The financial market impact of quantitative easing

Summary of Working Paper no. 393 Michael Joyce, Ana Lasaosa, Ibrahim Stevens and Matthew Tong

In response to the intensification of the global financial crisis towards the end of 2008, and a sharp downturn in domestic economic prospects, the Bank of England's Monetary Policy Committee (MPC) loosened monetary policy using both conventional and non-conventional means.

The MPC cut Bank Rate, the United Kingdom's policy rate, from 5% at the start of October 2008 to 0.5% in March 2009. But given the likelihood of undershooting the 2% CPI inflation target in the medium term, the Committee also decided it needed to ease monetary conditions further through a programme of asset purchases financed by the issuance of central bank reserves. This programme of large-scale asset purchases — commonly referred to as quantitative easing or QE — had resulted in the MPC making £200 billion of purchases, overwhelmingly of UK government securities (gilts), by February 2010; an amount equivalent to 14% of nominal GDP.

There are a number of ways through which injections of money into the economy via asset purchases funded by reserves might be expected to affect nominal spending growth. But one important route is through higher asset prices, which should reduce the cost of obtaining funding and increase the wealth of asset holders, thus boosting spending and increasing nominal demand. This paper assesses the impact of the Bank's QE policy on financial markets — the first leg in this transmission mechanism. We attempt to quantify how QE has affected gilt markets and how it has also fed through more widely into other financial asset prices.

There are three main channels through which QE might affect asset prices. First, the announcement of QE purchases may itself provide information to economic agents about the state of the economy and about how the MPC might be likely to react to future developments. This is a macro/policy news channel. Second, in general, provided different financial assets are not viewed as perfect substitutes by investors, QE will also have an effect through a portfolio rebalancing channel. The increase in demand for gilts resulting from the Bank's purchases will raise their prices and lower their yields. And the impact of the purchases should be felt across a range of assets, as sellers of gilts to the Bank use their new money balances to bid up the prices of other assets. Finally, the presence of a central bank in the market may improve market functioning and reduce the extra compensation ('liquidity premium') that investors demand for buying assets that risk being more difficult to sell in the future.

Asset prices in the United Kingdom recovered substantially during 2009, but not all of the improvement can be attributed to QE. A range of policies at home and abroad and other influences will have also affected asset prices. In order to isolate the impact that is directly attributable to QE, we use several approaches. We first examine the reaction of market prices over a relatively short interval around each QE announcement. To the extent that financial markets incorporate information efficiently, we would expect market prices to react to new information about the impact of QE within a short period. This method suggests that gilt yields are about 100 basis points lower than they would otherwise have been without QE, with the majority of the effect coming through the portfolio rebalancing channel.

Looking at immediate announcement reactions is less suited to examining the impact on other assets, since it may take time for investors to change the composition of their portfolios and for the effects of portfolio rebalancing to be fully incorporated into asset prices. Corporate bond yields, probably the closest sterling-denominated substitute for gilts, fell significantly following QE announcements. But further falls in corporate yields also occurred in subsequent months. Equity prices fell immediately after the initial QE announcements but strengthened significantly thereafter, and the balance of risks perceived by market participants around equity prices implied by option prices became less negative. We also find there were improvements in liquidity in corporate bond markets, and substantial increases in net equity and corporate bond issuance during 2009, which may be at least partly related to QE.

As an alternative approach, we try to infer what historical experience would imply about the effects of a QE-like policy. We do this by simulating its impact using two econometric models based on a portfolio balance framework. This exercise suggests an impact through the portfolio balance channel on gilts and corporate bonds that is broadly similar to that observed using our analysis of announcement reactions. The impact on equity prices, however, is subject to more uncertainty, though potentially large.

The effectiveness of the MPC's asset purchases will ultimately be judged by their impact on the wider macroeconomy. Our analysis suggests that the purchases have had a significant impact on financial markets and particularly gilt yields, but there is clearly more to learn about the transmission of those effects to the wider economy.

How do individual UK producer prices behave?

