Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Designing a Main Street Lending Facility

Alexandros P. Vardoulakis

2020-052

Please cite this paper as:
Vardoulakis, Alexandros P. (2020). "Designing a Main Street Lending Facility," Finance and Economics Discussion Series 2020-052. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2020.052.

Designing a Main Street Lending Facility*

Alexandros P. Vardoulakis†

June 25, 2020

Abstract

Banks add value by monitoring borrowers. High funding costs make banks reluctant to lend. A central bank can ease funding by purchasing loans, but cannot distinguish which loans require more or less monitoring, exposing it to adverse selection. A multi-tier loan pricing facility arises as the optimal institutional design setting both the purchase price and banks' risk retention for given loan characteristics. This design dominates uniform (flat) structure for loan purchases, provides the right incentives to banks and achieves maximum lending at lower rates to businesses. Both the multi-tier and flat structures deliver welfare gains compared to no intervention, but the relative gain between the two depends on three sufficient statistics: the share of loans requiring monitoring, the risk-retention ratio, and the liquidity premium.

Keywords: Main Street, central bank lending facilities, monitoring, small business, sufficient statistics, COVID-19

JEL Classification: E58, G01, G28

---

*I am thankful to seminar participants at the Federal Reserve Board, and to Levent Altinoglu, David Arseneau, Garth Baughman, Francesca Carapella, Jin-wook Chang, Matt Darst, Sebastian Infante, Elizabeth Klee, David Rappoport, and Skander Van den Heuvel for helpful comments and suggestions. All errors herein are mine. The views expressed in this paper are those of the author and do not necessarily represent those of Federal Reserve Board of Governors, or anyone in the Federal Reserve System.

†Board of Governors of the Federal Reserve System, United States; email:alexandros.vardoulakis@frb.gov

1 Introduction

A distinguishing feature of the economic consequences of the COVID-19 pandemic is the total shutdown of large parts of the economy. Small businesses are expected to be disproportionally affected given the inability to market their products and services remotely, but also due to a likely lack of established credit lines. Moreover, small businesses usually do not have good—tangible and fairly liquid—collateral that would allow them to borrow more easily, which also raises the importance of bank monitoring. In turn, the effective bank funding costs may be increasing rapidly and banks may be unwilling to extend credit even to good businesses with low probabilities of default.

The collapse in revenue and the inability to raise funds to cover working capital would likely force small businesses to closures and layoffs with adverse second-round effects for the economy as whole. Policy makers and commentators have recognized these issues and the need to support small businesses (see, for example, Bigio, 2020, Drechsel and Kalemli-Özcan, 2020). One way to provide liquidity support is for central banks to set up a lending facility that would buy from banks loans extended to small businesses. For example, the Federal Reserve has announced it is establishing a Main Street Lending Program.[1] The Program will operate through three facilities: the Main Street New Loan Facility (MSNLF), the Main Street Priority Loan Facility (MSPLF), and the Main Street Expanded Loan Facility (MSELF). The first two are designed to facilitate/stimulate new loans to eligible businesses, while the third covers the upsizing of existing loans. The facilities will purchase at par value *a uniform/flat participation* for each category of Eligible loans, which also have a uniform/flat interest rate (LIBOR+300 basis points). [2] Another alternative would be to have publicly owned development banks extend the necessary lines of credit. For example, Germany's state owned development bank KfW provides unlimited access to loans.[3] A central bank lending facility may

[1]The lending facilities established under section 13(3) of the Federal Reserve Act are meant to support liquidity funding and the flow of credit in response to the COVID-19. All of these facilities have been undertaken with the approval of the Treasury Secretary, and many of them are supported by funding from the CARES Act. As such, the Federal Reserve is primarily providing the necessary liquidity support without intentionally undertaking credit risk. These features are present in the model herein.

[2]One difference between the MSNLF and the MSPLF is that the latter will purchase loans that have priority over other debt and that Eligible firms may have higher debt to EBITDA ratios (earnings before interest, taxes, depreciation, and amortization) at origination, which is an observable firm characteristic driving its riskiness. Both facilities will purchase a 95 percent (uniform/flat) participation in all Eligible Loans at par value. See https://www.federalreserve.gov/monetarypolicy/mainstreetlending.htm for details.

[3]See https://www.bundesfinanzministerium.de/Content/DE/Pressemitteilungen/Finanzpolitik/2020/03/2020-03-13-download-en.pdf for details.

provide more flexibility and can be quickly deployed (and rolled back when not needed), but its design should be such that banks maintain the incentives to continue monitoring borrowers after selling their loans to the central bank. As mentioned, this is particularly important for small business loans.[4]

This paper studies the design of a lending facility with a special focus on unobservable heterogeneity across firms rather than simply different risk profiles or other observable firm characteristics that can be easily accounted for in the pricing scheme of the facility.[5] To study the optimal facility design, consider a stylized model where banks choose either to extend loans to entrepreneurs with no own capital and no collateral, or to invest in a storage technology. [6] Entrepreneurs have access to projects that deliver a positive net present value, but only if banks engage in monitoring for the whole term of the loan akin to Diamond (1984) and Holmström and Tirole (1997). Sufi (2007) shows that monitoring is important for syndicated loans. Gustafson, Ivanov and Meisenzahl (2020) provide further empirical evidence about bank monitoring for syndicated loans and show that about 20 percent of loans involve active monitoring, which is soft information similar to what this model assumes. They also show that monitoring takes place throughout the life of the loan with 55 percent of loans being monitored monthly or daily, and 30 percent annually. Arguably, the need for monitoring may be as high or even higher for the smaller businesses covered in the lending facility.

Monitoring is valuable, but it is costly for the bank and cannot be observed by outsiders. Banks value liquidity and, thus, may be unwilling to extend loans unless they anticipate to sell them in a secondary market at a price that covers the premium for liquidity. After origination, loan sales are possible if banks maintain a portion of the loan on their balance sheet such that they can credibly

[4]The paper focuses on central bank facilities that provide funding for private lenders to induce them to lend to small businesses, but abstracts for other types of policies advocated, such as fiscal transfers (Saez and Zucman, 2020). See Dreyer, McNamara, Nye, Nygaard and Sankar (2020) for a summary of the various types of policies employed worldwide to support small businesses.

[5]The model herein also derives the pricing of loans and risk-retention requirements given observable, or other easy to infer, firm characteristics. But does not not use estimated distributions of default or other data to fully compute the loan terms that lending facilities should set. The focus is on computing the welfare gains from optimally designing a facility that tackles the adverse selection arising from unobserved monitoring intensity. As it will be clear, just three sufficient statistics are adequate for this. Nevertheless, the model, and its extension in the appendix , could also be useful for calculating all loan terms using as inputs data about probabilities of default, loss given default, and other observable firm and bank characteristics. See English and Liang (2020) for an attempt to derive loan pricing terms using observable firm characterstics and inferred probabilities of default, but with no mention of the role of monitoring, which is central herein.

