BANK OF ENGLAND

Staff Working Paper No. 905

The more the merrier? Evidence from the global financial crisis on the value of multiple requirements in bank regulation

Marcus Buckmann, Paula Gallego Marquez, Mariana Gimpelewicz, Sujit Kapadia and Katie Rismanchi

January 2021

BANK OF ENGLAND

Staff Working Paper No. 905

The more the merrier? Evidence from the global financial crisis on the value of multiple requirements in bank regulation

Marcus Buckmann,[1] Paula Gallego Marquez,[2] Mariana Gimpelewicz,[3] Sujit Kapadia[4] and Katie Rismanchi[5]

Abstract

This paper assesses the value of multiple requirements in bank regulation using a novel empirical rule-based methodology. Exploiting a dataset of capital and liquidity ratios for a sample of global banks in 2005 and 2006, we apply simple threshold-based rules to assess how different regulations individually and in combination might have identified banks that subsequently failed during the global financial crisis. Our results generally support the case for a small portfolio of different regulatory metrics. Under the objective of correctly identifying a high proportion of banks which subsequently failed, we find that a portfolio of a leverage ratio, a risk-weighted capital ratio, and a net stable funding ratio yields fewer false alarms than any of these metrics individually – and at less stringent calibrations of each individual regulatory metric. We also discuss how these results apply in different robustness exercises, including out-of-sample evaluations. Finally, we consider the potential role of market-based measures of bank capitalisation, showing that they provide complementary value to their accounting-based counterparts.

Key words: Banking regulation, Basel III, bank failure, global financial crisis, marketbased metrics, regulatory complexity.

JEL classification: G01, G18, G21, G28.

(1) Bank of England. Email: marcus.buckmann@bankofengland.co.uk
(2) Bank of England. Email: paula.gallegomarquez@bankofengland.co.uk
(3) Bank of England. Email: mariana.gimpelewicz@bankofengland.co.uk
(4) European Central Bank. Email: sujit.kapadia@ecb.europa.eu
(5) Work undertaken while working at the Bank of England

The views expressed in this paper are those of the authors and do not reflect those of the Bank of England or the European Central Bank. We are grateful to David Aikman, Dorian Henricot, Alison Scott, Özgür Şimşek, Anton Van der Kraaij, Matthew Willison and seminar participants at the 4th annual workshop for the ESCB Research Cluster (1–2 October 2020) and the Bank of England (29 April 2019) for helpful comments and suggestions.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email enquiries@bankofengland.co.uk

ISSN 1749-9135 (on-line)

1 Introduction

The global financial crisis of 2007–08 (hereafter referred to as 'the crisis') revealed fundamental vulnerabilities in the banking system that had severe consequences for financial stability and the wider economy. Authorities had to step in to bail out banks, which were on the cusp of failing for a variety for reasons, such as excessive levels of leverage and unstable funding structures. Regulators responded to the crisis with a comprehensive set of reforms to international banking regulation, known as the Basel III package. This package introduced additional regulatory requirements to supplement the existing risk-weighted capital ratio (RWCR): a simple leverage ratio, LR, and two liquidity ratios designed to control the maturity mismatch between a bank's assets and its liabilities (the net stable funding ratio, NSFR, and the liquidity coverage ratio, LCR) (BCBS, 2011).

While each additional requirement is designed to mitigate specific vulnerabilities, they also aim to complement each other in the overall regulatory framework by working together to reduce the probability of bank failure. But there is an active debate about whether the shift to multiple regulatory requirements actually increases the resilience of the financial system and reduces regulatory arbitrage (Aikman et al., 2019; Carletti et al., 2020) or rather creates inefficient redundancies and unnecessary constraints on banks' activities (Admati and Hellwig, 2011; Cecchetti and Kashyap, 2018; Greenwood et al., 2017).

This paper takes an empirical perspective towards this debate, developing the brief analysis by Aikman et al. (2019). We assess whether a regulatory system with multiple capital and liquidity requirements prior to the crisis would have been better in identifying banks that subsequently failed compared to a system with fewer regulatory constraints, in an environment where the threshold calibration of each individual metric is free to adjust to its optimal setting for the given portfolio of metrics deployed. In particular, we apply different combinations of regulatory capital and liquidity constraints, modelled via threshold rules, to a snapshot of banks' balance sheets in both 2005 and 2006 and information about whether or not they subsequently failed during the crisis. Specifically, we set target objectives for the proportion of banks correctly identified via such rules as subsequently failing (i.e. for the 'hit rate'). We then use mixed-integer programming to minimise false alarms over these different target hit rates under different portfolios of metrics, where the threshold calibration of each individual metric is determined by the optimisation problem. For example, we determine the calibration of the LR threshold that would have minimised false alarms while correctly identifying 80% of failed banks (an 80% hit rate) under the condition that this is the only regulatory metric, and then compute this threshold over different hit rates. And then we repeat the exercise but additionally allow for the RWCR and/or NSFR to be part of the portfolio of regulatory

metrics, which affects both the false alarm rate that can be achieved and the calibration of the threshold for each metric. Due to data unavailability, the LCR is outside the scope of metrics we consider in the paper.

Our methodology has two valuable advantages over more commonly used regression-based approaches (for example logit models) to gauge the potential benefits of regulation in reducing the risk of bank failure. First, our rules are akin to a bank supervisor monitoring whether a bank meets all going concern regulatory thresholds. As such, these rules more closely resemble how supervisors assess banks against different regulations in practice than more complex models (such as logistic regression or machine learning). Therefore, we can assess the benefits of the Basel III system of regulations for bank supervision more directly. Second, our approach is better suited to capturing outliers than parametric models. To illustrate this point, a logistic regression would calculate an estimated probability of failure by weighting a bank's LR, RWCR and NSFR, so a weakness in one measure could be compensated by strength in another. In doing so, the logistic regression is more likely to miss failures that result from the crystallisation of more specific risks along one dimension —for example if a bank looks strong in terms of capital adequacy but has a very unstable funding structure. And it is often the case in practice that banks fail due to a major vulnerability in one area rather than above-average vulnerability across the board.

Our results generally support the case for a portfolio of metrics. We start by exploiting end-2006 balance sheet data on a sample of the largest global banks at that time. We find that a portfolio of three metrics proxying Basel III's RWCR, LR and NSFR typically performs better than individual metrics or different pairs of metrics in correctly identifying banks that failed in the crisis (i.e. banks that did not meet at least one of the threshold calibrations and subsequently failed) while yielding fewer false alarms (i.e. banks which survived despite not meeting at least one of the threshold calibrations). This is particularly pronounced at higher hit rates, which policymakers are likely to aim for due to the potentially significant systemic costs of bank failure, especially for larger banks (Boyd and Heitz, 2016), compared to, for example, the costs of unnecessarily imposing extra scrutiny on a bank incorrectly identified as vulnerable. We also find that the optimal threshold calibrations for each individual metric are lower in the portfolio than when metrics are calibrated individually. All of these results continue to hold when using 2005 data and thus introducing an additional lag of one year between measuring a bank's balance sheet position and the crisis date. They also broadly hold we apply different classifications for defining bank failure during the global financial crisis and when we split our sample into two by bank balance sheet size. Taken together, the results illustrate how the use of a portfolio of regulatory metrics may be beneficial in signalling bank vulnerability well in advance while both minimising regulatory false alarms and reducing required levels of

capital and liquidity.

The performance of the portfolio is mostly driven by failed banks falling short of the thresholds of the LR—the most powerful individual predictor of bank failure—and the NSFR. Consistent with this, in out-of-sample tests, the NSFR - LR pair slightly outperforms the portfolio of three metrics in terms of false alarms, although it requires marginally higher threshold calibrations. The out-of-sample tests suggest there is some over-fitting when we use all three metrics. At the same time, the primary objective of this paper is not to achieve the best out-of-sample prediction of bank failure, but rather to provide some assessment of the overall value of the Basel III system of regulatory metrics.

Recognising that we only examine a relatively small sample of banks during the global financial crisis, a slightly wider portfolio of metrics may be more robust to uncertainties which may be inherent in different episodes of banking sector distress. Importantly, it may also be less vulnerable to regulatory arbitrage or concerns that particular indicators may become less useful once they are the focus of regulation (Goodhart, 1975). Consistent with that, the RWCR – the only metric regulated internationally before the crisis – performs markedly worse in and out of sample. We therefore test our regulatory metrics on the subset of North American banks in our sample, where the LR, as well as the RWCR, was a regulatory requirement before the crisis. We find that the RWCR is more useful, and the LR less useful, for identifying failed banks in North America, suggesting that Goodhart's Law may play a role in the value of regulatory metrics. This provides further support for using a relatively wider portfolio of metrics to regulate banks rather than relying solely on the out-of-sample prediction results

We also test the simpler loan-to-deposit and market-based equivalents of the Basel III capital metrics that are not currently regulated. We find that the loan-to-deposit ratio slightly outperforms the more complex NSFR, suggesting that there is value in monitoring this metric. In addition, market-based capital ratios tend to outperform their balance-sheet counterparts, although they also appear to be more procylical. From a portfolio perspective, however, we find that combinations featuring both market-based and balance sheet measures of capital (as opposed to all market-based or all balance sheet) perform best. This may both reflect the value of combining market and balance sheet insights and that using different types of measures may be more robust, for example by better catching outlier banks without producing false alarms. While there may be difficulties associated with implementing regulations based on market prices, for example investors underpriced bank credit risk before the crisis (BIS, 2018), these results reinforce the evidence for using multiple complementary metrics to assess the vulnerability of banks (Bongini et al., 2002).

1.1 Literature review and contribution

Our paper is related to a range of literature which suggests several arguments for using multiple regulatory requirements (see also (Aikman et al., 2019)). First, the Tinbergen rule (Tinbergen, 1952) dictates that every market failure should be addressed by a distinct policy instrument. Historically, banks have failed for a variety of reasons, such as being over-leveraged, undercapitalised relative to the riskiness of their assets, or having excessive maturity mismatches. Thus, each distinct regulatory requirement can help to insure against one of these vulnerabilities. For example, the rationale for a non-risk-weighted capital adequacy ratio, the leverage ratio (LR), is that it may be difficult to estimate risk weights with a sufficient degree of certainty given that some risks in the financial system are likely to be unknowable and subject to Knightian uncertainty (Aikman et al., 2021). The LR is arguably also simpler and more transparent than the risk-weighted capital ratio (Haldane and Madouros, 2012). The two additional requirements of Basel III—the LCR and the NSFR—were introduced to harmonise liquidity standards in response to wholesale liquidity runs during the global financial crisis (BCBS, 2010). The LCR requires banks to have a minimum amount of highly liquid assets to meet a 30-day liquidity stress and seeks to protect the resilience of a bank's short-term liquidity risk profile. The NSFR requires banks to have a minimum amount of funding that is stable over a one-year period and therefore is concerned with the long-term stability of funding. Neither risk had previously been within the scope of international regulation.

Second, Bahaj and Foulis (2016) argue for a more active policy under uncertainty to avoid tail risks. There are multiple sources of uncertainty in banking regulation: difficulties in measuring and modelling risk, behavioural effects, and interconnectedness (Aikman et al., 2021). Regulation based on multiple metrics could therefore help to insure against risks in a world of uncertainty.

Third, using multiple regulatory measures could reduce the risk that banks arbitrage regulatory rules and might mitigate undesired effects generated by specific requirements. For example, the LR was introduced as a risk-neutral complement to the risk-weighted capital ratio. A regulatory regime using only a LR could incentivise banks to take on greater risk at the margin, whereas a regime that includes both the LR and RWCR can increase resilience whilst reducing incentives to risk up (Kiema and Jokivuolle, 2014; FPC, 2014; Acosta-Smith et al., 2018).

These theoretical arguments suggest that using multiple requirements in banking regulation is beneficial. However, increasing the number of regulatory requirements may increase market inefficiencies, as well as implementation and compliance costs for banks and monitoring costs for supervisors. In particular, some have argued that the portfolio of metrics in Basel III contains inefficient redundancies. For example, Greenwood et al.

(2017) and Moosa (2016) argue that the LR is redundant if risk weights are estimated correctly. Cecchetti and Kashyap (2018) use a stylised model to show that when banks are subject to LR, RWCR, LCR, and NSFR requirements, the overlaps between the LCR and the NSFR mean that one requirement is always slack and therefore redundant, although Behn et al. (2019) find that the two liquidity requirements are complementary and constrain different types of banks in different ways. More generally, Admati and Hellwig (2011) argue that additional restrictions on banks' activities can also increase risks when not designed properly. Consequently, several commentators have called for a simpler regulatory regime that relies only on a single high capital requirement, such as a leverage ratio of 10% (King, 2016) or 15%(Admati et al., 2010).

