibrium is the equilibrium where $\lambda_t^{f,l} > 0$ and all other Lagrange multipliers are zero.*

I will restrict the further analysis to the premium equilibrium. In Appendix A.2, I discuss other equilibria. The dynamics and intuition are the same as in the premium equilibrium.

I summarize the main result of this section as follows: For an equilibrium with a premium, some primary dealers need to be inventory constrained. The equilibrium always exists if the probability of finding the off-the-run assets is high enough such that buyers and sellers want to trade them.

3.4 Premium equilibrium

In this section I analyze and discuss the premium equilibrium. The first three subsections focus on the on-the-run phenomenon itself. The last subsection focuses on the policy discussion of broadening access to central bank facilities.

3.4.1 Graphical example with the two-period asset

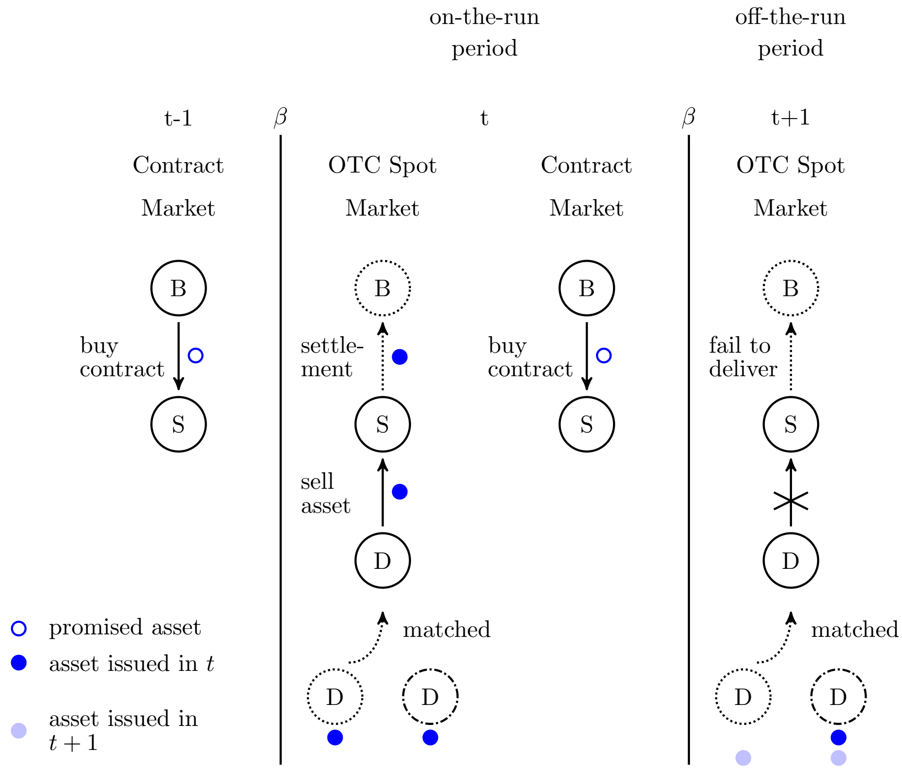
To discuss the relevant dynamics in the premium equilibrium, I illustrate the life cycle of two two-period assets issued in period t, in Figure 6 below (dark blue dots).

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 their inventory. Here I restrict the inventory to one asset for illustrative purposes. On the spot market, the seller is matched with one of the two primary

[44] 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.

dealers, buys the asset and delivers it to the buyer against a payment in the form of settlement good (not illustrated).[45] The other primary dealer is not matched with a seller and keeps their asset in their inventory. After the spot market, the contract market takes place in the second subperiod. The buyer again buys one contract, promising the delivery of this asset. In period $t+1$, the assets are off-the-run (as new assets are issued).[46] On the OTC spot market, the seller is matched with the primary dealer who was able to sell their asset in the prior period. A settlement fail occurs. Nevertheless, it is optimal for the buyer to initially buy one contract. The buyer takes the probability of a settlement fail into account when taking their decision.

Figure 6: Dynamics



Notes: The figure illustrates the life cycle and trading of two two-period assets issued in period t.

The figure shows only the two-period asset. But the model also contains the one-period on-the-run asset. It is easy to see that settlement fails occur for the 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. As I show in the next subsection, in equilibrium 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.

To summarize, the time since issuance is the only feature that distinguishes the two assets. And it is precisely this difference, combined with OTC market frictions and delivery constraints, that leads to

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

[46] In my model, the assets are always off-the-run after one period because new assets are issued (here depicted by the light blue dots). For the sake of simplicity, the buyer does not buy any contract, promising the delivery of these new assets, in this example.

settlement fails and hence the premium.[47] Finally, it is worth pointing out that off-the-run assets are scarcer (see Figure 6). However, it is uncertainty, not scarcity per se, that causes the premium. A scarce asset that could be bought without uncertainty would not lead to fails and the premium. The next subsection discusses this further.

3.4.2 On-the-run premium and inventory uncertainty

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

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

and

$$\begin{aligned} A_t^n &= \kappa'^{-1}\left[\beta(1+r_t)(\delta+g)-\delta\right] \\ A_t^{f,h} &= \kappa'^{-1}\left[\frac{\sigma}{1-(1-\sigma)(1+r_t)\beta}\beta(1+r_t)(\delta+g)-\delta\right] \\ A_t^{f,l} &= \mathcal{I}-A_{t-1}^2 \\ \kappa'(A_t^2) &= W+\beta(1-\sigma)\kappa'(\mathcal{I}-A_t^2) \end{aligned}$$

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

I observe from the equations that the spot prices directly depend on the buyer's "valuation" of the assets $(\delta+g)$. 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 with respect to the premium 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 former is that when dealers sell these two-period assets, they take into account that less can be sold tomorrow due to the binding inventory constraint. This binding inventory constraint is also why the amount of off-the-run assets traded by constrained primary dealers depends on $\mathcal{I}$.

Lastly, I derive the on-the-run premium. I deduct p_t^f from p_t^n. This gives rise to the premium according to the following proposition:

Proposition 3. *The on-the-run premium is given by* $\Delta = \left[1-\frac{\sigma}{1-(1-\sigma)\beta(1+r_t)}\right]\beta(1+r_t)(\delta+g) > 0$. *If* $\sigma \to 1$, *then* $\Delta \to 0$.

[47] Without settlement fails, both assets would be priced only according to the marginal utility that the coupons (including the additional utility g) 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.

The on-the-run premium depends on the buyer's asset "valuation" $(\delta + g)$, the probability of finding the off-the-run assets σ, and $\beta(1 + r_t)$, which is the discount factor cost for the seller for binding collateral. If assets would be found with certainty in the second period, i.e., $\sigma \to 1$, then the premium would vanish.[48]

It is important to point out that key for a positive on-the-run premium to arise is

$$I_t^{f,h} \neq I_t^{f,l}.$$

In the premium equilibrium, unconstrained primary dealers have a full inventory of off-the-run assets, i.e., $I_t^{f,h} = \mathcal{I}$, and constrained ones have a reduced inventory, i.e., $I_t^{f,l} = \mathcal{I} - A_{t-1}^2$ with $A_{t-1}^2 > 0$. Therefore, $I_t^{f,h} > I_t^{f,l}$. It is this difference in inventories that leads to the uncertainty that implies settlement fails. The fails themselves imply the premium. Compared to Vayanos and Weill (2008), the premium is always on the on-the-run asset.

To make the distinction between uncertainty and scarcity clearer, let us discuss what would happen if in the premium equilibrium $I_t^{f,h} = I_t^{f,l}$ (although this case never occurs). This means that there is no uncertainty on off-the-run inventories as they are all the same.

Proposition 4. *If $I_t^{f,h} = I_t^{f,l}$ would hold in the premium equilibrium (no inventory uncertainty), then $\Delta = 0$ independent of the size of $I_t^{f,h}$ and $I_t^{f,l}$ (i.e., asset scarcity).*

Proof. See Appendix A.6.3. □

Since all off-the-run inventories are equal under my assumption, the seller never sells and the buyer never buys more contracts of off-the-run assets than the amount available in the inventory. Otherwise, fails would be certain. At the same time, all promised assets are delivered with certainty because all primary dealers can supply them given their equal inventories. This implies that there are no fails and, therefore, there is no premium. Importantly, this result holds regardless of the size of the inventory. Even if the asset is very scarce, i.e., inventories are very small, the assets are found with certainty. Therefore, uncertainty is necessary for the existence of the premium. At the same time, we have seen that under uncertainty, which is always the case in my premium equilibrium, there is always a positive premium (see Proposition 3) independent of scarcity, since I have made no assumption about the size of $\mathcal{I}$. Nevertheless, it is important to point out that the occurrence of uncertainty and scarcity correlate positively in the model.

3.4.3 Matching the stylized facts

In the introduction I described four stylized facts: First, on-the-run assets are more expensive than off-the-run assets (positive on-the-run premium). Second, they trade in larger volumes. Third, settlement fails occur, and lastly, off-the-run assets fail to settle more often. To check if my model matches those facts, I first have to define the settlement failure rate. I define the settlement failure rate of asset i as

[48] The premium also vanishes if $\sigma = 0$, but this equilibrium is not a premium equilibrium, and the formula above does not hold.

the value of assets i involved in a fail divided by the overall amount of assets promised to be delivered:

$$f_t^i \equiv \frac{p_t^i \mathbb{I}_{a_t^i > I_t^i} \mathbb{P}_t^i (a_t^i - I_t^i)}{p_t^i a_t^i}$$

where $\mathbb{P}_t^i$ is the failure probability.[49] As $\lambda_t^{f,h} = 0$, it follows that $f_t^n = 0$ and as $\lambda_t^{f,l} > 0$, it follows that $f_t^f = \frac{(1-\sigma)p_t^f(a_t^f - I_t^f)}{p_t^f a_t^f} = (1-\sigma)\left(1 - \frac{\mathcal{I} - a_t^2}{a_t^f}\right)$.