[6]The baseline model covers the case of lending against firms' cash-flows with same observable characteristics and probabilities of default. See the Appendix for an extension of the model to firms with different default probabilities and loans secured by collateral with implications for asset-based lending facilities.

commit to continue monitoring borrowers, as shown in Gorton and Pennacchi (1995) and Pennacchi (1988).

In a systemic crisis, all banks may suffer from the same shock, which pushes up their cost of funding and deters them from lending. At the same time, selling loans to outside investor without the ability to monitor may not be feasible even if banks retain a big portion of the loan to signal their commitment to continue monitoring. A central bank can step in and provide the required funding/liquidity at a low cost, but also needs to make sure that banks continue to monitor to avoid bad behavior rendering loans non-performing down the road. In turn, this requires banks to retain a portion of loans and be adequately compensated for monitoring. If all small businesses require equal monitoring, this would dictate a flat structure specifying the same risk-retention requirements and interest rates for all loans to firms with the same observable characteristics (for example, the MSNLF and the MSPLF apply to firms with different debt to EBITDA, but they do not distinguish among firms with the same debt to EBITDA).

The novelty of this paper is that entrepreneurs are heterogeneous in the monitoring intensity required. Contrary to banks, the central bank cannot distinguish the entrepreneurs that require monitoring from those that do not. Importantly, the required monitoring intensity does not correlate in a systematic way with observable firms' characteristics and, thus, it cannot be inferred.[7] Hence, the central bank faces adverse selection if it negotiates the terms of loan purchases with each bank separately, or if it commits to uniform/flat terms for all loans. If it offers a lower price than what is required to compensate banks for monitoring, banks will be tempted to offload monitoring-intensive loans and stop monitoring thereafter. If it offers a price compensating for monitoring, banks will be tempted to also sell loans that do not require monitoring, extracting extra payment.

The optimal design consists of a multi-tier pricing scheme that the central bank facility commits to. The scheme sets differential purchase prices, loan rates and risk-retention ratios restoring the incentives of banks to differentiate between different types of entrepreneurs and continue monitoring. In other words, loans to firms with the same observable characteristics, such as the probability of default, could require differential risk-retention ratios. This design achieves the highest liquidity support to businesses at the lowest feasible rates maximizing the surplus to the real economy.

[7] Gustafson, Ivanov and Meisenzahl (2020) argue that there is little evidence of a significant tradeoff between their monitoring frequency measure and either loan amount or loan spreads despite the fact that, in their data, loans with higher credit spreads tend to be monitored less often.

Alternatively, the central bank may not attempt to account for the unobservable borrower heterogeneity in terms of monitoring intensity and, instead, offer a uniform/flat risk-retention requirement for all loans. The welfare gain of the optimal design compared to a uniform/flat scheme for all loans depends on three sufficient statistics: the liquidity premium, the (appropriately chosen) flat risk-retention ratio, and the share of loans that require monitoring. The risk-retention ratio is a proxy for the monitoring cost, while the liquidity premium proxies for the incremental cost of bank funding. Intuitively, the gain is increasing in the liquidity premium and the monitoring cost, and decreasing in the share of loans requiring monitoring. The reason is that the optimal design avoids compensating banks for loans that do not require monitoring, while at the same time all of these loans are sold to the central bank eliminating the liquidity premium banks would demand to hold a portion of them on their balance sheets. For empirically plausible value for these three sufficient statistics, the welfare loss from using a uniform/flat scheme with a flat risk-retention requirement ranges from 0.02% to 0.31% of the total face value of loan purchases by the central bank.

Nevertheless, the operational complexities of implementing the optimal design may be daunting. In that case offering a uniform/flat scheme may not provide all the right incentives to banks and may be relatively more costly for businesses, but it is still preferred than not providing any liquidity support. Moreover, for values for the sufficient statistics observed in the outbreak of the the COVID-19 crisis, the welfare loss is relatively small. Thus, uniform/flat pricing could appear as an attractive practical solution.

The rest of the paper proceeds as follows. Section 2 presents the model and the private equilibrium in the absence of central bank intervention. Section 4 derives the optimal institutional design, while section 5 estimates the welfare cost from using uniform/flat structure using a minimal set of sufficient statistics. Section 6 discusses some practical implementation issues and concludes.

2 Model

The economy has three dates, $t = 0, 1$, and 2, and is populated by a central bank and a continuum of three types of agents: entrepreneurs of type 1 with mass m_1, entrepreneurs of type 2 with mass m_2, and banks with mass 1. All agents are risk neutral and have a time discount factor of 1.

Each entrepreneur of either type 1 or type 2 is endowed with a project that requires an investment of 1 at $t = 0$ and generates at $t = 2$ a payoff $R > 0$ in the case of success and 0 in the case of failure. Projects yield nothing if there are liquidated early at $t = 1$. Entrepreneurs do not have funds of their own and need to borrow from banks. For simplicity, banks have a fixed weighted average cost of capital, normalized to 1, and total resources $e \geq m_1 + m_2$ at $t = 0$. The difference between type 1 and type 2 entrepreneurs accrues from a moral hazard problem. Type 1 entrepreneurs have a probability of success θ irrespective of being monitored or not, with $\theta R > 1$. Type 2 entrepreneurs may choose to shirk, which destroys the firms' value and reduces the probability of success from θ to $\tilde{\theta}$, with $\tilde{\theta} R < 1$, i.e., resulting in a negative net present value.[8] Shirking can be avoided by continuous (active) bank monitoring, which introduces an additional cost $X \in (0, \theta R - 1)$. Moreover, a pooling equilibrium with lending to all entrepreneurs and no monitoring for type 2 ones is not feasible, i.e., $(m_1\theta + m_2\tilde{\theta})R < 1$.[9] For simplicity, banks know what entrepreneurs are of type 1 and type 2, but the central bank does not. No one knows whether a project pays off until period 2.[10]

Apart from lending to entrepreneurs at $t = 0$, banks may choose to invest some resources in a riskless one-period storage technology with zero net yield. At $t = 1$ all banks receive with probability q, and at the same time, an investment opportunity that yields a riskless payoff $K \geq 1$ per unit of funds invested. Importantly, all banks receive this opportunity at the same time, so interbank lending is not helpful, and they are either unable to obtain outside funding or its cost is higher than K. Thus, this state resembles a systemic funding/liquidity shortage and the only way to invest in the new technology is to carry over funds from $t = 0$ using the storage technology. Instead of receiving an investment opportunity, banks could suffer liquidity shocks or face binding regulatory constraints, with K being the shadow cost of liquidity or the shadow value of relaxing the constraint. The exact nature of the demand for liquidity is not important; the systemic nature of the liquidity shortage is. The decision to lend to entrepreneurs or hoard liquidity to invest in the new technology will depend on their respective expected returns. Hence, the opportunity cost of lending a dollar at

[8]For simplicity, entrepreneurs are assumed to be identical aside from the moral hazard problem. In practice, entrepreneurs may differ in other dimensions as well. For example, they may have different probabilities of success θ. Such additional heterogeneity does not impact the results in the paper as shown in the model extension in the Appendix.