Summary of Working Paper no. 394 Philip Bunn and Colin Ellis

UK monetary policy is concerned with keeping inflation on target at 2% a year. So it is important for policymakers to consider how prices behave. In particular, the degree of nominal rigidity in the economy will influence the short-term impact of monetary policy on real activity and hence inflation. This paper uses a large database of individual producer price quotes for the United Kingdom to examine the behaviour of prices. The aim of this work is to improve our understanding about how prices are set. The results may help to shed light on which pricing theories most closely reflect how prices are determined in the real world.

There have been recent euro-area and US studies that use very large databases of individual price quotes underlying published aggregate inflation series to examine pricing behaviour. Using data that has been made available by the Office for National Statistics, this paper examines the behaviour of individual UK manufacturing output prices between 2003 and 2007 using the price quotes underlying the published Producer Price Index.

This paper uncovers a number of stylised facts about pricing behaviour. First, on average 26% of producer prices change each month. The total number of price changes is concentrated among a relatively small number of items that change price very frequently. Because a small number of items account for many price changes this means that price changes occur less frequently when measured by the average for individual products than the simple average would suggest.

UK producer prices appear slightly more flexible than in the euro area and they display a similar degree of flexibility to producer prices in the United States. There is substantial variation in the frequency of UK producer price changes between different sectors and product groups. The prices of energy products change the most often, with an average of 87% of prices changing in any given month. In general, prices appear to change more often in industries where a relatively high proportion of manufacturers' costs are accounted for by basic commodities. The prices of textile and clothing products change the least often.

The probability of price changes is not constant over time. The average frequency of UK producer prices changing increased every year between 2003 and 2007, but there is also some evidence of a correlation between the share of prices changing each month and the aggregate inflation rate. January is the most popular month for prices to change, followed by April. December is the month in which the lowest proportion of prices change. Producer prices are most likely to change one, four and twelve months after they were previously set.

There is little evidence to suggest that downward nominal rigidities are important in UK product markets since 40% of all price changes are decreases and a large proportion of those price cuts are small changes. The distribution of price changes is wide, although a significant number of changes are relatively small and close to zero. Just under 30% of all price changes are between -1% and 1%, and around 45% are between -2% and 2%. The distribution of producer price changes in the United Kingdom appears to be a little wider than in the euro area. There is substantial variation in the distribution of price changes between different industries. For periods of up to one year, the average size of price changes tends to be smaller for items that change price very frequently, although beyond one year there is little relationship between the frequency and magnitude of price changes.

UK producer price changes are less persistent at the disaggregated level than aggregate inflation data imply. Aggregate monthly inflation rates in UK producer prices are persistent, ie the change in prices in the current month is related to the change in the previous month. But we find no evidence of persistence in monthly inflation rates at the individual item level. Our results suggest that this persistence in aggregate producer price inflation rates may be a result of aggregation across heterogeneous products rather than that persistence in inflation rates at the individual item level is reflected in the aggregate data.

The notion of nominal rigidities is a feature of many economic models. A variety of mechanisms have been put forward to explain this assumption, which can have differing policy implications. The empirical evidence presented in this paper on UK producer prices is not consistent with any one pricing theory. There are pieces of evidence that can be used to both support and detract from different theories. Variation in the share of prices changing in different years and months, and differing probabilities of prices changing depending on the time since the previous price change, are not consistent with models that assume the probability of prices changing is constant over time. But the large number of small price changes that we see in the data are not consistent with models in which firms face small fixed costs to adjust their prices. Also, the significant number of large price changes we observe are not consistent with firms receiving disutility from making large price changes. The heterogeneity across industries and product groups implies that there may not be one theory that can explain pricing behaviour at the economy-wide level. Different models may better explain pricing behaviour in different sectors. The clear heterogeneity in the data would argue against the use of 'representative agent' models.

