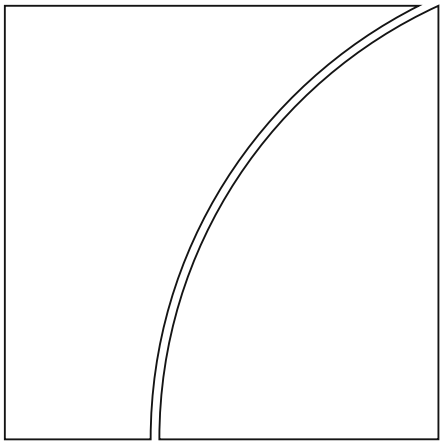
BANK FOR INTERNATIONAL SETTLEMENTS



BIS Working Papers

No 421

Evaluating early warning indicators of banking crises: Satisfying policy requirements

by Mathias Drehmann and Mikael Juselius

Monetary and Economic Department

August 2013

JEL classification: C40, G01

Keywords: EWIs, ROC, area under the curve, macroprudential policy

BIS Working Papers are written by members of the Monetary and Economic Department of the Bank for International Settlements, and from time to time by other economists, and are published by the Bank. The papers are on subjects of topical interest and are technical in character. The views expressed in them are those of their authors and not necessarily the views of the BIS.

This publication is available on the BIS website (www.bis.org).

ISSN 1020-0959 (print)

ISBN 1682-7678 (online)

Evaluating early warning indicators of banking crises: satisfying policy requirements

Mathias Drehmann and Mikael Juselius[1]

Abstract

Early warning indicators (EWIs) of banking crises should ideally be evaluated on the basis of their performance relative to the macroprudential policy maker's decision problem. We translate several practical aspects of this problem – such as difficulties in assessing the costs and benefits of various policy measures as well as requirements for the timing and stability of EWIs – into statistical evaluation criteria. Applying the criteria to a set of potential EWIs, we find that the credit-to-GDP gap and a new indicator, the debt service ratio (DSR), consistently outperform other measures. The credit-to-GDP gap is the best indicator at longer horizons, whereas the DSR dominates at shorter horizons.

JEL classification: C40, G01

Keywords: EWIs, ROC, area under the curve, macroprudential policy

[1] Bank for International Settlements.

The views expressed in the paper are those of the authors and do not necessarily represent the views of the BIS. We would like to thank the editor, an anonymous referee and our discussants James Wilcox and Fabio Fornari, as well as Ingo Fender, Andy Filardo, Oscar Jorda, Eric Schaaning and participants at the 9th International Institute of Forecasters' Workshop and the Norges Bank Macroprudential Regulation Workshop for helpful comments. Anamaria Illes provided excellent research assistance. All remaining errors are our own. Mathias Drehmann: Bank for International Settlements, Centralbahnplatz 2, CH-4002 Basel, Switzerland, mathias.drehmann@bis.org. Mikael Juselius: Bank for International Settlements, Centralbahnplatz 2, CH-4002 Basel, Switzerland, mikael.juselius@bis.org.

1. Introduction

Early warning indicators (EWIs) are an essential component for the implementation of time-varying macroprudential policies, such as countercyclical capital buffers, that can help reduce the high losses associated with banking crises. EWIs in this context must not only have sound statistical forecasting power, but also need to satisfy several additional requirements. For instance, signals need to arrive early enough, so that policy measures have enough time to be effective, and they need to be stable as policy makers tend to react on trends. In general, deriving optimal empirical models for forecasting requires detailed knowledge of the underlying decision problem (eg Granger and Machina (2006)). Such knowledge is, however, currently not available in the context of macroprudential policies, as there is limited experience from which expected costs and benefits could be estimated (CGFS (2012)). Nevertheless, it is still possible to incorporate the qualitative aspects of the policy maker's decision problem into the estimation and evaluation procedures for EWIs. Laying down such an approach and applying it to a range of EWIs is the main objective of this paper.

Given the difficulty of estimating the costs and benefits of macroprudential policies, the second best option is to evaluate EWIs over a range of possible utility functions. As the optimal decision under a specific utility function implies a specific trade-off between Type I and Type II errors, one way to achieve this is to consider the full mapping between such trade-offs that a given EWI generates. This mapping is called the receiver operating characteristic (ROC) curve.[2] Going back to World War II, the ROC curve has a long tradition in other sciences (eg Swets and Picket (1982)), but its applications to economics are more scarce. Recent exceptions include for instance Cohen et al (2009), Gorr and Schneider (2011), Berge and Jorda (2011) or Jorda et al (2011).

The ROC curve has several useful properties (eg Hsieh and Turnbull (1996)). In particular, the area under the curve (AUC) is a convenient and interpretable summary measure of the signalling quality of a binary signal. AUCs can also be estimated easily. Parametric and non-parametric estimators are available as well as confidence bands and Wald statistics for comparing the AUCs of two signals (eg Janes et al (2009) and Pepe et al (2009)).

Following this literature, we adopt AUC as the primary metric for assessing and comparing the classification ability of EWIs and use it to embed macroprudential policy requirements into the evaluation process. In particular, we specify three additional criteria related to the timing, stability and interpretability of ideal EWIs of banking crises.

The appropriate timing is a crucial requirement for EWIs. On the one hand, macroprudential policies need time before they become effective (eg Basel Committee (2010)). On the other hand, signals which arrive at very early stages can also be problematic as policy measures are costly (eg Caruana (2010)). We therefore require that signals should arrive at least one and a half years but no more than five

[2] The ROC curve is a mapping between the false positive rate (Type II errors) and true positive rate (the complement of Type I errors). The somewhat awkward name goes back to its original use in trying to differentiate noise from signals of radars. A parallel way of expressing the signalling quality of an EWI is the correct classification frontier (eg Jorda et al (2011)), which is more intuitive for optimal choice problems.

years ahead of a crisis. The stability of the signal is a second, largely overlooked, requirement. For one, policy makers tend to base decisions on trends rather than reacting to changes in signalling variables immediately (eg Bernanke (2004)). Gradual implementation of policy measures may also allow policy makers to affect market expectations more efficiently and deal with uncertainties in the transmission mechanism (CGFS (2012)). Since EWIs that issue stable and persistent signals reduce uncertainty regarding trends, they allow for more decisive policy actions. The final, less tangible, requirement is that EWI signals should be easy to interpret, as any forecasts, including EWIs, that do not "make sense" are likely to be ignored by policy makers (eg Önkal et al (2002), Lawrence et al (2006)).

In the empirical part of the paper, we apply our approach to assess the performance of 10 different EWIs. We mainly look at the EWIs individually, but at the end of the paper we also consider how to combine them. Our sample consists of 26 economies, covering quarterly time series starting in 1980. The set of potential EWIs includes more established indicators such as real credit growth, the credit-to-GDP gap, growth rates and gaps of property prices and equity prices (eg Drehmann et al (2011)) as well as the non-core liability ratio proposed by Hahm et al (2012). We also test two new measures: a country's history of financial crises and the debt service ratio (DSR). The DSR was first suggested in this context by Drehmann and Juselius (2012) and is defined as the proportion of interest payments and mandatory repayments of principal to income. An important data-related innovation of our analysis is that we use total credit to the private non-financial sector obtained from a new BIS database (Dembiermont et al (2013)).

We find that the credit-to-GDP gap and the DSR are the best performing EWIs in terms of our evaluation criteria. Their forecasting abilities dominate those of the other EWIs at all policy-relevant horizons. In addition, these two variables satisfy our criteria pertaining to the stability and interpretability of the signals. As the credit-to-GDP gap reflects the build-up of leverage of private sector borrowers and the DSR captures incipient liquidity constraints, their timing is somewhat different. While the credit-to-GDP gap performs consistently well, even over horizons of up to five years ahead of crises, the DSR becomes very precise two years ahead of crises. Using and combining the information of both indicators is therefore ideal from a policy perspective. Of the remaining indicators, only the non-core liability ratio fulfils our statistical criteria. But its AUC is always statistically smaller than the AUC of either the credit-to-GDP gap or the DSR. These results are robust with respect to different aspects of the estimation, such as the particular sample or the specific crisis classification used.

The remainder of the paper is organised as follows: Section 2 relates the procedures for evaluating EWIs to the decision problem of the macroprudential policy maker. In particular, it introduces ROC curves and translates various additional policy requirements into statistical evaluation criteria. Section 3 discusses data and introduces the potential EWIs. Section 4 evaluates and compares the signalling quality of the EWIs based on the criteria laid down in the previous sections and undertakes robustness checks. Section 5 concludes.

2. Evaluating EWIs based on policy requirements

When the purpose of a forecast is to guide a policy decision in an uncertain environment, the policy maker's preferences and constraints matter for the ex post

evaluation of its forecasting performance as these components define the loss function (eg Pesaran and Skouras (2002) and Granger and Machina (2006) and references therein). Equally, the preferences that implicitly correspond to standard statistical evaluation criteria rarely make sense in a specific policy context. For example, a comparison of alternative forecasts based on squared error loss will generally not, even as an approximation, capture the economically relevant trade-off and is therefore likely to be sub-optimal (Granger and Pesaran (2000) and Granger and Machina (2006)).

The close link between decisions under uncertainty and forecasts suggests that there are potentially substantial benefits to explicitly specifying the constraints and preferences of the policy maker. For example, Elliot and Lieli (2013) construct a utility-based forecast for binary outcomes and show that it obtains large gains over other existing methods. The main difficulty of such an approach, however, is that it is more information intensive and may require knowledge about preferences that are not directly observable.

In this section, we discuss the problem of evaluating EWIs of banking crises – ie forecasts of the likelihood that a banking crisis will occur given a set of covariates – from the perspective of macroprudential policy. We begin by discussing the difficulties of assessing the costs and benefits of such policies. In light of these difficulties, we introduce an evaluation metric that is consistent with the underlying decision problem but nevertheless robust over a wide range of preferences. We then discuss some additional requirements, for instance related to the timing and stability of the EWI signals, which are likely to be important for the successful implementation of macroprudential policies. We translate each of these requirements into clear statistical criteria for evaluating EWIs.

2.1 Macroprudential policy and the choice of evaluation metric for EWIs

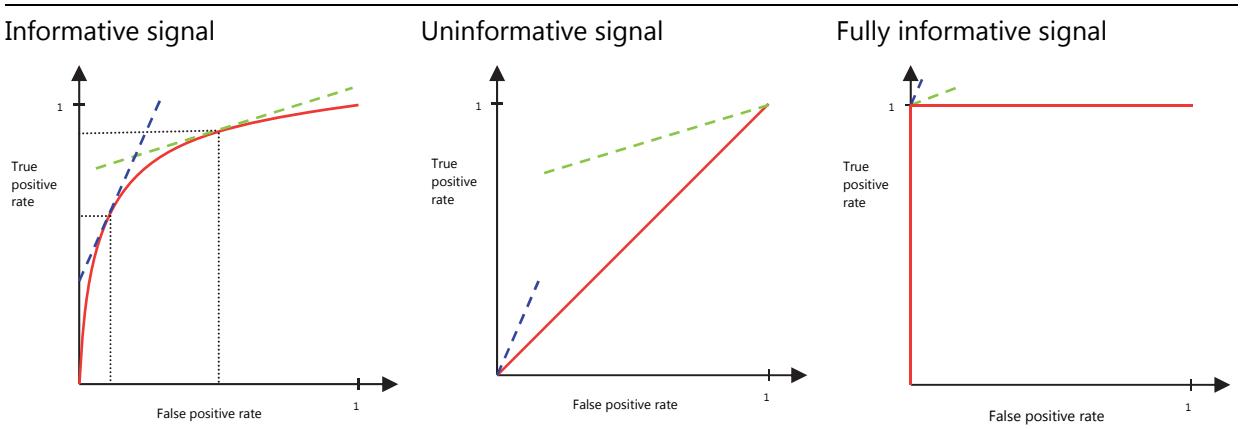
Following the 2008 financial crisis, the use of macroprudential policies has expanded rapidly (eg CGFS (2012) or IMF (2011)). Whilst tools and actual policies differ, the key objective of macroprudential policies is the reduction of systemic risk, defined as the risk of widespread disruptions to the provision of financial services that have serious negative consequences for the real economy (eg Borio (2009)). A crucial component of the macroprudential approach is to address the procyclicality of the financial system by, for example, stipulating the accumulation of buffers in "good times" so that these can be drawn down in "bad times". Tools which are already used in this regard include countercyclical capital buffers or dynamic provisioning. One key challenge for policy makers – and the focus of this paper – is the identification of the different states in real time, with particular emphasis on detecting unsustainable booms that may end up in a financial crisis.

To make matters more concrete and to see how the policy maker's utility affects the choice of an optimal EWI, assume a very simple economy that can be in three states: a normal state and a boom ("good times") that is inevitably followed by a crisis ("bad times"). Whilst the policy maker knows when there is a crisis, the true state in normal and boom times ($B=0$ and $B=1$ respectively) is not directly observable. In these states, the policy makers can either implement a policy ($P=1$) or not ($P=0$). Implementing a policy is costly, but it has the benefit of reducing economic losses in case there is a crisis. Let U_{PB} denote the utilities of choosing policy P in state B satisfying the natural assumptions $U_{11} > U_{01}$ and $U_{00} > U_{10}$.