While the paper primarily focuses on evaluating banking regulation, it also relates to the extensive literature on predicting bank failure. Traditional approaches rely on statistical models such as logistic regression and discriminant analysis applied to several possible predictor variables (Jordan et al., 2010; Cole and White, 2012; Mayes and Stremmel, 2014; Cleary and Hebb, 2016). This literature generally identifies a range of metrics, including capital and liquidity ratios, that used together can correctly signal bank failure and generally simple metrics, such as the LR, can outperform complex, risk-weighted approaches, particularly when tested out of sample. Another strand of the literature uses machine learning techniques and shows that these methods outperform standard statistical methods in predicting bank failure (Kumar and Ravi, 2007; Demyanyk and Hasan, 2010; Iturriaga and Sanz, 2015; Le and Viviani, 2018; Carmona et al., 2019). A related line of literature finds that market-based variables can contain information that complements accounting-based variables when predicting bank distress (Berger et al., 2000). But much of this literature either speaks to bank failure prior to the global financial crisis and/or abstracts from the suite of regulation introduced in the post-crisis framework when assessing the risk of failure. And methodologically, both machine learning methods and classical regression models common in most of this literature can be prone to over-fitting to the noise in the data when the available sample is small and there are few failures.

Our methodology instead falls within the field of simple decision heuristics. Heuristics aim to decrease over-fitting by reducing the complexity of the model (Mousavi and Gigerenzer, 2014). The fewer parameters a model has, the less likely it is to be influenced by random noise. Rather, it will pick up the meaningful signals in the data. There is ample evidence that heuristic rules that use few variables and integrate them in simple ways can perform as well as, or better than, complex models under uncertainty (Czerlinski et al., 1999; Gigerenzer and Brighton, 2009). For example, Aikman et al. (2021) show that a simple decision tree that sequentially tests regulatory metrics can predict bank failure as accurately out of sample as a logistic regression. Another advantage of simple rules is

that they are typically easier to understand, execute and explain than statistical models or black box machine learning models (Lipton, 2018; Rudin and Radin, 2019). This is particularly helpful in the context of bank regulation, where regulatory and supervisory approaches and decisions need to be explained to regulated institutions and the wider public.

This paper contributes to the literature on regulatory framework design in several ways. First, relative to the inconclusive theoretical literature on the costs and benefits of multiple regulatory requirements, our paper extracts clear empirical evidence from the global financial crisis which speaks to the effects of the shift from one regulatory requirement to multiple requirements. In particular, we can assess benefits in the form of correctly identifying banks that failed (i.e. the hit rate), and costs in the form of false alarms and the stringency of the calibration of each individual regulation.

Second, our paper is one of the first to empirically assess the overall going concern Basel III regulatory framework and related metrics in helping to reduce the risk of bank failure. Lallour and Mio (2016) consider the role of multiple regulatory requirements but focus primarily on the marginal contribution of the NSFR specifically. And Haldane and Madouros (2012) find that market-based capital ratios outperform measures based on accounting-based regulatory capital in predicting failure during the global financial crisis but in a comparison which abstracts from other variables which could be the subject of regulation. By contrast, we take a holistic approach to the system of Basel III regulations, developing the brief analysis by Aikman et al. (2019) in several important ways. In particular, we take a more systematic empirical approach which also assesses 2005 balance sheet data, explores sub-samples of the data, includes out-of-sample testing and considers additional metrics, including market-based capital metrics. In addition, we estimate Shapley values to proxy the marginal contribution of each metric to the performance of the overall portfolio.

Finally, we use a novel approach to evaluate bank regulation. Our analysis formalises and develops the threshold-based approach briefly discussed by Aikman et al. (2019). The modelling of regulatory rules closely resembles how supervisors assess banks against different regulations in practice, while also allowing us to assess the impact of individual metrics on the predictions of the portfolio. We do not impose assumptions about policymakers' preferences to derive the calibrations of regulatory thresholds. Instead, we analyse the optimal threshold over all possible hit rates to derive general results, regardless of policymaker preferences. Our methodology also differs from Aikman et al. (2021) in that we do not impose a rank ordering over the regulatory metrics. And it should be noted that the methodology can be used more generally for a wide range of other (economic) prediction problems with a binary outcome and multiple predictors, for example

to evaluate the risk of corporate failure as developments such as the Covid-19 pandemic unfold.

The rest of the paper is organised as follows. Section 2 describes the data and methodology. Section 3 presents our baseline results, shows robustness tests, and assesses a range of alternative metrics. Section 4 concludes.

2 Dataset and Methodology

2.1 Data

We use the dataset assembled by Aikman et al. (2021) and also used by Lallour and Mio (2016). It contains 116 banks, spanning 25 countries, that had more than $100 billion in assets at the end of 2006.[1] Of these, 76 banks have no missing values for the three core metrics: LR, RWCR and NSFR and are included in our baseline experiment.

The dataset consists primarily of Liquidatum data at the consolidated bank level, supplemented with data from Capital IQ, SNL and banks' annual reports. We use a range of balance sheet and market-based metrics, which were assessed at two points in time: end-2005 and end-2006.

The dataset includes a binary variable indicating whether a bank survived or failed between 2007 and the end of 2009. Very few banks technically defaulted during the crisis, but many would have without significant government intervention. Therefore, the definition of failure has been the subject of some discussion and a degree of judgment is necessary to categorise banks as having survived or failed. For the baseline we use the definition of failure by Laeven and Valencia (2010),[2] supplemented by the small number of adjustments applied by Aikman et al. (2021). This yields 35 banks which are classified as failing and 41 banks as surviving in our baseline sample. But we also check the robustness of our main results to adopting two alternative variants of the failure definition: a more restrictive definition where we classify US banks that participated in the Troubled

[1] The data contains banks, broker-dealers and building societies/mutuals. Federal institutions, diversified financials, speciality lenders and development banks were excluded. The firms were extracted from a list of top global banks provided by *The Banker* publication, supplemented by data from Capital IQ and SNL databases to extract broker-dealers and building societies/mutuals. Due to data coverage and quality issues, the National Agricultural Cooperative Federation (Korea), Daiwa Securities (Japan), and all Chinese banks were excluded from the sample.

[2] Beyond clear-cut cases of default or nationalisation, Laeven and Valencia (2010) define banks to have failed if at least three of the following six conditions were present: (i) extensive liquidity support (5 percent of deposits and liabilities to non-residents); (ii) bank restructuring costs (at least 3 percent of GDP); (iii) partial bank nationalisation (e.g. government recapitalisation); (iv) significant guarantees put in place; (v) significant asset purchases (at least 5 percent of GDP); (vi) deposit freezes and bank holidays.

Asset Relief Program (TARP) as part of the Capital Purchase Program (CPP) as failing, increasing the number of failing banks to 39; and a looser definition in which we classify all French banks, Bank of America and Lloyds Banking Group as surviving, reducing the number of failed banks to 27.[3]

TABLE I: Metric definitions.

Metric	Abbreviated references used in this paper	Definition
	-Baseline-	
Failure indicator (2007-2009)	Failed	{0,1}
Tier 1 Capital Ratio	RWCR	$\frac{Tier\ 1\ Capital}{Risk-weighted\ Assets} \times 100$
Tier 1 Leverage Ratio	LR	$\frac{Tier\ 1\ Capital}{Total\ Assets} \times 100$
Net Stable Funding Ratio[4]	NSFR	$\frac{Available\ stable\ funding}{Required\ stable\ funding}$
	-Extensions-	
Market-based Capital Ratio	RWCR MB	$\frac{Market\ Capitalisation}{Risk-weighted\ Assets} \times 100$
Market-based Leverage Ratio	LR MB	$\frac{Market\ Capitalisation}{Total\ Assets} \times 100$
Price-to-book Ratio	PTB	$\frac{Market\ Capitalisation}{Tier\ 1\ Capital}$
Loan-to-Deposit Ratio[5]	LTD	$\frac{Retail\ loans}{Retail\ deposits}$

Table I shows all metrics used in the analysis. Our baseline consists of the three measures that most closely proxy the key elements of Basel III: a Tier 1 capital ratio (RWCR), a simple leverage ratio (LR) and a net stable funding ratio (NSFR). Basel III also defines a liquidity coverage ratio (LCR) which compares weighted versions of banks' liquid assets and their short-term (i.e less than 30-day maturity) deposits (BCBS, 2013). Given

[3]In France, the State Shareholding Corporation (SPPE) was set up to inject capital both into financial institutions in difficulty and into sound financial institutions. All six major French banks benefited from the first round of injections and are therefore classified as failed in our baseline. Bank of America and Lloyds Banking Group are classified as failed although it could be contended that their distress was primarily due to their takeover of other failing banks—Merrill Lynch and HBOS respectively—which are separate observations in the data.

[4]The weighting scheme used to classify different assets and liabilities to determine the NSFR on the basis of Liquidatum data is the same as that used in Aikman et al. (2021) and Lallour and Mio (2016).

[5]For some banks, it is not possible to distinguish between retail deposits and deposits placed by non-bank financial corporations. In these instances, the loan-to-deposit ratio is proxied by (Customer loans / Customer deposits) in line with Aikman et al. (2021).

the complexities and data requirements of this metric, it is difficult to construct proxy measures of banks' historic LCRs, and so the LCR is therefore outside the scope of this paper. Similarly, gone concern capital requirements such as requirements on bail-inable debt, which are also an important aspect of the post-crisis package of regulatory reform, are also outside our scope.

Given the available data, we proxy the Basel III requirements by relying on data and definitions available before the global financial crisis. In particular, an important part of the Basel reforms since 2010 has been to change the definition of capital, the approach to calculating risk-weighted assets and to introduce a specific denominator to calculate leverage. Our empirical analysis cannot test the pre-crisis performance of these new definitions.

The RWCR and LR use regulatory capital as their numerator, which reflects the regulatory view of the amount of going-concern capital a bank has available to absorb losses. Given data availability, the RWCR is calculated on the basis of Basel I risk-weighted assets, which were widely used before the global financial crisis.[6] Basel I featured an 8% total capital requirement, of which half had to be met with Tier 1 capital as defined at that time—we define our RWCR in Tier 1 space. The LR uses total assets in the denominator as opposed to the Basel III leverage exposure measure (which captures off-balance sheet exposures, secured financing transactions and derivatives). But the measure of assets used does attempt to correct for different netting arrangements permitted under alternative accounting standards used by banks in different countries.[7] We use the NSFR, which requires banks to have a minimum amount of funding that is stable over a one-year period, as calculated by Lallour and Mio (2016). Their definition attempts to proxy the Basel III definition of the NSFR as closely as possible using data from Liquidatum.[8]

The remaining variables are considered in the extensions to the baseline analysis as simpler or market-based alternatives to Basel III measures. First, we consider the loan-to-deposit (LTD) ratio as an alternative liquidity measure to the NSFR. It is a simpler, non-regulated metric and so can help provide insights into the debate on regulatory com-

[6]Basel I assigned one of four risk categories to assets in a simple way. The risk weights ranged between 0 and 100% e.g. a retail mortgage portfolio was assigned a 50% risk weight, while a corporate loan portfolio was assigned a 100% risk weight. An advantage of using Basel I risk-weighted assets is that RWCR is more consistent and comparable across countries in our sample. Basel II reformed the risk weight methodology and allowed the use of internal models. Note that the implementation of Basel II was underway when the crisis hit.

[7]Broadly, the prevailing accounting standards are Generally Accepted Accounting Principles (GAAP), which can also vary between regions, and International Financial Reporting Standards (IFRS). The ability to offset under IFRS is limited in comparison with United States GAAP, especially for derivatives traded with the same counterparty under a Master Netting Agreement. So for US GAAP banks the nominal amount of derivatives is included in assets as a proxy for the IFRS treatment. See ISDA (2012).

[8]The NSFR is defined under Basel III, see BCBS (2014). Further details are available in Aikman et al. (2021) and Lallour and Mio (2016).

TABLE II: Summary statistics of the metrics used in the analysis.