I show that the equilibrium is consistent with the four stylized facts in the following proposition:

Proposition 5. *In the premium equilibrium*

$$\begin{aligned} &p_t^n > p_t^f \\ &A_t^n > \sigma A_t^{f,h} + (1-\sigma) A_t^{f,l} \\ &f_t^f > 0 \\ &f_t^f > f_t^n. \end{aligned}$$

Proof. See Appendix A.6.4. □

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 counter-intuitive 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 (U.S. Department of the Treasury)). The model can be generalized by having a non-uniform distribution across dealers (e.g., due to different auction allocations across primary dealers) with buyers being aware of the distribution but not the individual holdings. This would also generate fails for the on-the-run assets, as observed in the data. But as long as the off-the-run assets have a higher probability of failing (e.g., because of a growing dispersion over time), the results hold.

Before moving on to the policy discussion, I summarize the results of the first part on the on-the-run phenomenon as follows: The on-the-run premium is due to differences in inventories of off-the-run assets as they are in the market longer. The reason is as follows: Some of the off-the-run assets are locked up in buy-and-hold portfolios because they were 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 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 carry a premium, i.e., they are more expensive on the spot market and trade in larger quantities than off-the-run assets.

[49] An alternative way to analyze settlement fails would be to compare the overall value of assets involved in fails (see e.g., Fleming et al. (2014)).

Finally, it is worth pointing out that this basic model can be extended in such a way that it captures the entire life cycle of a Treasury (from the primary auction to going off-the-run). In Appendix A.3, I present an extension of the basic model that includes auction prices and explains two more stylized facts of the market.

3.4.4 Policy discussion: Central bank facility access

How does broadening access to a central bank's reverse repo or deposit facility to any type of intermediary affect the Treasury market (in normal times)? 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, even if the facility were to provide a specific Treasury, the Treasury would have to be returned the next day and would only temporarily ease availability.

I analyze the situation where all sellers have or gain access to the central bank facility. In Appendix Section A.4 I do the equivalent analysis for heterogeneous sellers, meaning that some already have access and others gain it. The facility provides liquidity between transactions. It can be accessed every first subperiod for one subperiod. Sellers can deposit settlement good and receive interest rate r_t on it. Sellers do this because they have a positive net settlement good position after the trades; they received the contract price q_t^i and had to pay (in case of delivery) the lower spot price p_t^i and could keep the collateral ω_t^i. 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 (in normal times) when any type of financial firm other than a primary dealer is given access to the facility.

I still look at the premium equilibrium. I assume a permanent unanticipated increase in the facility rate r_t at the beginning of the contract market in $t = \widetilde{t} - 1$. In addition to the situation where all sellers already have access and the facility rate increases, the increase can also represent the situation where all sellers gain access to the facility and because of the facility, now face a higher interest rate than before (with no interest or a lower market rate). I assume that there is still a positive premium after the sellers gain access. The results of the analysis can be summarized as follows:

Corollary 1. *An unanticipated and permanent increase in r_t in the second subperiod in $t = \widetilde{t} - 1$*

a) increases all spot prices, $\frac{dp_t^i}{dr_{\widetilde{t}-1}} > 0 \, \forall i$ and $\forall t \geq \widetilde{t}$,

b) decreases the on-the-run premium, $\frac{d\Delta_t}{dr_{\widetilde{t}-1}} < 0 \, \forall t \geq \widetilde{t}$,

c) increases the quantities of on-the-run and off-the-run assets traded if no inventory constraint binds, $\frac{dA_t^n}{dr_{\widetilde{t}-1}} > 0$ and $\frac{dA_t^{f,h}}{dr_{\widetilde{t}-1}} > 0 \, \forall t \geq \widetilde{t}$, *d) implies that the quantities of off-the-run assets traded initially, i.e., in* $t = \widetilde{t}$, *stay the same and then decrease if inventory constraints bind,* $\frac{dA_t^{f,l}}{dr_{\widetilde{t}-1}} = 0$ *for* $t = \widetilde{t}$ *and* $\frac{dA_t^{f,l}}{dr_{\widetilde{t}-1}} < 0$ $\forall t > \widetilde{t}$ *if* $\sigma > \widetilde{\sigma}$,

e) implies that overall more assets are offloaded from the inventories and end up in the portfolio of the buyer, i.e., the holder with the highest marginal asset valuation, $\frac{d(A_t^2 + A_t^n + \sigma A_t^{f,h} + (1-\sigma)A_t^{f,l})}{dr_{\widetilde{t}-1}} > 0$ *for* $t = \widetilde{t}$ *and* $\forall t > \widetilde{t}$ *if* $\sigma > \widetilde{\sigma}$,

f) increases the settlement failure rate of the off-the-run asset, $\frac{df_t^f}{dr_{\widetilde{t}-1}} > 0$ *for* $t = \widetilde{t}$ *and* $\forall t > \widetilde{t}$ *if* $\sigma > \widetilde{\sigma}$.

Proof. See Appendix A.6.6. □

The value $\widetilde{\sigma}$ is the value of σ above which always (but not only) $\frac{\partial A_t^2}{\partial r_{\widetilde{t}-1}} > 0$ $\forall t \geq \widetilde{t}$. I define it in Appendix A.6.6 and show that with a reasonable calibration it is an empirically very small value.

The intuition for this result is as follows: An increase in the facility rate 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. The higher the interest rate, the higher the profitability. 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 and quantities traded by unconstrained primary dealers increase.

Off-the-run prices and quantities react more strongly than their on-the-run equivalents, and the premium falls. The reason is that the case of non-delivery for the off-the-run asset is less costly than before because the collateral values increase due to the rise in prices. This reduces the spread between the real valuation of holding the asset and the collateral. This effect additionally triggers the demand for off-the-run assets. 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)).

The amount of off-the-run assets traded by constrained dealers remains the same in the first period after the rate hike because inventories are determined from the previous period. Later, it decreases because the inventories of the constrained dealers are smaller because more assets have already been sold during their on-the-run period (if search frictions are above a small minimum value of $\widetilde{\sigma}$). This is also the main reason for the increase in the settlement failure rate of off-the-run assets. The assets are not available later and more fails occur. Nevertheless, due to the interest rate increase, more assets end up in the portfolio of the buyer, the agent with the highest marginal asset valuation.

Putting all the above results in a broader context, I conclude that any kind of policy that lowers the costs of trade and intermediation can trigger the effects described. Access is one possibility.

The next result is about who benefits from an increase in the facility rate r_t. I look at the impact on the lifetime values of the buyer, the seller, and the primary dealer. There are two groups of primary dealers: One group of primary dealers has an inventory of $\mathcal{I}$ off-the-run assets and the other group has an inventory of $\mathcal{I} - A_t^2$. I take the lifetime value of both groups and average them according to their proportion of the population. To simplify the notation, I define $V_{\widetilde{t}-1}^{D,a} \equiv (1-\sigma)V_{\widetilde{t}-1}^{D}(\mathcal{I} - A_t^2) + \sigma V_{\widetilde{t}-1}^{D}(\mathcal{I})$.

Corollary 2. *An unanticipated and permanent increase in* r_t *in the second subperiod in* $t = \widetilde{t} - 1$

a) does not affect the lifetime value of sellers, $\frac{dV^s_{\widetilde{t}-1}}{dr_{\widetilde{t}-1}} = 0$,

b) decreases the lifetime value of buyers, $\frac{dV^b_{\widetilde{t}-1}}{dr_{\widetilde{t}-1}} < 0$,

c) increases the lifetime value of primary dealers, $\frac{dV^D_{\widetilde{t}-1}}{dr_{\widetilde{t}-1}} > 0$ *if* $\sigma > \widetilde{\sigma}$.

Proof. See Appendix A.6.7. □

As noted above, see Appendix A.6.6 for the definition of $\widetilde{\sigma}$.

The intuition for the result is as follows: Competition among sellers in the contract market erodes any positive profits for them to zero. The initial profitability increase 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 dealers' profits increase.

Buyers' utility falls. To gain intuition, I first explain how buyers' utility or benefit materializes 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.[50] But the first part of the assets bought, I^f_t, is found with certainty. The value the buyer places on these assets is therefore higher. Nevertheless, in equilibrium the seller pays the same price for all the assets and, therefore, they receive a small benefit.

An increase in the facility rate increases the off-the-run price and decreases the quantity of off-the-run assets 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 or utility 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 their inventory and takes into account that a two-period asset sold today cannot be sold in the next period.

I solved my baseline model under the assumption that the primary dealer has full bargaining power. I relax this assumption and prove that even when the seller has positive bargaining power, in equilibrium, the seller does not make a profit in contrast to the primary dealer.

Corollary 3. *Result 2 still holds even with the positive bargaining power of sellers.*

[50] A settlement fail implies that the seller has to give the buyer the collateral they had to build up during the contract market. Otherwise, the seller could have kept it.

When maximizing the joint surplus of both agents, the first-order conditions of the spot market problem of the primary dealers 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 decisions about the optimal number of contracts to sell and buy do not change. This is crucial because given the seller's first-order conditions and their value function, they make 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 they have no total profits. Otherwise, the seller would supply an infinite number of contracts or no contracts at all. I assume that this is not the case in equilibrium, which is a reasonable assumption. Therefore, even with positive bargaining power, the seller never profits from an increase in the facility rate if they have access.

4 Empirics

In this section I analyze whether my theory is consistent with the empirical evidence.