[9]All the results in the paper go through even if $(m_1\theta + m_2\tilde{\theta})R > 1$, but $(\theta - \tilde{\theta})R > X$, i.e., monitoring adds value.

[10]The model abstracts from other frictions that may impede the ability to sell loans. For example, Parlour and Plantin (2008) assume that monitoring loans before they are sold reveals information about their performance, which introduces an additional adverse selection problem. Arguably, this friction would be more relevant for existing loan relationships with banks (Main Street Expanded Loan Facility) compared to newly issued loans (Main Street New Loan Facility). Extending the model in this direction would be interesting. Note, however, that the central bank already faces adverse selection because of the heterogeneity in the required monitoring intensity even if banks do not have superior information about the probabilities of success of different borrowers.

$t = 0$ is $1+q(K-1)$ irrespective of the nature of the liquidity shock, where $q(K-1)$ is the liquidity premium.[11]

More generally, $1+q(K-1)$ could be thought as the shadow value of a dollar at $t = 0$, with the realized shadow value being 1 with probability $(1-q)$ and K with probability q. Under this general interpretation the model can be applied to a situation where the cost of capital increases to K with probability q inducing banks to require higher returns to extend loans. Thus, banks may be unwilling to extend credit to positive net present value projects, because of binding financial frictions, tighter balance sheet constraints, or elevated aversion to aggregate risk. Although the micro-foundations for these alternative frictions differ, the basic principle and qualitative results in the paper continue to hold, as long as the central bank can alleviate these constraints and/or has higher tolerance for aggregate risk. However, the quantitative results may be markedly different and the welfare gain from implementing the optimal design may be much higher, as discussed in section 5.

Finally, note that there are no restrictions on q and for $q \to 1$ the aggregate shock has materialized and the need for liquidity/central bank support is imminent, which many may believe has been the situation after the outbreak of the COVID-19 pandemic. The value of q is not important for the qualitative and quantitative results (given the use of sufficient statistics). However, it is important that banks know at $t = 0$, when setting the terms for the new loans, that the central bank will intervene with probability q. Also, time periods do not need to have the same length. For example, the time between $t = 0$ and $t = 1$ can be several months, while the time between $t = 1$ and $t = 2$ can be several years. For example, the Main Street Lending Facilities of the Federal Reserve will purchase loans that are originated between April 24, 2020 and September 30, 2020, and have a maturity of four years. Despite being stylized, this framework has the necessary ingredients to justify a central bank lending facility that buys bank loans. A more elaborate framework would endogenize the reasons why banks cannot raise enough funds, from example, similar to Holmström and Tirole (1998).

Figure 1 presents the timeline of the model where the role of the central bank is suppressed. Optimal central bank intervention is discussed in section 4 after the private equilibrium in section

[11] Note that, for simplicity, the occurrence of the systemic shock does not affect the probability that projects succeed at t=2. One might expect that the systemic shock would reduce this probability from θ to some θ'. As long as the project continues to have positive net present value, i.e., $\theta' R > 1$, all the results in the paper go through. Naturally, if $\theta' R < 1$, the central bank would not intervene as it would violate its no loss constraint by buying loans of insolvent firms.

3. Note that the monitoring takes place between $t = 1$ and $t = 2$ to capture the finding of continuous active monitoring in Gustafson et al. (2020).

$t = 0$	$t = 1$	$t = 2$
Banks decide whether to lend to type 1 and/or type 2 entrepreneurs Banks set type 1 and 2 loan rates Banks store remaining funds	Liquidity shock materializes or not Banks invest in new technology if shock hits Banks use storage if shock doesn't hit Banks monitor type 2 loans	Projects succeed or fail Entrepreneurs repay their loans All agents consume

Figure 1: Model Timeline without Central Bank Facility

3 Private Equilibrium

Suppose that $\theta R \geq 1 + q(K-1) + X$, i.e., the payoff from lending to either type 1 or type 2 entrepreneurs is higher than the payoff from hoarding liquidity and investing in the new technology if it arrives at $t = 1$, accounting for any monitoring costs. Given perfect competition, banks will lend to both types at competitive rates $\hat{R}_1 = (1 + q(K-1))/\theta$ and $\hat{R}_2 = (1 + q(K-1) + X)/\theta$, respectively. Moreover, they will invest the remaining resources $e - m_1 - m_2$ in the storage technology. Equivalently, for $\theta R \in [1 + q(K-1), 1 + q(K-1) + X)$, banks will voluntarily lend only to type 1 entrepreneurs and invest the remaining resources, $e - m_1$, in the storage technology. For $\theta R < 1 + q(K-1)$ banks are unwilling to lend to either type 1 or type 2 entrepreneurs. This case represent a systemic liquidity crisis in the sense that there are positive net present value projects available at $t = 0$, but banks choose to hoard their existing funds. The next section derives the optimal institutional design to deal with this extreme situation.

4 Central Bank Lending Facility

The central bank can set up a facility to provide banks with liquidity in order to support lending to businesses. In particular, the central bank sets a Special Purpose Vehicle (SPV) that buys business loans from banks at $t = 1$ if and when the liquidity shock hits, which is common knowledge.

The objective of the central bank is to maximize the surplus accruing to entrepreneurs, without incurring any losses in its loan portfolio, making sure that banks are willing to intermediate, and maintaining bank incentives for monitoring.[12] The pricing of the loans by the SPV is important to provide the right incentives to banks such that the SPV, first, achieves its purpose of supporting business lending for positive net present value projects, second, it does not overpay *on aggregate*, which would constitute a direct transfer of surplus from businesses to banks, and, third, that monitoring takes place. The main difficulty that the SPV faces is that it does not know whether it will be buying type 1 or type 2 loans, since this is private information to banks. Hence, the central bank faces adverse selection.

Figure 2 presents the augmented timeline with the central bank.

$t = 0$	$t = 1$	$t = 2$
SPV sets terms for purchasing loans Banks decide to lend to type 1 and type 2 entrepreneurs Banks store remaining funds	Liquidity shock materializes or not Banks sells loans to SPV and invest in new technology if shock hits Banks use storage if shock doesn't hit Banks monitor type 2 loans given adequate incentives	Projects succeed or fail Entrepreneurs repay their loans All agents consume

Figure 2: Model Timeline with Central Bank Facility

There are three problems that the SPV needs to resolve. First, consider that the SPV does not condition pricing terms to loan interest rates that can help distinguish type 1 and type 2 loans—or, equivalently, commits to a uniform/flat structure for all loans. If banks are not compensated for monitoring, they would be tempted to sell to the SPV type 2 loans pretending that they are type 1, and not engage in monitoring thereafter. This is undesirable because the aggregate loan pool has a negative net present value if type 2 loans are not monitored. If, instead, banks are compensated for monitoring type 2 loans, they will be tempted to sell type 1 loans at those terms as well, extracting a higher surplus. Second, there is no guarantee that banks will actually monitor type 2 loans even if they are offered a payment that compensates them for it. Third, the central bank needs to price loans such that it does not lose money on its aggregate portfolio.