New insights into price-setting behaviour in the United Kingdom

Summary of Working Paper no. 395 Jennifer Greenslade and Miles Parker

It is important to understand how companies set prices, since price-setting behaviour plays a key role in the monetary transmission mechanism. Part of the reason why monetary policy may affect the real economy, at least in the short run, is that some prices adjust sluggishly. Many of the economic models that are frequently used for monetary policy analysis assume that there are constraints on price adjustment, often called 'nominal rigidities'. Many surveys have taken place to try to improve our understanding of the extent of price rigidity, and the reasons underlying it, by asking firms directly. Examples include surveys for the United States, Canada, the euro area, as well as an earlier survey for the United Kingdom in 1995. The advantage of surveys over econometric techniques is that by asking firms directly we can obtain qualitative information, such as the factors taken into consideration by firms when reviewing the prices charged for their products.

This paper analyses the results of a new survey of the price-setting behaviour of nearly 700 UK firms. It was carried out between December 2007 and February 2008, before the onset of the recent recession. Consequently, the results need to be interpreted as applying to a period of relative macroeconomic stability, although there had been a background of volatile commodity prices and a very recent tightening in credit conditions. Our survey suggested that in the United Kingdom, many firms reviewed their prices at regular intervals, but it was also common for firms to review prices in response to specific events. This was similar to the findings of a recent euro-area survey. Overall, the median UK firm reviewed its price twice a year, although there were notable differences between sectors.

When determining the optimal price, around a fifth of firms used a rule of thumb. Around one third of firms set their prices based on their expectations of the near future. In terms of how companies set prices, survey evidence supported the use of the mark-up over costs form of pricing. Firms reviewed prices more frequently than actually changing them, with the median firm changing price only once per year. But there were marked differences between sectors — for example, UK construction and retail companies changed their prices more often than companies in the manufacturing and other services sectors. And large firms often changed prices on a more frequent basis. So, there were important heterogeneities at work.

Different factors influenced price rises and price falls. Higher costs — in particular, labour costs and raw materials — were the most important driver behind price rises, whereas lower demand and competitors' prices were the main factors resulting in price falls. The survey also considered the speed of response to changes in cost and demand conditions. Nearly half of companies changed their prices within a quarter following an increase in costs or a fall in demand.

When asked which factors were most important in causing price stickiness, the existence of implicit and explicit contracts between firms and customers and 'co-ordination failure' (where firms felt constrained because they were acting individually) were viewed as the most important. Pure menu costs (time, effort, reprinting etc) were not widely cited, in keeping with previous survey results.

Looking at how price-setting had changed over time, a substantial number of firms increased the frequency of price resets over the decade preceding the survey. Firms mainly attributed that to an increase in competition over the period, which increased the cost to the firm from deviating from the optimal price, and higher variability of input prices. Yet the more stable macroeconomic environment then in place also resulted in some firms decreasing the frequency of price changes.

Using estimated models to assess nominal and real rigidities in the United Kingdom

Summary of Working Paper no. 396 Gunes Kamber and Stephen Millard

Most monetary policy makers focus on achieving price stability: typically defined as low and stable inflation. But in order to achieve price stability, it is important to understand what drives prices and inflation and how monetary policy fits in, ie how the monetary transmission mechanism works. The standard framework for understanding inflation is the 'New Keynesian' Phillips Curve that relates inflation this period to expected inflation in the next period, and to the deviation of real marginal cost from trend. This framework has proved to be useful for thinking about the monetary transmission mechanism and inflation. But, in order to use it to provide quantitative predictions, it is necessary to embed it within a quantitative general equilibrium framework, which takes account of the dynamic relationships in the economy and the constant arrival of shocks to the system. Estimating the key parameters of such a model, allows us to assess the uncertainty around the parameters themselves and, hence, predictions made using the framework.

In this paper, we estimate two such models using UK data. In both cases, we use a 'minimum distance' technique which estimates the parameter values as those that make the theoretical responses of variables to particular shocks as close as possible to those same responses in the data. In our case, motivated by our particular interest in understanding inflation dynamics within the United Kingdom and how monetary policy makers can use interest rates and other means so as to achieve their inflation target, we set the parameters so as to match the responses of variables to movements in interest rates, the tool used by the Monetary Policy Committee over our sample period.