Further, suppose that the policy maker observes a real valued signal *S*, which carries imperfect information about the current state. The signal can be anything from a probability prediction about *B* from a statistical model to an observable economic variable. For simplicity, we assume that the higher the value of *S*, the more likely it is that the economy is booming, but any variable which becomes lower in a boom will have this property if multiplied by -1. The decision problem for the policy maker is to assign a threshold, θ, for *S* above which the probability of being in the boom state is high enough for the cost-benefit trade-off of corrective policy actions to be optimal. Setting such a threshold effectively turns *S* into a binary EWI for the crisis state.

In an ideal situation, the chosen threshold for *S* would signal the state with certainty. However, some noise will be associated with the signal in practice. This implies that there is a trade-off between the rate of true-positives, $TPR_S(\theta) = P(S > \theta \mid B = 1)$, and the rate of false positives, $FPR_S(\theta) = P(S > \theta \mid B = 0)$.[3] For very low values of the threshold, for instance, *TPR* will be close to one, but the same will also hold for *FPR*. When the threshold is high, the opposite occurs. For intermediate values of the threshold, the trade-offs between the *TPR* and *FPR* rates will move close to the upper left boundary of a unit square if *S* is highly informative, and along a 45° line if it is uninformative. The mapping from *FPR* to *TPR* for all possible thresholds is called the receiver operating characteristic (ROC) and defined by *TPR = ROC(FPR)*. The trade-offs of three hypothetical variables are depicted by the red lines in Graph 1.

Signal quality and welfare — Graph 1



Note: Red line: ROC curve. Dotted lines: preferences of a policy maker who weights the expected costs and expected benefits linearly. The blue (green) line indicates high (low) costs relative to benefits.

Given a trade-off between true and false positives, how should policy makers set the threshold for action? It is straightforward to show (eg Baker and Kramer (2007) and Cohen et al (2009)) that the policy maker should choose the threshold in such a way that the expected marginal rate of substitution between the net marginal utilities of accurate prediction in the normal and boom states equals the slope of the ROC curve, that is

[3] The FPR and the complement of the TPR correspond to the familiar Type II and Type I errors from classical statistics.

$$\frac{dROC}{dFPR} = \frac{(U_{00} - U_{10})(1 - \pi)}{(U_{11} - U_{01})\pi},$$

where π is the unconditional probability of a crisis. For example, if implementing the policy measure is costly compared to its expected benefits, the policy maker is relatively averse to high *FPR*. This is illustrated by the steep blue line in Graph 1. The opposite holds if the cost of a crisis is relatively high, as indicated by the flat green line in the graph. The optimal threshold is then the one that corresponds to the tangent points between the red and green or blue curves in Graph 1.

Unfortunately, it is difficult to assess the expected costs and benefits of macroprudential policy and hence to specify the optimal trade-off between the TPR and FPR of different signals. For example, some studies have attempted to quantify the expected costs of tighter capital requirements. But estimates of the impact of a 1 percentage point increase in capital requirements on output range from close to 0 to a reduction of 0.35%, depending on modelling assumptions and the time horizon considered. Even less is known about the expected costs and benefits of other macroprudential policies such as countercyclical liquidity or loan-to-value requirements (eg CGFS (2012)). For this reason, Drehmann (2012) simulates the policy trade-offs for reasonable parameterisations of the costs and benefits of macroprudential policy measures. He finds that the scope for policy makers' relative trade-offs is surprisingly wide and comprises even extreme cases, when policy makers essentially care only about true or false positives.

The question, therefore, is: how to evaluate the quality of different signals in the absence of knowledge about the costs and benefits of policy actions? A possible solution is to look at the entire ROC curve, which essentially amounts to evaluating the signal over the full range of possible utility functions (Elliot and Lieli (2013)). The ROC curve has several convenient properties (eg Hsieh and Turnbull (1996)): (i) it is invariant under monotone increasing transformations of the measurement scale; (ii) it lies above the diagonal in the unit square if the distribution of *S* during the boom is stochastically larger than during the normal state; (iii) it is concave if the densities of *S* associated with the two states have a monotone likelihood ratio; and (iv) the area under the curve can be interpreted as the likelihood that the distribution of *S* during the boom is stochastically larger than during normal times.

The last property suggests that the area under the curve (AUC) provides a convenient and interpretable summary measure of the signalling quality of *S*. Formally, the AUC of signal *S* is given by

$$AUC(S) = \int_0^1 ROC(FPR(S))dFPR(S)$$

AUC is increasing with the predictive power of the indicator across all possible thresholds θ and lies between 0 and 1. It takes the value 0.5 for uninformative indicators. AUC is larger than 0.5 if *S* is informative and stochastically larger in booms than in normal times. Conversely, AUC is smaller than 0.5 if *S* is informative and stochastically smaller in booms than in normal times.

Given its useful properties and the absence of detailed knowledge about costs and benefits of macroprudential regulation, we adopt the AUC to assess the relative performance of different EWIs in this paper. That said, the property that $AUC(S_1) > AUC(S_2)$ for two different signals does not guarantee that the ROC curve of S_1 is larger than the ROC curve of S_2 for all FPRs. This implies that the AUC does not necessarily lead to the same optimal ordering of different EWIs that could be

obtained if the policy maker's preferences were known. But for the variables assessed here, this problem is of limited practical relevance.

Pepe et al (2009) and Janes et al (2009) discuss both parametric and non-parametric estimators of AUC. They also provide confidence bands and discuss a Wald statistic for comparing the AUC of two signals. A complication arising in the panel context is that the data are likely to be correlated over time within individual countries. In this case, the comparison of the AUCs of two different variables can be based on bootstrapped standard errors (Janes et al (2009) and Gorr and Schneider (2011)).

We next discuss requirements that the implementation of macroprudential policies places on EWIs and express them in terms of the AUC.

2.2 Timing, stability, and relative performance

In this section we highlight two important characteristics of an ideal EWI and formally state the criterion for choosing one indicator above an alternative. Throughout the discussion we implicitly assume that the indicators increase with the probability of a crisis. In general, decreasing indicators can be accommodated either by reversing their interpretation (ie multiplying by -1) or the inequalities in the criteria.

The appropriate timing of an ideal EWI is crucial for policy makers and has two dimensions. First, EWIs need to signal crisis early enough so that policy actions can be implemented in time to be effective. The timeframe required to do so depends inter alia on the lead-lag relationship between changing a specific macroprudential tool and the impact on the policy objective (CGFS (2012)). In contrast to monetary policy, where it is well known that it takes at least a year for interest rates to impact on inflation, this relationship is less well understood for macroprudential instruments. Yet, it is likely to be at least as long. For instance, banks have one year to comply with increased capital requirements under the countercyclical capital buffer framework of Basel III (Basel Committee (2010)). In addition, data are reported with lags and policy makers generally do not act immediately on data developments but observe trends for some time before they change policies (eg Bernanke (2004)). This suggests that EWIs should start issuing signals at least 6 quarters before a crisis.

Second, ideal EWIs should not signal crises too early as there are costs to macroprudential policies. This can undermine the support for adopted measures if they are implemented too early (eg Caruana (2010)). For instance, after Spain introduced dynamic provisions in 2000, the provisioning system was weakened in 2004, because of pressures from banks and uncertainties by the authorities over the correct calibration (Fernández De Lis and Garcia-Herrero (2012)). Judging what is "too early" for an ideal EWI is difficult. But to be conservative, we use a 5 year horizon for our empirical analysis.

To assess the appropriate timing of an indicator S_i, we compute $AUC(S_{i,h})$ for all horizons h within a 5 year window before a crisis, ie h runs from -20 to -1 quarters.[4]

[4] By looking at each horizon separately we do not want to suggest that the revealed average time pattern should be used to anchor policy very specifically. Rather, our aim is to broadly document the temporal stability of EWIs, which is important for policy making.

When we compute $AUC(S_{i,h})$, we ignore signals in all other quarters than h in the window. For example, at horizon -6, $TPR(S_{i,-6})$ is solely determined by signals issued 6 quarters before crises. The $FPR(S_{i,-6})$, on the other hand, is based on all signals issued outside the five year window before crises.

We define S_i to have the right timing, if $AUC(S_{i,h})$ is statistically greater than 0.5 for some $h \in [-20,-6]$, or:

<u>Criterion 1:</u> *An EWI S_i has the right timing if*

$$AUC(S_{i,h}) > 0.5 \quad \text{for some horizon } h \in [-20,-6].$$

A special difficulty related to Criterion 1 can arise if the direction of an indicator reverses at different time horizons. For example, suppose that high values of an indicator S_i signal a boom at $h = 16$ (ie $AUC(S_{i,16}) > 0.5$), whereas low values do the same at $h = 8$ (ie $AUC(S_{i,8}) < 0.5$). Since such a pattern is informative in its own right, we use $AUC(S_{i,h}) \neq 0.5$ in Criterion 1 in these cases, rather than multiplying S by -1 at the problematic horizons. This problem is also connected to the stability of the signal, which is the next issue discussed.

The stability of signals is an important additional requirement that has been largely overlooked in the literature so far. As already discussed, policy makers do not react immediately on data developments in practice, but base policy decisions on trends (eg Bernanke (2004)). Such behaviour can be optimal when information is noisy (see eg Orphanides (2003) in the context of monetary policy). Gradualism can similarly be a useful strategy for macroprudential policy makers and may allow them to better affect the expectations of market participants and deal with uncertainties in the transmission mechanism (CGFS (2012)). Since EWIs that issue stable or persistent signals reduce uncertainty regarding trends, they allow for more decisive policy actions.

To judge the stability of an indicator S_i, we therefore assess whether the signalling quality of S_i is deteriorating when the forecast horizon becomes shorter. Given Criterion 1, we use $AUC(S_{i,-6})$ as the comparator for all signals with different horizons.[5] Thus, our second policy requirement is:

<u>Criterion 2:</u> *An EWI S_i is stable if*

$$AUC(S_{i,-6-j}) \leq AUC(S_{i,-6}) \leq AUC(S_{i,-6+k}) \quad \text{for } j = 1,...,14 \text{ and } k = 1,...,5.$$

By definition, any informative signal that reverses direction during the policy relevant horizons is deemed not stable.

The stability criterion is also connected to the persistence of the underlying conditioning variables. For instance, Park and Phillips (2000) show that binary choice models tend to generate long periods where no signals are issued followed by episodes of intensive signals when the explanatory variables are difference stationary. This suggests that variables that fulfil Criterion 2 may display a high degree of persistence. We investigate this property for a set of potential EWIs and discuss its implications for estimation and inference in Section 3.

[5] Comparing the stability requirement across all horizon pairs within the policy interval would involve $2.4e^{18}$ computations.

Finally, as indicated above, we use the AUC to rank different indicators. Writing this as a formal criterion we get:

Criterion 3: EWI S_i outperforms EWI S_j for horizon h if

$$AUC(S_{i,h}) > AUC(S_{j,h}) .$$

Note that to compare an increasing indicator, $S_{i,h}$, with a decreasing indicator, $S_{j,h}$ say, we would have to multiply the latter by -1 or replace AUC($S_{i,h}$) by 1 – AUC($S_{i,h}$) in Criterion 3.

2.3 Other policy requirements: robustness and interpretability

Beyond the three policy requirements formalised above, policy makers place additional demands on EWIs. One obvious requirement is robustness. For example, the signalling quality of an EWI should remain intact over different samples and not be overly sensitive to the specific crises dating employed. While this seems self-evident, we nevertheless stress the importance of such testing for EWIs, as the financial sector has a tendency to undergo rapid changes. Of course, while robustness checks allow us to find prevalent features in past data, it is never possible to assess the future stability of EWIs.

A second additional policy requirement is that the EWI signals should be easy to interpret, ie an ideal EWI should not only fulfil the aforementioned statistical criteria, but also needs to "make sense". Otherwise an EWI will not be used, as practitioners typically value the interpretability of forecasts more than accuracy (Önkal et al (2002)) and adjust forecasts if they lack justifiable explanations (Gönül et al (2009)). In addition, if EWIs have sound conceptual underpinnings, they are better suited for clear communication – an important aspect of macroprudential policy making (CGFS (2012)) – and will increase the confidence in the future ability of the EWI to signal crises.[6]

This suggests that purely statistical EWIs, for example obtained from various data-mining exercises, are not very appealing from a policy perspective. Ideally, the analytical framework supporting EWIs would be based on one or several well established theoretical models. Unfortunately, most state-of-the-art macro models do not yet convincingly account for financial crises – the main event we are interested in – despite growing work in this area. For this reason, we rely on EWIs that reflect the tradition of Kindleberger (2000) and Minsky (1982), who see financial crises as the result of mutually reinforcing processes between the financial and real sides of the economy. For instance, as the economy grows, cash flows, incomes and asset prices rise, risk appetite increases and external funding constraints weaken, thereby generating potentially large financial imbalances. At some point, these imbalances have to unwind, potentially causing a crisis, characterised by large losses, liquidity squeezes and possibly a credit crunch.

[6] If EWIs are used to guide policy decisions, they will become subject to the usual Lucas critique, implying that their leading properties might disappear. For instance, if banks are forced to build up buffers based on signals issued by well specified EWIs, they would be more resilient toward busts, which in turn could make crises less likely. As Drehmann et al (2011) argue, however, the loss of predictive content per se would be no reason to abandon the scheme – it would be rather a sign of its success.