4.1 Hypothesis development

In my theory, primary dealer inventory risk is key to explaining the on-the-run premium. The reason is that inventory uncertainty leads to settlement fails. They are more frequent for off-the-run assets, which makes them unattractive compared to on-the-run assets, leading to a premium on the latter. My first two hypotheses follow this reasoning.

Hypothesis 1. *A higher failure rate of off-the-run Treasuries leads to a higher on-the-run premium.*

Hypothesis 2. *Lower primary dealer inventories lead to a higher failure rate.*

The first hypothesis follows directly from Proposition 3. There I show that if the probability of finding the off-the-run assets, σ, goes to 1 (no fails occur), then the premium vanishes. Otherwise, there are settlement fails of off-the-run assets that lead to the premium. Second, I conjecture that lower primary dealer inventories lead to settlement fails. Lower primary dealer inventories are expected to capture and correlate with inventory uncertainty, which is key for failures to arise (see the first three subsections of Section 3.4).

To develop my third hypothesis, I follow Corollary 1 where I show that an increase in the facility rate implies a decrease in the on-the-run premium. I therefore make the following hypothesis:

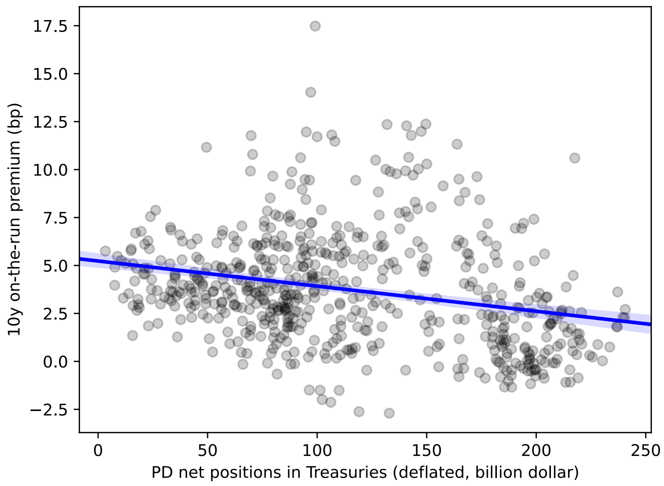
Hypothesis 3. *A higher reverse repo facility rate leads to a lower on-the-run premium.*

In the following two subsections I test the hypotheses. An overview of the data sources and a summary statistic of the data can be found in Appendix A.7.

4.2 Empirical evidence 1: On-the-run premium, fails and inventories

To advance the testing of my first two hypotheses, I first show the scatter plot between the 10-year on-the-run premium and the net outright positions of primary dealers in Treasuries, in Figure 7.[51] I observe that there is a negative correlation between the on-the-run premium and the net outright positions of the primary dealers (with tight 95% confidence bands). This aligns with my explanation in the theoretic part that inventory uncertainty leads to settlement fails, which imply the premium.

Figure 7: The correlation between the premium and the net positions



Notes: The figure plots the on-the-run premium on 10-year Treasury notes against the net positions of primary dealers in Treasuries. Source: FED and author's calculations.

I first test the first hypothesis and examine the relationship between the failure rate and the on-the-run premium. For the outcome variable in the first regression (see Table 1), I use data of the 1-, 2-, 3-, 5-, 7-, 10-, 20-, and 30-year on-the-run premia. Each maturity, denoted by m, and day, denoted by t, in the sample is a separate observation. The explanatory variable, the failure rate of off-the-run Treasuries, is not available for different maturities. The explanatory variable is therefore the same for each on-the-run premium maturity. I run three regressions. In the first, I only regress on the failure rate. In the second, I add other control variables: the logarithm of the VIX, the 10-year to 2-year yield spread, and the general collateral financing repo rate. In the third regression, I add additional crisis and maturity fixed effects. The crisis fixed effects capture the months October 2014, September 2019, and March 2020. In these months there was a crisis in the Treasury market (see U.S. Department of the Treasury et al. (2015),

[51]The data sources for the on-the-run premium are the same as in Figure 2a. The net positions of primary dealers 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 time horizon starts in April 2013.

Anbil et al. (2020), Schrimpf et al. (2020)). As I use weekly data, the crisis dummies are equal to one in each week within these months. The maturity fixed effects capture each maturity m of the on-the-run premia. I use the time horizon from April 2013 to the end of 2024. The reason for this is that I want to use the same time horizon to test the first two hypotheses, as they are related, and the net positions of primary dealers (used in the second regression) are only available in the preferred granularity since April 2013. The regression equation (3) is given by

$$\begin{aligned}\text{on-the-run premium}_{t,m} = \alpha_m &+ \beta_1 \text{ failure rate off-the-run Treasuries}_t + \beta_2\, ln(\text{VIX}_t) \\ &+ \beta_3 \text{ 10-year to 2-year yield spread}_t \\ &+ \beta_4 \text{ general collateral financing repo rate}_t \\ &+ \beta_6\, \mathbb{I}_{10/14,t} + \beta_7\, \mathbb{I}_{09/19,t} + \beta_8\, \mathbb{I}_{03/20,t} + u_t.\end{aligned}$$

The results are shown in Table 1. As expected, the coefficient of the failure rate is positive. It is significant at the 1% level (with and without additional control variables and fixed effects). If I do not control for any other variables, then an increase in the failure rate by 1 percentage point increases the premium by 2.1 basis points, which is about half a standard deviation. The size of the effect is almost the same if I add the additional control variables.

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

	(1) Premium	(2) Premium	(3) Premium
Failure rate off-the-run Treasuries	2.0652***	2.1925***	2.1763***
	(0.280)	(0.295)	(0.279)
lnVIX		1.2233***	0.9016***
		(0.295)	(0.232)
10-year to 2-year yield spread		0.3104***	0.2724**
		(0.126)	(0.125)
General collateral financing repo rate		0.1276**	0.1163**
		(0.061)	(0.056)
Constant	1.3041***	-2.6594***	-2.0432**
	(0.192)	(0.963)	(0.827)
Crisis and maturity fixed effects	No	No	Yes
No. Observations:	4560	4560	4560
Adj. R-squared:	0.012	0.017	0.227

Notes: 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 2013–2024, 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 notes 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/nominal-yield-curve.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. The crisis fixed effects capture the months October 2014, September 2019, and March 2020, and the maturity fixed effects each maturity of the on-the-run premia (1y, 2y, 3y, 5y, 7y, 10y, 20y, and 30y). The constant in the last column captures the fixed effect of the 1-year maturity.

Next, I test the second hypothesis and examine the relationship between the failure rate and dealer inventories. I regress the failure rate of off-the-run Treasuries on the net outright positions of primary dealers in Treasuries of different maturity baskets (see Table 2). The maturity of the baskets is denoted by k. Specifically, I use the following baskets for the net outright positions in Treasuries: below or equal to 2 years, above 2 years to 3 years, above 3 years to 6 years, above 6 years to 11 years, above 11 years. Each basket and day is a separate observation. I run the regression first without additional control variables and then with. The control variables are the same as in the previous regression (see Table 1). In the third regression, I add crisis fixed effects the same way as before. Regression equation (3) is given by

$$\begin{aligned} \text{off-the-run failure rate}_t = \alpha &+ \beta_1 \text{ primary dealer net positions}_{t,k} + \beta_2 \, ln(\text{VIX}_t) \\ &+ \beta_3 \text{ 10-year to 2-year yield spread}_t \\ &+ \beta_4 \text{ general collateral financing repo rate}_t \\ &+ \beta_6 \, \mathbb{I}_{10/14,t} + \beta_7 \, \mathbb{I}_{09/19,t} + \beta_8 \, \mathbb{I}_{03/20,t} + u_t. \end{aligned}$$

The results are shown in Table 2. As expected, the coefficient of the primary dealers' net positions is negative and significant at the 1% level.[52] Adding the additional control variables and fixed effects does not change the result. As the net positions are denominated in dollars, the coefficient is best interpreted by first multiplying it by the standard deviation of the net outright positions. If the net positions increase by one standard deviation, then the failure rate reduces by 2 (regression 1) and 3 (regressions 2 and 3) basis points, respectively.[53]

In Appendix A.5, I provide further evidence consistent with my theory. I examine the volatility of the net positions of primary dealers in Treasuries. I show that primary dealer inventories in on-the-run Treasuries are much less volatile than inventories including many kinds of Treasuries.

4.3 Empirical evidence 2: On-the-run premium and facility rate

To test my third hypothesis, I first show the scatter plot between the 10-year on-the-run premium and the reverse repo facility rate of the FED, in Figure 8.[54] The time horizon is 23 September 2013 to 31 December 2020 (as in regression 3), and the frequency is weekly. I observe that there is a negative correlation between the 10-year on-the-run premium and the reverse repo facility rate (with tight 95% confidence bands) as expected.

I regress on-the-run premia data on the reverse repo facility rate (see Table 3).[55] My theory predicts a negative coefficient. I use data on the 1-, 2-, 3-, 5-, 7-, 10-, 20-, and 30-year on-the-run premia. Each maturity m and each day t in the sample is a separate observation. The explanatory variable is the reverse repo facility rate. The first observation I use is from 23 September 2013, as this was the first day

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

[53] This corresponds to 0.09 (regression 1) and 0.14 (regressions 2 and 3) standard deviations, respectively.

[54] The data source for the premium is the same as in Figure 2a. The reverse repo facility rate data can be downloaded from FRED.

[55] I use the reverse repo facility and not any deposit facility data as the access discussion centres around this one. An increasing number of institutions already access this facility (Frost et al. (2015), Baklanova et al. (2015), Marte (2021)).