An SPV design featuring two pricing schemes for loan purchases can resolve the three aforementioned problems and achieve maximum lending support to businesses. The pricing schemes define the price paid for loans with certain loan rates as well as the risk-retention ratio. Importantly,

[12] Recall that the projects of both entrepreneurial types have positive net present value (if monitored), so lending to both types is socially optimal.

the SPV pre-commits to this pricing scheme at $t=0$.

The first pricing scheme sets a price P_1 that the bank is willing to pay for type 1 loans with rate R_1 and zero risk-retention. The price offered to banks should make them willing to extend credit instead of hoarding liquidity at $t=0$, i.e., the following participation constraint should hold:

$$(1-q)\theta R_1 + qP_1K \geq 1 + q(K-1). \tag{1}$$

As shown later, the loan rate R_1 will be chosen such that the SPV does not lose money on its aggregate loan portfolio.

The second pricing scheme sets a price P_2 and risk-retention ratios γ_2 for type 2 loans with rate R_2. The SPV chooses the terms such that banks (i) use this scheme to sell only type 2 loans, (ii) have an incentive to continue monitoring after selling some loans to the SPV, and (iii) have an incentive to extend lending to type 2 entrepreneurs. We show how these conditions determine the terms in reverse order.

Banks will lend to good type 2 entrepreneurs at $t=0$, anticipating that they can sell a portion $1-\gamma_2$ to the SPV at price P_2 at $t=1$ with probability q, if the following participation constraint is satisfied:[13]

$$(1-q)\theta R_2 + q\left(\gamma_2\theta R_2 + (1-\gamma_2)P_2K\right) \geq 1 + q(K-1) + X. \tag{2}$$

Banks will have the incentive to continue monitoring after selling a portion of their type 2 loan portfolio to the SPV at $t=1$ if the following incentive compatibility constraint is satisfied:

$$\gamma_2\theta R_2 \geq X, \tag{3}$$

Banks should not have an incentive to misrepresent type 2 loans for type 1 loans (and vice versa). Recall that the SPV pays P_1 and set zero risk retention for loans with observable rate R_1. So, banks could restructure type 2 loans before selling them to the SPV to change the rate from R_2 to R_1, which entrepreneurs would be happy to accept, and sell the whole loan to the SPV at price P_1.

[13] Multiplying (1) by m_1 and (2) by m_2, and adding them, one can obtain the aggregate participation constraint.

Banks will truthfully reveal type 2 loans if the following constraint holds:[14]

$$\gamma_2 \theta R_2 + (1-\gamma_2) P_2 K - X \geq P_1 K. \tag{4}$$

Banks may also be tempted to misrepresent type 1 loans to receive a higher payment accounting for a monitoring cost that they do not need to incur. But, they are unable to do so because this would require restructuring type 1 loans to change the rate from R_1 to R_2, which entrepreneurs would not accept (recall the SPV offers P_2 for loans with observable rate R_2). Equivalently, banks will not be able to issue type 1 loans at rate R_2 at $t = 0$, because there is a profitable deviation of offering R_1 and attracting all type 1 entrepreneurs.

Finally, the SPV needs to guarantee that its aggregate portfolio does not lose money. Under the two pricing schemes described above, banks will extend loans to all type 1 and type 2 entrepreneurs, i.e., the requirement becomes

$$m_1(\theta R_1 - P_1) + m_2(1-\gamma_2)(\theta R_2 - P_2) \geq 0. \tag{5}$$

Formally, the SPV chooses terms (R_1, P_1) and (R_2, P_2, γ_2) to maximize an objective function

$$W = m_1 \theta (R - R_1) + m_2 \theta (R - R_2) - L, \tag{6}$$

subject to constraints (1)-(5). $L > 0$ represents an unmodeled loss to the central bank and could potentially be associated to reputational concerns from intervening in credit markets. Thus, the central bank will intervene only if the surplus to the real economy is high enough to justify incurring the loss and would definitely not intervene when the liquidity premium is low and banks voluntarily lend to entrepreneurs. Note that L is not a pecuniary loss from the loan portfolio given constraint (5).[15]

Because of perfect competition in the banking sector, the SPV chooses P_1 and P_2 such that (1)

[14] Note that banks will not want to issue type 2 loans at rate R_1 given that they can restructure at $t = 1$, because for $q < 1$, there is a chance that they don't sell them to the SPV. But, even for $q = 1$, they would not issue type 2 loans at rate R_1, and they will truthfully reveal their type, if (4) holds

[15] None of the results depend on the loss L, which can been easily set to zero. Though in that case the model would predict that the central bank should always intervene, even when the liquidity premium is very low, because, contrary to banks, the central bank does not require compensation for it and can offer better loan rates (see corollary 1 below).

and (2) hold with equality, i.e., it pays the lowest possible price to banks. Moreover, (3) and (4) hold with equality because higher risk retention makes banks require a higher payment without improving further monitoring incentives. Finally, (5) also bind because the SPV would rather offer entrepreneurs better rates to maximize their surplus than making a positive profit. Altogether these yield a closed-form solution for the the optimal pricing schemes characterized in the following proposition.

Proposition 1. *For $\theta R < 1 + q(K-1)$ and $(m_1+m_2)(\theta R-1) - m_2 X > L$ the central bank intervenes. The optimal SPV design involves two distinct pricing schemes: A price $P_1 = 1$ for loans with rate $R_1 = 1/\theta$ and zero risk-retention requirements, and a price $P_2 = 1+X$ for loans with rate $R_2 = (1+X)/\theta$ and risk-retention requirement $\gamma_2 = X/(1+X)$. Banks will choose to sell type 1 and type 2 loans to the first and second schemes, respectively.*

This design provides the right incentive to banks to truthfully report the type of loan and achieves maximum lending to entrepreneurs. Banks sell type 2 loans at the second scheme, retain a portion, and maintain the incentive to monitor, while they fully sell type 1 loan at the first scheme. But, as the following corollary shows, the central bank facility is also allowing businesses to borrow at the lowest rates possible maximizing their surplus, because banks are compensated for the liquidity premium.

Corollary 1. *Loan rates to both type 1 and type 2 entrepreneurs are lower under the optimal SPV design compared to the competitive rates in normal times by an amount equal to the liquidity premium $q(K-1)$.*

Now, suppose that the SPV does not condition pricing terms to loan interest rates. Instead, it sets at $t = 0$ a uniform/flat pricing scheme, P^*, for all loans with the same loan rate, R^*, along with a uniform/flat risk-retention ratio, γ^*, to maintain the incentives for monitoring. Given that monitoring is necessary only for type 2 entrepreneurs, banks will be willing to lend at $t = 0$ to a pool of type 1 and type 2 entrepreneurs at the same rate if the following participation (aggregate) constraint is satisfied:

$$(1-q)\theta R^* + q[\gamma^* \theta R^* + (1-\gamma^*)P^* K] \geq 1 + q(K-1) + m_2/(m_1+m_2)X. \tag{7}$$

Moreover, banks will maintain incentives to monitor type 2 loans after selling a portion at $t = 1$ if the following incentive compatibility constraint is satisfied:

$$\gamma^{*}\theta R^{*} \geq X. \tag{8}$$

Finally, the SPV pays banks a price that results in no losses on the loan portfolio, i.e.,

$$(m_1 + m_2)(1 - \gamma^{*})(\theta R^{*} - P^{*}) \geq 0 \tag{9}$$

The SPV chooses $(R^{*}, P^{*}, \gamma^{*})$ to maximize

$$W^{*} = (m_1 + m_2)\theta(R - R^{*}) - L, \tag{10}$$

subject to constraints (7)-(9).