3. EWIs and data

In the remainder of the paper, we test whether a range of EWIs fulfil the discussed policy requirements. Rather than looking at a wide range of potential indicators, we focus on those that have a clear economic interpretation, are available across time and countries, and other studies found to be successful in this context. In total we assess 10 different variables.

3.1 EWIs

Drehmann et al (2011) analyse a wide range of potential indicators covering macroeconomic variables, indicators of banking sector conditions and market indicators. They find the latter two groups do not perform well as EWIs of systemic banking crises. We therefore focus more narrowly on a small set of macroeconomic indicator variables, which have greater potential for capturing the build-up of domestic financial vulnerabilities.

In line with the predictions of Minsky (1982), Drehmann et al (2011) find that indicators of excessive credit and asset price booms generally perform well as EWIs. The authors show that the credit-to-GDP gap, which measures deviations of credit-to-GDP from a long run trend, is the single best indicator. This variable also acts as the starting point of discussions about the level of countercyclical capital buffer charges according to Basel Committee (2010). The importance of booming credit conditions is also in line with Reinhart and Rogoff (2009), Gourinchas and Obstfeld (2012) and Jorda et al (2011), among others. We therefore include the credit-to-GDP gap and the change in real credit in the analysis. As alternative indicators of such financial booms, we also include the change in real residential property and equity prices and their respective gaps in the analysis.

More recently, Drehmann and Juselius (2012) propose the aggregate debt service ratio (DSR) as a useful early warning indicator. The DSR is a measure of the proportion of interest payments and mandatory repayments of principals relative to income for the private non-financial sector as a whole and can be interpreted as capturing incipient liquidity constraints of private sector borrowers. If DSRs are high, it is a clear sign that households and firms are overextended, so that even small income shortfalls prevent them from smoothing consumption or making new investments. Larger shortfalls could even trigger a rise in defaults and ultimately a crisis. If both lending rates and maturities are constant, the DSR and the credit-to-GDP gap provide the same information. Yet, Drehmann and Juselius (2012) show that this condition – in particular for lending rates – is not fulfilled, so that the DSR captures the burden that debt imposes on borrowers more accurately.

Hahm et al (2012) argue that lending booms can only be sustained if banks are able to fund assets with non-core liabilities, in particular wholesale and cross-border funding. The reason is that traditional retail deposits (core liabilities) adjust only sluggishly. In their paper, they assess a range of proxies for core and non-core liabilities. They find empirically that cross-border liabilities plus M3 minus M2 (proxy for non-core liabilities) divided by M2 (proxy for core liabilities) works best as an EWI for crises. In line with their findings, we include this variable as the non-core liability ratio in our analysis.

We analyse two further variables as potential benchmarks. First, given its use as an indicator for the business cycle, we assess the signalling quality of real GDP

growth, even though Drehmann et al (2011) have already shown that it performs quite poorly as an EWI for systemic banking crisis. Second, as a naïve benchmark, we assess if a country's history of financial crises is informative. If some countries are more prone to crises than others, knowing a country's name and history would already provide beneficial information for policy makers. We call this variable History and it contains the number of financial crises a country has experienced since World War II up to each point in time.

3.2 Data

We analyse quarterly time-series data from 26 countries. The sample starts in 1980 for most countries, and at the earliest available date for the rest. It ends in 2012 Q2. Table A1 in Annex 2 provides an overview of the sample.

With respect to the dating of systemic banking crises we follow Laeven and Valencia (2012), but ignore three crises that were primarily driven by cross-border exposures. This is because our indicators are based on domestic data and are geared towards capturing the build-up of domestic vulnerabilities.[7] In addition, we adjust the exact crisis dates in a few instances following discussions with central banks. In Section 4.2, we undertake robustness checks with respect to alternative crisis dates based on Reinhart and Rogoff (2009), Borio and Drehmann (2009), and the inclusion of crises driven by cross-border exposures. A list of all crisis dates is included in Table A1.

We use a balanced sample for the main part of the paper, ie we only consider a subsample for which all indicator variables are available. In addition, we require that all variables are available for the full five-year forecast horizon before any crisis is included in the sample so that the estimated time profile of AUCs is not changing because of differences in the number of countries captured. We also drop the crisis quarter and two years afterwards, as binary EWIs become biased if the post-crisis period is included in the analysis (Bussiere and Fratzscher (2006)). We then have around 2,500 quarterly observations and 19 systemic crises. Using the unbalanced sample and the more lenient crisis definition, we cover at most 31 crises and approximately 3,300 quarterly observations in our robustness checks.

Macroeconomic variables are taken from national data sources and the IMF *International Financial Statistics* (IMF-IFS). Residential real estate property prices are based on BIS statistics which are only available for a subset of countries and generally do not cover the full sample period. Data on M2 and M3 are also difficult to obtain from a single source. We therefore merge data from national authorities and Datastream. Cross-border liabilities are taken from the IMF-IFS.

An important data-related feature of our analysis is that we use a measure of total credit to the private non-financial sector obtained from a new BIS database (Dembiermont et al (2013)).[8] The past literature has generally relied on proxies for this measure, such as bank credit to the private-non financial sector reported in the IMF-IFS. However, this can be misleading as it excludes important sources of credit,

[7] Crises of the latter type were identified for three countries (Germany, Sweden and Switzerland in 2007 and 2008) through information provided to us by national central banks.

[8] This database is available on the BIS website (http://www.bis.org/statistics/credtopriv.htm). New Zealand is not yet covered by these data, so we continue to use bank credit for this country.

such as bond markets or cross-border loans. Dembiermont et al (2013) show that across countries and time banks provide on average 70% of credit to the private non-financial sector. But this varies considerably. For example, in the United States, banks provided more than 50% of credit in the 1950s, but they extend just above 30% currently. In Australia, on the other hand, the ratio of bank credit to total credit has increased steadily, from around 35% in the 1970s to more than 70% in 2012.[9]

Several of our variables are expressed as deviations from trend – ie gaps. We derive gap measures by subtracting a one-sided Hodrick-Prescott filtered trend from the level of a series.[10] This is achieved by recursively extending the sample by one period and retaining the difference between the actual value of the variable and the value of the trend at the new point. Thus, a property price trend calculated in, say, 1988 Q1 only takes account of information until that date. To ensure that trends are stable enough, we require at least 10 years of information.[11]

The calculation of the Hodrick-Prescott filter involves a key smoothing parameter λ. It has become standard to set the smoothing parameter λ to 1,600 for quarterly data. Ravn and Uhlig (2002) show that for series of other frequencies (daily, annual etc) it is optimal to set λ equal to 1,600 multiplied by the fourth power of the observation frequency ratio. We set lambda for all the gaps to 400,000, implying that financial cycles are four times longer than standard business cycles. This seems appropriate, as crises occur on average once in 20 to 25 years in our sample. Equally, we could have used a time trend such as Gourinchas and Obstfeld (2012), but our approach is in line with the suggestion for calculating credit-to-GDP gaps as indicator variables for the countercyclical capital buffers in Basel III (Basel Committee (2010)).

Debt service ratios (DSRs) are taken from Drehmann and Juselius (2012). Even though levels are surprisingly similar across countries and time despite different levels of financial development, some country differences persist, for example, due to different rates of homeownership or different industrial structures. To account for this, we subtract 15-year rolling averages from the DSRs for our analysis. For similar reasons, we also subtract 15-year rolling averages from the non-core liabilities ratio.

3.3 Persistency and real-time considerations

In this section, we discuss two issues which are related to the temporal dimension of the data. The first concerns the persistency of the data and its implications for estimation and inference. The second relates to the need to make evaluations primarily based on information available in real time.

Section 2.2 discussed the potential connection between the stability Criterion 2 and the persistence of the underlying conditioning variables. While this type of

[9] Developments in Australia were driven by the dismantling of tight regulation that had led to the emergence of a large shadow banking sector. In addition, the substantial increase in household borrowing was mainly satisfied by the banking sector (see Dembiermont et al (2013)).

[10] For asset price gaps, the difference between the actual data and the trend is normalised by the trend.

[11] Given the limited availability of property price data, this assumption is relaxed for the residential property price gap. For a few countries, we only require a minimum of six years of data before we calculate the trend for the first four years in the sample. This is most important for Korea, as we are then able to include the Asian financial crisis in our baseline sample.

persistence may be beneficial for policy makers, it can nevertheless have some important implications for the econometric approach.

To assess the persistence of the potential indicator variables described in Section 3.1, we estimate AR(k) processes for each variable and apply standard unit root tests. We also calculate the sum of autoregressive coefficients, which is often used as a summary measure for the persistence of a series.[12] Table 1 summarises the results, highlighting that credit growth, property price growth and, to a lesser extent, GDP growth all display a high level of persistency. This persistence is in many cases statistically indistinguishable from unit-root dynamics. This would imply that the *levels* of these variables are near double unit-root processes. In contrast, equity price growth is less persistent and the unit-root hypothesis is rejected in most cases. Gap transformations of these variables are even more persistent than the growth rates but otherwise follow the same internal ordering in terms of persistency. Note also that both the DSR and non-core liabilities display a very high degree of persistency. The high levels of persistency can be problematic for standard regression-based models of binary choice, for which a statistical theory is generally unavailable (Park and Phillips (2000)). Inference, in particular, can become very misleading.

Descriptive statistics on persistence — Table 1

	GDP growth	Credit growth	Prop. pr. growth	Equity pr. growth	Credit-to-GDP gap	Property pr. gap	Equity pr. gap	DSR	Non-core liability ratio
Stationary	13/26	4/26	10/23	23/26	2/25	4/23	8/25	3/26	2/26
AR Sum	0.85	0.90	0.90	0.75	0.95	0.97	0.93	0.97	0.96

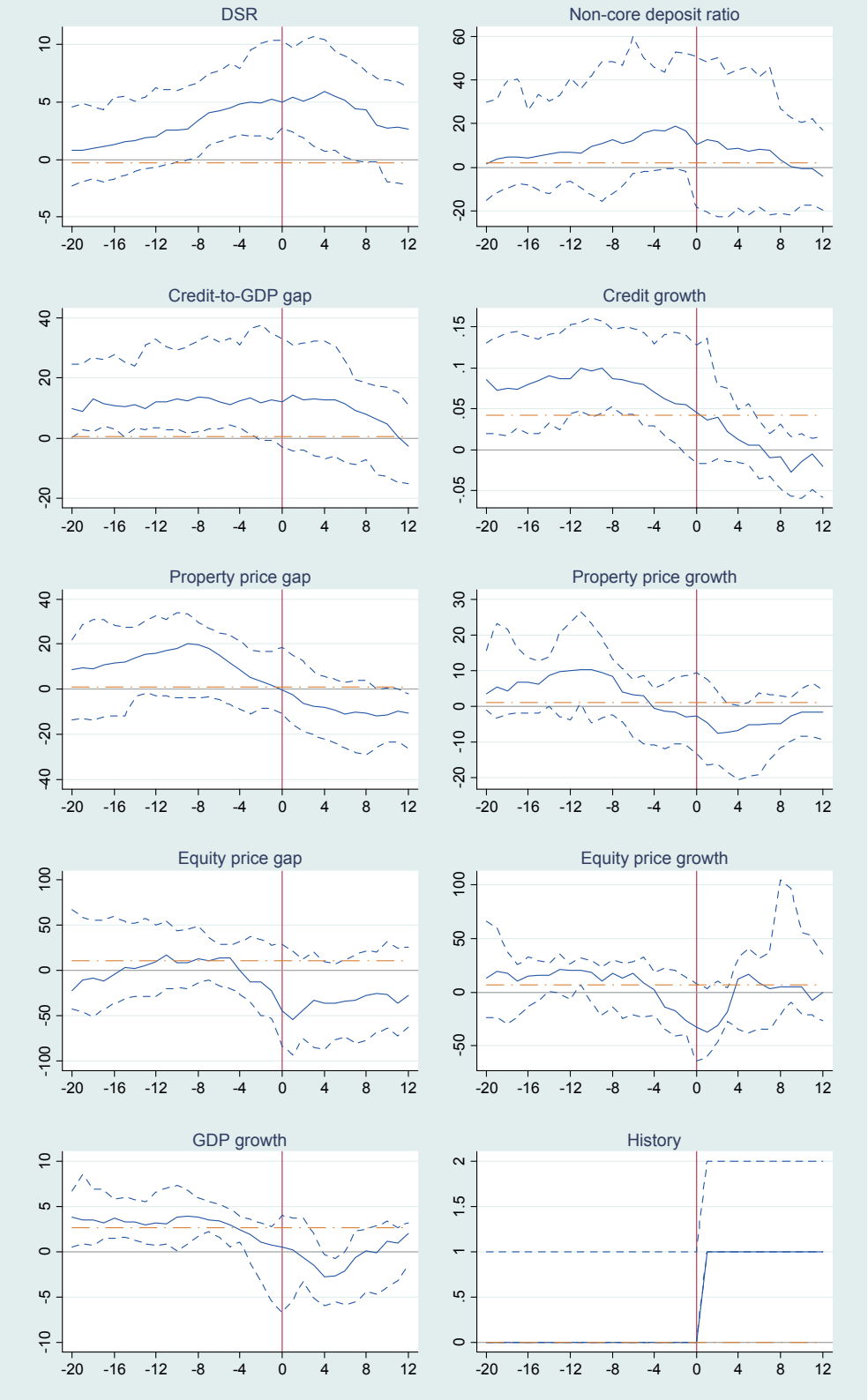
Note: The row labelled "Stationary" reports the number of countries for each variable (columns) for which a standard ADF-test is rejected at the 5% significance level. Lag-lengths were chosen based on the AIC information criterion from a maximum of 5 lags. The row labelled "AR Sum" reports the average sum of AR coefficients across countries for each variable. Countries with less than 10 years of time-series observations on the variables are excluded.