Table 2: Failure rate on net positions regression

	(1) Off-the-run failure rate	(2) Off-the-run failure rate	(3) Off-the-run failure rate
PD net positions (deflated baskets)	-0.0012***	-0.0020***	-0.0020***
	(0.000)	(0.000)	(0.000)
lnVIX		-0.0805**	-0.0846**
		(0.036)	(0.037)
10-year to 2-year yield spread		-0.1717***	-0.1745***
		(0.017)	(0.017)
General collateral financing repo rate		-0.0676***	-0.0683***
		(0.008)	(0.008)
Constant	0.6990***	1.1684***	1.1815***
	(0.011)	(0.114)	(0.117)
Crisis fixed effects	No	No	Yes
No. Observations:	3402	3402	3402
Adj. R-squared:	0.008	0.124	0.126

Notes: 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 2013–2024 and the frequency is weekly.
The data source for the net outright positions (in billion USD) is the same as in Figure 7. The data are also deflated in the same way (with the index set to 1 when the time horizon 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 notes 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/nominal-yield-curve.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. The crisis fixed effects capture the months October 2014, September 2019, and March 2020.

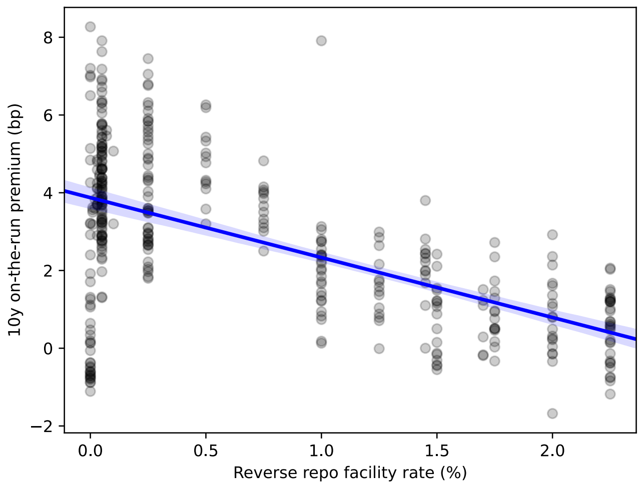
the facility was available on a large scale.[56] I run my regressions with and without additional control variables. The additional control variables are the logarithm of the VIX, the 10-year to 2-year yield spread, and the general collateral financing repo rate. I also add crisis and maturity fixed effects in a third regression, as in regression 1. The regression equation (3) is given by

$$\begin{aligned}\text{on-the-run premium}_{t,m} = \alpha_m &+ \beta_1 \text{ reverse repo facility rate}_t + \beta_2 \, ln(\text{VIX}_t) \\ &+ \beta_3 \text{ 10-year to 2-year yield spread}_t \\ &+ \beta_4 \text{ general collateral financing repo rate}_t \\ &+ \beta_5 \, \mathbb{I}_{10/14,t} + \beta_7 \, \mathbb{I}_{09/19,t} + \beta_8 \, \mathbb{I}_{03/20,t} + u_t.\end{aligned}$$

From Table 3, as expected, the coefficient of the reverse repo facility rate is negative. A 1 percentage point increase in the reverse repo facility rate reduces the on-the-run premium by 0.9 basis points (regression 1) and 2.1 basis points[57] (regressions 2 and 3), respectively. The coefficients are significant at the 1% level. When I extend the time horizon beyond 2020, the evidence is less consistent across all

[56] See https://www.newyorkfed.org/markets/opolicy/operating_policy_130920.html for more information.
[57] This corresponds to around 0.3 (regression 1) and 0.6 (regressions 2 and 3) standard deviations, respectively.

Figure 8: The correlation between the premium and the reverse repo facility rate



Notes: The figure plots the on-the-run premium on 10-year Treasury notes against the rate of the FED's reverse repo facility. Source: FED and author's calculations.

three columns.[58]

[58]The sign and significance of the main coefficient vary depending on the time horizon and the control variables.

Table 3: Premium on reverse repo facility rate regression

	(1) Premium	(2) Premium	(3) Premium
Reverse repo facility rate	-0.8621***	-2.0726***	-2.0795***
	(0.038)	(0.293)	(0.282)
lnVIX		0.8779***	-0.0456
		(0.213)	(0.105)
10-year to 2-year yield spread		-0.6460***	-0.6785***
		(0.086)	(0.075)
General collateral financing repo rate		0.7575***	0.7142***
		(0.283)	(0.275)
Constant	3.5027***	1.9599***	5.6443***
	(0.053)	(0.614)	(0.313)
Crisis and maturity fixed effects	No	No	Yes
No. Observations:	13984	13984	13984
Adj. R-squared:	0.044	0.071	0.466

Notes: 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 23 September 2013 to 31 December 2020, and the frequency is daily. The data source for the premium (in basis points) is the same as in Figure 2a. The reverse repo facility rate data can be downloaded from FRED. The rate is in percent. The VIX data are taken from FRED. The 10-year and 2-year yields (par yields of seasoned notes 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/nominal-yield-curve.htm. The general collateral financing repo rate (in percent) is provided by the Depository Trust and Clearing Corporation (DTCC) here: https://www.dtcc.com/charts/dtcc-gcf-repo-index. The crisis fixed effects capture the months October 2014, September 2019, and March 2020, and the maturity fixed effects each maturity of the on-the-run premia (1y, 2y, 3y, 5y, 7y, 10y, 20y, and 30y). The constant in the last column captures the fixed effect of the 1y maturity.

5 Conclusion

I develop a model of the U.S. Treasury market to explain the on-the-run phenomenon. I provide a novel explanation using inventory uncertainty that I also test empirically. The model is then used to discuss broad access to central bank facilities. The analysis is motivated by the current discussion on how to reform the market (see, e.g., Duffie (2023)) and the increase in intermediation and participation by non-bank financial institutions (Eren and Wooldridge (2021)). The latter raises the question of why only a limited number of participants currently have access to central bank facilities.[59] I analyze the implications of providing broad access to a reverse repo or deposit facility.

In summary, the on-the-run premium reflects market frictions and higher inventory risk for off-the-run Treasuries, making them less attractive. Off-the-run Treasuries fail to settle more often, consistent with the data. Adding to the seminal work by Vayanos and Weill (2008), I provide a fundamental reason why there is a premium on the on-the-run Treasuries and not the off-the-run ones. I explicitly take into account the time elapsed since the assets' issuance, as this also defines the difference between on- and off-the-run. The longer the assets are on the market post-issuance, the higher the dispersion of their inventories across primary dealers, leading to inventory uncertainty, which implies the on-the-run premium.

The second part of the study examines the impact of facility access. An increase in the facility rate reduces trading costs, stimulates trading, and raises prices. The on-the-run premium decreases as shown empirically. However, the rate increase also increases settlement fails for off-the-run Treasuries, as they are more frequently traded but less available. Additionally, only primary dealers benefit from increased access to the facility.

Future extensions of the model could account for the characteristics of non-bank financial institutions, such as leverage by hedge-funds, and explore the effects of broad access to a repo facility, especially in light of the ongoing discussions about Treasury market restructuring.

[59] See, e.g., 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.

References

Anbil, S., A. Anderson, and Z. Senyuz (2020). What happened in money markets in September 2019? *FEDS Notes*.

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. *FRB 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*, 79–94.

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*.

Chaboud, A., E. Correia-Golay, C. Cox, M. J. Fleming, Y. Huh, F. M. Keane, K. Lee, K. Schwarz, C. Vega, and C. Windover (2025). All-to-all trading in the U.S. Treasury market. *Economic Policy Review 31*(2), 1–27.

Christensen, J. H. E., J. A. Lopez, and P. Shultz (2017). Do all new Treasuries trade at a premium? *FRB of San Francisco 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).

Cohen, A., M. Kargar, B. Lester, and P.-O. Weill (2024). Inventory, market making, and liquidity in OTC markets. *Journal of Economic Theory 222*, 105917.

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.

Doerr, S., E. Eren, and S. Malamud (2023). Money market funds and the pricing of near-money assets. *BIS Working Paper*.

Duffie, D. (1996). Special repo rates. *The Journal of Finance 51*(2), 493–526.

Duffie, D. (2020). Still the world's safe haven? Redesigning the U.S. Treasury market after the COVID-19 crisis. *Hutchins Center on Fiscal and Monetary Policy at Brookings*. https://www.brookings.edu/research/still-the-worlds-safe-haven.

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

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 (2023). *The Research Handbook of Financial Markets*, Chapter "The Treasury and when-issued markets". Edward Elgar.

Eisfeldt, A. L., B. Herskovic, and S. Liu (2023). Interdealer price dispersion. *Available at SSRN*.

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. *FRB of New York Staff Reports* (964).

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

Fleming, M. J. and K. D. Garbade (2002). When the back office moved to the front burner: Settlement fails in the Treasury market after 9/11. *FRB of New York Economic Policy Review*.

Fleming, M. J. and K. D. Garbade (2004). Repurchase agreements with negative interest rates. *Current Issues in Economics and Finance 10*(5), 1–7.

Fleming, M. J. and K. D. Garbade (2005). Explaining settlement fails. *Current Issues in Economics and Finance 11*(9), 1–7.

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. *FRB of New York Economic Policy Review*.

Gârleanu, 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.

Graveline, J. J. and M. R. McBrady (2011). Who makes on-the-run Treasuries special? *Journal of Financial Intermediation 20*(4), 620–632.

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.

He, Z., S. Nagel, and Z. Song (2022). Treasury inconvenience yields during the COVID-19 crisis. *Journal of Financial Economics 143*(1), 57–79.

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

Lagos, R., G. Rocheteau, and P.-O. Weill (2011). Crises and liquidity in over-the-counter markets. *Journal of Economic Theory 146*(6), 2169–2205.