As above, all constraints bind in equilibrium yielding the uniform/flat pricing scheme described in the following proposition.

Proposition 2. *A uniform/flat pricing scheme sets price* $P^{*} = 1 + X[m_2/(m_1 + m_2) + q(K-1)]/[1 + q(K-1)]$ *for loans with rate* $R^{*} = P^{*}/\theta$*, and requires banks to retain* $\gamma^{*} = X/P^{*}$.

The following proposition compares the optimal two-tier and uniform/flat pricing schemes.

Proposition 3. *Compared to the optimal multi-tier pricing scheme,* $0 < (m_1 + m_2)\theta(R - R^{*}) < m_1\theta(R - R_1) + m_2\theta(R - R_2)$*, i.e., the surplus to entrepreneurs is smaller under the uniform/flat pricing scheme. The surplus gain is equal to* $m_1 q(K-1)X/(1 + q(K-1))$.

Proposition 3 shows that the optimal two-tier pricing system is preferred to uniform/flat pricing. The gain in surplus/welfare is increasing in the number of type 1 entrepreneurs, the cost of monitoring, and the liquidity premium. The reason is that the optimal design avoids compensating banks for loans that do not require monitoring, while at the same time all of these loans are sold to the central bank eliminating the liquidity premium banks would demand to hold a portion of them on their balance sheets.

Uniform pricing also delivers economic benefits and should be preferred to taking no action. Note, however, that uniform pricing also needs to be accompanied by a positive risk-retention requirement; otherwise, monitoring for type 2 loans will not take place and the overall pool of loans purchased by the SPV will have negative net present value as $(m_1\theta + m_B\tilde{\theta})R < 1$.[16]

5 Sufficient Statistics for the Welfare Gain

In practice, central banks may not know what is the cost of monitoring for all the different banks using the facility, which is a key parameter in the model. However, it could be potentially inferred by a relationship of the type described in condition (3) or (8). In other words, policymakers may use the voluntary risk-retention ratios on past loans to gauge the shadow cost of monitoring accounting for all other loan characteristics. For example, Sufi (2007) shows that the lead share—share of a loan held by the lead bank in syndicate—is a proxy for monitoring, and is on average 28.5 percent (23.5 percent median) in his database. Alternatively, the 5 percent risk-retention requirement for issuers of Asset Backed Securities in Dodd-Frank could be used to calibrate γ in the model.[17]

Other parameters could also be calibrated using existing empirical studies. The share of type 2 entrepreneurs could be calibrated to 20 percent, i.e., equal the percentage of syndicated loans actively monitored in Gustafson et al. (2020). Moreover, the liquidity premium $q(K-1)$ captures the additional return from investing the available fund in the risk-free, short-term, storage technology and having them available for use at $t = 1$. In other words, the price of a risk-free, long-term (two period), discount bond would be $1/(1+q(K-1))$. As such, the liquidity premium can be approximated by the spread between the one-month overnight index swap (OIS) and the four-week Treasury bill (T-bill)—a spread known as the convenience yield (see, Infante, 2020; Cashin et al. 2017). Given that T-bills are publicly produced short-term safe assets, trading with very narrow bid-ask spreads, and the OIS is merely a contractual agreement promising a risk-free payoff, the spread measures the convenience of holding safe assets. Alternatively, the liquidity premium could be approximated by the spread between the General Collateral (GC) repo rate and the (4-week) T-bill (Nagel, 2016) or the spread between the 1-week Treasury repo rate and the (4-week) T-bill, because these repo rates are free of risk as they are backed by safe collateral, but investing in them is illiquid during the term of the loan.[18] Figure 3 plots the monthly averages for these three proxies

[16]Naturally, if monitoring is not essential but useful, i.e., $(m_1\theta + m_B\tilde{\theta})R > 1$ and $(\theta - \tilde{\theta})R > X$, then uniform pricing with no risk-retention requirements is viable, yet still not optimal.

[17]https://www.govinfo.gov/content/pkg/FR-2014-12-24/pdf/2014-29256.pdf.

[18]Nagel (2016) discussed additional proxies such as the spread between the CDs rate and the T-bill, the spread between

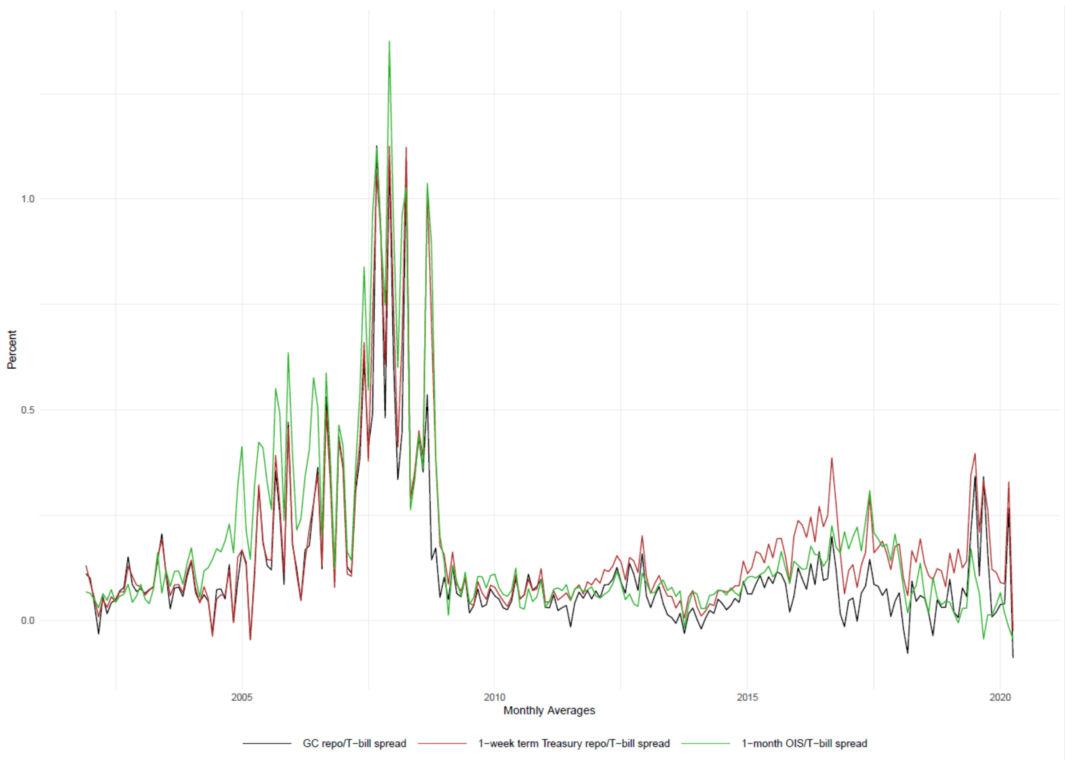

Figure 3: Estimates for the Liquidity Premium

for the liquidity premium.