The temporal dimension of the data poses an additional complication, namely the need to evaluate EWIs based on information available in real time only. This is an important practical constraint as policy makers do not have the benefit of hindsight. For this reason, all of our indicators are quasi real–time. For instance, as pointed out above, we calculate trends by using only information which was available up to each specific point in time. A particular problem arises when the signal, S_i, is formed as a predicted value from a binary regression of the crisis indicator on a set of covariates. In this case, S_i will contain full-sample information (the estimated coefficients) even if all covariates only reflect real-time information individually. In Annex 1 we show that relying on such full sample estimates can make a substantial difference for the AUCs of some indicators, even though they could never be applied for real policy decisions. Ideally, we would also use vintage data. Unfortunately, such data are not available for the large set of countries and the historical period we need to cover in order to establish robust indicators. It has

[12] It is well known that the sum of autoregressive coefficients can be significantly downward biased in small samples, in particular when the model includes a deterministic trend (Andrews and Chen (1994)). Although none of the estimated AR(k) models included deterministic trends, we caution that the actual persistence may be even larger than what is reported in Table 1.

Indicator variables around crises

Graph 2



Note: The horizontal axis depicts minus 20/plus 12 quarters around a crisis (time zero, vertical red line). The historical dispersion (median (solid line), 25th and 75th percentiles (dashed lines)) of the relevant variable is taken at the specific quarter across crisis episodes in the balanced sample. Brown lines (dash dots) show the median value outside the minus 20/plus 12 window around crises.

also been shown that data revisions, at least for credit and GDP series in the US, are not of first-order importance for our type of analysis (Edge and Meisenzahl (2011)).

We circumvent both the problems related to persistency and (full-sample) parameter estimates by adopting a non-parametric estimation method discussed for instance by Pepe et al (2009). This is also similar to the signal extraction approach popularised by Kaminsky and Reinhart (1999) for EWIs of banking crises. An additional benefit of the non-parametric approach is that it is robust to potential specification errors. Binary models are estimated to maximise a specific likelihood function that, to the extent it is subject to misspecification, can perform arbitrarily badly at specific points of the policy maker's loss function (Elliott and Lieli (2013)).

3.4 The behaviour of indicator variables around systemic crises

Before conducting our statistical tests, we look at the time profile for all indicator variables around systemic banking crises. Graph 2 summarises the behaviour of the variables during a window of 20 quarters before and 12 quarters after the onset of a crisis (time 0). For each variable, we show the median (solid line) as well as the 25th and 75th percentiles (dashed lines) of the distribution across episodes. As a benchmark, we include the median value of the variable in other periods (brown dash-dotted line).

The graph reveals that several variables could be useful indicators for signalling impeding crises. In particular, the DSR looks promising. In normal periods, the median DSR is roughly zero. In the four years before a crisis, though, it more than triples, peaking shortly afterwards. The credit-to-GDP gap is also substantially different before a crisis than in other periods, as it is highly elevated during the entire five-year run-up to a crisis.

For the other variables, the median value for normal periods generally falls within the 25th and 75th pre-crisis percentiles, suggesting that these variables are likely to deliver noisy early warning signals. Nonetheless, several variables, such as total credit growth, the non-core deposit ratio and the property price variables, show clear tendencies to rise well in advance of a crisis and fall just before or directly after its onset.

By construction, the History variable jumps directly after a crisis. The lines for its percentiles reflect the important fact that banking crises are rare events: nine out of the 26 countries in the balanced sample experience no crises, 15 countries experience only one crisis and the remaining two experience two crises.

4. The signalling quality of different EWIs

In this section, we report the results from testing whether the 10 proposed EWIs fulfil the three statistical criteria. As outlined in Section 3.3, ROC curves are estimated non-parametrically. We use trapezoid approximations to smooth the estimated curves when calculating the AUC values. In line with Janes et al (2009), we

also correct for potential clustering along the country dimension and derive standard errors via bootstraps using 1,000 replications.[13]

4.1 EWIs and policy requirements – AUCs over time

Graph 3 summarises the main results, which are also reported numerically in Table A2 in Annex 2 including all statistical tests whether the criteria are fulfilled or not. The graph shows the estimated AUCs and their 95% confidence intervals (dashed lines) for all indicator variables and forecast horizons.[14] Horizon -6 is indicated by a black vertical line and the different line specifications reflect which criteria are fulfilled:

- Dashed blue line: Criterion 1 is not fulfilled, ie the AUC at horizon h is not statistically different from an uninformative indicator (AUC=0.5 denoted by grey vertical lines).
- Solid blue lines: Criterion 1 is fulfilled, ie the AUC is significantly different from 0.5.
- Hollow blue circles: Criterion 2 is satisfied, ie the signal is stable, conditional on Criterion 1 being fulfilled.
- Blue diamonds: The AUC of the indicator is not statistically different from the indicator with the highest AUC at horizon h, conditional on Criteria 1 and 2 being fulfilled.
- Red diamonds: The indicator has the highest AUC at horizon h.

The message from Graph 3 is clear: the best performing EWIs are the DSR and the credit-to-GDP gap.

The DSR's early warning properties are especially strong in the last two years preceding crises. In the last four quarters before crises, the DSR is even a nearly perfect indicator: its AUCs are around 0.94% with the upper confidence intervals close to 100% (see Table A2, Annex 2). Given the discussion in Section 2.1, however, this may be too late for countervailing policy actions. But AUCs for horizons -6 to -10 are still impressive with values between 0.82 and 0.91.[15] AUCs then drop continuously the longer the forecast horizon gets.

The credit-to-GDP gap shows a markedly different pattern. Its AUCs are highly stable. In the first four years, they fluctuate between 0.83 and 0.85. Only in the last four quarters (ie quarter -16 to -20) do AUCs start to drop to 0.8. Given the length of the forecast horizon, this performance is remarkable, showing that the credit-to-GDP gap provides reliable EWIs well in advance of systemic banking crises.

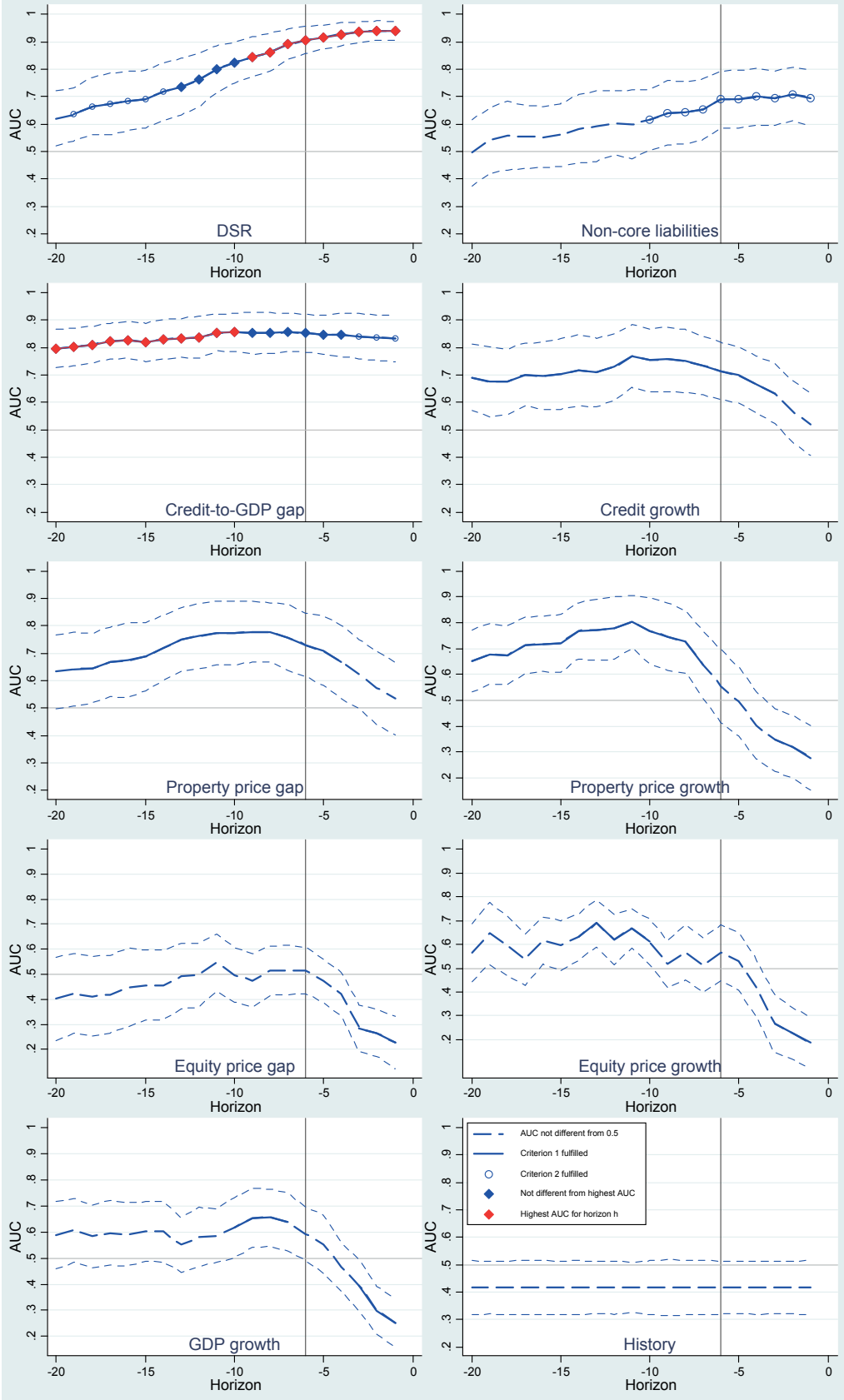
[13] The results are not significantly altered if we do not cluster and/or use trapezoid approximations for the ROC curve. The latter only affects the AUC of History as it is discrete valued, taking only the values 0, 1 and 2 before crises. Thus, there are only four true positive rates 0, 0.05, 0.21 and 1 and a step-wise ROC curve becomes highly misleading (see Graph A1 in Annex 1).

[14] As an example, ROC curves for horizon -8 are shown in Graph A2 in Annex 2.

[15] AUC values of 0.85 are high relative to other empirical findings. For instance, Jorda (2011) cites studies showing that a widely used prostate-specific antigen (PSA) blood test has an AUC of around 0.8 and that the S&P 500 has an AUC of 0.86 for detecting whether the economy is in recession or not in real time.

EWIs and policy requirements – AUCs over time

Graph 3



Note: Horizon: quarters before crises. Black vertical line: Horizon -6. Grey horizontal line: 0.5. Blue dashed lines: 95% confidence intervals. Blue diamonds are only shown if criteria 1 and 2 are fulfilled.

Beyond the DSR and the credit-to-GDP gap, some of the other indicators also provide useful signals for policy makers, in particular non-core liabilities. This is the only indicator that satisfies Criteria 1 and 2. However, the AUC of non-core liabilities is statistically smaller than the highest AUCs of either the DSR or the credit-to-GDP gap across all horizons. In contrast, the property price variables have AUCs between 0.7 and 0.8 three to two years before a crisis, but their informational content decreases in the run-up, implying that they do not satisfy Criterion 2. This is unsurprising, given previous research showing that property prices peak well ahead of crises (Borio and McGuire (2004)). In fact, property price growth reverses direction approximately one year before a crisis, so that low growth becomes a significant early warning indicator two quarters ahead of crises. Whilst these signals are informative, they are issued potentially too late for policy actions.[16] Similarly, real credit growth satisfies Criterion 1 and shows AUC values that are not significantly lower than those of the DSR or the credit-to-GDP gap, but it becomes uninformative in the immediate run-up to a crisis and thus does not issue stable signals.

Turning to the remaining variables, it is clear that they these are not very well suited as EWIs of banking crises. Whilst there are some horizons when equity indicators or GDP growth are informative, these signals are not very consistent or are issued too close to crises to be useful. A general observation that emerges from Graph 3 is that growth rates, while having the advantage of not requiring statistical pre-filtering, are dominated by the more persistent variables.

Finally, it appears from Graph 3 that History nearly provides a useful negative early warning signal, ie the more crises a country has experienced up to some point in time, the *less* likely it is to experience another. This can be seen from its AUC value of 0.42, which is not statistically different from 0.5 at the 5% confidence level, but just so. However, this result is purely due to the fact that banking crises are rare events in the sample. The vast majority of countries (16 out of 19) that experience a crisis do not experience another. For these countries, History takes the value 0 before the crisis and 1 thereafter. Hence, when History jumps to 1, the likelihood of another crisis *within the sample* becomes very low. If we drop post-crisis signals after the last crisis of each country, History becomes totally uninformative.[17] Given these results, we exclude History from our further discussion.

4.2 Robustness

In this section we check the robustness of our results with respect to changes in the sampling period, in the country coverage and in the adopted crisis dating. Throughout the section, we refrain from testing whether AUC values are statistically different from each other as the samples analysed are generally not nested.