		All	Survived	Failed	α
		BASELINE ANALYSIS			
Banks	Number[9]	76	41	35	
	% of baseline total	100%	54%	46%	
LR	Mean (SD)	4.01 (1.42)	4.36 (1.37)	3.59 (1.39)	**
	Median (25%–75%)	3.89 (2.88–4.82)	4.36 (3.51–4.87)	3.40 (2.54–4.08)	
RWCR	Mean (SD)	8.35 (1.67)	8.50 (1.95)	8.18 (1.26)	
	Median (25%–75%)	7.96 (7.15–8.95)	8.19 (6.90–9.89)	7.82 (7.43–8.64)	
NSFR	Mean (SD)	0.94 (0.21)	0.95 (0.23)	0.91 (0.18)	
	Median (25%–75%)	0.93 (0.78–1.02)	0.96 (0.79–1.05)	0.93 (0.78–1.00)	
LTD	Mean (SD)	1.74 (1.87)	1.32 (0.60)	2.23 (2.62)	*
	Median (25%–75%)	1.42 (0.97–2.09)	1.27 (0.91–1.90)	1.69 (1.25–2.16)	
		ANALYSES INCLUDING MARKET-BASED METRICS			
Banks	Number	59	32	27	
	% of baseline total	100%	54%	46%	
LR MB	Mean (SD)	10.59 (5.49)	12.84 (5.79)	7.92 (3.71)	***
	Median (25%–75%)	9.45 (7.20–13.21)	12.18 (8.82–14.85)	7.26 (4.97–9.64)	
RWCR MB	Mean (SD)	21.28 (7.81)	23.60 (7.91)	18.53 (6.85)	**
	Median (25%–75%)	19.74 (15.84–25.31)	21.45 (18.25–26.80)	16.78 (15.16–21.00)	
PTB	Mean (SD)	2.54 (0.77)	2.78 (0.78)	2.26 (0.67)	***
	Median (25%–75%)	2.48 (2.07–2.95)	2.63 (2.28–3.05)	2.11 (1.81–2.78)	

The asterisks (rightmost column) indicate whether the metrics' mean differences between failed and survived banks are statistically significant. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

plexity. The LTD has sometimes been used as an indicator in practice and proxies can be constructed for a larger sample of banks than the NSFR.

Second, we look at market-based capital measures. The market-based capital ratio (RWCR MB) and leverage ratio (LR MB) are analogous to the first two measures except that they use the market value of a bank's equity in the numerator, i.e. the bank's stock market capitalisation.

Finally, we also consider the price-to-book (PTB), the ratio between the market value and book value of a bank's equity. PTB ratios can be a useful additional indicator of bank resilience; some regulators do monitor them even though they are not part of formal bank prudential regulation (IMF, 2019; FPC, 2020). If PTB ratios are persistently below unity, they could indicate investor concerns about the realisable value of a bank's assets and future bank profitability (Haldane, 2011; IMF, 2018).

Table II shows summary statistics across our baseline sample, which comprises the 76 banks that have no missing values for the three core metrics: LR, RWCR and NSFR.

[9]For the baseline, this only includes banks for which we have complete information across all metrics as at 2006. For the extensions, it only includes banks for which we have complete data on all variables and we can create a comparable sample using the baseline portfolio.

It shows that banks that failed had, on average, lower capital and liquidity ratios. This is also the case for market-based capital ratios. Using Welch's unequal variances t-test (Welch, 1947), we find that the difference in the average values of metrics between failed and survived banks is statistically significant for the LR, LTD and market-based metrics, but not for the RWCR and NSFR.

2.2 Methodology

We assess the different regulatory requirements and metrics against evidence from the global financial crisis by developing the methodology briefly discussed by Aikman et al. (2019). We suppose that, prior to the crisis, banks would have been required to meet either one or multiple capital and liquidity ratios that we refer to as 'thresholds'. This reflects how supervisors and policymakers use capital and liquidity standards in practice to assess and regulate bank safety and soundness.

In essence, we are interested in identifying the combination of thresholds that produces the best performing rule. For example, assume we want to achieve an 80% hit rate (the share of failed banks that would have been constrained by the metrics). For a portfolio of LR, RWCR and NSFR, we optimise to find the combination of threshold calibrations for these metrics that would have identified at least 80% of the failed banks, whilst minimising false alarms (the share of survived banks that would have been constrained by the metrics). We then repeat this over different hit rates to plot the share of false alarms over the different target hit rates.

TABLE III: Confusion matrix for different outcomes

	Failed	Survived
Below threshold(s)	hits	false alarms
Above threshold(s)	misses	correct rejections

Formally, we evaluate effectiveness of the regulatory requirements using a confusion matrix (Table III). All metrics are assigned a negative direction, meaning that a lower threshold t implies a higher probability of failure. Banks with metrics that fall below a threshold t and subsequently fail are correctly identified and referred to as 'hits', banks with metrics that fall below t and survive are 'false alarms'. Survived banks above t are 'correct rejections' and failed banks above t are 'misses'.

As regulatory thresholds tighten (e.g. the assumed capital requirement increases), the hit rate achieved by each metric automatically increases. This means that the metric could have prompted the bank or the regulator to take pre-emptive action. However, a tighter threshold also creates more false alarms which can reduce the usefulness of the metric as

a signal for bank riskiness and may increase the need for supervisory action which would not have been necessary.

Policymakers therefore need to trade off false alarms against hits when calibrating regulatory thresholds. One novel aspect of our approach is that we do not rely on any assumptions about policy makers' preferences. Instead, we assess the false alarm rate at all possible hit rates (0-100%) for each requirement. When we use a portfolio of multiple regulatory requirements, we calibrate this to find the optimal set of thresholds that minimises false alarms at all possible hit rates.

Individual metrics

We test an individual metric against a threshold t. For example:

$$\text{IF } leverage\,ratio < 3\% \text{ THEN predict } failure$$

Varying t over different threshold values $t = (t_1, \ldots, t_N)$, we assess the performance of a metric at each t by considering the hit rate (the share of failed banks that were constrained by the metrics) and the false alarm rate (the share of survived banks that were constrained). Let the hit rate be denoted by $r_{hit}(t) = \frac{\text{hits}_t}{\text{hits}_t + \text{misses}_t}$ and the false alarm rate be denoted by $r_{fa}(t) = \frac{\text{false alarms}_t}{\text{false alarms}_t + \text{correct rejections}_t}$. With an increase of t, the hit rate increases—at the cost of an increased false alarm rate. We plot the resulting pairs of false alarm rate and hit rate to build a receiver operator characteristic (ROC) curve, which is frequently used in the literature for policy evaluation (Berge and Jorda, 2011; Jorda and Taylor, 2011).

Optimising a portfolio of combined metrics

We also investigate rules with portfolios of metrics. For example, a portfolio of three metrics could comprise:

$$\text{IF } leverage\,ratio < 3\% \text{ OR } capital\,ratio < 8\% \text{ OR } net\,stable\,funding\,ratio < 100\%$$
$$\text{THEN predict } failure$$

Let $\mathbf{t} = \{t_1, \ldots t_D\}$ denote the vector of numeric thresholds of the D regulatory metrics. This is the control variable. Without loss of generality, we assume that the metrics have been normalized to values between 0 and 1. We optimise to find $\mathbf{t}$ that minimises false alarms for a given target hit rate, r_{hit}^{min}. Mathematically, this is expressed as:

$$\min_{\mathbf{t}} \{r_{fa}(\mathbf{t})\} \;\; s.t. \;\; r_{hit}(\mathbf{t}) \geq r_{hit}^{min}. \tag{1}$$

We vary the target hit rate r_{hit}^{min} over the range [0.05,0.95] in steps of 0.05 so that we can plot the resulting ROC curve. That is, starting with a target hit rate of 95%, the optimisation would find the combination of LR, RWCR and NSFR calibrations that would

have identified at least 95% of the failed banks, whilst minimising false alarms. We then reduce the target hit rate to 90% and repeat the exercise, and so on.

We add two additional criteria into the optimisation that are decisive when more than one solution for $\mathbf{t}$ produces the same false alarm rate for a given target hit rate. Specifically, let w_1 be the weight put on solutions that provide higher hit rates and let w_2 be the penalty to solutions that require higher thresholds (if we assume regulation is not costless, then solutions that require less stringent calibrations would be preferred).[10] The full optimisation is:

$$\min_{\mathbf{t}}\{r_{fa}(\mathbf{t}) - w_1 \times r_{hit}(\mathbf{t}) + w_2 \times |\mathbf{t}|_1\} \;\; s.t. \;\; r_{hit}(\mathbf{t}) \geq r_{hit}^{min} \tag{2}$$

The parameters w_1 and w_2 are set to small values such that they only play a role when solutions have an equivalent false alarm rate r_{fa}. We implement the optimisation with a mixed integer program which is described in detail in Appendix A.1.

Note that the statistical power of our empirical analysis is limited due to the small size of the dataset. Therefore we do not report statistical tests when comparing the performance of the metrics.[11] One may argue that significance tests are less important for our analysis in the first place, as our dataset does not reflect a random sample of a population to which we want to generalise. Rather the data approximately represent the population of the world's largest banks prior to the global financial crisis.

Framework for comparing results

We assess which metric or combination of metrics performs best on average and also look at performance at specific hit rates. We therefore plot the entire ROC curve and report the area under the curve (AUROC) and Shapley values (Shapley, 1953), which are useful tools to evaluate average performance:

- A ROC curve shows a series of rules, obtained by varying the target hit rate. A rule outperforms another if, for a given hit rate, it leads to fewer false alarms. Better rules will be closest to the top-left of the chart.
- The AUROC gives an indication of average performance of a classification rule over the entire set of hit rates. A metric that can perfectly discriminate between failed and survived banks would have an AUROC of 1, while a value of 0.5 reflects random

[10] The second criteria that penalises solutions with higher thresholds is an innovation relative to the optimisation used in Aikman et al. (2019); it does not change the baseline results.

[11] For example, assuming a true AUROC score of 0.65 in the population, we would have a power of 0.64 to find that the AUROC is significantly different (α = 0.05) from a random model in our baseline sample of 35 failed and 41 survived banks.

performance.

- Shapley values provide a good proxy for the marginal contribution of a metric to the performance of a portfolio. We measure the Shapley value of a metric as the average increase in AUROC when adding the metric to all possible combinations of metrics. For example, in our baseline portfolio of LR, RWCR and NSFR, the Shapley value of LR is calculated by considering LR's contribution on its own (compared to a random baseline), in all possible portfolios of two metrics, and in the portfolio of three metrics. By definition, the sum of all Shapley values and the random performance (AUROC = 0.5) adds up to the AUROC of the portfolio.

Metrics that measure average performance are agnostic about policymakers' preference for hits and false alarms. In practice, the policymaker might only be interested in certain areas of the ROC space. Following Demirgüç-Kunt and Detragiache (1998), a stylised way to think about policymakers' preference is to minimise a loss function with differential weights for misses and false alarms. Let the policymaker loss function be:

$$L = \theta \times \text{miss rate} + (1 - \theta) \times \text{false alarm rate}, \tag{3}$$

where θ measures the policymaker's preference. Given that the miss rate = 1− hit rate, we can redefine the loss as:

$$L = \theta \times (1 - \text{hit rate}) + (1 - \theta) \times \text{false alarm rate} \tag{4}$$

The true preference of policymakers is unknown. However, given the severe social, economic, and political consequences of financial crises, policymakers are more likely to be concerned about misses than false alarms (Demirgüç-Kunt and Detragiache, 1998; Borio and Lowe, 2002; Borio and Drehman, 2009; Detken et al., 2014). Although our focus is on individual bank failure, most banks in the sample are large and systemically important, so bank failure is likely to be strongly linked to a systemic financial crisis. So policymaker preferences are likely to be reflected in a θ greater than 0.5.