The first model we consider is the model of Smets and Wouters, which has become a 'workhorse' model. However, in this model, there is no distinction made between employment and hours: firms hire 'total hours' in a spot labour market. But a long tradition in monetary economics has assigned labour market frictions and, in particular wage-setting frictions, a central role in explaining inflation dynamics. So we also estimate the model of Gertler, Sala and Trigari, in which the labour market is modelled more explicitly within the New Keynesian framework. More specifically, it assumes that it takes time for unemployed workers to find jobs and for vacant jobs to find workers, and that both activities are costly.

We first use a structural vector autoregression approach to obtain estimates of the effects of interest rate changes on some important macroeconomic variables in the United Kingdom. This approach, based on a set of equations explaining each variable in terms of the same set of lagged variables, allows us to identify the effect of interest rates with only minimal restrictions on the theory. We find that output, consumption, investment and capacity utilisation all fall in response to rises in interest rates and that the responses of all these variables are 'hump shaped' with the peak response of output occurring five quarters after the initial rise in rates. Inflation rises on impact before falling to a trough two years after the initial rise in rates. The effect on inflation dies out after three years. Changes in interest rates have little effect on the relative price of capital and real wages. Productivity responds quickly, suggesting that movements in employment occur with a lag relative to movements in output.

In terms of the models, we find that both are able to explain these responses reasonably well. In addition, they are able to do this without relying on too much price or wage stickiness. In particular, our estimates imply that wages are reset about once every three quarters, and prices every year and a half. Having said that, the results also imply a large degree of indexation in price and wage-setting. It is not clear that this result is in line with our intuition for what actually happens in the United Kingdom. Neither model is able satisfactorily to explain the response of productivity to interest rate movements. An implication of this is that they are unable to explain the response of employment, given that they can explain the response of output. This suggests that it may be worth thinking more about the costs of adjusting labour input if we are to explain movements in employment as well as we can explain movements in output. We leave this for future research.

Evolving macroeconomic dynamics in a small open economy: an estimated Markov-switching DSGE model for the United Kingdom

Summary of Working Paper no. 397 Philip Liu and Haroon Mumtaz

The United Kingdom has experienced major structural and economic changes over the past three decades. In a large empirical literature, researchers have argued that these changes have manifested themselves as shifts in the dynamics of macroeconomic variables, with a number of papers focusing on documenting these changes. An understanding of what lies behind and the consequences of these changes is obviously important for the conduct of monetary policy.

However, much of the work on the UK economy is subject to a number of criticisms. Among these, first, studies are typically formulated in a closed economy setting. This is surprising given the fact that the United Kingdom is a small open economy and international developments have become increasingly important. Second, they typically employ vector autoregressions (VARs), systems of regression equations which simply specify each variable of interest as a function of past values of all variables included in the model. Although VARs have the distinct advantage of simplicity and flexibility, they do not always deliver a clear economic interpretation of shocks hitting the economy.

The aim of this paper is to investigate structural changes in the United Kingdom using a model where these criticisms are mitigated. We examine the evolving structure using an estimated open economy dynamic stochastic general equilibrium model (DSGE) where the parameters of key structural equations are allowed to change periodically over time. DSGEs are models where all the dynamic linkages between variables are transparently explained in terms of the behaviour of firms, households or the policymaker. The 'stochastic' part means that unexpected shocks continually hit the economy. So unlike VARs, the DSGE model explicitly incorporates expectations of agents (for example, the public and the central bank) into the modelling process and provides a clear interpretation of shocks that are assumed to hit the economy at any given time. We estimate several different versions of this model — ie versions that allow parameters of different structural equations to change over time. We then use statistical criteria to test how well each version of the model fits UK data. The changing dynamics of the UK economy are examined using the best-fitting model.

This turns out to be a very plausible one. One feature is that periods of turbulence come and go, but were infrequent between 1992 and the recent past, although the results towards the end of our sample in 2007/08 and early 2009 are characterised by high volatility. Moreover, these estimates from the chosen model suggest that the mid-1970s were characterised by small reactions by the monetary authorities to inflation. As a consequence, output, inflation and the real exchange rate were more volatile then than the recent past.