[16] In addition, the informational content of the DSR in these periods is significantly higher, even if the AUC is determined in such a way that low growth rates are used as crises signals. Such an AUC peaks at 0.72 for horizon -1 (see Table A2 in Annex 2).

[17] Dropping post-crisis periods has no significant effect on AUCs of all other variables except for the equity gap, which becomes significant for some policy-relevant horizons. Results are available on request.

4.2.1 Stability across time

We first assess robustness over time. We split the sample into two roughly equal parts consisting of data pre- and post-2000 Q1.

Graph 4 shows that the results are very similar across the two samples. The AUCs of the DSR are virtually identical in the two subsamples, except for forecast horizons beyond three years. The signalling qualities of the credit-to-GDP gap and credit growth are somewhat reduced in the more recent sample. The reason for this is that several countries, such as Ireland, Spain and Portugal, had prolonged periods of very high real credit growth (implying high credit-to-GDP gaps) starting already before or in the early 2000s. Given the five-year forecast horizon, credit variables in these cases provide many false positives, and proportionally more so if the samples are split, even though the credit booms resulted in crises in the end. Thus, from a policy perspective these signals ultimately proved to be correct and an earlier intervention could have reduced some of the high-cost crises experienced by these countries.

The only important differences arise for the equity price gap. It has some predictive power in the pre-2000 sample, in line with earlier findings (eg Borio and Lowe (2002)). Yet, this fades in the more recent period, possibly reflecting the fact that equity and property price cycles were more synchronised then compared to now, so that equity prices acted as a proxy for property prices (Borio and McGuire (2004)).

4.2.2 Stability across countries

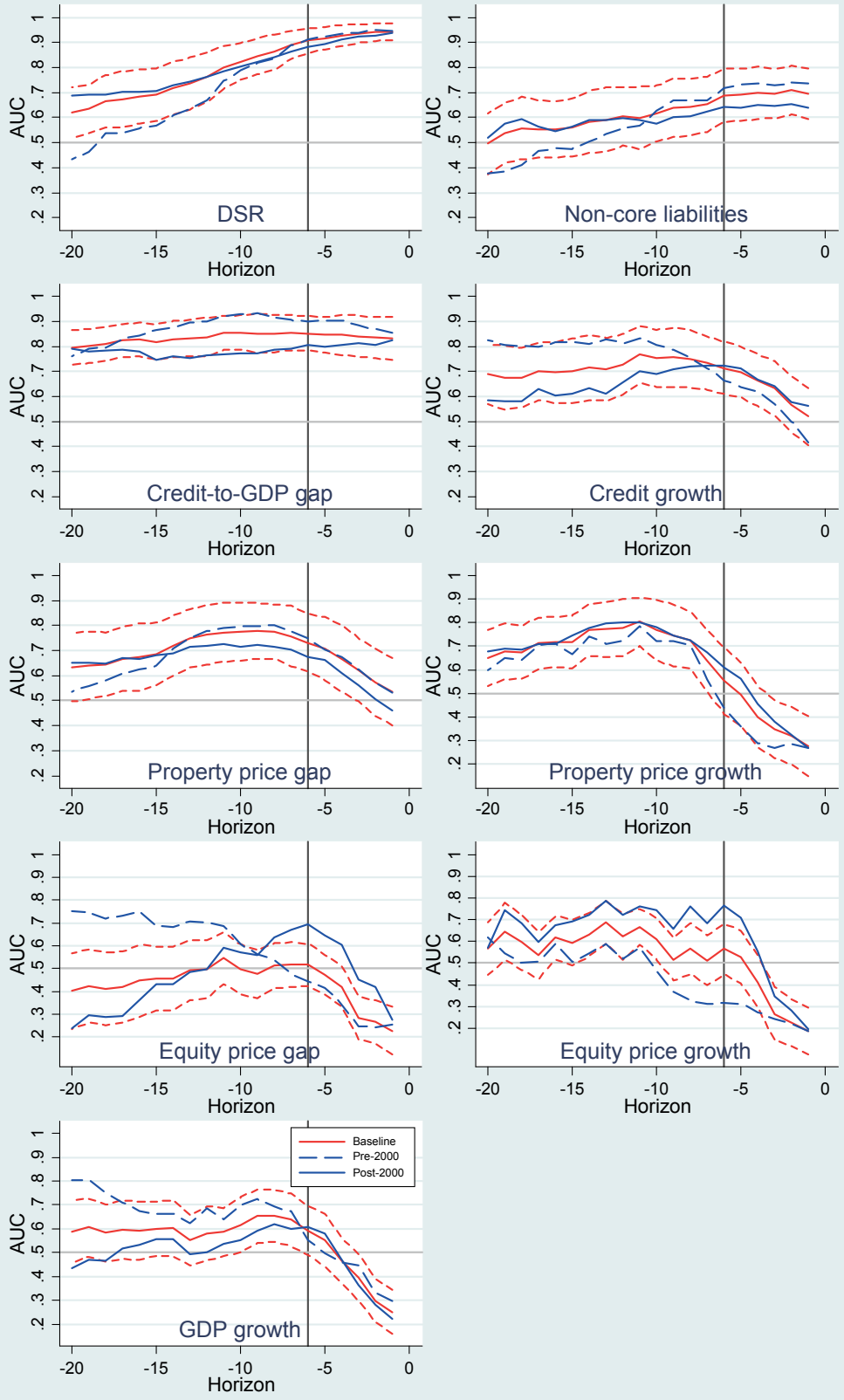
Our main results are based on a sample covering a broad range of countries, such as the United States, Japan, Thailand and Portugal. To test robustness with respect to country coverage, we split the sample into those countries that are members of the G20 and those that are not, ie large and small economies (G20 membership is indicated in Table A1 in Annex 2). In contrast to the previous sections, we use an unbalanced panel, to ensure that we have enough crises in the two samples.[18]

The blue lines (solid for the G20 countries and dashed for non-G20 countries) in Graph 5 show that the AUCs of the two subsamples are essentially encapsulated by the confidence bands corresponding to the baseline results (red lines). In this regard, the results are robust. However, the rank ordering of the credit-to-GDP gap and the DSR changes. Whilst the credit-to-GDP gap is more informative for countries which are part of the G20, the opposite holds for the DSR. In the G20 subsample, the AUC for the credit-to-GDP gap is close to or above 0.9 from horizon -16 onwards. The AUCs of the DSR in this sample, on the other hand, reach these levels only in the last year before crises. For the non-G20 countries, however, the AUC of the DSR is extremely high: From horizon -6 onwards even the lower confidence band is at a minimum 0.94 and the upper one is always 1.

[18] The results of the baseline case (Section 4) are unaffected regardless of whether a balanced or unbalanced sample is used. Results are available on request.

AUCs in different time periods

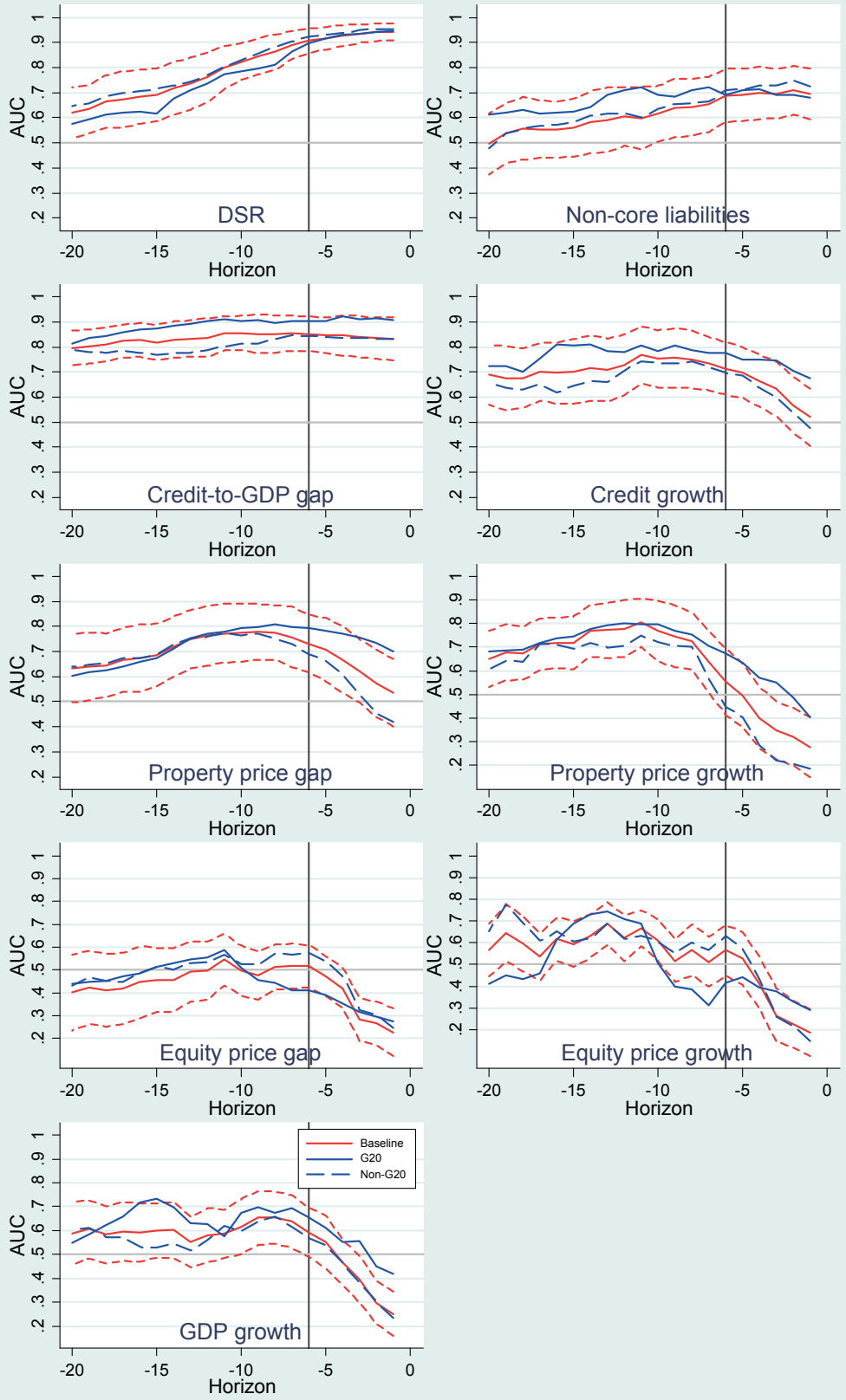
Graph 4



Note: Dotted red lines: 95% confidence intervals of AUC for the baseline estimation. Horizon: quarters before a crisis.

AUCs for different countries

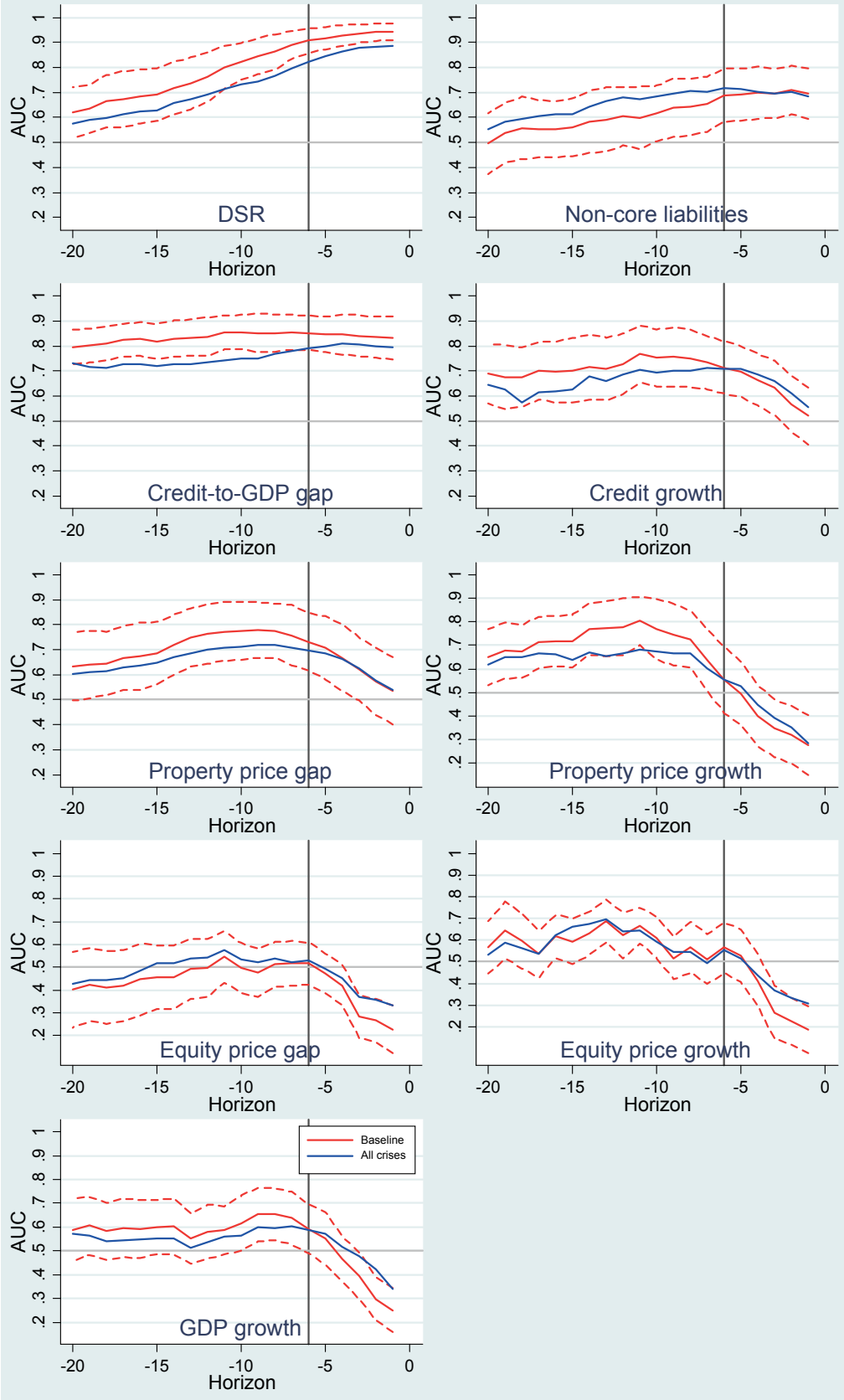
Graph 5



Note: Dotted red lines: 95% confidence intervals of AUC for the baseline estimation. Horizon: quarters before a crisis.