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.

Schrimpf, A., H. S. Shin, and V. Sushko (2020). Leverage and margin spirals in fixed income markets during the COVID-19 crisis. *BIS Bulletin* (2).

Spindt, P. A. and R. W. Stolz (1992). Are U.S. 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. 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.

Vissing-Jorgensen, A. (2021). The Treasury market in spring 2020 and the response of the Federal Reserve. *NBER Working Paper Series* (29128).

Weill, P.-O. (2007). Leaning against the wind. *The Review of Economic Studies 74*(4), 1329–1354.

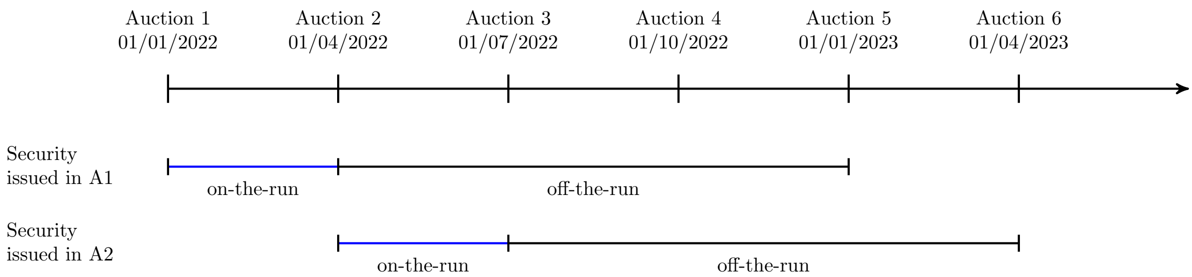
Wells, M. (2023). Averting a Treasury market crisis. *Econ Focus*, 14–17.

A Appendix

A.1 On-the-run cycle

The following figure illustrates the on-the-run cycle. An auction is held every quarter. The newly issued security is on-the-run until the next auction of assets with the same maturity. It then becomes off-the-run.

Figure 9: On-the-run cycle



Notes: The figure illustrates that an asset is initially on-the-run and becomes off-the-run when the next generation of assets with the same maturity are auctioned and issued.

A.2 Other equilibria

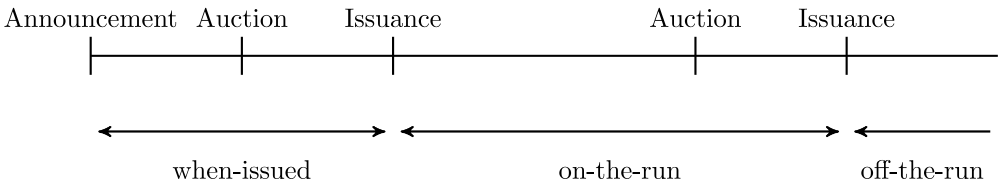
In the main body of the paper, I focused on the case where $\lambda_t^2 = \lambda_t^n = \lambda_t^{f,h} = 0$. This means that primary dealers are unconstrained if they still have the full inventory of assets i. I showed that I can always find issuance sizes $\mathcal{I}$ where such an equilibrium exists. For the sake of completeness I also considered 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 described in Section 3.4, as long as $\lambda_t^{f,l} > 0$, which I 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 any further, 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.

A.3 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. The extension also allows me to explain two more stylized facts, which I describe after introducing the life cycle.

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 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 (2023)). 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 their 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.

Figure 10: Life cycle



Notes: This figures shows the life cycle of an asset, including its when-issued, on-the-run and off-the-run periods.

In addition to the stylized fact that the on-the-run price is above the off-the-run price, I show below that the model explains two other facts about prices.

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

II The primary market prices are lower than the secondary market prices.

First, a detailed analysis by Durham and Perli (2023) showed that when-issued prices carry a premium compared to already issued Treasuries. In their analysis, they include both on- and off-the-run Treasuries in their comparison with the when-issued ones. Second, the secondary market price is known to be higher than the auction price. Direct comparisons of primary and secondary market prices at the time of the auction have 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.[60]

Before I explain the two stylized facts with my model, I derive the auction price. I still look at the premium equilibrium. 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 they are 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:

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

[60]Fleming et al. (2022) document that Treasury dealers appear to be compensated for taking inventory risks at the auction by price increases in subsequent weeks. This is also evidence that auction prices are lower than secondary market prices after the auction.

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

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

I 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, I now summarize and compare all the prices. In the case of the one-period asset, these are

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

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

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

Next I show that the two stylized facts introduced above hold in the premium equilibrium.

Proposition 6. *In the premium equilibrium*

$$q_t^2 > p_t^2$$
$$q_t^n > p_t^n$$
$$q_t^n > p_t^f.$$

and

$$p_t^2 > p_t^{2,A}$$
$$p_t^n > p_t^{1,A}.$$

Proof. See Appendix A.6.5. □

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 the income. This income has to cover the spot price they have to pay tomorrow, which in equilibrium is equal to the collateral value they have 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 between the auction and 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, who 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 others are sold. This is consistent with the model 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 the model, the buyer is the ultimate owner of the asset, so their 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.

In summary, the on-the-run prices are 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 latter valuation is lower than the former. Therefore, the primary market prices lie below the secondary market prices. The when-issued prices are higher than the spot prices due to the cost of collateralization.

A.4 Heterogeneous sellers

In Section 3.4.4 I assumed that all sellers have or gain access to the facility. In this section, I assume that there is a measure of sellers who have access to the facility, ξ, and a measure who do not have access,

$(1-\xi)$. Sellers who have access are denoted by a and the ones who do not by na. The facility rate is given by r_t and the market rate by r_t^m. I assume that $r_t > r_t^m$. Agents with no access face the market rate, while the others will use the facility. Again, I will look at the impact of sellers gaining access. But first I present the equilibrium equations.

I again consider the premium equilibrium. I first show the prices and quantities of the newly issued assets. Analogously to Section 3.4, the contract and spot prices are given by

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

and

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

The traded quantities are determined by the following equations:

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

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

Next, I show the off-the-run prices and quantities. Let me denote the probability of a seller with access finding the off-the-run assets by $\mathbb{P}^a$. Analogously, I define $\mathbb{P}^{na}$ as the probability for the sellers without access. Analogously to Section 3.4, the off-the-run prices are given by

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

As in Section 3.3, I assume that the primary dealers with a full inventory are unconstrained. Therefore,

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

Lastly, I 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 1, they either equal σ or $\sigma + (1-\sigma)(1-\xi)$, given that $(\mathcal{I} - A_{t-1}^{2,na}) > (\mathcal{I} - A_{t-1}^{2,a})$, or $A_{t-1}^{2,a} > A_{t-1}^{2,na}$.[61]

Next, I analyze the effect of sellers without access gaining access. The analysis is done analogously to Section 3.4.4. As in Section 3.4.4, I assume that there is still a positive premium after all sellers have access. The results are summarized below.

Corollary 4. *If sellers without access gain access, meaning that their interest rate unanticipated permanently increases from r_t^m to r_t in the second subperiod in $t = \widetilde{t} - 1$, then*

a) all on-the-run spot prices charged by the sellers gaining access increase and all sellers charge $p_t^{j,a}$ in $t \geq \widetilde{t}$ for $j = 2$ and n,

b) off-the-run spot prices charged by the sellers gaining access initially, i.e., in $t = \widetilde{t}$, increase and all sellers charge $p_{\widetilde{t}}^{f,a}$,

c) the on-the-run premium Δ_t initially, i.e., in $t = \widetilde{t}$, decreases,

d) all quantities of on-the-run assets traded by the sellers gaining access increase and all sellers trade $A_t^{j,a}$ in $t \geq \widetilde{t}$ for $j = 2$ and n,

e) quantities of off-the-run assets traded by the sellers gaining access initially, i.e., in $t = \widetilde{t}$, increase if no inventory constraints bind and all sellers trade $A_{\widetilde{t}}^{f,a,h}$,

f) quantities of off-the-run assets traded by the sellers gaining access initially, i.e., in $t = \widetilde{t}$, stay the same if inventory constraints bind and equal $A_{\widetilde{t}-1}^{f,na,l}$.

Points b), c) and e) also hold for $t > \widetilde{t}$ if the probabilities to find the off-the-run assets do not change. Otherwise the effects are ambiguous.

Proof. See Appendix A.6.8. □

As noted in Section 3.4.4, the value $\widetilde{\sigma}$ is the value of σ above which always (but not only) $\frac{\partial A_t^2}{\partial r_{\widetilde{t}-1}} > 0 \; \forall t \geq \widetilde{t}$. I define it in Appendix A.6.6. I also show that with a reasonable calibration, it is an empirically very small value.

[61] For the inventories I do not only need to distinguish if a primary dealer has met a seller the previous period or not (as before) 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 in 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 in the previous period.

The result is self-explanatory. The direction of the effects on the sellers who gain access is (initially at $t = \widetilde{t}$ and also later if the probabilities of finding the off-the-run assets do not change) analogous to my baseline scenario with homogeneous sellers, in Section 3.4.4.

A.5 Volatilities

To provide further evidence consistent with my theory, I examine the volatility of the net positions of primary dealers in Treasuries.[62] I find evidence in the data consistent with my theory that off-the-run inventories are much more affected by search and matching frictions than on-the run inventories. Table 4 shows the volatilities. Primary dealer inventories in on-the-run Treasuries are much less volatile than inventories including different types of Treasuries. This holds across all maturity baskets. This is consistent with my theory if I assume that there is a slight variation in σ (the search and matching friction) over time, which is the case in real life. 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.