The following corollary recasts the surplus/welfare (percentage) gain in terms of the aforementioned sufficient statistics: the risk-retention requirement, γ^*, the share of firms requiring monitoring $\bar{m} \equiv m_2/(m_1+m_2)$, and the liquidity premium, $r_\ell \equiv q(K-1)$, using the equilibrium conditions.

Corollary 2. *The percentage welfare gain of using the optimal design compared to a uniform/flat pricing scheme is given by*

$$\%\mathit{Welfare\ Gain} = \frac{(1-\bar{m})\gamma^* r_\ell}{1+(1-\gamma^*)r_\ell - \gamma^* \bar{m}}.$$

Note that the flat risk-retention requirement, γ^*, is a sufficient statistic, which means that the corollary can be used to evaluate the welfare loss from the uniform/flat pricing scheme observed in actual lending facilities.

the off-the-run and on-the-run Treasure notes, and the spread between the off-the-run Treasury notes and the T-bill. These measures additionally capture either some credit risk (CDs) or a premium for imperfect market liquidity (Treasury notes).

Table 1 reports the welfare gain for different levels of the liquidity premium and risk-retention requirement when the share of loans that require monitoring is set to 20 percent. For a 5 percent flat risk-retention requirement and a liquidity premium close to the one suggested by the COVID-19 spike in the three liquidity measures, the gain is minuscule. But even for a higher liquidity premium close to what was observed during the Financial Crisis of 2007-2008 and a flat risk-retention requirement calibrated close to the lead share in Sufi (2007), the welfare gains are relatively small. Hence, uniform/flat pricing may not be optimal, but it does not imply a big welfare loss.

	γ		
r_ℓ	0.05	0.15	0.25
0.50%	0.02%	0.06%	0.10%
1.00%	0.04%	0.12%	0.21%
1.50%	0.06%	0.18%	0.31%

Table 1: Welfare gain of optimal design relative to uniform/flat pricing for different levels of risk-retention requirements and liquidity premium. The share of loans that required monitoring is set to 20 percent.

The model interprets the rise in banks' funding costs as an increase in the liquidity premium. Alternatively, one could consider that funding costs increase for additional reasons during the crisis. As mentioned, r_ℓ could more generally be interpreted as the incremental increase in the shadow cost of bank funding, which may be higher than the increase in the liquidity premium due to, for example, binding balance sheet constraints and a higher cost of bank equity. In other words, banks would need to commit equity capital to fund loans to businesses, the price of which may increase a lot during a crisis translating into a high r_ℓ in the model. Naturally, this may result in higher welfare losses than the ones reported in Table 1, which can be seen as a reasonable lower bound, rendering the implementation of the optimal design more important. Given estimates for the increase in the shadow cost of bank equity capital during crises, one could derive a different estimate for r_ℓ and use corollary 2 to compute the losses from implementing a uniform/flat structure.[19]

Note that these relative welfare-gain calculations presume that the SPV has set the risk retention at the appropriate level to resolve the moral hazard problem, or in other words, the risk-retention ratio is a sufficient statistic for the monitoring cost. Both the optimal and flat-pricing designs would

[19] Of note, regulatory actions allowing the exclusion of new loans to small businesses from the calculation of regulatory capital requirements could complement central bank purchases in order to alleviate the pressure from binding balance sheet constraints. This combined policy response is interesting, but beyond the scope of this paper.

result in negative net present value if the risk-retention requirement is not high enough to satisfy the incentive compatibility constraint for monitoring.

6 Conclusion

The model captures a key difficulty central banks may face when purchasing loans extended to small businesses; namely that bank monitoring may be necessary, but not observable/contractible. Establishing a lending facility with tiered pricing schemes, for given observable borrower characteristics, is optimal to provide the right incentives to banks, while maximizing lending to businesses at the lowest possible rates. The model focuses on small business loans because they are likely candidates for requiring continuous monitoring. But, the mechanism and conclusions in this paper could apply to other cases where active monitoring is important, such as the syndicated loan market (Gustafson et al., 2020). However, being stylized, the model leaves a number of implementation issues unanswered.

For instance, loans may have different characteristics and covenants mandating when ownership should be transferred or other actions are to be taken. Not all these covenants are present to address moral hazard issues by borrowers, but some of them may act as an ex ante screening mechanism. The baseline model is extended in the appendix to incorporate not only ex post monitoring, but also ex ante screening to separate entrepreneurs with good projects from entrepreneurs with bad projects. To the extend that the various covenants cannot tackle perfectly the moral hazard problem and monitoring is still useful, the two-tier pricing scheme along with differential risk-retention ratios is still the optimal institutional design for the central bank lending facility. In turn, this is reasonable for small business loans given that they do not have a long track record and are typically unrated. Similarly, small businesses may lack tangible and liquid collateral, and they usually collateralize cash flows or have covenants that mandate the transfer of control rights, which may be only valuable if entrepreneurs do not engage in destructive behavior.

Moreover, the operational complexity of implementing a tiered pricing facility may be high. Taking such important concerns into consideration, the SPV may decide to offer a uniform/flat pricing scheme similar to the one described in proposition 2. That would be still preferable to not providing the necessary liquidity to the economy, while the welfare cost is not expected to be that high.

Overall, practical implementation can be daunting. Nevertheless, the main principle of having pricing schemes that depend not only on observable borrower characteristics, but also on monitoring intensity and monitoring cost should continue to hold in more elaborate environments. It is also true that the central bank lending facility in this model mainly delivers benefits, because central banks do not need to be compensated for the shadow value of liquidity. Central banks will weigh the benefits to the real economy from intervening to the relevant costs, which may be reputational or operational, given that the terms are chosen such that the loan portfolio does not lose money in expectation. Only when the shadow value of liquidity is high enough, and in accordance to their mandate, would lending facilities being employed. Such considerations do not affect the message of the paper, as the shadow value of liquidity can be made arbitrarily high to represent systemic liquidity shortages and justify central bank intervention only in crisis periods.

References

Cashin, David, Erin Syron Ferris, Beth Klee and Cailey Stevens (2017), 'Take it to the limit: The debt ceiling and treasury yields', *Finance and Economics Discussion Series 2017-052* .

Diamond, Douglas W. (1984), 'Financial intermediation and delegated monitoring', *Review of Economic Studies* **51**(3), 393–414.

Drechsel, Thomas and Sebnem Kalemli-Özcan (2020), 'Are standard macro policies enough to deal with the economic fallout from a global pandemic?'.
URL: *https://econfip.org/policy-brief/are-standard-macro-policies-enough-to-deal-with-the-economic-fallout-from-a-global-pandemic/*

Dreyer, Mallory, Christian McNamara, Alexander Nye, Kaleb Nygaard and Priya Sankar (2020), 'Large-scale assistance programs for small businesses'.
URL: *https://som.yale.edu/blog/large-scale-assistance-programs-for-small-businesses*

English, William B. and Nellie Liang (2020), 'Designing the Main Street Lending Program: Challenges and Options', *Hutchins Center Working Paper # 64* .