AUCs over time with different crisis dates

Graph 6



Note: Dotted red lines: 95% confidence intervals of AUC for the baseline estimation. Horizon: quarters before a crisis.

4.2.3 Stability with respect to crisis dating

The analysis so far has relied on the stringent crisis dating of Laeven and Valencia (2012), modified in a few cases to reflect information provided by national central banks. According to these authors, two conditions must be met before an event is classified as a crisis: First, there are significant signs of financial distress in the banking system (as indicated by significant bank runs, losses in the banking system and/or bank liquidations), and second, there are significant policy interventions in response to significant losses in the banking system.

Other authors have suggested weaker crisis definitions, declaring a crisis for example when one major bank gets into trouble. If we continue to use the unbalanced sample and rely on the crisis dating by Reinhart and Rogoff (2009), as well as the weakest definition in Borio and Drehmann (2009) for the most recent episode, we obtain a maximum of 31 crises. We also include crises which were primarily driven by cross-border exposures (Germany, Sweden and Switzerland during the recent global crisis episode) although such events do not strictly fall within the scope of our indicators. For example, we should not expect the Swiss credit-to-GDP gap – an indicator of domestic financial vulnerabilities – to predict the near failure of one systemically relevant bank in Switzerland due to the bank's business in the United States.

Graph 6 highlights that our results are very robust to changing the crisis dating. As we include cross-border crises, it is unsurprising that the predictive ability of many indicators decreases somewhat. However, even for the DSR, where this is most apparent, the results remain extremely strong. The AUCs are still well above 0.8 from horizon -7 onwards, peaking at 0.89 one quarter before the crisis.

A couple of other interesting results stand out from Graph 6. First, the differences for the residential property price and equity price indicators are smallest across crisis dating methodologies. One explanation could be that asset prices were exuberant even in the less severe crises, whereas truly systemic events also involved excessive credit as measured by the credit-to-GDP gap or the DSR. Second, the informational content of non-core liabilities increases for the weaker crisis definition. It is unclear why. But it is certainly the case that banks become more vulnerable the more they rely on non-core liabilities as these are less stable in stressed conditions.

4.3 Combining indicators

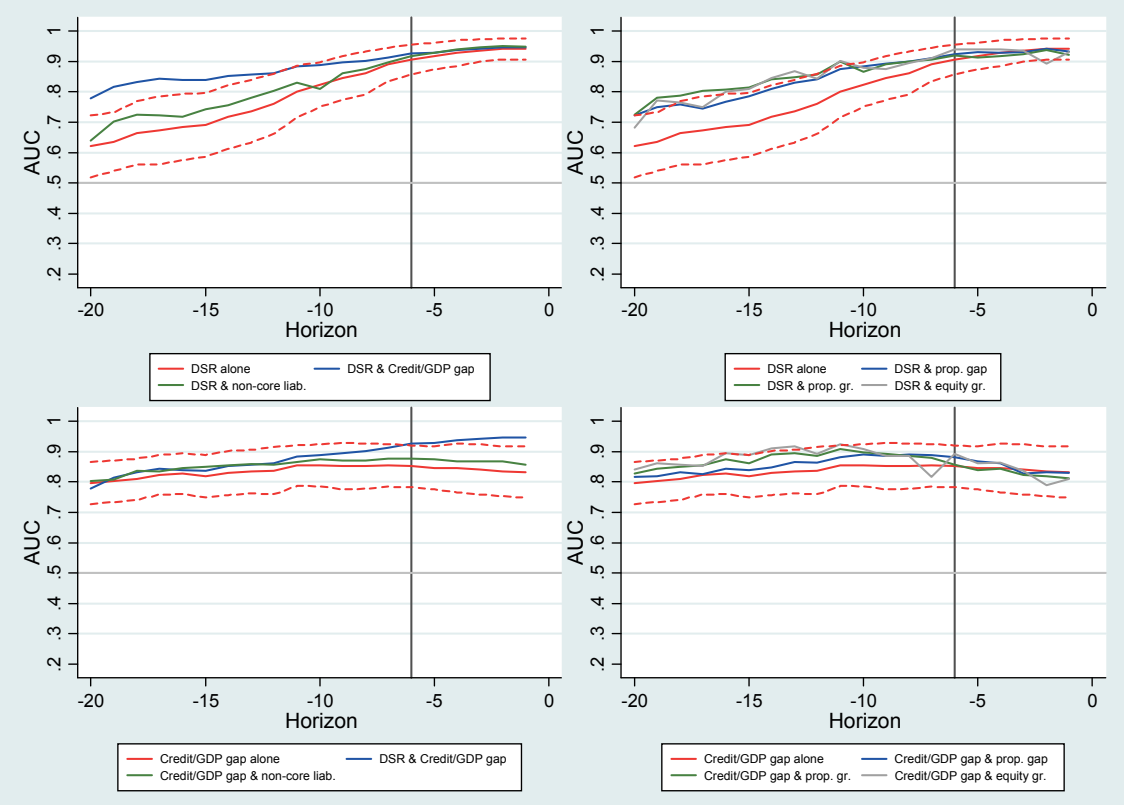
The different time profiles of the analysed EWIs suggest that there may be gains to combining them. In this section, we discuss problems related to doing this in practice. And we analyse the performance of bivariate combinations of our EWIs.

In general, there is no unique and optimal way to combine information from different indicators. For example, given two indicator variables S_i and S_j one could create a new variable $S_k = \beta_1 S_i + \beta_2 S_j$ and evaluate it using the standard approach outlined in Section 2. However, a function such as $min\{I(S_i > \theta_i), I(S_j > \theta_j)\}$, where $I(\cdot)$ is a standard indicator function, or any other function could potentially serve better or equally well as an EWI. A particularly tempting approach would be to fit a regression model of the indicator variables to the crisis date variable and use the fitted values as a combined variable. For example, Su and Liu (1993) derive an optimal linear estimator in the sense that it maximises the AUC among all possible linear combinations. This indicator coincides with Fisher's linear discriminant under some

particular assumptions. In the context of cancer screening, McIntosh and Pepe (2002) show that rules based on prediction probabilities of diseases (which compare with crises in our case) are optimal in the sense that the ROC curve is maximised at every point. A probit or logit model can, for example, be employed to estimate these probabilities.

There are at least two difficulties with applying a regression-based approach in the present context. First, the statistical properties of binary regression models are largely unknown under the high levels of persistency of our indicator variables (see Section 3.3). Second, to ensure that the evaluation is truly real-time, the estimated coefficients underlying the combined indicator at time *t* should only reflect information available at that time. For example, consider what happens if the fitted values from a full-sample regression of the crisis dates on some indicator variables and country-specific dummies are used as an indicator variable. In this case, the coefficients on the country-specific dummies reflect the number of crises that each country has in the full sample. As we show in Annex 1, the full-sample information about the number of crises by itself is already highly informative but amounts to assuming that a policy maker knows how many crises a country is going to have in the future. One solution to this problem is to do rolling regressions starting from an initial "training" sample, as in Berge and Jorda (2011) for the case of predicting business cycle turning points. However, in the present rare-event context with only 19 crises, this seems less appealing.

AUCs over time for bivariate combinations of the indicator variables — Graph 7



Note: Dotted red lines: 95% confidence intervals of AUC for the baseline estimation for the individual variable. Horizon: quarters before a crisis.

An alternative to a regression-based approach is to non-parametrically calculate the TPRs and FPRs associated with a function such as $min\{I(S_i > \theta_i), I(S_j >$

$\theta_j)\}$ for all possible combinations of θ_i and θ_j. For a given TPR, this generates a range of FPRs. The relevant ROC curve is then the one that minimises the FPRs at each point. While this approach is straightforward, it quickly becomes computationally infeasible as the number of indicator variables and the range of reasonable threshold values increase. Baker (2000) suggests several ways of reducing this computational problem in the context of general mappings from the indicator variables to regions of positive signals. Nevertheless, the problem persists.

We illustrate the effects of combining indicator variables by using the *min{·}* function defined above for all binary combinations of our indicator variables. We apply a non-parametric approach in line with Baker (2000). Not surprisingly, the highest AUCs are found when at least one of either the DSR or the credit-to-GDP gap is included. To save space, Graph 7 only reports the results from these combinations.[19] While the AUCs of the combinations are higher than or equal to the AUCs associated with the DSR or the credit-to-GDP gap individually, the increases are marginal in most cases and fall well within the confidence bands of the individual indicators. The only real exception to this pattern results from combining the DSR and the credit-to-GDP gap, in which case their complementing time profiles can be exploited. Other potentially beneficial combinations for early horizons are the DSR and asset price growth rates, but these combinations are outperformed by the combination of the DSR and the credit-to-GDP gap.

5. Conclusions

We argue that assessments of EWIs of banking crises should be based on the underlying decision problem of the policy maker. Several aspects of this problem have implications for the choice of statistical evaluation procedures. For instance, uncertainty about the costs and benefits of macroprudential policies imply that evaluations need to be robust over a wide range of the policy maker's preferences. Also, the EWI signals should have the right timing, and their quality should not deteriorate in the run-up to a crisis. We embed these requirements in a statistical evaluation procedure for EWIs.

Applying our approach to several EWIs, we find that the credit-to-GDP gap, the DSR and the non-core liability ratio satisfy the policy requirements. The first two variables consistently outperform the third, with the credit-to-GDP gap dominating at longer horizons and the DSR at shorter horizons. Our results are robust with respect to changes in both sample and crisis classification.

A distinguishing feature of our analysis is that we pay greater attention to the temporal dimension of the EWI signals. Our findings reveal that the signalling quality of different EWIs can fluctuate sharply over time. This suggests that greater reliance should be placed on EWIs that continuously perform well within the policy-relevant forecasting interval.

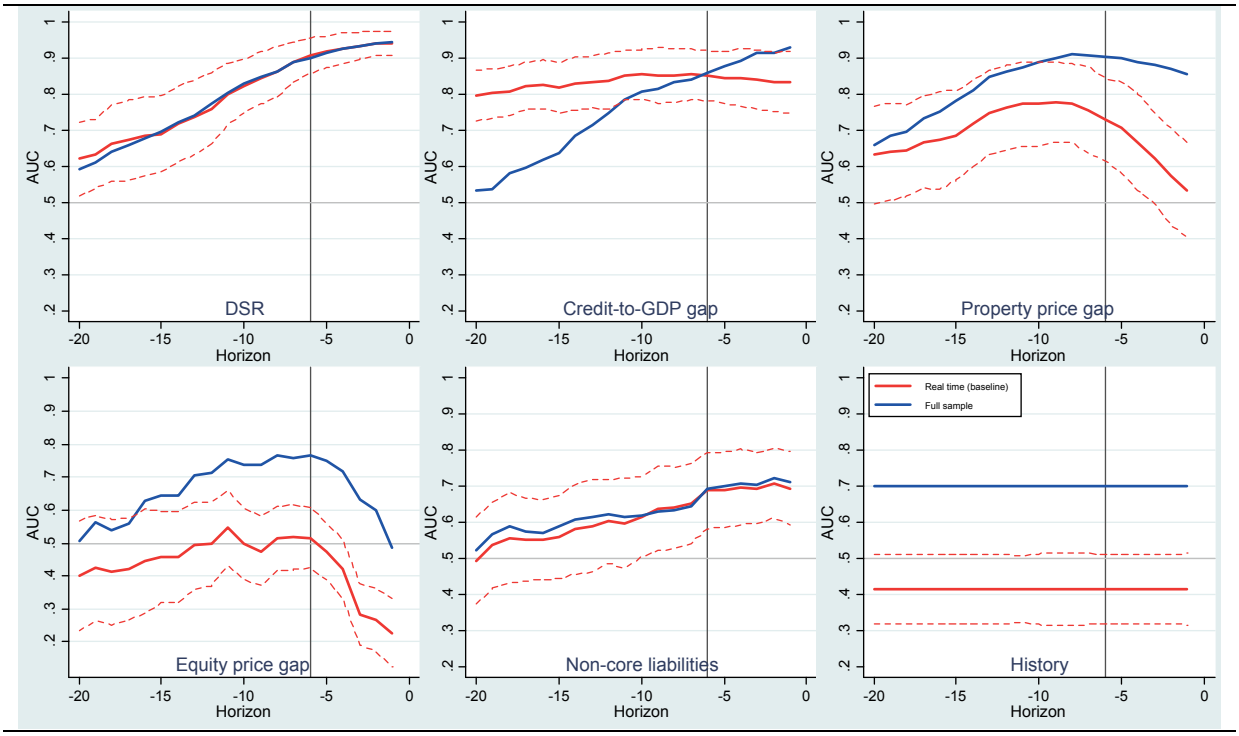
[19] All other results are available on request.