3 Results

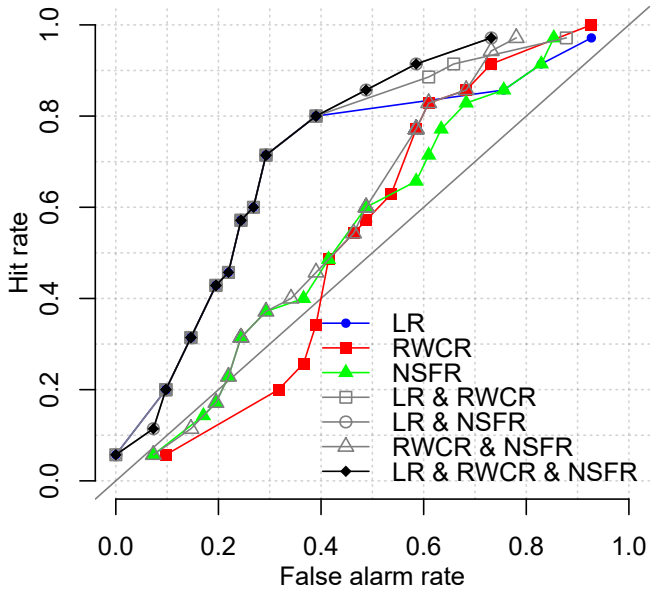
3.1 Baseline result

Our baseline consists of the three proxies for the RWCR, LR, and NSFR. Results in this section use the data for the 76 banks without any missing values on these three metrics in the year 2006. The benefit of using this consistent sample is that all rules are directly comparable, but our findings qualitatively do not change when we include banks with

partial data. We also repeat the results for values of the three metrics in 2005. For this we use the 78 banks without any missing values on the three metrics in the year 2005.

Figure I compares the performance of the individual metrics and the portfolios of two and three metrics. The rules are calibrated and evaluated on the complete (in-sample) dataset. The corresponding AUROCs are summarised in Table IV. The first finding is that a portfolio of three metrics outperforms individual metrics, on average. The portfolio of three also outperforms all other pairs of metrics apart from the NSFR-LR pair to which it is comparable. Figure II shows how each metric in the portfolio of three is contributing to hits and false alarms at the different target hit rates. Most failed banks fail to meet the LR threshold, and either the NSFR or LR threshold at higher hit rates (sometimes both together). The Shapley values in Table V show that the LR makes the most significant contribution to the AUROC. This is in line with Aikman et al. (2021), who find the LR on its own would have outperformed the RWCR and NSFR in identifying banks that subsequently failed in the crisis.

FIGURE I: ROC curves for the individual metrics (LR, RWCR, NSFR), pairs, and the portfolio of all three.



Looking at average performance is helpful but might overlook findings more relevant for policymakers who may be interested in specific outcomes. As discussed above, while we do not know the exact policymaker preference, we expect that policymakers put more weight on not missing a failure (i.e. $\theta > 0.5$). For $\theta = 0.5$, the policymaker loss function is minimised at a hit rate of 71%. For $\theta = 0.7$, the loss function is minimised at a hit rate of 91%. This illustrates that policy-makers are more likely to care about performance at

TABLE IV: Baseline performance results. AUROC of individual metrics and their combinations of rules calibrated in 2006 (n = 76). The highest AUROC are in **bold**.

LR	RWCR	NSFR	LR & RWCR	LR & NSFR	RWCR & NSFR	LR & RWCR & NSFR
0.70	0.55	0.56	0.73	**0.74**	0.59	**0.74**

FIGURE II: Contribution of each metric to false alarms and hit rates in the portfolio.

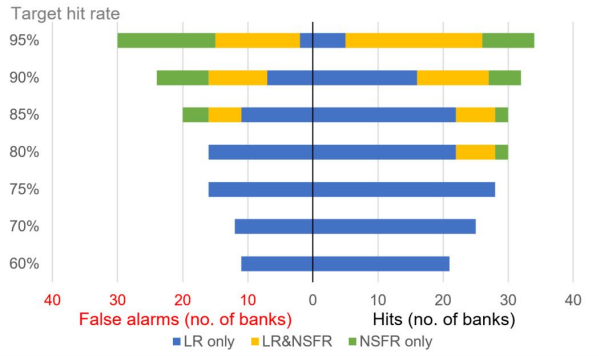


TABLE V: Shapley values (n=76).

Measure	**Shapley**
LR	0.18
RWCR	0.02
NSFR	0.04
Combined AUROC ($\sum$+.5)	0.74

hit rates above 70%.

Crucially, the portfolio performs best at high hit rates. Up to a target hit rate of 80%, the LR does as well on its own as the portfolio of three in constraining banks that subsequently failed. But for hit rates above 80%, the portfolio of three does better than any individual metric. Specifically, it yields fewer false alarms for any given hit rate above this level. This is driven by the NSFR identifying additional banks that failed (for example, Countrywide Financial Corporation, which failed, had a LR of 7.68% but an NSFR of only 76%).

Table VI shows the optimised thresholds for the individual metrics and Table VII shows the thresholds for the portfolio of all three metrics. They show that we require less stringent thresholds for each metric in the portfolio than when using metrics individually. For example, to achieve a hit rate of 85%, the portfolio of three would require the LR to be set at 4.15%, the RWCR at 5.52%, and NSFR at 76%. To achieve the same hit rate, the LR on its own would need to be set at 4.99%; the RWCR on its own would need to be set at 9.04%; and the NSFR at 109%; and all of these would also come at the cost of significantly more false alarms. This result indicates that there may be synergies between capital and liquidity requirements, which could also affect their optimal calibration (Brooke et al., 2015; BCBS, 2016; Cecchetti and Kashyap, 2018). However, the RWCR does not contribute at any hit rate, as shown in Figure II. This explains why the RWCR threshold stays constant (Table VII) and why the LR - NSFR pair performs

TABLE VI: Baseline results for individual metrics (n=76).

	LR			RWCR			NSFR		
Target hit rate	**False alarm**	**Hit rate**	**Thresh-old**	**False alarm**	**Hit rate**	**Thresh-old**	**False alarm**	**Hit rate**	**Thresh-old**
0.50	0.24	0.57	3.48	0.46	0.54	7.95	0.49	0.60	0.95
0.60	0.27	0.60	3.55	0.54	0.63	8.25	0.49	0.60	0.95
0.70	0.29	0.71	3.82	0.59	0.77	8.65	0.61	0.71	1.00
0.75	0.39	0.80	4.15	0.59	0.77	8.65	0.63	0.77	1.01
0.80	0.39	0.80	4.15	0.61	0.83	8.74	0.68	0.83	1.03
0.85	0.76	0.86	4.99	0.68	0.86	9.04	0.76	0.86	1.09
0.90	0.83	0.91	5.66	0.73	0.91	9.83	0.83	0.91	1.20

as well as the portfolio of three.

In summary, we observe two general benefits of combining a small number of regulatory metrics in assessing the health of banks: we can achieve the same hit rate as any individual metric while yielding fewer false alarms; and we require less stringent calibrations.

Early warning metrics—calibrating models in 2005

Although our study is predominantly an evaluation of regulatory metrics, the rules could also be early-warning indicators for bank failure. The data allows the early-warning interpretation as the regulatory metrics are estimated in 2006, while the bank failure is captured in subsequent years (2007–2009). If indicators also work over longer lags between the signal and failure, this would allow for more time to take mitigating action. We therefore also test how well the metrics in 2005 predict bank failure in 2007–2009.

The results are qualitatively similar to the baseline results. The portfolio of three outperforms individual and paired metrics (Table VIII). Again, this is most pronounced at higher hit rates (Figure III). The result is driven primarily by the LR, while the RWCR

TABLE VII: Baseline results for the portfolio optimisation (n=76).

Target hit rate	**LR threshold**	**RWCR threshold**	**NSFR threshold**	**False alarm**	**Hit rate**
0.50	3.48	5.52	0.50	0.24	0.57
0.60	3.55	5.52	0.50	0.27	0.60
0.70	3.82	5.52	0.50	0.29	0.71
0.75	4.15	5.52	0.50	0.39	0.80
0.80	4.15	5.52	0.50	0.39	0.80
0.85	4.15	5.52	0.76	0.49	0.86
0.90	4.14	5.52	0.90	0.59	0.91

and NSFR provide similar average contributions, as indicated by the Shapley values (Table IX). In contrast to the 2006 results, the portfolio of metrics also outperforms the LR-NSFR combination when using the 2005 data, highlighting the likely greater robustness of the portfolio of the three metrics.

FIGURE III: ROC curves of the individual metrics (LR, RWCR, NSFR) pairs, and the portfolio of all three calibrated in 2005.

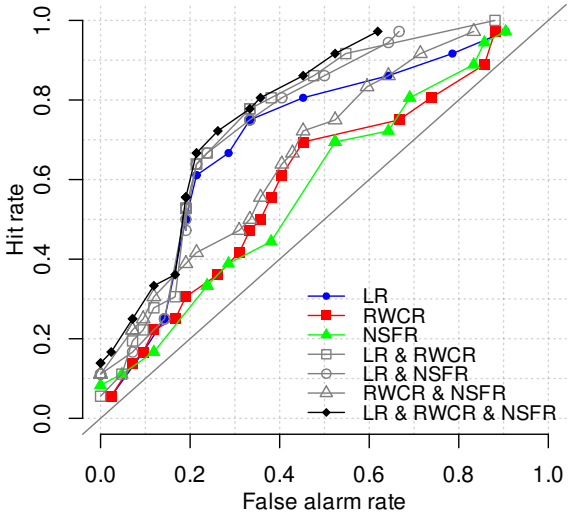


TABLE VIII: AUROC of individual metrics and their combinations of rules calibrated in 2005 (n = 78). The highest AUROC are in bold.

	LR	RWCR	NSFR	LR & RWCR	LR & NSFR	RWCR & NSFR	LR & RWCR & NSFR
AUROC	0.71	0.60	0.59	0.76	0.76	0.68	**0.78**

TABLE IX: Shapley values of rules calibrated in 2005 (n = 78).

	LR	RWCR	NSFR
2005	0.16	0.07	0.06

3.2 Robustness tests and complementary exercises

Out-of-sample tests

The results presented so far are based on in-sample experiments. In other words, the regulatory thresholds are calibrated and evaluated on the same set of banks. Using this approach, the rules might *overfit* the sample data, i.e. fit to the specific noise in the data, and might not generalise well to unseen data points. To test the predictive

performance of the regulatory thresholds further, we use *out-of-sample* testing. This entails calibrating thresholds on a randomly selected subsample of banks and testing them on the remaining banks. We use 90% of the observations for training and test the models on the remaining 10%. To obtain stable results, we repeat the out-of-sample experiments 1,000 times.[12]

FIGURE IV: Out-of-sample performance on the 2006 and 2005 data.

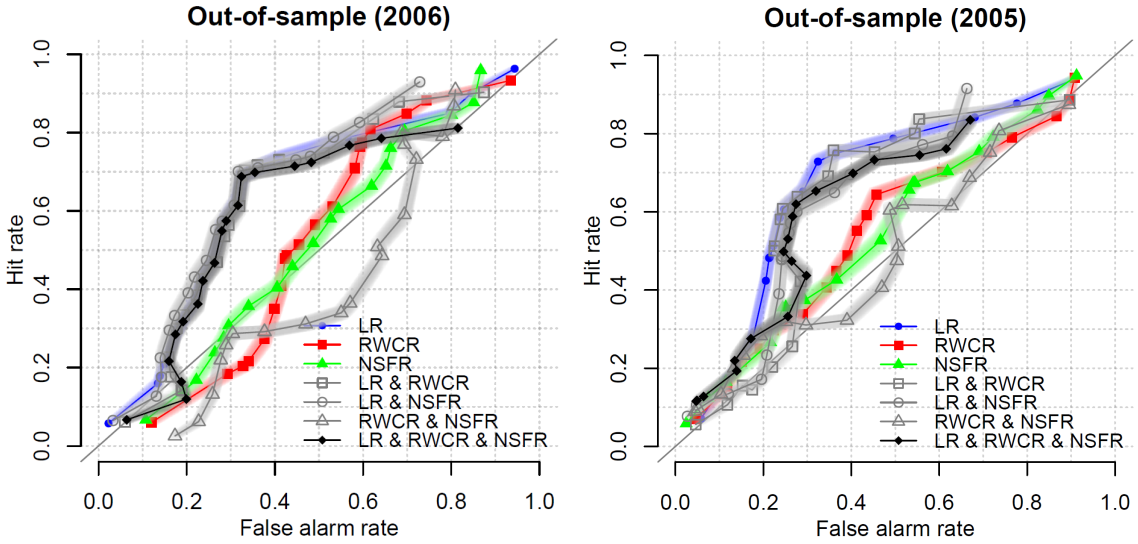


TABLE X: Out-of-sample AUROC of individual metrics and their combinations of rules calibrated in 2006 (n = 76) and 2005 (n = 78). The highest AUROC are in bold.