The sterling unsecured loan market during 2006–08: insights from network theory

Summary of Working Paper no. 398 Anne Wetherilt, Peter Zimmerman and Kimmo Soramäki

Financial markets in general can be viewed as networks, where buyers and sellers engage in repeated interactions. In particular, this analogy can be applied to money markets, as borrowers and lenders rely on each other for their daily funding needs. This paper examines the unsecured sterling overnight money market during a period which covers the crisis of 2007–08. A unique data set of individual trades in the UK CHAPS interbank payment system is used to construct a network of lending relationships between banks in the overnight market.

Network analysis of the overnight money market indicates that the structure of relationships between banks changed as the crisis unfolded. First, the data show that there is a core of a small number of banks which account for a large portion of overnight relationships. But, when concerns about counterparty risk increased, banks in the network diversified their relationships, reducing their dependence on the core. A possible explanation is that banks attempted to reduce funding liquidity risk by establishing more funding relationships.

Second, the analysis indicates that some of the observed changes in the network are asymmetric, in that they affected borrowers more than lenders. The paper argues that this asymmetry may be unique to the overnight market where increased counterparty risk is a concern for borrowers, but perhaps less so for lenders. This may be because many borrowers hope to roll overnight loans for an extended period. Thus borrowers may be keen to establish a relationship with one or more core counterparties, who are more likely to be able to provide this funding on a daily basis.

Third, the paper also suggests that changes to the reserve regime in September 2007 made liquidity management more straightforward, because banks had less strict end-of-day targets to meet. Banks therefore had much more discretion about whether to participate in the overnight market, and who to trade with. The network data show a drop in the probability of forming a relationship at this time.

The paper does not attempt to measure whether the impact of market events was greater or less than the impact of policy events. This question could be important when attempting to gauge the effect of central bank actions.

The analysis is confined to the overnight unsecured market, reflecting data availability. It does not examine to what extent this market was affected by changes in the term markets and in the secured markets. Hence, this research does not permit conclusions about the resilience of liquidity in the money markets in general, or the case for any changes in the underlying infrastructure. These issues are left for future research.

Liquidity costs and tiering in large-value payment systems

Summary of Working Paper no. 399 Mark Adams, Marco Galbiati and Simone Giansante

Interbank payment networks (ie the channels through which banks execute payments), differ widely across countries. In some countries, these networks have a 'star' shape: all (or most) banks are directly connected to a central node, a piece of infrastructure where all payments are executed. In other countries one instead observes 'tiered' structures: a few banks (*first-tier banks*) are directly connected to the central processor, while all other banks are connected to first-tier banks and channel payments through them. This paper studies the forces behind the formation of 'stars' versus 'trees' in payment networks; what it does *not* consider instead is the question of *which structure* is more desirable. This work has therefore a purely explanatory aim, rather than a normative one.

These forces stem from the nature of modern *large-value* payment systems (LVPSs). Most LVPSs today work in real-time gross settlement (RTGS) mode, whereby each payment must be settled individually by transferring the corresponding value from payer to payee. The main advantage of RTGS is that it eliminates credit risk. However, as payments must be settled in *gross* amounts, the RTGS mode requires large amounts of liquidity — a shortcoming which can however be reduced by co-ordinating payments, so liquidity is 'recycled' between banks.

Another reason why central banks pushed for the adoption of RTGS is that in practice, although not by necessity, RTGS systems use central bank money as medium of settlement. That is, the funds used to settle payments are held in accounts at the central bank. This brings about two benefits: first, the safekeeper of these funds cannot default; second, the central bank is able to monitor and possibly regulate the payment activity.