Annex 1: Real-time versus full-sample results

In this annex, we show the importance of using real-time indicators when assessing the usefulness of EWIs for policy makers. To replicate the conditions that policy makers face as closely as possible we used quasi-real-time indicators in the main text (see Section 3.3), ie indicators for which each observation is constructed using only information up to that point in time. In particular, the three gaps, the DSR, the non-core liability ratio and History were constructed in this way. In this annex, we compare our previous results with those obtained from using two-sided de-trending, ie the HP filter (with lambda 400,000) for the gaps and full-sample averages for the other variables. History now reflects the total number of crises in a country from World War II until 2012.

Graph A1 highlights that full-sample versus real-time EWIs can make substantial differences, except for the DSR and the non-core liability ratio. The gaps become more informative at shorter and less informative at longer horizons. This reflects the high emphasis the one-sided HP filter with λ=400,000 places on past growth trends, whereas the two-sided filter already 'anticipates' future changes in growth rates, which are particularly stark around crises (see Graph 2 in the main text).

Real time versus full sample — Graph A1



Note: Horizon: quarters before crises. Black vertical line: Horizon -6. Vertical grey line: 0.5. Light dashed line: 95% confidence intervals of baseline estimation.

In the full-sample case, History also becomes quite informative (AUC=0.7), which is significantly different from 0.5. This is intuitive, as knowing that a country has one crisis helps to differentiate this country from others which do not experience one. It is interesting to note that the full-sample History delivers the same AUC as the predicted values from a regression with only country-specific

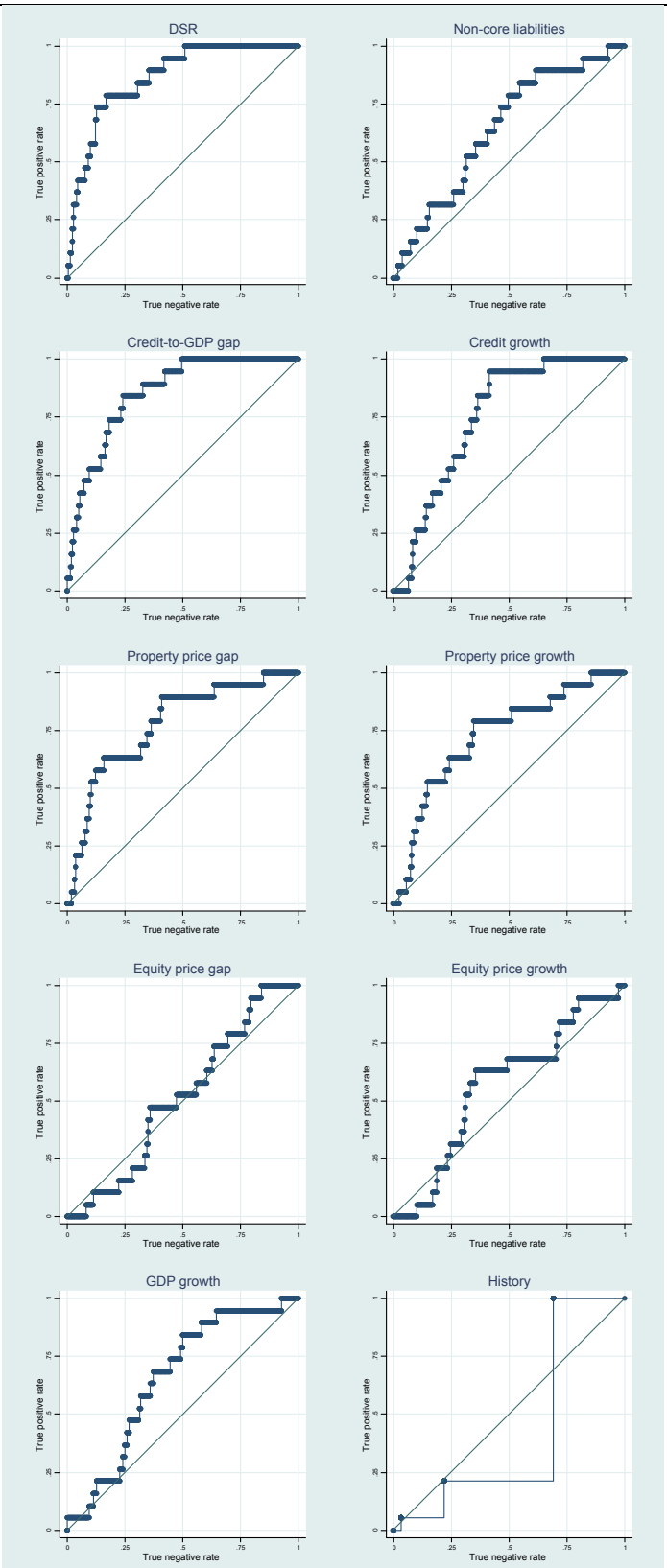
dummies.[20] This in turn highlights that researchers have to be careful how they specify potential models for EWIs.

[20] This can easily be seen for the linear case, where estimated coefficients on country-specific dummies equal the average number of crises experienced by a particular country over the sample.

Annex 2: Additional graphs and tables

ROC curves for horizon -8

Graph A2



The sample

Table A1

	Credit-to-GDP gap		Credit growth		DSR		Equity gap		Equity growth		GDP growth		Non-core liability ratio		Property price gap		Property price growth		Crisis[1]	
	start	end	start	end	start	end	start	end	start	end	start	end	start	end	start	end	start	end	systemic	additional
Australia[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	84q2	12q2	80q1	12q1	80q1	12q1		89q4, 08q4
Belgium	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q2	80q1	11q4	80q1	11q4	08q4	
Canada[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	01q3	80q1	11q2	80q1	11q2		
Czech Republic	04q4	12q2	98q2	12q2	98q2	12q2	02q2	12q2	98q2	12q2	98q2	12q1	98q2	12q2	05q1	10q4	00q1	10q4		
Denmark	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q2	80q1	11q2	80q1	11q2	08q4	87q4
Finland	80q4	12q2	80q1	12q2	80q2	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q2	80q1	12q1	80q1	12q1	91q3	
France[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	11q3	80q1	11q3	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q1	08q4	94q1
Germany[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	11q4	80q1	11q4		07q3
Greece	88q4	12q2	80q1	12q2	80q4	12q2	80q1	12q2	80q1	12q2	81q4	12q2	80q1	12q2	99q4	12q1	94q4	12q1	08q4	
Ireland	81q2	12q2	80q1	12q2	80q1	12q2	91q1	12q2	84q1	12q2	80q1	12q1	99q1	12q2	80q1	12q2	80q1	12q2	08q4	
Italy[3]	82q4	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q2	80q1	11q4	80q1	11q4	92q3, 08q4	
Japan[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	11q4	80q1	11q4	92q4	
Korea[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q2	92q1	11q2	87q1	11q2	97q3	
Malaysia	88q4	12q2	80q1	12q2	81q1	12q2	81q4	12q2	80q1	12q2	81q4	12q2	80q1	12q2	05q1	12q1	00q1	12q1	97q3[2]	
Netherlands	80q1	12q2	80q1	12q2	80q1	12q2	80q1	11q3	80q1	11q3	80q1	12q2	82q4	12q2	80q1	12q2	80q1	12q2	08q4	
New Zealand	80q1	12q1	80q1	12q1	80q1	12q1	80q1	11q3	80q1	11q3	80q1	12q1	80q1	11q2	80q1	11q4	80q1	11q4		87q1
Norway	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	06q4	80q1	12q1	80q1	12q1	90q4	
Poland	02q1	12q2	93q1	12q2	92q4	12q2	99q2	12q2	92q2	12q2	81q4	12q2	90q2	12q2	08q4	12q1	03q4	12q1		
Portugal	80q1	12q2	80q1	12q2	80q1	12q2	80q1	11q3	80q1	11q3	80q1	12q1	80q1	12q2	94q1	12q2	89q1	12q2	08q4	
South Africa[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q2	80q1	12q2	80q1	12q2		89q4
Spain	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q2	80q1	12q1	80q1	12q1	08q4	
Sweden	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q1	91q3	08q4
Switzerland	85q2	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q1	80q1	12q2	80q1	12q1	80q1	12q1	91q3	07q3
Thailand	86q1	12q2	86q1	12q2	86q1	12q2	86q1	12q2	86q1	12q2	86q1	12q2	86q1	12q2	97q1	11q1	92q1	11q1	97q3[2]	
UK[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	87q1	12q2	80q1	12q1	80q1	12q1	90q2[2], 07q3	
USA[3]	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	12q2	80q1	11q4	80q1	12q1	80q1	12q1	90q2, 07q3	

Note: [(1)] History covers all crises since WWII. In addition to the listed crises, this includes systemic crises in the UK (1973), Spain (1977), the Czech Republic (1996) and Thailand (1983) as well as additional crises in Thailand (1979) and South Africa (1977). [(2)] Not part of the balanced sample. [(3)] Part of the G20.

AUCs for different horizons

Table A2

		Horizon																			
		-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	-19	-20
Credit-to-GDP gap	**AUC**	**0.83**	**0.84**	**0.84**	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**	**0.86**	**0.85**	**0.84**	**0.83**	**0.83**	**0.82**	**0.83**	**0.82**	**0.81**	**0.80**	**0.80**
	high	0.92	0.92	0.92	0.93	0.92	0.92	0.93	0.93	0.93	0.92	0.92	0.92	0.91	0.90	0.89	0.90	0.89	0.88	0.87	0.87
	low	0.75	0.75	0.76	0.77	0.78	0.78	0.78	0.78	0.78	0.79	0.79	0.76	0.76	0.76	0.75	0.76	0.76	0.74	0.73	0.73
	Sig -6	0.57	0.54	0.62	0.78	0.55		0.79	0.96	0.98	0.79	0.89	0.54	0.35	0.27	0.12	0.40	0.35	0.19	0.15	0.17
	Sig top	0.01	0.01	0.03	0.06	0.09	0.16	0.36	0.82												
	std	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.04	0.04
Credit growth	**AUC**	**0.52**	**0.57**	**0.63**	**0.66**	**0.70**	**0.71**	**0.73**	**0.75**	**0.76**	**0.75**	**0.77**	**0.73**	**0.71**	**0.72**	**0.70**	**0.70**	**0.70**	**0.68**	**0.67**	**0.69**
	high	0.64	0.68	0.74	0.77	0.80	0.82	0.84	0.87	0.88	0.87	0.88	0.85	0.83	0.85	0.83	0.82	0.82	0.79	0.80	0.81
	low	0.41	0.46	0.52	0.56	0.60	0.61	0.63	0.64	0.64	0.64	0.65	0.61	0.58	0.59	0.57	0.57	0.59	0.56	0.55	0.57
	Sig -6	0.00	0.00	0.04	0.08	0.36		0.13	0.13	0.13	0.24	0.11	0.72	0.91	0.94	0.82	0.72	0.77	0.48	0.45	0.66
	Sig top	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.11	0.08	0.17	0.04	0.02	0.06	0.03	0.03	0.02	0.02	0.04	0.09
	std	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.07	0.06	0.06	0.06	0.07	0.06
DSR	**AUC**	**0.94**	**0.94**	**0.94**	**0.93**	**0.92**	**0.91**	**0.89**	**0.86**	**0.85**	**0.82**	**0.80**	**0.76**	**0.74**	**0.72**	**0.69**	**0.68**	**0.67**	**0.66**	**0.64**	**0.62**
	high	0.98	0.98	0.97	0.97	0.96	0.96	0.95	0.93	0.92	0.90	0.89	0.86	0.84	0.82	0.80	0.79	0.79	0.77	0.73	0.72
	low	0.91	0.91	0.90	0.88	0.87	0.86	0.84	0.79	0.77	0.75	0.71	0.66	0.63	0.61	0.59	0.58	0.56	0.56	0.54	0.52
	Sig -6	0.01	0.00	0.00	0.01	0.01		0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sig top									0.88	0.50	0.26	0.12	0.06	0.04	0.03	0.01	0.00	0.00	0.00	0.00
	std	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.06	0.05	0.05	0.05
Equity price gap	**AUC**	**0.23**	**0.27**	**0.28**	**0.42**	**0.47**	**0.52**	**0.52**	**0.51**	**0.48**	**0.50**	**0.55**	**0.50**	**0.49**	**0.46**	**0.46**	**0.45**	**0.42**	**0.41**	**0.42**	**0.40**
	high	0.33	0.36	0.38	0.51	0.56	0.61	0.61	0.61	0.58	0.61	0.66	0.62	0.62	0.60	0.60	0.60	0.57	0.57	0.58	0.57
	low	0.12	0.17	0.19	0.33	0.39	0.42	0.42	0.42	0.37	0.39	0.43	0.37	0.36	0.32	0.32	0.29	0.27	0.25	0.26	0.24
	Sig -6	0.00	0.00	0.00	0.00	0.00		0.94	0.94	0.32	0.69	0.59	0.74	0.69	0.35	0.35	0.32	0.18	0.13	0.18	0.13
	Sig top	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	std	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.06	0.06	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.09
Equity price growth	**AUC**	**0.19**	**0.23**	**0.27**	**0.41**	**0.53**	**0.56**	**0.51**	**0.57**	**0.52**	**0.61**	**0.67**	**0.62**	**0.69**	**0.63**	**0.59**	**0.62**	**0.54**	**0.60**	**0.65**	**0.57**
	high	0.30	0.33	0.39	0.53	0.65	0.68	0.63	0.68	0.62	0.71	0.75	0.73	0.79	0.73	0.70	0.72	0.64	0.72	0.78	0.69
	low	0.08	0.12	0.15	0.29	0.41	0.45	0.40	0.45	0.42	0.52	0.59	0.51	0.59	0.53	0.49	0.52	0.43	0.47	0.52	0.45
	Sig -6	0.00	0.00	0.00	0.00	0.11		0.05	0.96	0.49	0.47	0.14	0.50	0.09	0.42	0.74	0.55	0.72	0.69	0.27	0.99
	Sig top	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.00
	std	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.07	0.06
GDP growth	**AUC**	**0.25**	**0.30**	**0.40**	**0.47**	**0.55**	**0.59**	**0.64**	**0.65**	**0.65**	**0.62**	**0.59**	**0.58**	**0.55**	**0.60**	**0.60**	**0.59**	**0.60**	**0.58**	**0.61**	**0.59**
	high	0.34	0.39	0.49	0.56	0.66	0.69	0.75	0.76	0.77	0.73	0.69	0.69	0.66	0.72	0.72	0.71	0.72	0.70	0.73	0.72
	low	0.16	0.21	0.30	0.37	0.44	0.49	0.53	0.55	0.54	0.50	0.48	0.47	0.45	0.49	0.49	0.47	0.47	0.46	0.48	0.46
	Sig -6	0.00	0.00	0.00	0.00	0.08		0.11	0.05	0.09	0.58	0.90	0.83	0.40	0.85	0.85	0.98	0.94	0.87	0.83	0.95
	Sig top	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	std	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.06	0.06	0.06	0.05	0.06	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.07