	LR	RWCR	NSFR	LR & RWCR	LR & NSFR	RWCR & NSFR	LR & RWCR & NSFR
2006	0.65	0.51	0.54	0.65	**0.68**	0.43	0.63
2005	**0.67**	0.56	0.55	0.65	**0.67**	0.52	0.65

Figure IV shows out-of-sample ROC curves and Table X summarises the corresponding AUROCs. Out-of-sample performance declines across all rules compared to the in-sample baseline, as shown by the lower AUROCs. The best performing portfolio in terms of false alarms in both 2005 and 2006 is the LR-NSFR pair. The portfolio of all three metrics performs better than the RWCR and NSFR individually but slightly worse than the LR. Shapley values confirm the centrality of the LR in terms of average performance (Table XI). But at high hit rates, the LR-NSFR pair performs best, especially using the 2006 data.

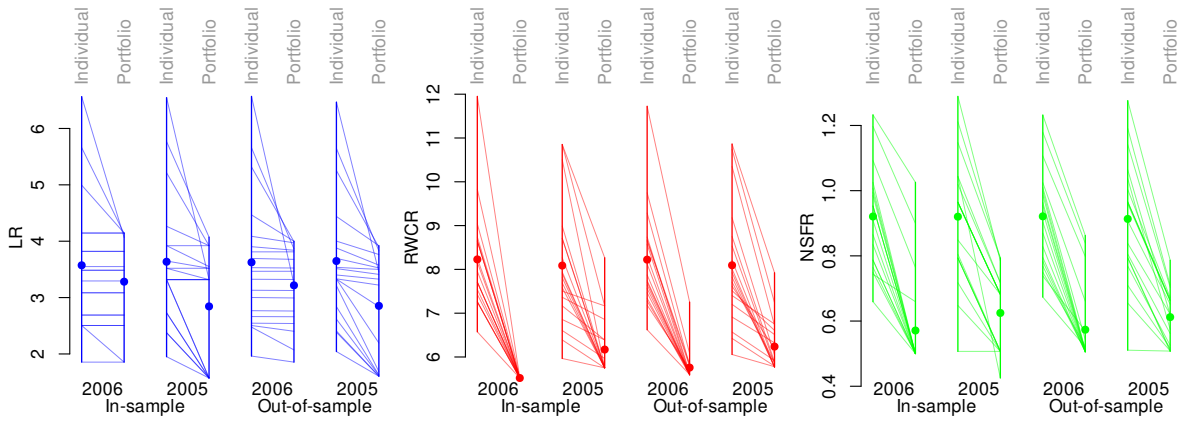
Figure V shows the threshold calibrations for individual metrics and the portfolio of three in sample and out of sample for both 2005 and 2006. Each line maps the calibrations

[12] Appendix B describes the procedure in more detail.

TABLE XI: Out-of-sample Shapley values of rules in 2006 (n = 76) and 2005 (n = 78).

	LR	RWCR	NSFR
2006	0.16	-0.03	0.00
2005	0.13	0.01	0.01

FIGURE V: Threshold for individual vs portfolios over different tests



for a given target hit rate, and the dot shows the mean thresholds across hit rates. We can see that calibrations tend to be lower for the portfolio than when used individually, most markedly for the RWCR and NSFR. This is true across years and in both in- and out-of-sample tests.

Overall, the discrepancy between in-sample and out-of-sample performance indicates some degree of over-fitting when testing a portfolio of three metrics, which is to be expected given the small sample. Out of sample, the LR-NSFR pair yields slightly fewer false alarms than the portfolio of three, albeit a marginally higher threshold calibrations. And we consistently find that using more than one metric outperforms individual metrics.

Alternative classifications of bank failure

Our headline results also broadly hold when we test their sensitivity to alternative classifications of bank failure. As shown in Table XII, in-sample the portfolio performs best across the different failure definitions. This again is particularly true for high hit rates (see Figure VI). Out-of-sample, the LR-NSFR continues to perform best for the restrictive definition of failure, where all US banks taking part in the TARP-CPP program are classified as failed. For the looser definition of failure, where all French banks, Bank of America (BoA) and Lloyds Banking Group (LBG) are classified as survived, the LR on its own just performs best. But focusing only on high hit rates across the two out-of-sample ROC figures, it is evident that the portfolio of three metrics or the LR-NSFR pair are

TABLE XII: AUROCs using alternative definitions of failure in 2006 (n=76).

	LR	RWCR	NSFR	LR & RWCR	LR & NSFR	RWCR & NSFR	LR & RWCR & NSFR
In-sample							
Baseline	0.70	0.55	0.56	0.73	**0.74**	0.59	**0.74**
+ TARP-CPP	0.61	0.49	0.60	0.67	**0.71**	0.61	**0.71**
– French banks, BoA, LBG	0.66	0.56	0.50	0.69	0.69	0.57	**0.70**
Out-of-sample							
Baseline	0.65	0.51	0.54	0.65	**0.68**	0.43	0.63
+TARP-CPP	0.58	0.47	0.56	0.59	**0.62**	0.49	0.59
–French banks, BoA, LBG	**0.63**	0.52	0.46	0.61	0.61	0.43	0.58

the most consistent strong performers.

Testing within subsamples: asset size

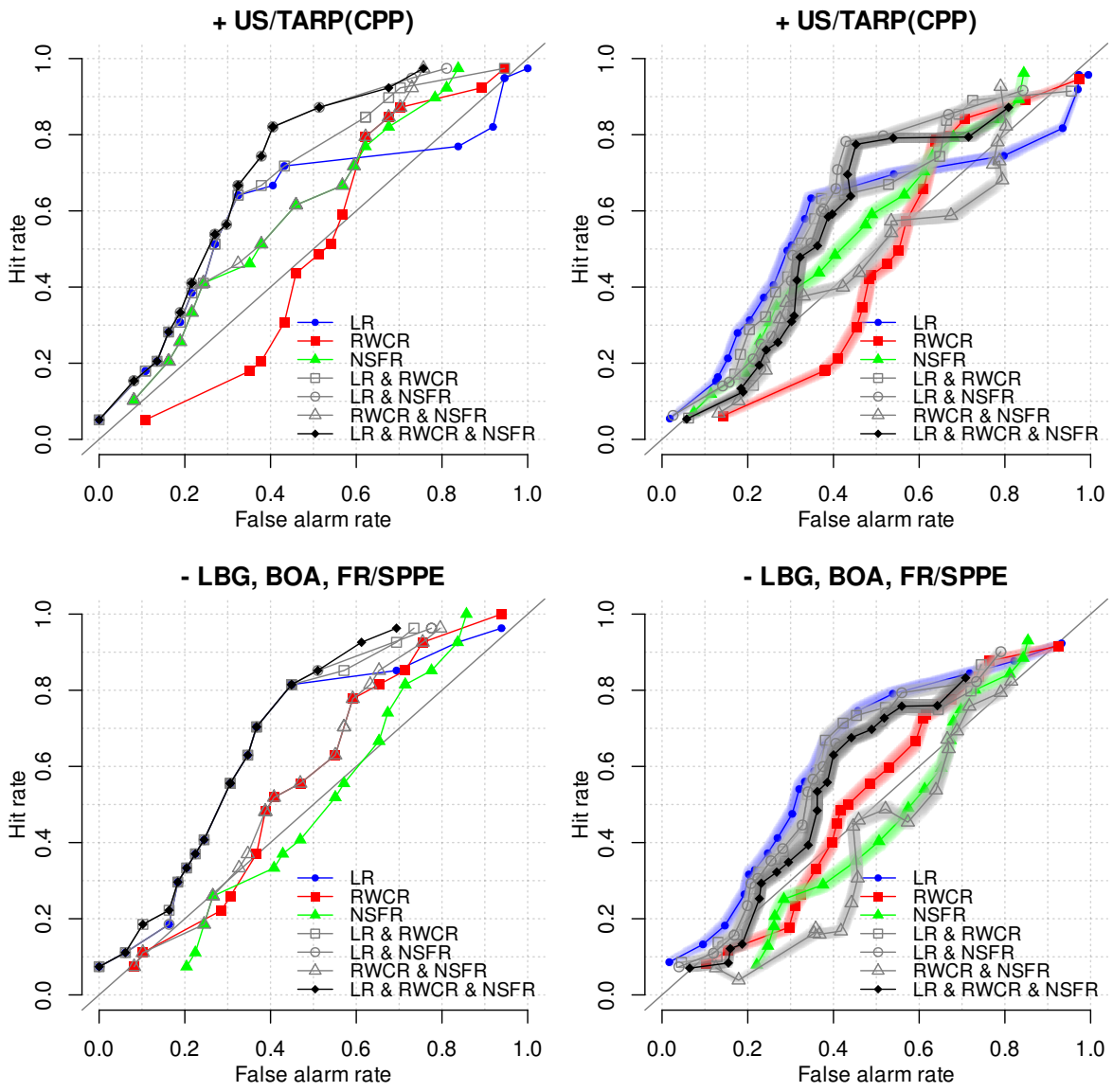
Our sample contains only large institutions globally. Nevertheless, we split the sample by size into two groups: banks with total assets below and above the sample median ($350bn). We find that our baseline result that a portfolio of metrics outperforms individual metrics generally holds for both groups (see Appendix B.1). But since all banks in our sample are relatively large, we caution against drawing conclusions from this exercise for the regulation of smaller banks. Further, we do not conduct out-of-sample tests on the two groups given the limited sample size.

Testing within subsamples: North America

The RWCR performs relatively poorly in both in-sample and out-of-sample experiments, which might suggest that it is not a useful indicator of bank failure. One potential explanation for this result is that the RWCR was part of Basel I and Basel II before the crisis, unlike the NSFR and LR. Its low predictive power could thus be a result of Goodhart's Law (Goodhart, 1975), which says that a metric becomes less useful once it is a target, as agents adjust their behaviour in response. Applied to banking regulation Goodhart's law suggests that a metric that has been a good indicator for bank failure may lose its power to predict distress once it becomes an object of regulation, perhaps due to regulatory arbitrage (Chrystal and Mizen, 2003; Haldane and Madouros, 2012).

To examine this more closely, we look at the example of the LR in North America by focusing on the sub-sample of North American banks in comparison to other banks. While the LR was not part of Basel I and Basel II regulation, it was regulated in North America before the global financial crisis alongside the RWCR. The latter was reasonably

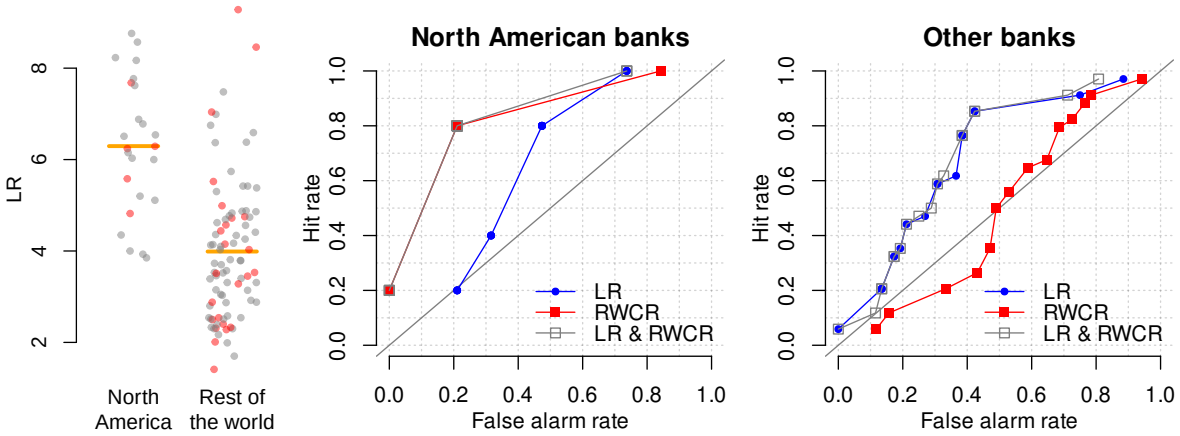
FIGURE VI: In-sample and out-of-sample ROC curves for alternative failure classifications.



comparable globally, given the RWCR uses a Basel I measure of risk-weighted assets which relied on regulatory standardised approaches. We consider all banks that have data on both the LR and RWCR—the NSFR is omitted as it reduces the sample size too much for the North American subset. We should note that the North American sample only has 5 banks that failed, all in the US.

Figure VII shows the results for the North American subsample. The left panel compares the LR of North American banks and the remaining banks. The orange bar indicates the mean across banks and failed banks are highlighted in red. As the LR was regulated in North America, it is not surprising that the LR is, on average, substantially higher for

FIGURE VII: North America vs. rest of the world. Left panel: Distribution of LR, middle panel: metrics calibrated on North American banks (n = 24), right panel: metrics calibrated on the other banks (n = 85).