Table 4: Primary dealer net positions volatilities

On-the-run	Maturity	2y	3y	5y	7y	10y	30y
	Volatility	4.1	3.9	3.4	2.7	4.1	2.2
All	Maturity	≤2y	(2y,3y]	(3y,6y]	(6y,7y]	(7y,11y]	>11y
	Volatility	20.7	6.9	13.9	7.6	8.1	12.1

Notes: The table presents the volatilities of net positions held by primary dealers in Treasuries across various maturities. In addition to the category encompassing all Treasuries, on-the-run Treasuries are also displayed separately for comparison. Source: FED.

A.6 Proofs

A.6.1 Proof of Proposition 1

Proof. Suppose it is not the case that $\mathcal{I} < A^2_{t-1} + A^{f,h}_t$ or $\lambda^{f,l}_t > 0$. Then $\lambda^{f,l}_t = \widetilde{\lambda}^{f,l}_t = 0$. From the first-order conditions of the seller and the buyer and $\omega^f_t = p^f_t$, it follows that $p^n_t = p^f_t = \beta^2(1+r_t)(\delta+g)$. Therefore, there is no premium, $\Delta_t = 0$. $\lambda^{f,l}_t = 0$ and $\widetilde{\lambda}^{f,l}_t > 0$ is not possible. If $\lambda^{f,l}_t = 0$, then it must be that $\widetilde{\lambda}^{f,l}_t = \widetilde{\lambda}^{f,h}_t = 0$, given the optimal behavior of the seller, which takes into account the profitable amount to be sold by the primary dealer. Also $\lambda^{f,l}_{t+1} > 0$ and $\widetilde{\lambda}^{f,l}_{t+1} > 0$ is not possible. If $\lambda^{f,h}_{t+1} = 0$ and $\lambda^{f,l}_{t+1} > 0$, then $a^f_t > A^{f,l}_t$, and it follows that $\widetilde{\lambda}^{f,l}_{t+1} = 0$. □

[62] 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 my on-the-run net positions time series starts 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.

A.6.2 Proof of Proposition 2

Proof. First, I show that for an equilibrium with trade in all assets to exist, I need $\sigma > \frac{1-\beta(1+r_t)}{\beta(1+r_t)}\frac{\delta}{g}$. In equilibrium $p_t^f = \kappa'(A_t^{f,h}) + \delta = \frac{\sigma}{1-(1-\sigma)\beta(1+r_t)}\beta(1+r_t)(\delta+g)$. For $\kappa'(A_t^{f,h}) > 0$ and $A_t^{f,h} > 0$, $\frac{\sigma}{1-(1-\sigma)\beta(1+r_t)}\beta(1+r_t)(\delta+g) - \delta > 0$ or $\sigma > \frac{1-\beta(1+r_t)}{\beta(1+r_t)}\frac{\delta}{g}$ must hold.

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 $p_t^f = \kappa'(A_t^{f,h}) + \delta$. As $p_t^n > p_t^f$ and as $\kappa(A_t^i)$ is a strictly convex function, it follows that $A_t^n > A_t^{f,h}$. 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 $(A_{t-1}^2 + A_t^{f,h}) > \mathcal{I} > max(A_t^2, A_t^n, A_t^{f,h})$. I know that $A_t^2 > A_t^n > A_t^{f,h}$. Therefore, it follows that in equilibrium $(A_{t-1}^2 + A_t^{f,h}) > \mathcal{I} > A_t^2$. From the equilibrium conditions it follows that

$$\begin{aligned}
A_t^n &= \kappa'^{-1}\left[\beta(1+r_t)(\delta+g) - \delta\right] \\
A_t^{f,h} &= \kappa'^{-1}\left[\frac{\sigma}{1-(1-\sigma)\beta(1+r_t)}\beta(1+r_t)(\delta+g) - \delta\right] \\
\kappa'(A_{t-1}^2) &= W + \beta(1-\sigma)\kappa'(\mathcal{I} - A_{t-1}^2)
\end{aligned}$$

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

The left-hand side of the third equation, $f_1(x) = \kappa'(x)$, is a strictly increasing function with $x \in [0, \mathcal{I}]$, $min(f_1(x)) = 0$ and $max(f_1(x)) = \kappa'(\mathcal{I})$. The right-hand side 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 $\kappa'(\mathcal{I}) > W$ or $\mathcal{I} > \kappa'^{-1}(W)$. By continuity it then follows that $\mathcal{I} > A_{t-1}^2$. As $\mathcal{I} - A_{t-1}^2(\mathcal{I})$ is a continuous function (as $\kappa'(A_t^i)$ and $\kappa'^{-1}(A_t^i)$ are both continuous) with minimal value 0 for $\mathcal{I} = \kappa'^{-1}(W)$, I can always find an $\mathcal{I}$ where $\mathcal{I} - A_{t-1}^2(\mathcal{I}) < A_t^{f,h}$ for any $A_t^{f,h} > 0$. □

A.6.3 Proof of Proposition 4

Proof. If $I_t^{f,h} = I_t^{f,l}$, then $\lambda_t^{f,h} = \lambda_t^{f,l}$. It is not possible that $\lambda_t^{f,h} = \lambda_t^{f,l} > 0$ as by assumption $\lambda_t^{f,h} = 0$. Therefore, $\lambda_t^{f,h} = \lambda_t^{f,l} = 0$. But the buyer's binding first-order conditions (4) imply that $q_{t-1}^n = q_{t-1}^f = (\delta+g)$. The seller's binding first-order conditions (5) in combination with the delivery constraints (1) then imply that $p_t^n = p_t^f$ and, therefore, $\Delta = 0$. □

A.6.4 Proof of Proposition 5

Proof. First, I know that in equilibrium $f_t^n = 0$ and $f_t^f = \frac{(1-\sigma)p_t^f(a_t^f - I_t^f)}{p_t^f a_t^f} = (1-\sigma)\left(1 - \frac{\mathcal{I}-a_t^2}{a_t^f}\right)$. Therefore, trivially, $f_t^f > 0 = f_t^n$. Second, by comparing $p_t^n = \beta(1+r_t)(\delta+g)$ and $p_t^f = \frac{\sigma}{1-(1-\sigma)\beta(1+r_t)}\beta(1+r_t)\beta(\delta+g)$, it follows immediately that $p_t^n > p_t^f$. Third, $p_t^n > p_t^f$ also implies that $A_t^n > A_t^{f,h}$ as $A_t^n = \kappa'^{-1}(p_t^n - \delta)$ and $A_t^{f,h} = \kappa'^{-1}(p_t^f - \delta)$. Lastly, as $\lambda_t^{f,h} = 0$ and $\lambda_t^{f,l} > 0$, it follows that $A_t^{f,h} > A_t^{f,l}$. □

A.6.5 Proof of Proposition 6

Proof. First, as $\beta(1+r_t) > 0$, it follows that $q_t^2 > p_t^2$ and $q_t^n > p_t^n$. Also, from Proposition 5, I know that $p_t^n > p_t^f$ and, therefore, $q_t^n > p_t^f$. Second, from the first-order condition of the primary dealer (see Section 3.2.1) I know that $p_t^2 = \kappa'(A_t^2) + \delta + \beta\delta + \beta(1-\sigma)\lambda_{t+1}^f$ where $\lambda_{t+1}^f = p_{t+1}^f - \kappa'(A^{f,l}) - \delta$. It follows immediately that $p_t^2 > p_t^{2,A}$. Also from the first-order condition of the primary dealer, I know that $p_t^n = \kappa'(A_t^n) + \delta$. It follows immediately that $p_t^n > p_t^{1,A}$. □

A.6.6 Proof of Corollary 1

Proof. I assume an unanticipated permanent increase in the interest rate that occurs at the beginning of the contract market in $t = \widetilde{t} - 1$. The stock of available off-the-run assets in the next spot market is given, and it can reach a new steady state only after a period.

First, I show that an increase in the facility rate r_t raises all prices. I know that in equilibrium, prices are given by $p_t^n = \beta(1+r_t)(\delta+g)$, $p_t^2 = \beta(1+r_t)(1+\beta)(\delta+g)$, $p_t^f = \frac{\sigma}{1-(1-\sigma)\beta(1+r_t)}\beta(1+r_t)(\delta+g)$. It follows that

$$\frac{dp_t^n}{dr_{\widetilde{t}-1}} = \beta(\delta+g) > 0$$

$$\frac{dp_t^2}{dr_{\widetilde{t}-1}} = \beta(1+\beta)(\delta+g) > 0$$

$$\frac{dp_t^f}{dr_{\widetilde{t}-1}} = \left[\frac{\sigma}{1-(1-\sigma)\beta(1+r_t)} + \frac{\sigma(1-\sigma)\beta(1+r_t)}{[1-(1-\sigma)\beta(1+r_t)]^2}\right]\beta(\delta+g) > 0$$

for all $t \geq \widetilde{t}$.

Second, I show that the on-the-run premium decreases given an increase in r_t. In equilibrium $\frac{dp_t^n}{dr_{\widetilde{t}-1}} < \frac{dp_t^f}{dr_{\widetilde{t}-1}}$ if $\frac{\sigma}{1-(1-\sigma)\beta(1+r_t)} + \frac{\sigma(1-\sigma)\beta(1+r_t)}{[1-(1-\sigma)\beta(1+r_t)]^2} > 1$, which is the case if $\sigma > \frac{[1-\beta(1+r_t)]^2}{[\beta(1+r_t)]^2} \approx 0$.[63] The premium is given by $\Delta_t = p_t^n - p_t^f$. Therefore, $\frac{d\Delta_t}{dr_{\widetilde{t}-1}} = \frac{dp_t^n}{dr_{\widetilde{t}-1}} - \frac{dp_t^f}{dr_{\widetilde{t}-1}} < 0\ \forall t \geq \widetilde{t}$.