However, in some countries (including the United Kingdom), many banks are *not* direct members of the national RTGS system, and their payments are *not* settled on the RTGS system. These are the 'tiered' systems mentioned above, where second-tier banks execute payments via *correspondents* in the first tier. Payments between correspondents (due to the correspondents' proprietary and/or client operations) settle on the official RTGS system. But payments between banks with a common correspondent are made on the books of the correspondent itself. *Internalised* by the correspondent banks, these payments thus do not transit across the RTGS system. As a consequence, they are neither subject to the RTGS rules, nor can they be easily monitored by the authorities.

Surveys of UK correspondent banks indicate that internalised payments are a significant fraction — around one third by value — of all interbank payments. The value of payments which correspondents make through the RTGS system on behalf of clients is also large. These latter payments may also create risks, as they are often not pre-funded. That is, correspondents often agree to make them by extending credit to the client. So, when present, tiering is an important feature of a payment system which may have an important bearing on the system's functioning, and on the risks therein.

As mentioned above, one shortcoming of RTGS systems is their potentially high liquidity need. Tiering can be seen as a spontaneous response to this, because a major effect of tiering is to reduce liquidity costs. This is for two reasons. First, internalised payments can be made without liquidity (the *internalisation effect*). Second, by pooling own and client payment flows, the correspondents may face smoother, better manageable and therefore less costly liquidity needs (the *pooling effect*).

We build a model of tiering choices, with two 'inputs': the cost of liquidity, and an exogenous pattern of payment flows. Starting from these, we formally model the internalisation and liquidity pooling effects. We then show that even such a parsimonious model, when calibrated on real data, generates realistic payment networks. This ability to reproduce some stylised facts suggests that the cost of liquidity *is* an important driver of tiering. This is ultimately controlled by the central bank, so we conclude that a central bank has powerful policy levers to influence tiering patterns. However again: this paper sheds light on how these policy levers can affect tiering, but is silent on *how they should* be used to this aim. Such a judgement cannot be expressed here, because several consequences of tiering are not considered in this work. Above all, we disregard any 'risk' to individual institutions and to the system as a whole.

More precisely, our model features a fixed number of banks sending payments to each other. During a day, each bank receives a random stream of payment instructions at a constant rate. Each instruction requires payment of a single unit of currency to another bank. Intraday banks act mechanically: payments are executed as soon as payment instructions are received. Banks instead make decisions about where they want to sit in the 'payment network'. To be more precise, one bank is randomly picked in each period, and is given the choice between becoming a direct member of the RTGS system, or to arrange for a correspondent to execute their payments. If a bank joins the RTGS system, its payment activity generates liquidity costs. If instead it becomes a client of a correspondent, the client bank incurs no liquidity costs, but pays a fee to the correspondent for its service. The correspondent's payment activity changes as a result of taking on a client, and hence so does its liquidity cost. We specify a stylised but realistic 'protocol' for the negotiation of these fees.

By virtue of the internalisation and liquidity pooling effects, total liquidity costs for a correspondent and its customer together are no larger than the sum of the standalone costs, thus giving incentives to tier. On the other hand, banks make their decisions sequentially and, depending on their payment activity, they may find it convenient to join different correspondents. Hence, more than one correspondent bank may coexist. After a possibly long (but finite) number of 'days', the system reaches a steady state where a non-trivial network of client-correspondent relationships is formed. We simulate this model, calibrating it to data on the UK CHAPS system, and we look at the resulting networks. As mentioned above, the model produces networks which reproduce some features of the real CHAPS client-correspondent network. We perform some comparative statics exercises, suggesting how the payment network would change, if the central bank changed the price of liquidity.

Liquidity-saving mechanisms and bank behaviour

Summary of Working Paper no. 400 Marco Galbiati and Kimmo Soramäki

Interbank payment systems form the backbone of financial architecture; their safety and efficiency are of great importance to the whole economy. Most large-value interbank payment systems work in RTGS (real-time gross settlement) mode: each payment must be settled individually by transferring the corresponding value from payer to payee in central bank money. As such, all settlement risk is eliminated.

But an RTGS structure may incentivise free-riding. A bank may find it convenient to delay its outgoing payments (placing it in an internal queue) and wait for incoming funds, in order to avoid the burden of acquiring expensive liquidity in the first place. As banks fail to 'internalise' the systemic benefits of acquiring liquidity, RTGS systems may suffer from inefficient liquidity underprovision.