North American banks.[13]

The middle panel compares the ROC curves of the LR and RWCR when calibrated on the 24 North American banks; the right panel shows the ROC curves calibrated on the remaining banks. The LR performs worse than the RWCR in identifying banks that failed in North America, while it is the other way round across the rest of the world. In both subsamples the pair of metrics performs better than the individual metrics.

Our result is consistent with Goodhart's law at play. One explanation could be that banks that were constrained by the LR in North America may have been incentivised to risk up, which could be captured by the RWCR. However, we are not able to determine whether this is the only (or correct) explanation. For example, it could be that there were different risks crystallising in North America in the global financial crisis which were better picked up by the RWCR. Banks' business models may also vary by region. For example, European banks have substantially more residential mortgage risk on the balance sheet than North American banks, which tend to securitise and take those exposures off balance sheet. The nature of the risks is therefore different, which could affect which metrics are useful for predicting failure in different jurisdictions. At the same time, using the two complementary metrics always does best in identifying banks that failed, providing further evidence that a portfolio approach to regulation could make the regime more robust.

[13] Our optimisation adapts to this and uses correspondingly higher LR thresholds for the North American sub-sample.

3.3 Assessing alternative metrics

Loan-to-deposit ratios

We investigate the performance of the loan-to-deposit (LTD) ratio, a similar metric to the NSFR. The LTD ratio and the NSFR both measure a bank's reliance on short-term funding, computing the ratio of stable funding to funding needs — the LTD is simpler, proxying these using deposits and loans. Compared to the NSFR, the LTD is available for a wider set of banks (n=96), as it does not require information on the maturity element of assets and liabilities. We look at a comparable sample to our baseline (with the same 76 observations) and also the larger sample, conducting both in-sample and out-of-sample analysis.

The top-left panel of Figure VIII shows that the performance of the LTD is not markedly

FIGURE VIII: Comparison of LTD samples (n=96 vs 76) and LTD in-sample and out-of-sample (n=96).

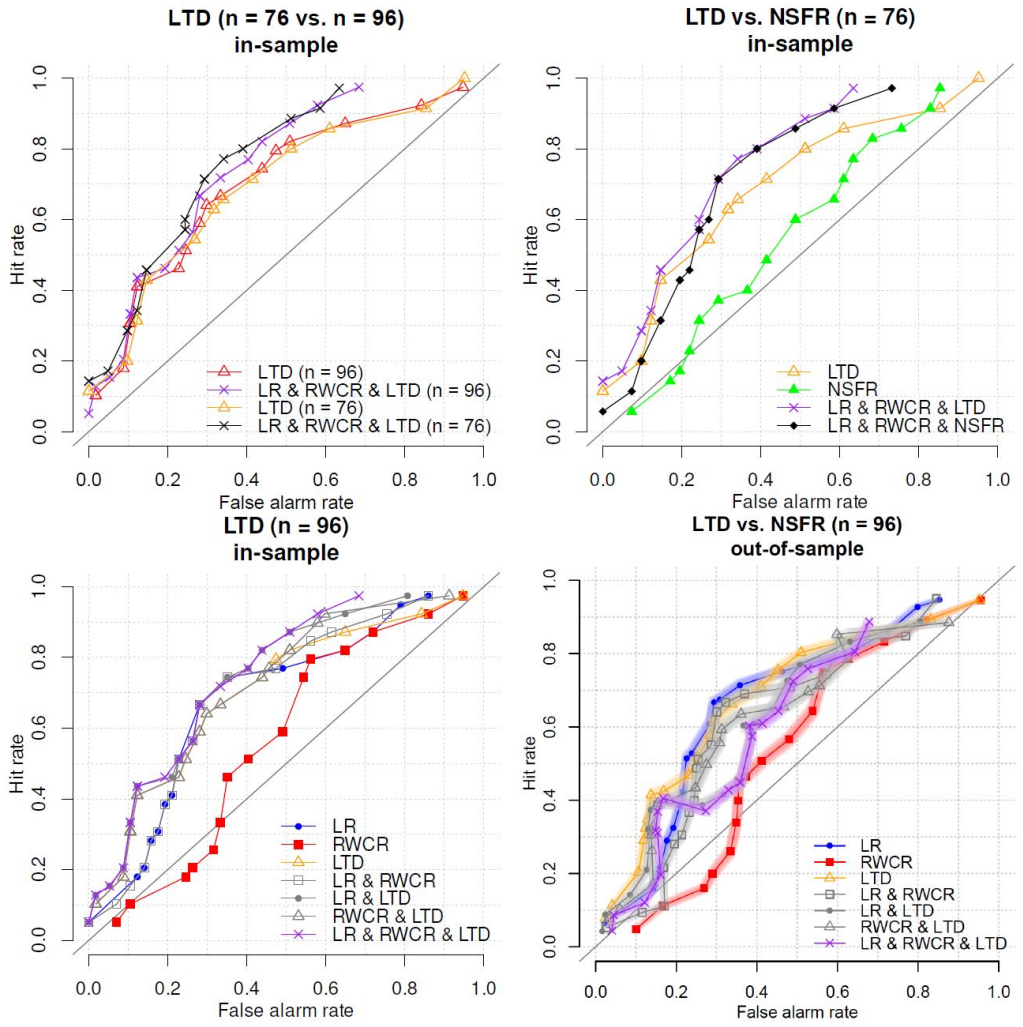


TABLE XIII: Predictions based on the LTD, measured by the AUROC. The top two rows compare the LTD and the NSFR on the baseline sample (n = 76). The bottom two rows show the performance of LTD on a bigger sample (n = 96).

		LTD	NSFR	LR	RWCR	LR & RWCR	LTD & LR	LTD & RWCR	LR & RWCR & LTD	LR & RWCR & NSFR
n = 76	In-sample	0.70	0.56	0.70	0.55	0.73	0.76	0.72	**0.77**	0.74
	Out-of-sample	**0.66**	0.54	0.65	0.52	0.65	0.64	0.62	0.59	0.63
n = 96	In-sample	0.70	N/A	0.69	0.57	0.70	0.74	0.72	**0.75**	N/A
	Out-of-sample	**0.68**	N/A	0.66	0.54	0.64	0.63	0.63	0.62	N/A

different between our baseline sample of banks (on which we also have NSFR data) and the larger sample. Figure VIII (top-left panel) shows that the LTD outperforms the NSFR in our baseline sample, both on an individual basis and as part of a three-metric portfolio.

The bottom-left panel shows the performance of our rules using LTD rather than NSFR. Mirroring the earlier results when using the NSFR, the portfolio of three metrics and the LR-LTD pair perform best. The bottom-right panel of Figure VIII shows that—out-of-sample—the portfolio rules containing LTD perform worse than the LR and LTD individually except at very high hit rates when the LR - RWCR - LTD portfolio appears to do best. The LTD alone performs remarkably well out-of-sample and is, on average, better than any other rule or individual metric, as also confirmed by the AUROCS in Table XIII. Overall, the results highlight that there may be complementary value in monitoring LTD ratios alongside the Basel 3 metrics.

Market-based metrics

We investigate how market-based equivalents of the LR and RWCR perform relative to their regulatory counterparts and also test PTB ratios. In Figure IX we compare in-sample performance using a consistent sample of 59 banks for which all metrics (balance sheet and marked-based) are available. Table XIV shows the AUROCs, also including results from out-of-sample tests.

The left panel of Figure IX and the 'Individual' row in Table XIV show that the market-based RWCR substantially outperforms its balance-sheet counterpart in and out of sample. The two estimates of the LR perform comparably, but the chart shows that the market-based one performs slightly better at very high and low hit rates and Table XIV also points to the slightly better out-of-sample performance. PTB ratios in isolation only outperform the RWCR on average; they are most informative at low hit rates but fall behind the RWCR at high hit rates.

TABLE XIV: Comparison of in-sample AUROC of the balance sheet and marked-based metrics. Individual metrics and portfolios are fitted to a consistent sample of 59 banks.

		In sample	Out of sample
Individual	LR	0.75	0.70
	LR (MB)	0.79	**0.74**
	RWCR	0.55	0.52
	RWCR (MB)	0.74	0.68
	PTB	0.71	0.68
Pairs	LR & RWCR	0.78	0.70
	LR & RWCR (MB)	0.84	0.73
	NSFR & RWCR (MB)	0.77	0.66
	LR (MB) & RWCR	0.79	0.70
	LR (MB) & RWCR (MB)	0.82	0.70
	LR (MB) & NSFR	0.80	0.72
Portfolio of 3	LR & RWCR & NSFR	0.79	0.68
	LR & RWCR (MB) & NSFR	0.85	0.71
	LR (MB) & RWCR & NSFR	0.80	0.69
	LR (MB) & RWCR (MB) & NSFR	0.83	0.70
Portfolio of 4	LR & RWCR & NSFR & PTB	0.87	0.69
	LR & RWCR (MB) & NSFR & PTB	**0.89**	**0.74**
	LR (MB) & RWCR & NSFR & PTB	0.86	0.72
	LR (MB) & RWCR (MB) & NSFR & PTB	0.87	**0.74**

Focusing on pairs and portfolios, we find that using market-based metrics generally improves performance, both in and out of sample. The balance sheet LR and market-based RWCR is the best performing pair and, similarly, the combination of the balance sheet LR, market-based RWCR and NSFR is the best portfolio of 3 metrics. The black lines in the right panel of Figure IX show that the market-based portfolios are more informative across all hit rates than the balance sheet portfolios, but the mixed portfolio performs best.

Adding PTB to create portfolios of four metrics improves performance further. Looking across all possible combinations in Table XIV, we find that the portfolio of four featuring the balance sheet LR and market-based RWCR is the best performing rule in sample. The right panel of Figure IX shows this portfolio, where we can see that adding PTB increases performance across all hit rates relative to the portfolio of three. It is also one of the best rules out of sample, together with the market-based LR on its own and the market-based portfolio of four.

Market-based metrics might add value because they contain up-to-date information on the firm that is not reflected in balance-sheet regulatory metrics, which are, by construction, backward looking. The fact that the market-based RWCR significantly outperforms its balance sheet counterpart may also partly reflect that the latter was subject to regulation pre-crisis, consistent with our findings in relation to the North American sub-sample.

FIGURE IX: Comparison of balance sheet and market-based metrics (n=59).