Third, I discuss the impact on the quantities. I know that in equilibrium quantities are determined by $\kappa'(A_t^n) = p_t^n - \delta$, $\kappa'(A_t^2) - \beta(1-\sigma)\kappa'(I_t^2 - A_t^2) = p_t^2 - \beta(1-\sigma)p_{t+1}^f - (1+\beta\sigma)\delta$, and $\kappa'(A_t^{f,h}) = p_t^f - \delta$.

[63] Let me define $k \equiv (1-\sigma)\beta(1+r_t)$. It follows: $\frac{\sigma}{1-k} + \frac{\sigma k}{(1-k)^2} > 1$ or $\sigma > (1-k)^2$. Further rearranging yields $0 > 1 - 2\beta(1+r_t) + (1-\sigma)\left[\beta(1+r_t)\right]^2$ or $\sigma > \frac{[1-\beta(1+r_t)]^2}{[\beta(1+r_t)]^2} \approx 0$.

It follows that

$$\frac{d\kappa'(A_t^n)}{dr_{\widetilde{t}-1}} = \frac{dp_t^n}{dr_{\widetilde{t}-1}} > 0$$
$$\frac{d\kappa'(A_t^{f,h})}{dr_{\widetilde{t}-1}} = \frac{dp_t^f}{dr_{\widetilde{t}-1}} > 0$$

for all $t \geq \widetilde{t}$. Therefore, $\frac{dA_t^n}{dr_{\widetilde{t}-1}} > 0$ and $\frac{dA_t^{f,h}}{dr_{\widetilde{t}-1}} > 0$ for all $t \geq \widetilde{t}$. Also

$$\frac{d\kappa'(A_t^2)}{dr_{\widetilde{t}-1}} - \beta(1-\sigma)\frac{d\kappa'(I_t^2 - A_t^2)}{dr_{\widetilde{t}-1}} = \frac{dp_t^2}{dr_{\widetilde{t}-1}} - \beta(1-\sigma)\frac{dp_{t+1}^f}{dr_{\widetilde{t}-1}}.$$

It follows that $\frac{dA_t^2}{dr_{\widetilde{t}-1}} > 0$ for all $t \geq \widetilde{t}$ iff $\frac{dp_t^2}{dr_{\widetilde{t}-1}} > \beta(1-\sigma)\frac{dp_{t+1}^f}{dr_{\widetilde{t}-1}}$. $\frac{dp_t^2}{dr_{\widetilde{t}-1}} > \beta(1-\sigma)\frac{dp_{t+1}^f}{dr_{\widetilde{t}-1}}$ if $\sigma > \widetilde{\sigma}$.

The value of $\widetilde{\sigma}$ is derived as follows:

$\frac{dp_t^2}{dr_{\widetilde{t}-1}} > \beta(1-\sigma)\frac{dp_{t+1}^f}{dr_{\widetilde{t}-1}}$ iff $(1+\beta) > \beta(1-\sigma)\left[\frac{\sigma}{1-(1-\sigma)\beta(1+r_t)} + \frac{\sigma(1-\sigma)\beta(1+r_t)}{[1-(1-\sigma)\beta(1+r_t)]^2}\right]$ or $(1-\beta(1+r_t))^2 > \frac{\sigma}{(1-\sigma)}\left[(1-\sigma)\left(\frac{\beta}{(1+\beta)} + (\beta(1+r_t))^2\right) - 1\right]$, where $(1-\beta(1+r_t))^2 \approx 0$. There are two values of σ that solve the equation $(1-\beta(1+r_t))^2 = \frac{\sigma}{(1-\sigma)}\left[(1-\sigma)\left(\frac{\beta}{(1+\beta)} + (\beta(1+r_t))^2\right) - 1\right]$. Below the lower root and above the upper root, the right-hand side is lower than the left-hand side. I define $\widetilde{\sigma}$ as the larger root. Note that using empirically reasonable values, $\frac{dp_t^2}{dr_{\widetilde{t}-1}} > \frac{dp_{t+1}^f}{dr_{\widetilde{t}-1}}$ not holding, can only be for a limited range of small values of σ. For example, if $\beta = 0.96$ and $r_{\widetilde{t}-1} = 0.01$, then it follows only for $\sigma \in (0.003, 0.25)$ that $\frac{dp_t^2}{dr_{\widetilde{t}-1}} < \frac{dp_{t+1}^f}{dr_{\widetilde{t}-1}}$.

Next I look at the quantity of off-the-run assets traded by constrained dealers, which is given by $A_t^{f,l} = \mathcal{I} - A_{t-1}^2$. It follows that $\frac{dA_t^{f,l}}{dr_{\widetilde{t}-1}} = 0$ for $t = \widetilde{t}$ because $\frac{dA_{\widetilde{t}-1}^2}{dr_{\widetilde{t}-1}} = 0$. For all $t > \widetilde{t}$, it follows that $\frac{dA_t^{f,l}}{dr_{\widetilde{t}-1}} < 0$ if $\frac{dA_t^2}{dr_{\widetilde{t}-1}} > 0$, which is true if $\sigma > \widetilde{\sigma}$. From the above results, it directly follows that $\frac{d(A_t^2 + A_t^n + \sigma A_t^{f,h} + (1-\sigma)A_t^{f,l})}{dr_{\widetilde{t}-1}} > 0$ for $t = \widetilde{t}$ and $\forall t > \widetilde{t}$ if $\sigma > \widetilde{\sigma}$. Decreases in the trading of off-the-run assets by constrained primary dealers are due to equivalent increases in the trading of the asset in the on-the-run period.

Lastly, the off-the-run settlement failure rate is given by $f_t^f = \frac{(1-\sigma)p_t^f(a_t^f - I_t^f)}{p_t^f a_t^f} = (1-\sigma)\left(1 - \frac{\mathcal{I} - a_{t-1}^2}{a_t^f}\right)$. It follows that $\frac{df_t^f}{dr_{\widetilde{t}-1}} > 0$ for $t = \widetilde{t}$ because $\frac{dp_t^f}{dr_{\widetilde{t}-1}} > 0$ for $t \geq \widetilde{t}$ and $\frac{\partial A_{\widetilde{t}-1}^2}{\partial r_{\widetilde{t}-1}} = \frac{\partial a_{\widetilde{t}-1}^2}{\partial r_{\widetilde{t}-1}} = 0$. For $\forall t > \widetilde{t}$ it follows that $\frac{df_t^f}{dr_{\widetilde{t}-1}} > 0$ if $\frac{\partial A_{\widetilde{t}-1}^2}{\partial r_{\widetilde{t}-1}} = \frac{\partial a_{\widetilde{t}-1}^2}{\partial r_{\widetilde{t}-1}} > 0$, which is the case if $\sigma > \widetilde{\sigma}$. □

A.6.7 Proof of Corollary 2

Proof. I look at the impact of an increase in the interest rate r_t (see details below) on the lifetime values of the buyer, the seller, and the primary dealer. There are two groups of primary dealers: One group of primary dealers has an inventory of $\mathcal{I}$ off-the-run assets and another has an inventory of $\mathcal{I} - A_t^2$. I take the lifetime value of both groups and average according to their share in the population. To ease notation, let me define $V^{D,a}_{\tilde{t}-1} \equiv (1-\sigma)V^{D}_{\tilde{t}-1}(\mathcal{I}-A^2_{\tilde{t}}) + \sigma V^{D}_{\tilde{t}-1}(\mathcal{I})$. The lifetime values are given by

$$\begin{aligned}
V^b_{\tilde{t}-1} &= \sum_i \left[\beta(\delta+g) - \beta q^i_{\tilde{t}-1}\right] a^i_{\tilde{t}} + \beta^2(\delta+g)a^2_{\tilde{t}} - \beta(1-\sigma)[(\delta+g) - \omega^f_{\tilde{t}-1}](a^f_{\tilde{t}} - I^f_{\tilde{t}}) + \beta V^b_{\tilde{t}} \\
V^s_{\tilde{t}-1} &= \sum_i \left\{-\omega^i_{\tilde{t}-1} + \beta\left[q^i_{\tilde{t}-1} - p^i_{\tilde{t}} + \omega^i_{\tilde{t}-1}\right](1+r_t)a^i_{\tilde{t}}\right\} - \beta(1-\sigma)[\omega^f_{\tilde{t}-1} - p^f_{\tilde{t}}](1+r_t)(a^f_{\tilde{t}} - I^n_{\tilde{t}}) \\
&\quad + \beta V^s_{\tilde{t}-1} \\
V^{D,a}_{\tilde{t}-1} &= \beta\{(1-\sigma)^2[p^f_{\tilde{t}}(I^2_{\tilde{t}-1} - A^2_{\tilde{t}-1}) - \kappa(I^2_{\tilde{t}-1} - A^2_{\tilde{t}-1}) - \delta(I^2_{\tilde{t}-1} - A^2_{\tilde{t}-1})] \\
&\quad + (1-\sigma)\sigma[p^f_{\tilde{t}}A^{f,h}_{\tilde{t}} - \kappa(A^{f,h}_{\tilde{t}}) - \delta A^{f,h}_{\tilde{t}}] + (1-\sigma)\sum_{j\in\{n,2\}}\left[p^j_{\tilde{t}}A^j_{\tilde{t}} - \kappa(A^j_{\tilde{t}}) - \delta A^j_{\tilde{t}}\right] \\
&\quad + \delta(I^n_{\tilde{t}} + I^2_{\tilde{t}} + I^2_{\tilde{t}-1} - (1-\sigma)A^2_{\tilde{t}-1})\} + \beta V^{D,a}_{\tilde{t}}.
\end{aligned}$$