Gorton, Gary B. and George G. Pennacchi (1995), 'Banks and loan sales marketing nonmarketable assets', *Journal of Monetary Economics* **35**(3), 389–411.

Gustafson, Matthew T., Ivan T. Ivanov and Ralf R. Meisenzahl (2020), 'Bank monitoring: Evidence from syndicated loans', *Journal of Financial Economics, forthcoming* .

Holmström, Bengt and Jean Tirole (1997), 'Financial intermediation, loanable funds, and the real sector', *The Quarterly Journal of Economics* **112**(3), 663–691.

Holmström, Bengt and Jean Tirole (1998), 'Private and public supply of liquidity', *The Journal of Political Economy* **106**(1), 1–40.

Infante, Sebastian (2020), 'Private money creation with safe assets and term premia', *Journal of Financial Economics* **136**(3), 828 – 856.

Nagel, Stefan (2016), 'The Liquidity Premium of Near-Money Assets', *The Quarterly Journal of Economics* **131**(4), 1927–1971.

Parlour, Christine A. and Guillaume Plantin (2008), 'Loan sales and relationship banking', *The Journal of Finance* **63**(3), 1291–1314.

Pennacchi, George G. (1988), 'Loan sales and the cost of bank capital', *The Journal of Finance* **43**(2), 375–396.

Saki, Bigio (2020), 'A concrete policy response to the Covid-19 crisis'.
URL: *http://saki-bigio.squarespace.com/policy-recommendation*

Sufi, Amir (2007), 'Information asymmetry and financing arrangements: Evidence from syndicated loans', *The Journal of Finance* **62**(2), 629–668.

Appendix

The appendix extends the baseline model to account for adverse selection. Within each type of entrepreneurs, there are two sub-types: good and bad. Good entrepreneurs of either type 1 or type 2 have a probability of success θ_G, while bad entrepreneurs of either type have a probability of success θ_B satisfying $\theta_G R > 1+X > 1 > \theta_B R$. In other words, only good entrepreneurs have a positive net present value project, even if monitoring is not required. Denote by α the relative percentage of good entrepreneurs, which, for simplicity and without loss of generality, is the same for both types 1 and 2. Then, the average payoff for a pool of type 1 or 2 entrepreneurs is $\bar{R} = \bar{R}_1 = \bar{R}_2 = (\alpha\theta_G + (1-\alpha)\theta_B)R$, and the average payoff for the whole population is $(m_1+m_2)\bar{R}$, where $\bar{R} > 1+X$. Hence, the average project is similar to the project in the baseline model.

Before extending credit, banks can choose to screen entrepreneurs by setting non-price terms along with the loan rate. For simplicity, we assume that the non-price loan terms take the form of a payment to banks in the case of failure. Thus, banks can offer a two-dimensional loan contract (R_s, C_s), where R_s is the loan rate and C_s are the non-price loan terms. This characterization is meant to proxy for collateral posted or for other covenants mandating, for example, the transfer of ownership to banks or other restrictions decreasing the value to entrepreneurs and increasing the value to banks. Entrepreneurs lose C_s in the event of failure, while banks gain κC_s, where $1-\kappa \geq 0$ is the per unit bank cost of enforcing/servicing the non-price terms. For conciseness, C_s will be referred to as collateral.[20] Instead of offering a two-dimensional contract, banks may choose to pool all borrowers together by offering a one-dimensional loan contract defining only the loan rate R_p, as is the case in the baseline model.

With respect to the moral hazard friction, monitoring type 2 entrepreneurs—either good or bad—guarantees both that their project will not fail with certainty and that the collateral will not be useless. The latter is meant to capture the fact that the value of the various non-price terms depends on entrepreneurial effort. The reason could be that small businesses can only post intangible collateral in the form of early payments, accounts receivable (trade credit), or transfer of ownership rights, which may be worthless if the entrepreneur engages in destructive behavior (for example

[20] In reality, small business loans are multidimensional characterized by many covenants. The simplified structure of the model extension is meant to capture the idea that the appropriate set of covenants can be used to separate good from bad borrowers.

damaging organizational structures, or using them to establish other businesses). In sum, this assumption implies that non-price loan terms are not sufficient to discourage bad behavior, and that bank monitoring can effectively put a stop to this agency problem.[21]

Consider type 1 entrepreneurs. The separating contract, $(\hat{R}_s^1, \hat{C}_s^1)$, should separate good from bad entrepreneurs and be profitable for each bank. The former condition requires that $p_B(R - \hat{R}_1^s) - (1-p_B)\hat{C}_1^s \leq 0$, while the second requires $p_G\hat{R}_1^s + \kappa(1-p_G)\hat{C}_1^s \geq 1 + q(K-1)$. Because collateral enforcement is costly and because of perfect competition in the banking sector, these two constraints hold with equality $\hat{R}_1^s = ((1+q(K-1))(1-p_B) - \kappa p_B(1-p_G)R)(p_G(1-p_B) - \kappa p_B(1-p_G))$ and $\hat{C}_s^1 = p_B/(1-p_B) \cdot ((p_G R - (1+q(K-1)))(1-p_B))(p_G(1-p_B) - \kappa p_B(1-p_G))$. Similarly, the separating contract terms for type 2 entrepreneurs are $\hat{R}_2^s = ((1+q(K-1)+X)(1-p_B) - \kappa p_B(1-p_G)R)(p_G(1-p_B) - \kappa p_B(1-p_G))$ and $\hat{C}_s^2 = p_B/(1-p_B) \cdot ((p_G R - (1+q(K-1)+X))(1-p_B))(p_G(1-p_B) - \kappa p_B(1-p_G))$. Instead, the pooling contract sets only the loan rate equal to $\hat{R}_1^p = (1+q(K-1))/(\alpha p_G + (1-\alpha)p_B)$ and $\hat{R}_1^p = (1+q(K-1))/(\alpha p_G + (1-\alpha)p_B)$ for type 1 and type 2 entrepreneurs, respectively.

The following proposition derives which contract will be chosen in equilibrium.