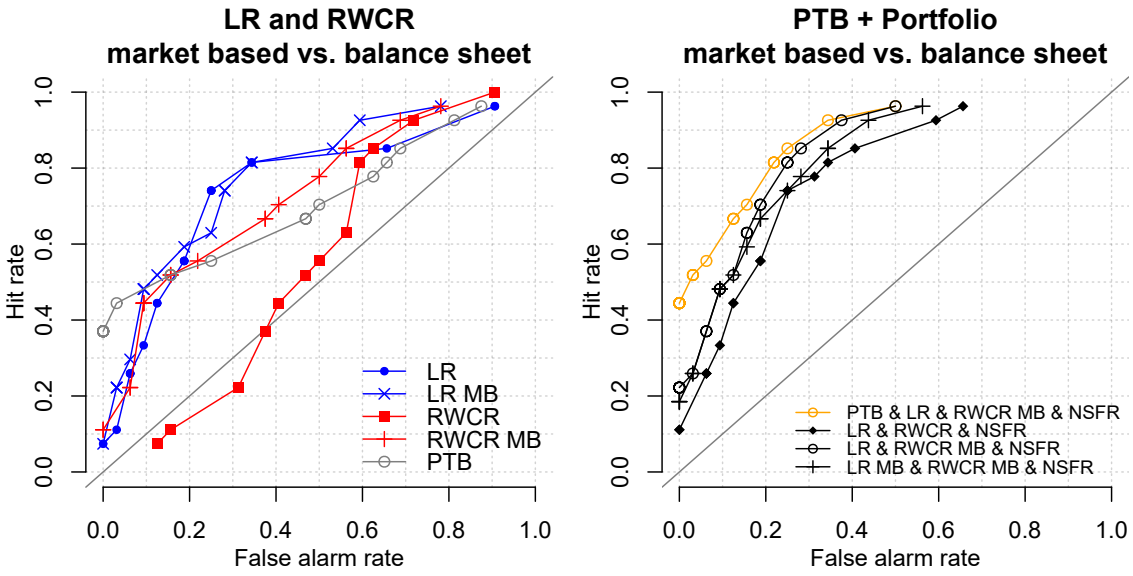
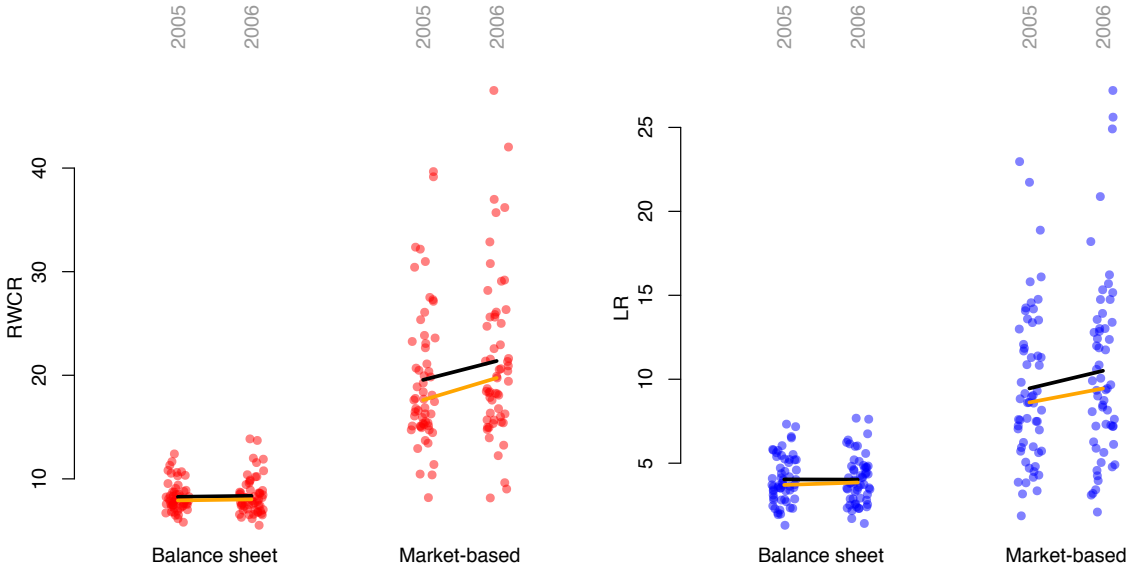


FIGURE X: Individual bank market-based and balance sheet capital ratios in 2005 and 2006. Orange bars show the median, black bars show the mean.



It is interesting that the best performing rule in sample and joint best out of sample (the portfolio of 4 metrics with the book-based LR and market-based RWCR) features balance sheet and market-based measures, and also a risk-based and a non-risk measure. It suggests that such a combination makes the rule more robust, and reinforces the potential benefits of using multiple complementary measures which focus on different dimensions of risk that a bank might face.

At the same time market-based metrics have the drawback that they may be more sensitive to market sentiment and other factors that move procyclically. If they were to be used as fixed regulatory requirements over time, during periods of financial sector exuberance, when the market typically under-prices risk and overvalues firm equity, they might be prone to underestimating the share of banks that might become stressed when the cycle turns. We see evidence consistent with this in Figure X, where we compare the 2005 and 2006 data: on average, the market-based RWCR and LR were higher in 2006 than in 2005, despite increasing risks in the lead-up to the crisis (Ye and Douady, 2018). By comparison, the regulatory metrics based on banks' balance sheets were more stable. Our findings therefore support using market-based measures as complements to existing requirements.

4 Conclusion

This paper uses a novel methodology to assess empirically the potential benefits of regulating banks using multiple capital and liquidity requirements. We proxy key elements of the Basel III framework and analyse how they would have performed individually or jointly in identifying banks that subsequently failed during the global financial crisis. We also consider what value loan-to-deposit ratios or market-based capital metrics might have in helping to gauge banking sector vulnerabilities.

Exploiting both end-2005 and end-2006 balance sheet data, we find that a small portfolio of metrics comprising the leverage ratio, the risk-weighted capital ratio and the NSFR generally outperforms individual metrics and pairs of metrics, while also requiring less stringent calibrations. Out-of-sample, the NSFR - LR pair performs best, pointing towards some overfitting of the portfolio of three metrics. Our findings are robust to two alternative definitions of bank failure and also hold when we split our sample into two by bank balance sheet size.

Since our main focus is on assessing the value of the Basel III system of regulatory metrics rather than on out-of-sample prediction of bank failure, other considerations are also relevant to the appropriate size of the portfolio. In particular a wider portfolio might be more robust across different types of banking crises and may be less vulnerable to regulatory arbitrage or Goodhart's law. Analysing the subset of North American banks provides support for this argument. The LR is the best individual indicator of failure in the whole sample but performs poorly in North America, where it was regulated before the crisis. Conversely, the RWCR performs rather poorly across all banks but is a valuable metric for North American banks. We are not able to distinguish whether this is due to Goodhart's law or due to different risks in the two subsamples. But in either case, we

find that using both metrics together consistently yields superior performance.

Overall, these results suggest that policymakers face trade offs between increasing the number of regulatory requirements, how tightly to calibrate them and the rate of false alarms. At the same time, our results indicate that the benefits of additional metrics diminish as more are added, highlighting the importance of finding a balance between those benefits and the greater complexity and costs associated with additional regulatory requirements.

We also find that there may be complementary value in monitoring loan-to-deposit ratios and market-based capital metrics alongside the Basel III requirements. Among portfolios including market-based measures, we find that the best performing setup comprises four metrics: the leverage ratio, the market-based capital ratio, the NSFR and the price-to-book ratio. It is striking that this combination features balance sheet and market-based measures and also a risk-based measure of capital and a non-risk-based measure. This may help to provide a degree of robustness and again highlights the potential benefits of using multiple, complementary measures which focus on different dimensions of risk that a bank might face. At the same time, it should be noted that market-based metrics appear to be more procyclical than balance-sheet metrics, suggesting that they could better serve as additional monitoring devices rather than regulatory standards.

Evidently, our analysis only captures some dimensions of the benefits and costs of using multiple regulatory metrics. Due to data limitations, we are also unable to assess the role of the LCR or gone-concern capital requirements alongside other metrics or test the performance of the final Basel III definitions of capital, risk-weighted assets and the leverage exposure measure.

Despite these caveats, while our results do not rule out that some simplification of the Basel III framework could be beneficial on balance, they do clearly suggest that recent calls for simplicity in regulation should not be equated to reducing the regulatory framework to a single metric. They also indicate that there may be synergies between the calibration of capital and liquidity requirements, whereby a less stringent capital requirement may be compensated with a stricter liquidity calibration to achieve a given level of resilience. As such, the results point towards complementarities between the different Basel III standards in supporting banking sector resilience rather than redundancies and inefficiencies from the multiple constraints.

A Detailed Methodology

A.1 Mixed Integer Program

To calibrate D metrics at once, we implement a mixed integer program that constrains a bank if it breaches a given threshold on at least one metric. We optimise the objective function in Equation A.1. It gives most weight to minimising the false alarm rate r_{fa}, the share of survived banks that are constrained because they breach at least one threshold. If there exist several solutions with the same false alarm rate, the program chooses the one with a higher hit rate r_{hit} (the share of failed banks that were constrained by the metrics) and lower thresholds t_d. Equation A.2 constraints the solution to have a hit rate greater than or equal to the minimum hit rate r_{hit}^{min}. Equations A.3 and A.4 define the hit rate and false alarm rate, where the outcome variable y_i indicates whether bank i failed (1) or survived (-1) and the error ϵ_i indicates whether the model's prediction is wrong (1) or correct (0). The total number of banks in the sample is denoted by N and N_+ and N_- respectively denote the number of banks in the sample that failed and survived.

All metrics are assigned a negative direction, meaning that a lower threshold t_d implies a higher probability of failure. Also, the metrics have been normalised to values between 0 and 1. Equations A.5 and A.6 linearise the matrix Z, whose entries are 1 for banks having a lower metric value x_{id} than the threshold t_d, and 0 otherwise. For the linearisation we require a vector of constants δ defined as follows: we order the values of metric d in the data set in increasing order $u_1, u_2, \ldots, u_n$ and set $\delta_d = \min\{u_{l+1} - u_l | u_{l+1} \neq u_l, l = 1, \ldots, n-1\}$. Additionally, we define the constant $M = 1 + \max \delta$. Equations A.7 and A.8 linearise the definition of the error ϵ_i as a function of the outcome y and the prediction of the model. The model predicts 1, if $\sum_{d=1}^{D} Z_{id} \geq 1$, i.e. bank i has a value lower than the threshold on at least one metric.

$$\max 100(1 - r_{fa}) + 1 r_{hit} - 0.01 \sum_{d=1}^{D} t_d \tag{A.1}$$

s.t.

$$r_{hit} \geq r_{hit}^{min} \tag{A.2}$$

$$r_{hit} \times N_+ = N_+ - \sum_{i=1}^{N} \epsilon_i \frac{y_i + 1}{2} \tag{A.3}$$

$$(1 - r_{fa}) \times N_- = N_- - \sum_{i=1}^{N} \epsilon_i (1 - \frac{y_i + 1}{2}) \tag{A.4}$$

$$M Z_{id} \geq t_d - x_{id} \qquad i = 1, ..., N, d = 1, ..., D \tag{A.5}$$

$$M Z_{id} \leq t_d - (x_{id} + \delta_d) + M \qquad i = 1, ..., N, d = 1, ..., D \tag{A.6}$$

$$D\epsilon_i \geq 0.5y_i - \sum_{d=1}^{D} Z_{id}y_i \qquad i = 1, ..., N \tag{A.7}$$

$$D\epsilon_i \leq 0.5y_i - \sum_{d=1}^{D} Z_{id}y_i + D \qquad i = 1, ..., N \tag{A.8}$$

$$Z_{id} \in \{0, 1\} \qquad i = 1, ..., N, d = 1, ..., D \tag{A.9}$$

$$\epsilon_i \in \{0, 1\} \qquad i = 1, ..., N \tag{A.10}$$

$$t_d \in [0, 1] \qquad d = 1, ..., D \tag{A.11}$$

A.2 Description of the out-of-sample testing procedure

We split the dataset into a randomly selected training set containing a fraction k of the observations and a test set containing the remaining $1 - k$ observations, where $k \in (0,1)$. Rules are calibrated in the training set for each target hit rate (r_{hit}^{min}). In the test set, the hit rate and false alarm rate pairs (r_{hit}, r_{fa}) are evaluated. This is repeated z times so that we obtain z hit rate and false alarm rate pairs $(r_{hit}^{(1)}, r_{fa}^{(1)}), (r_{hit}^{(2)}, r_{fa}^{(2)}), \ldots, (r_{hit}^{(z)}, r_{fa}^{(z)})$ for each target hit rate (r_{hit}^{min}). The average hit rate and false alarm rate across all z pairs ($\overline{r_{hit}}, \overline{r_f}$) are calculated for each target hit rate. Confidence intervals for $\overline{r_{hit}}$ and $\overline{r_{fa}}$ are estimated by ±2 standard errors of the mean across the z iterations.

B Robustness testing

B.1 Testing by bank size

We split the sample of 76 banks by balance sheet size into two groups: banks with total assets below and above the sample median ($350bn). Figure B.I and Table B.I show the in-sample ROC curves and AUROC scores for the two groups.

FIGURE B.I: In-sample ROC curves for banks with total assets above and below the median.

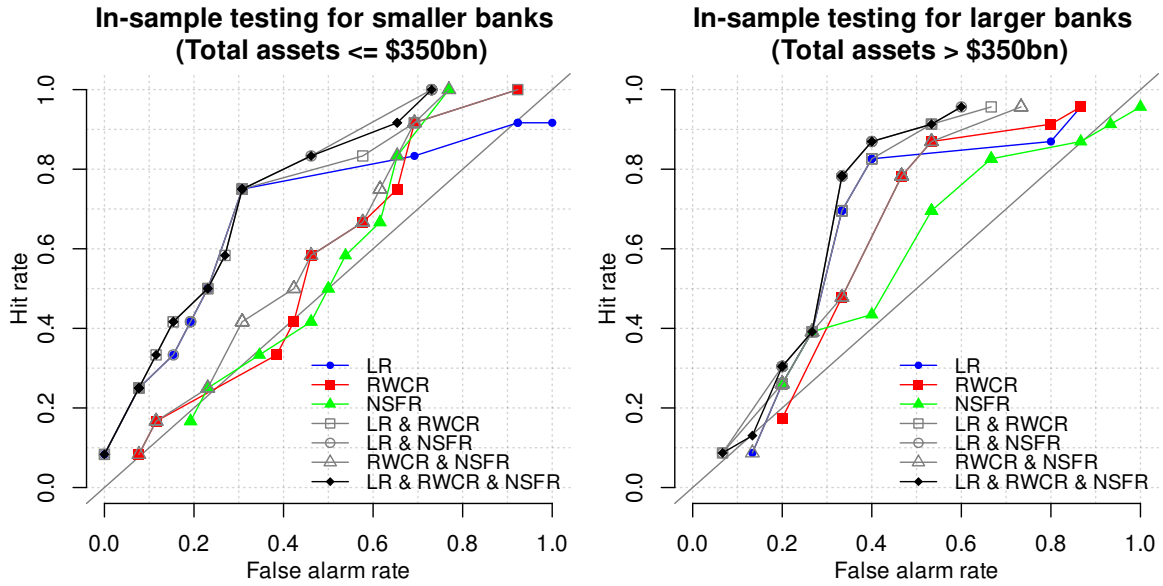


TABLE B.I: In-sample AUROCs for banks with total assets below and above the median.