I assume that there is an unanticipated permanent increase in the interest rate at the beginning of the contract market in $t = \tilde{t}-1$. 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 simplify the buyer's and seller's profits using the first-order conditions in equilibrium. This yields

$$\begin{aligned}
V^b_{\tilde{t}-1} &= \beta(1-\sigma)[(\delta+g) - p^f_{\tilde{t}}]I^f_{\tilde{t}} + \frac{\beta^2}{1-\beta}(1-\sigma)\left[(\delta+g) - p^f\right]I^f \\
V^s_{\tilde{t}-1} &= 0.
\end{aligned}$$

The derivatives are

$$\begin{aligned}
\frac{dV^b_{\tilde{t}-1}}{dr_{\tilde{t}-1}} &= -\beta(1-\sigma)\frac{dp^f_{\tilde{t}}}{dr_{\tilde{t}-1}}(\mathcal{I} - a^2_{\tilde{t}-1}) - \frac{\beta^2}{1-\beta}(1-\sigma)\left\{\frac{dp^f}{dr_{\tilde{t}-1}}(\mathcal{I} - a^2) + \left[(\delta+g) - p^f\right]\frac{da^2}{dr_{\tilde{t}-1}}\right\} \\
\frac{dV^d_{\tilde{t}-1}}{dr_{\tilde{t}-1}} &= 0
\end{aligned}$$

$$\begin{aligned}
\frac{dV^{D,a}_{\tilde{t}-1}}{dr_{\tilde{t}-1}} &= \beta\Bigg\{(1-\sigma)^2\frac{dp^f_{\tilde{t}}}{dr_{\tilde{t}-1}}(\mathcal{I} - A^2_{\tilde{t}-1}) + (1-\sigma)\sigma\frac{d[\kappa'(A^{f,h}_{\tilde{t}})A^{f,h}_{\tilde{t}} - \kappa(A^{f,h}_{\tilde{t}})]}{dr_{\tilde{t}-1}} \\
&\quad + (1-\sigma)\sum_{j\in\{n,2\}}\frac{d[\kappa'(A^j_{\tilde{t}})A^j_{\tilde{t}} - \kappa(A^j_{\tilde{t}})]}{dr_{\tilde{t}-1}}\Bigg\} \\
&\quad + \frac{\beta^2}{1-\beta}\Bigg\{(1-\sigma)^2\frac{d[\kappa'(\mathcal{I}-A^2)(\mathcal{I}-A^2) - \kappa(\mathcal{I}-A^2) - \beta(1-\sigma)\kappa'(\mathcal{I}-A^2_{\tilde{t}-1})\mathcal{I}]}{dr_{\tilde{t}-1}} \\
&\quad + (1-\sigma)\sigma\frac{d[\kappa'(A^{f,h})A^{f,h} - \kappa(A^{f,h})]}{dr_{\tilde{t}-1}} + (1-\sigma)\sum_{j\in\{n,2\}}\frac{d[\kappa'(A^j)A^j - \kappa(A^j)]}{dr_{\tilde{t}-1}}\Bigg\}.
\end{aligned}$$

I simplify the last equation. This yields

$$\begin{aligned}
\frac{dV^{D,a}_{\tilde{t}-1}}{dr_{\tilde{t}-1}} =& \beta(1-\sigma)^2\frac{dp^f_{\tilde{t}}}{dr_{\tilde{t}-1}}(\mathcal{I} - A^2_{\tilde{t}-1}) + \frac{\beta}{1-\beta}\Bigg\{\beta(1-\sigma)^2\kappa''(\mathcal{I}-A^2)A^2\frac{\partial A^2}{\partial r_{\tilde{t}-1}} \\
&+ (1-\sigma)\sigma\kappa''(A^{f,h})A^{f,h}\frac{\partial A^{f,h}}{\partial r_{\tilde{t}-1}} + (1-\sigma)\sum_{j\in\{n,2\}}\kappa''(A^j)A^j\frac{\partial A^j}{\partial r_{\tilde{t}-1}}\Bigg\}.
\end{aligned}$$

Regarding the buyer's profits, note that $(\delta + g) - p^f > 0$. If $\frac{\partial A^2}{\partial r_{\widetilde{t}-1}} = \frac{\partial a^2}{\partial r_{\widetilde{t}-1}} > 0$, which is the case if $\sigma > \widetilde{\sigma}$, then, $\frac{dV^b_{\widetilde{t}-1}}{dr_{\widetilde{t}-1}} < 0$, $\frac{dV^s_{\widetilde{t}-1}}{dr_{\widetilde{t}-1}} = 0$, and $\frac{dV^{D,a}_{\widetilde{t}-1}}{dr_{\widetilde{t}-1}} > 0$. □

A.6.8 Proof of Corollary 4

Proof. I analyze the impact of sellers without access gaining access, meaning that their interest rate permanently and unanticipated increases from r^m_t to r_t in the second subperiod in $t = \widetilde{t} - 1$. This implies that all sellers now trade at the same prices $p^{j,a}_t$ and quantities $A^{j,a}_t$ for $j = 2$ and n, i.e., all the newly issued assets. Therefore, the spot prices and quantities of all on-the-run assets traded by sellers who gain access increase as $p^{2,a}_t > p^{2,na}_t$, $p^{n,a}_t > p^{n,na}_t$, $A^{2,a}_t > A^{2,na}_t$, $A^{n,a}_t > A^{n,na}_t$, while the prices and quantities for those who already had access stay the same.

Following the same reasoning, in the first period after the shock in $t = \widetilde{t}$, the same result holds for the off-the-run prices p^f_t as well as off-the-run quantities if no inventory constraint binds, i.e., for $A^{f,h}_t$. For this result, it is important to note that the probabilities of finding the assets have not yet changed because inventories are predetermined from the previous period. The latter also implies that given inventory constraints bind, the off-the-run quantities which are sold, $A^{f,l}_t = \mathcal{I} - A^2_{t-1}$, stay the same in this first period. The on-the-run premium initially, i.e., in $t = \widetilde{t}$, decreases following the same reasoning as in Corollary 2.

To analyze the off-the-run prices and quantities traded in the periods $t > \widetilde{t}$, I take into account how the probabilities to find the assets change after the first period after the shock. Otherwise, the reasoning is analogous. I assume that settlement fails still occur, so that there is still a positive premium (as in Section 3.4.4). The probability to find the assets, for all sellers is σ. This is weakly lower than beforehand. If already before access $\mathbb{P}^a = \mathbb{P}^{na} = \sigma$, then the effects are the same as in the first period after the shock. Otherwise, the effects on the prices, premium, and quantities traded are ambiguous. This is because, as implied by the equations for p^f_t and $A^{f,h}_t$, the effects of the interest rate change and the probability change work in opposite directions. □

A.7 Data sources and summary statistics

This section summarizes all data sources used for the regressions and provides summary statistics for the variables.

To calculate the on-the-run premium, the on-the-run yield is subtracted from the par yield of seasoned bills and notes, as in Christensen et al. (2017). 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))[64] and the FRB-H15 Tables.[65] The failure-rate data is calculated based on data from the FED's Primary Dealer Statistics. It includes outright and financing fails.[66] The net positions of primary dealers can be downloaded from the FED's Primary Dealer Statistics as well.[67] The net positions are deflated using

[64]The data can be downloaded here: https://www.federalreserve.gov/data/nominal-yield-curve.htm
[65]The data can be downloaded here: https://www.federalreserve.gov/releases/h15/.
[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. Each series is outlier adjusted, where an outlier is defined as being below the 2.5% percentile or above the 97.5% percentile.
[67]The data can be downloaded here: https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics.

the Consumer Price Index for All Urban Consumers: All Items in U.S. City Average, which can be downloaded from FRED.[68] The reverse repo facility rate and VIX data can be downloaded from FRED. The 10-year and 2-year yields used to calculate the corresponding spread are also taken from the FED yield curve noted above.[69] The general collateral financing repo rate is provided by the Depository Trust and Clearing Corporation (DTCC).[70]

For the regressions where I adjust the frequency from daily to weekly, I use the same observation days for all series.

Table 5: Summary statistics

Data	Maturity	Mean	Median	St. Dev.
On-the-run premium	1y	2.81	2.46	4.17
in bp	2y	2.09	1.62	2.22
	3y	2.6	2.47	1.72
	5y	2.39	2.33	1.71
	7y	-0.68	-0.69	1.44
	10y	3.82	3.57	2.93
	20y	2.53	3.86	6.63
	30y	6.89	6.69	6.29
Primary dealer	(0y, 1y]	30.75	25.11	20.09
net positions (defl.)	(1y, 2y]	26.65	21.25	20.73
in bil USD	(2y, 3y]	1.3	1.44	6.93
	(3y, 6y]	19.03	17.41	13.94
	(6y, 7y]	8.23	9.02	7.56
	(7y, 11y]	1.33	1.11	8.06
	(11y, 30y]	27.4	30.58	12.08
Failure rate of off-the-run Treasuries in %		0.68	0.64	0.23
Reverse repo rate in %		1.52	0.75	1.86
ln VIX		2.8	2.74	0.32
10-year to 2-year yield spread in %		0.74	0.69	0.86
General collateral financing repo rate in %		1.57	0.56	1.84

Notes: The time horizon used for this summary statistics table is April 2013 to 2024. If the frequency is not weekly, I change it to weekly and use the last value of the week.

[68] I set the index to 1 at the beginning of April 2013 as this is the start of the time horizon.
[69] The data can be downloaded here: https://www.federalreserve.gov/data/nominal-yield-curve.htm.
[70] The data can be downloaded here: https://www.dtcc.com/charts/dtcc-gcf-repo-index.