Proposition A.1. *Consider $\theta_G R \geq 1 + q(K-1)$. For $\alpha > \bar{\alpha}_1 \in (((1+q(K-1))/R - p_B)/(p_G - p_B), 1)$, there exists $\bar{\kappa}_1 \in (0,1)$ such that banks choose the pooling contract for type 1 entrepreneurs for $\kappa < \bar{\kappa}_1(\alpha)$ and they choose the separating contract with terms otherwise. For $\alpha \in ((1+q(K-1))/R - p_B)/(p_G - p_B)$ only the separating contract is chosen.*

Proof. The separating contract constitutes an equilibrium if no individual bank has an incentive to deviate and offer the pooling contract. In turn, this is true if the effective payment for good entrepreneurs, i.e., $\hat{R}_s^1 + (1-p_G)/p_G\hat{C}_s^1$ is lower that the pooling rate, i.e., $\hat{R}_1^p$. First, note that for $\alpha \leq ((1+q(K-1))/R - p_B)/(p_G - p_B)$, $\hat{R}_p^1$ is higher that R and, thus, no entrepreneur will choose to borrow. Hence, only the separating contract would be viable. Next, observe that: $d(\hat{R}_s^1 + (1-p_G)/p_G\hat{C}_s^1)/d\kappa < 0$; that $\hat{R}_s^1 + (1-p_G)/p_G\hat{C}_s^1 < \hat{R}_p^1$ for $\kappa = 1$ irrespective of the value of α; and that for $\kappa = 0$ $\hat{R}_s^1 + (1-p_G)/p_G\hat{C}_s^1 > \hat{R}_p^1$ for $\alpha \to 1$, while $\hat{R}_s^1 + (1-p_G)/p_G\hat{C}_s^1 < \hat{R}_p^1$ for $\alpha \to ((1+q(K-1))/R - p_B)/(p_G - p_B)$. All these together imply, by continuity, that there is an

[21] This is not an unreasonable assumption. Gustafson, Ivanov and Meisenzahl (2020) show that bank monitoring for syndicated loans can either complement or substitute for covenant-based monitoring, depending on whether the monitoring informs covenant compliance.

$\bar{\alpha}_1$ such that $\hat{R}_s^1 + (1-p_G)/p_G\hat{C}_s^1 > \hat{R}_p^1$ for $\alpha > \bar{\alpha}_1$ and κ. In turn, this implies that there exists a $\bar{\kappa}_1(\alpha)$ as a function α such that the pooling contract is the equilibrium contract for $\kappa < \bar{\kappa}_1(\alpha)$ (and the separating otherwise). $\square$

Proposition A.2. *Consider that* $\theta_G R \geq 1 + q(K-1) + X$. *For* $\alpha > \bar{\alpha}_2 \in (((1+q(K-1)+X)/R - p_B)/(p_G - p_B), 1)$, *there exists* $\bar{\kappa}_2 \in (0,1)$ *such that banks choose the pooling contract for type 2 entrepreneurs for* $\kappa < \bar{\kappa}_2(\alpha)$ *and they choose the separating contract with terms otherwise. For* $\alpha \leq (((1+q(K-1)+X)/R - p_B)/(p_G - p_B)$ *only the separating contract is chosen.*

Proof. The proof is the same as for Proposition (A.1) above where $1+q(K-1)$ is replaced by $1+q(K-1)+X$. $\square$

Corollary A.1. *The separating contract for type 2 entrepreneurs carries a higher loan rate and lower collateral than for type 1 entrepreneurs, while the same is true for the loan rate in the respective pooling contracts. Moreover, separating contracts are easier to obtain for type 2 entrepreneurs* $\bar{\kappa}_2 < \bar{\kappa}_1$.

Proof. The first part is trivial given that $X > 0$. To prove the second part, observe that, for $\kappa = 1$, $\hat{R}_s^2 + (1-p_G)/p_G\hat{C}_2^s - \hat{R}_2^p < \hat{R}_s^1 + (1-p_G)/p_G\hat{C}_1^s - \hat{R}_1^p < 0$ and $d(\hat{R}_s^1 + (1-p_G)/p_G\hat{C}_1^s)/d\kappa < d(\hat{R}_s^2 + (1-p_G)/p_G\hat{C}_2^s)/d\kappa < 0$. $\square$

For $1+q(K-1) > \theta_G R$, banks do not engage in lending and, instead, hoard liquidity. As in the baseline, model the central bank can set an SPV to provide the necessary liquidity and improve outcomes. In the case of pooling contracts, the optimal design is the same one described in proposition 1 for $\theta = \alpha\theta_G + (1-\alpha)\theta_B$. The two-tier pricing system is also optimal for the case of separating contracts given that collateral cannot address the moral hazard problem. The solution is obtained by the set of conditions (1)-(5) by replacing θR_1 and θR_2 with effective expected loan payments $\theta_G R_1^s + (1-\theta_G)C_1^s$ and $\theta_G R_2^s + (1-\theta_G)C_2^s$, respectively, with $C_1^s = p_B/(1-p_B)(R - R_1^s)$ and $C_2^s = p_B/(1-p_B)(R - R_2^s)$.

Proposition A.3. *In the presence of both ex ante screening and ex post monitoring, the optimal SPV design involves two distinct pricing schemes consisting of two-dimensional and/or one-dimensional contracts depending on the level of* κ.

- *For $\kappa < \kappa'$, the optimal design sets a price $P_1 = 1$ for loans with rate $R_1 = 1/(\alpha\theta_G + (1-\alpha)\theta_B)$, zero collateral, and zero risk-retention requirements, and a price $P_2 = 1+X$ for loans with rate $R_2 = (1+X)/(\alpha\theta_G + (1-\alpha)\theta_B)$, zero collateral and risk-retention requirement $\gamma_2 = X/(1+X)$.*

- *For $\kappa \in [\kappa', \kappa'')$, the optimal design sets a price $P_1 = 1$ for loans with rate $R_1 = ((1-p_B) - \kappa p_B(1-p_G)R)(p_G(1-p_B) - \kappa p_B(1-p_G))$, collateral $C_1 = p_B/(1-p_B) \cdot ((p_G R - 1)(1-p_B))(p_G(1-p_B) - \kappa p_B(1-p_G))$, and zero risk-retention requirements, and a price $P_2 = 1+X$ for loans with rate $R_2 = (1+X)/(\alpha\theta_G + (1-\alpha)\theta_B)$, zero collateral, and risk-retention requirement $\gamma_2 = X/(1+X)$.*

- *For $\kappa \in [\kappa'', 1)$, the optimal design sets a price $P_1 = 1$ for loans with rate $R_1 = ((1-p_B) - \kappa p_B(1-p_G)R)(p_G(1-p_B) - \kappa p_B(1-p_G))$, collateral $C_1 = p_B/(1-p_B) \cdot ((p_G R - 1)(1-p_B))(p_G(1-p_B) - \kappa p_B(1-p_G))$ and zero risk-retention requirements, and a price $P_2 = 1+X$ for loans with rate $R_2 = ((1+X)(1-p_B) - \kappa p_B(1-p_G)R)(p_G(1-p_B) - \kappa p_B(1-p_G))$, collateral $C_2 = p_B/(1-p_B) \cdot ((p_G R - (1+X))(1-p_B))(p_G(1-p_B) - \kappa p_B(1-p_G))$ and risk-retention requirement $\gamma_2 = X/(1+X)$.*

Proof. The prices, contract terms, and risk-retention ratios can be derived using the same steps as in proposition 1 and realizing that for $\kappa < \kappa'$ only the pooling contracts will be chosen by banks (and only the separating ones for $\kappa \geq \kappa''$). There thresholds are derived as the thresholds $\bar{\kappa}_1$ and $\bar{\kappa}_2$ in propositions A.1 and A.2. □