	LR	RWCR	NSFR	LR & RWCR	LR & NSFR	RWCR & NSFR	LR & RWCR & NSFR
Assets ≤ $350bn	0.69	0.57	0.55	0.73	0.75	0.60	0.75
Assets > $350bn	0.66	0.63	0.57	0.70	0.72	0.65	0.71

References

Acosta-Smith, Jonathan, Michael Grill, and Jan Hannes Lang (2018) "The leverage ratio, risk-taking and bank stability," *Bank of England Staff Working Paper Series*, No. 766.

Admati, Anat, Franklin Allen, Richard Brealey, Michael Brennan, Markus K. Brunnermeier, Arnoud Boot, John H. Cochrane, Peter M. DeMarzo, Eugene F. Fama, Michael Fishman et al. (2010) "Healthy banking system is the goal, not profitable banks," *Financial Times*, Vol. 9.

Admati, Anat and Martin Hellwig (2011) "Good banking regulation needs clear focus, sensible tools, and political will," *International Centre for Financial Regulation Research Paper*, December.

Aikman, David, Mirta Galesic, Gerd Gigerenzer, Sujit Kapadia, Katsikopoulos Katsikopoulos, Amit Kothiyal, Emma Murphy, and Tobias Neumann (2021) "Taking uncertainty seriously - simplicity versus complexity in financial regulation," *Industrial and Corporate Change,* forthcoming.

Aikman, David, Andy G. Haldane, Marc Hinterschweiger, and Sujit Kapadia (2019) "Rethinking financial stability," in Olivier Blanchard and Lawrence. H. Summers eds. *Evolution or Revolution? Rethinking Macroeconomic Policy after the Great Recession*: MIT Press, pp. 143–193.

Bahaj, Saleem and Angus Foulis (2016) "Macroprudential policy under uncertainty," *Bank of England Working Paper*, No. 584.

BCBS (2010) "Basel III: International framework for liquidity risk measurement, standards and monitoring," Basel Committee on Banking Supervision.

——— (2011) "Basel III: A global regulatory framework for more resilient banks and banking systems - revised version June 2011," Basel Committee on Banking Supervision.

——— (2013) "The liquidity coverage ratio and liquidity risk monitoring tools," Basel Committee on Banking Supervision.

——— (2014) "Basel III: the net stable funding ratio," Basel Committee on Banking Supervision.

——— (2016) "Literature review on integration of regulatory capital and liquidity instruments," *Basel Committee on Banking Supervision Working Paper Series*, No. 30.

Behn, Markus, Renzo Corrias, and Magdalena Rola-Janicka (2019) "On the interaction between different bank liquidity requirements," *European Central Bank Macroprudential Bulletin*, Vol. 9.

Berge, Travis J. and Oscar Jorda (2011) "Evaluating the classification of economic activity into recessions and expansions," *American Economic Journal: Macroeconomics*, Vol. 3, No. 2, pp. 246–77.

Berger, Allen N., Sally M. Davies, and Mark J. Flannery (2000) "Comparing market and supervisory assessments of bank performance: Who knows what when?" *Journal of Money, Credit and Banking*, Vol. 32, No. 2, pp. 641–667.

BIS (2018) *Structural changes in banking after the crisis* in , CGFS Papers, No. 60: Bank for International Settlements.

Bongini, Paola, Luc Laeven, and Giovanni Majnoni (2002) "How good is the market at assessing bank fragility? a horse race between different indicators," *Journal of Banking & Finance*, Vol. 26, pp. 1011–1028.

Borio, Claudio and Mathias Drehman (2009) "Assessing the risk of banking crises–revisited," *BIS Quarterly Review*, pp. 29–46.

Borio, Claudio and Philip Lowe (2002) "Assessing the risk of banking crises," *BIS Quarterly Review*, pp. 43–54.

Boyd, John H. and Amanda Heitz (2016) "The social costs and benefits of too-big-to-fail banks: A "bounding" exercise," *Journal of Banking & Finance*, Vol. 68, pp. 251 – 265.

Brooke, Martin, Oliver Bush, Robert Edwards, Jas Ellis, Bill Francis, Rashmi Harimohan, Katharine Neiss, and Caspar Siegert (2015) "Measuring the macroeconomic costs and benefits of higher United Kingdom bank capital requirements," *Bank of England Financial Stability Paper*, No. 35.

Carletti, Elena, Itay Goldstein, and Agnese Leonello (2020) "The interdependence of bank capital and liquidity," *BAFFI CAREFIN Centre Research Paper*, No. 2020-128.

Carmona, Pedro, Francisco Climent, and Alexandre Momparler (2019) "Predicting failure in the us banking sector: An extreme gradient boosting approach," *International Review of Economics & Finance*, Vol. 61, pp. 304–323.

Cecchetti, Stephen and Anil Kashyap (2018) "What binds? interactions between bank capital and liquidity regulations," in Philipp Hartmann, Haizhou Huang, and Dirk Schoenmaker eds. *The Changing Fortunes of Central Banking*: Cambridge University Press, pp. 192–202.

Chrystal, Alec K. and Paul D. Mizen (2003) "Goodhart's law: its origins, meaning and implications for monetary policy," in Paul Mizen ed. *Central banking, monetary theory and practice: Essays in honour of Charles Goodhart*: Edward Elgar Publishing, pp. 221–243.

Cleary, Sean and Greg Hebb (2016) "An efficient and functional model for predicting bank distress: In and out of sample evidence," *Journal of Banking & Finance*, Vol. 64, pp. 101–111.

Cole, Rebel A. and Lawrence J. White (2012) "Déjà vu all over again: The causes of us commercial bank failures this time around," *Journal of Financial Services Research*, Vol. 42, No. 1-2, pp. 5–29.

Czerlinski, Jean, Gerd Gigerenzer, and Daniel G Goldstein (1999) "How good are simple heuristics?" in *Simple heuristics that make us smart*: Oxford University Press, pp. 97–118.

Demirgüç-Kunt, Asli and Enrica Detragiache (1998) "The determinants of banking crises: Evidence from industrial and developing countries," *IMF Staff Papers*, Vol. 45, No. 1, pp. 81–109.

Demyanyk, Yuliya and Iftekhar Hasan (2010) "Financial crises and bank failures: A review of prediction methods," *Omega*, Vol. 38, No. 5, pp. 315–324.

Detken, Carsten, Olaf Weeken, Lucia Alessi, Diana Bonfim, Miguel Boucinha, Christian Castro, Sebastian Frontczak, Gaston Giordana, Julia Giese, Nadya Jahn, Jan Kakes, Benjamin Klaus, Jan Hannes Lang, Natalia Puzanova, and Peter Welz (2014) "Operationalising the countercyclical capital buffer: indicator selection, threshold identification and calibration options," *ESRB Occasional Paper Series*, Vol. 5.

FPC (2014) "The financial policy committee's review of the leverage ratio," Financial Policy Committee, Bank of England.

——— (2020) "Financial Stability Report, August 2020," The Financial Policy Committee, Bank of England.

Gigerenzer, Gerd and Henry Brighton (2009) "Homo heuristicus: Why biased minds make better inferences," *Topics in Cognitive Science*, Vol. 1, No. 1, pp. 107–143.

Goodhart, Charles A. E. (1975) *Monetary Theory and Practice: The UK experience*, Chap. Problems of monetary management : the UK experience, pp. 91–121: RSpringer.

Greenwood, Robin, Jeremy C. Stein, Samuel G. Hanson, and Adi Sunderam (2017) "Strengthening and streamlining bank capital regulation," *Brookings Papers on Economic Activity*, Vol. 2017, No. 2, p. 479 —565.

Haldane, Andy G. (2011) "Capital discipline." Speech given at the American Economic Association, Denver, January 9 2011.

Haldane, Andy G. and Vasileios Madouros (2012) "The dog and the frisbee." Speech given at the Federal Reserve Board of Kansas City's 36th Economic Policy Symposium, Jackson Hole, 31 August 2012.

IMF (2018) "Global financial stability report: A decade after the global financial crisis: Are we safer?" International Monetary Fund.

——— (2019) "Global financial stability report: Vulnerabilities in a maturing credit cycle," International Monetary Fund.

ISDA (2012) "Netting and offsetting: Reporting derivatives under U.S. GAAP and under IFRS," International Swaps and Derivatives Association.

Iturriaga, Félix J. López and Iván Pastor Sanz (2015) "Bankruptcy visualization and prediction using neural networks: A study of US commercial banks," *Expert Systems With Applications*, Vol. 42, No. 6, pp. 2857–2869.

Jorda, Oscar and Alan M. Taylor (2011) "Performance evaluation of zero net-investment strategies," *National Bureau of Economic Research Working Papers Series*, No. 17150.

Jordan, Dan J., Douglas Rice, Jacques Sanchez, Christopher Walker, and Donald H Wort (2010) "Predicting bank failures: Evidence from 2007 to 2010," *European Central Bank Working Paper Series*, No. 1597.

Kiema, Ilkka and Esa Jokivuolle (2014) "Does a leverage ratio requirement increase bank stability?" *Journal of Banking and Finance*, Vol. 39, pp. 240–254.

King, Mervyn (2016) *The end of alchemy: Money, banking, and the future of the global economy*: W. W. Norton & Company.

Kumar, P. Ravi and Vadlamani Ravi (2007) "Bankruptcy prediction in banks and firms via statistical and intelligent techniques–a review," *European Journal of Operational Research*, Vol. 180, No. 1, pp. 1–28.

Laeven, Luc and Fabian Valencia (2010) "Resolution of banking crises: The good, the bad, and the ugly," *IMF Working Papers*, Vol. 10, No. 146.

Lallour, Antoine and Hitoshi Mio (2016) "Do we need a stable funding ratio? banks' funding in the global financial crisis," *Bank of England Staff Working Paper Series*, No. 602.

Le, Hong Hanh and Jean-Laurent Viviani (2018) "Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios," *Research in International Business and Finance*, Vol. 44, pp. 16–25.

Lipton, Zachary C. (2018) "The mythos of model interpretability," *Queue*, Vol. 16, No. 3, pp. 31–57.

Mayes, David G. and Hanno Stremmel (2014) "The effectiveness of capital adequacy measures in predicting bank distress," *SUERF Studies*, Vol. 1.

Moosa, Imad A. (2016) *Good Regulation, Bad Regulation: The Anatomy of Financial Regulation*: Palgrave Macmillan.

Mousavi, Shabnam and Gerd Gigerenzer (2014) "Risk, uncertainty, and heuristics," *Journal of Business Research*, Vol. 67, No. 8, pp. 1671–1678.

Rudin, Cynthia and Joanna Radin (2019) "Why are we using black box models in AI when we don't need to? a lesson from an explainable AI competition," *Harvard Data Science Review*, Vol. 1, No. 2.

Shapley, Lloyd S. (1953) "A value for n-person games," *Contributions to the Theory of Games*, Vol. 2, No. 28, pp. 307–317.

Tinbergen, Jan (1952) *On the theory of economic policy*: North-Holland Publishing Company.

Welch, Bernard L. (1947) "The generalization of student's problem when several different population variances are involved," *Biometrika*, Vol. 34, No. 1–2, pp. 28–35.

Ye, Xingxing and Raphael Douady (2018) "Systemic risk indicators based on nonlinear polymodel," *Journal of Risk and Financial Management*, Vol. 12, No. 2, pp. 1–24.