AUC for different horizons (continued)

Table A2

		Horizon																			
		-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18	-19	-20
History	**AUC**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**	**0.42**
	high	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.52	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
	low	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.31	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
	Sig -6																				
	Sig top	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	std	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Non-core liability ratio	**AUC**	**0.70**	**0.71**	**0.70**	**0.70**	**0.69**	**0.69**	**0.65**	**0.64**	**0.64**	**0.62**	**0.60**	**0.60**	**0.59**	**0.58**	**0.56**	**0.55**	**0.55**	**0.56**	**0.54**	**0.50**
	high	0.80	0.81	0.79	0.80	0.80	0.79	0.76	0.75	0.76	0.73	0.72	0.72	0.72	0.71	0.68	0.66	0.67	0.68	0.66	0.62
	low	0.60	0.61	0.60	0.60	0.59	0.58	0.54	0.53	0.52	0.51	0.47	0.49	0.46	0.46	0.45	0.44	0.44	0.43	0.42	0.37
	Sig -6	0.71	0.16	0.68	0.30	0.93		0.04	0.02	0.03	0.01	0.03	0.04	0.05	0.04	0.01	0.01	0.01	0.02	0.01	0.00
	Sig top	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	std	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Property price gap	**AUC**	**0.54**	**0.57**	**0.62**	**0.67**	**0.71**	**0.73**	**0.76**	**0.78**	**0.78**	**0.77**	**0.77**	**0.76**	**0.75**	**0.72**	**0.69**	**0.67**	**0.67**	**0.65**	**0.64**	**0.63**
	high	0.67	0.71	0.75	0.80	0.83	0.85	0.88	0.88	0.89	0.89	0.89	0.88	0.87	0.84	0.81	0.81	0.79	0.77	0.78	0.77
	low	0.40	0.44	0.50	0.53	0.58	0.62	0.64	0.67	0.67	0.66	0.66	0.64	0.63	0.60	0.56	0.54	0.54	0.52	0.51	0.50
	Sig -6	0.00	0.00	0.00	0.00	0.05		0.00	0.00	0.01	0.06	0.12	0.27	0.55	0.80	0.38	0.30	0.28	0.17	0.17	0.16
	Sig top	0.00	0.00	0.00	0.00	0.01	0.01	0.07	0.26	0.22	0.15	0.17	0.26	0.19	0.09	0.05	0.02	0.03	0.02	0.02	0.02
	std	0.07	0.07	0.06	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.07
Property price growth	**AUC**	**0.28**	**0.32**	**0.35**	**0.40**	**0.49**	**0.55**	**0.64**	**0.73**	**0.75**	**0.77**	**0.80**	**0.78**	**0.77**	**0.77**	**0.72**	**0.72**	**0.71**	**0.67**	**0.68**	**0.65**
	high	0.40	0.44	0.47	0.53	0.63	0.70	0.77	0.85	0.88	0.90	0.91	0.90	0.89	0.88	0.83	0.83	0.82	0.79	0.80	0.77
	low	0.15	0.20	0.23	0.27	0.36	0.41	0.51	0.61	0.62	0.64	0.70	0.66	0.65	0.66	0.61	0.61	0.60	0.56	0.56	0.53
	Sig -6	0.00	0.00	0.00	0.00	0.01		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04	0.11	0.10	0.21
	Sig top	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.12	0.22	0.40	0.45	0.38	0.33	0.14	0.07	0.07	0.02	0.05	0.02
	std	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.06	0.07	0.07	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06

Note: High\low: upper and lower 95% confidence interval. Sig -6: Significance test whether different to AUC at horizon -6. Sig top: Significance test whether different to highest AUC at horizon h. Highest AUC for horizon -1 to -10 is DSR and from -11 to -20 the credit-to-GDP gap.

Bibliography

Andrews, D W K and H-Y Chen (1994): "Approximately median-unbiased estimation of autoregressive models", *Journal of Business and Economic Statistics,* 12, 187-204.

Baker, S (2000): "Identifying combinations of cancer markers for further study as triggers of early intervention", *Biometrics*, 56, 1082-1087.

Baker, S and B Kramer (2007): "Peirce, Youden, and Receiver Operating Characteristic Curves", *The American Statistician*, 61, 343-34.

Basel Committee on Banking Supervision (2010): *Guidance for national authorities operating the countercyclical capital buffer*.

Berge, T J and O Jorda (2011): "Evaluating the classification of economic activity into recessions and expansions", *American Economic Journal: Macroeconomics,* 3, 246-277.

Bernanke, B S (2004): "Gradualism", Remarks at an economics luncheon co-sponsored by the Federal Reserve Bank of San Francisco (Seattle Branch) and the University of Washington, Seattle, 20 May 2004.

Borio, C (2009): "Implementing the macroprudential approach to financial regulation and supervision", *Banque de France Financial Stability Review,* September.

Borio, C and M Drehmann (2009): "Assessing the risk of banking crises - revisited", *BIS Quarterly Review,* March, 29-46.

Borio, C and P Lowe (2002): "Assessing the risk of banking crises", *BIS Quarterly Review,* December, 43-54.

Borio, C and P McGuire (2004): "Twin peaks in equity and housing prices?", *BIS Quarterly Review*, 79-93.

Bussiere, M and M Fratzscher (2006): "Towards a new early warning system of financial crises", *Journal of International Money and Finance,* 25, 953-973.

Caruana, J (2010): "The challenge of taking macroprudential decisions: Who will press which button(s)?". Speech at the 13th Annual International Banking Conference, Federal Reserve Bank of Chicago, in cooperation with the International Monetary Fund, Chicago, 24 September 2010.

Cohen, J, S Garman and W Gorr (2009): "Empirical calibration of time series monitoring methods using receiver operating characteristic curves", *International Journal of Forecasting,* 25, 484-497.

Committee on the Global Financial System (CGFS) (2012): "Operationalising the selection and application of macroprudential instruments", *CGFS Publications* No. 48.

Dembiermont, C, M Drehmann, and S Muksakunratana (2013): "How much does the private sector really borrow - a new database for total credit to the private non-financial sector", *BIS Quarterly Review,* March.

Drehmann, M (2012): "How often should macroprudential tools be used?", *Mimeo*.

Drehmann, M, C Borio and K Tsatsaronis (2011): "Anchoring countercyclical capital buffers: The role of credit aggregates", *International Journal of Central Banking,* 7.

Drehmann, M and M Juselius (2012): "Do debt service costs affect macroeconomic and financial stability?", *BIS Quarterly Review,* September, 21-34.

Edge, R M and R R Meisenzahl (2011): "The unreliability of credit-to-gdp ratio gaps in real-time: Implications for countercyclical capital buffers", *International Journal of Central Banking,* 7.

Elliott, G and R P Lieli (2013): "Predicting binary outcomes", *Journal of Econometrics*, 174, 15-26.

Fernández de Lis, S and A Garcia-Herrero (2012): "Dynamic provisioning: A buffer rather than a countercyclical tool?", *BBVA Research Working Paper* 12/22.

Gönül, S, D Önkal and P Goodwin (2009): "Expectations, use and judgmental adjustment of external financial and economic forecasts: An empirical investigation", *Journal of Forecasting,* 28, 19-37.

Gorr, W and M J Schneider (2011): "Large-change forecast accuracy: Reanalysis of m3-competition data using receiver operating characteristic analysis", *International Journal of Forecasting,* 29, 274–281.

Gourinchas, P-O and M Obstfeld (2012): "Stories of the twentieth century for the twenty-first", *American Economic Journal: Macroeconomics,* 4, 226-265.

Granger, C and M Machina (2006): "Forecasting and Decision Theory," In Elliott, G, C Granger and A Timmermann (Eds.), *Handbook of Economic Forecasting*, Elsevier.

Granger, C and M H Pesaran (2000): "Economic and statistical measures of forecast accuracy", *Journal of Forecasting*, 19, 537–560.

International Monetary Fund (IMF). (2011). *Macroprudential policy: An organizing framework*.

Hahm, J-H, H S Shin and K Shin (2012): "Non-core bank liabilities and financial vulnerability", forthcoming in *Journal of Money, Credit and Banking*.

Hsieh, F and B Turnbull (1996): "Nonparametric and semiparametric estimation of the receiver operating characteristic curve", *Annals of Statistics,* 24, 25-40.

Janes, H, G Longton and M Pepe (2009): "Accommodating covariates in ROC analysis", *Stata Journal,* 9.

Jorda, O (2011): "Anchoring countercyclical capital buffers: The role of credit aggregates: Discussion", *International Journal of Central Banking,* 7, 241-259.

Jorda, O, M Schularick and A M Taylor (2011): "Financial crises, credit booms, and external imbalances: 140 years of lessons", *IMF Economic Review,* 59, 340-378.

Kaminsky, G L and C M Reinhart (1999): "The twin crises: The causes of banking and balance-of-payments problems", *American Economic Review,* 89, 473-500.

Kindleberger, C (2000): *Maniacs, panics and crashes*, Cambridge University Press, Cambridge.

Laeven, L and F Valencia (2012): "Systemic banking crises database: An update", *IMF Working Paper* WP/12/163.

Lawrence, M, P Goodwin, M O'Connor and D Onkal (2006): "Judgmental forecasting: A review of progress over the last 25years", *International Journal of Forecasting,* 22, 493-518.

McIntosh, M and M Pepe (2002): "Combining Several Screening Tests: Optimality of the Risk Score", *Biometrics*, 58, 657-664.

Minsky, H P (1982): *Can it happen again? Essays on instability and finance*, M E Sharpe, Armonk.

Önkal, D, M E Thomson and A A C Pollock (2002): "Judgmental forecasting", in Clements, M P and Hendry, D F (eds), *A companion to economic forecasting*, Blackwell Publishers, Malden and Oxford.

Orphanides, A (2003): "Monetary policy evaluation with noisy information", *Journal of Monetary Economics,* 50, 605-631.

Park, J Y and C B Phillips (2000): "Nonstationary binary choice", *Econometrica,* 68, 1249-1280.

Pepe, M, H Janes and G Longton (2009): "Estimation and comparison of receiver operating characteristic curves", *Stata Journal,* 9.

Pesaran, M H and S Skouras (2002): "Decision-based methods for forecast evaluation". In: Clements, M P and D F Hendry (Eds.), *A Companion to Economic Forecasting*, Blackwell, Malden and Oxford.

Ravn, M O and H Uhlig (2002): "On adjusting the hodrick-prescott filter for the frequency of observations", *Review of Economics and Statistics,* 84, 371-376.

Reinhart, C M and K S Rogoff (2009): *This time is different: Eight centuries of financial folly*, Princeton University Press, Princeton and Oxford.

Su, J Q and S Liu (1993): "Linear combinations of multiple diagnostic markers", *Journal of the American Statistical Association*, 88, 1350-1355.

Swets, J A and R M Picket (1982): *Evaluation of diagnostic systems: Methods from signal detection theory*, Academic Press, New York.