Inefficiencies may also emerge for a second reason. Payments queued internally in segregated queues are kept out of the settlement process and do not contribute to 'recycling' liquidity. A tempting idea is therefore to pool these pending payments together in a central processor, which could look for cycles of offsetting payments and settle them as soon as they appear. This would save liquidity, and might also reduce settlement time: payments could settle as soon as it is technically possible to do so. Segregated queues may instead hold each other up for a long time, not 'paying to each other' because none is doing so.

Such central queues are called 'liquidity-saving mechanisms' (LSMs). There are a number of studies on plain RTGS systems, but only a few on RTGS systems augmented with LSMs. Our work contributes to this line of research.

We first model a benchmark system, ie a plain RTGS system where each bank decides: (i) the amount of liquidity to use; and (ii) which payments to delay in an internal queue (payments are made as banks randomly receive payment orders, which need be executed with different 'urgency'). The benchmark model is then compared to an RTGS-plus-LSM system, where banks decide: (i) the amount of liquidity to use in RTGS as above; and (ii) which payments to submit to the LSM stream, where payments are settled as soon as offsetting cycles form.

A necessary caveat is that we consider a specific LSM, comparing it to a specific model of internal queues. Other LSMs, perhaps associated with different settlement rules, may yield different outcomes. For example, one could think of a system where *all* payments (even those sent to the RTGS stream) are first passed through the LSM. Then, if LSM settlement does not happen instantly because a cycle has not formed, the urgent RTGS payments are immediately settled by transferring liquidity. This is another way of interacting between the LSM and RTGS streams — one of the many possible ones not considered here.

We first look at the liquidity/routing choices of a social planner willing to minimise overall costs, defined as the sum of liquidity costs and delay costs. In the plain RTGS system, the planner's choice is dichotomous: if the price of liquidity exceeds a certain threshold, the planner delays all payments in the internal queues. Otherwise, it delays none, while asking banks to provide some liquidity. In this case, payments could still be queued in the RTGS stream for a while, if banks run out of liquidity. A similar dichotomy appears in the system with an LSM: the planner uses either only the LSM (when liquidity costs exceed a given threshold), or only the RTGS stream, increasing liquidity in RTGS as the liquidity price falls. Thus, from a central planner perspective, the LSM enhances the operation of the system only in extreme circumstances.

However, payment systems are not run by a 'central planner', but are populated by independent banks interacting strategically. We therefore look at the equilibrium liquidity/routing choices. A typical equilibrium here has banks routing part of their payments to RTGS, and part into the LSM, with the reliance on the LSM increasing with the price of liquidity. Despite the fact that such an outcome is inefficient (the planner would choose either of the two streams, never both), it can still be better than the one emerging without the LSM. So, an LSM may lead to a 'second-best' outcome, improving on the vanilla RTGS system.

The system with an LSM however also possesses some 'bad' equilibria. These feature the somehow paradoxical mix of high liquidity usage, intense use of the LSM, and costs which exceed those of the vanilla RTGS system. The reason behind the existence of such equilibria is probably the following: if many payments are sent in the LSM, this can be self-sustaining, in the sense that each bank finds it convenient to do so. However, the RTGS stream may become less expedite (as fewer payments are processed there), which may in turn imply that the equilibrium level of liquidity is also large. This suggests that LSMs can be useful, but they may need some co-ordination device, to ensure that banks arrive at a 'good' equilibrium.

Most of our results (above all, the ability of an LSM to improve on a vanilla RTGS system) depend on a key parameter: the price of liquidity. We do not perform any calibration of the model's parameters, so we cannot say if our LSM is advisable for any specific system. However, LSMs in general are likely to become increasingly desirable. Indeed, in the wake of the recent financial crisis, banks are likely to be required to hold larger amounts of liquid assets relative to their payment obligations. This may increase their interest in mechanisms that reduce the liquidity required to process a given value of payments.