Kiel
Working Papers

Kiel Institute
for the World Economy

An empirical evaluation of macroeconomic surveillance in the European Union

Jens Boysen-Hogrefe, Nils Jannsen, Martin Plödt und Tim Schwarzmüller

No. 2014 | January 2016

Web: www.ifw-kiel.de

Kiel Working Paper No. 2014 | First version: December 2015. This version: January 2016

An empirical evaluation of macroeconomic surveillance in the European Union

Jens Boysen-Hogrefe, Nils Jannsen, Martin Plödt und Tim Schwarzmüller

Abstract:

The macroeconomic surveillance mechanism of the European Union, namely the Macroeconomic Imbalance Procedure (MIP), is based on the Scoreboard, which comprises of a set of indicators that serve as a signaling device for potentially harmful macroeconomic developments. We evaluate the early warning properties of the Scoreboard indicators with regard to financial crises. Thereafter, we analyze the role of emerging crisis signals from the Scoreboard for the subsequent step of the MIP, in which the gravity of macroeconomic imbalances is specified. The results of our study help to identify ways to improve the current set-up and ultimately to deliver more transparent and effective policy advice.

Keywords: Macroeconomic Imbalance Procedure, early warning indicators, signals approach, financial crises.

JEL classification: E02, E61, C25.

Jens Boysen-Hogrefe
Kiel Institute for the World Economy
24105 Kiel, Germany
Telephone: +49 431 8814 210
E-mail: jens.hogrefe@ifw-kiel.de

Martin Plödt
Kiel Institute for the World Economy
24105 Kiel, Germany
Telephone: +49 431 8814 604
E-mail: martin.ploedt@ifw-kiel.de

Nils Jannsen
Kiel Institute for the World Economy
24105 Kiel, Germany
Telephone: +49 431 8814 210
E-mail: nils.jannsen@ifw-kiel.de

Tim Schwarzmüller
Swiss National Bank
Börsenstrasse 15
CH-8022 Zürich
E-mail: Tim.Schwarzmueller@snb.ch

An empirical evaluation of macroeconomic surveillance in the European Union*

Jens Boysen-Hogrefe† Nils Jannsen† Martin Plödt†
Tim Schwarzmüller‡

This version: January 11, 2016

Abstract

The macroeconomic surveillance mechanism of the European Union, namely the Macroeconomic Imbalance Procedure (MIP), is based on the Scoreboard, which comprises of a set of indicators that serve as a signaling device for potentially harmful macroeconomic developments. We evaluate the early warning properties of the Scoreboard indicators with regard to financial crises. Thereafter, we analyze the role of emerging crisis signals from the Scoreboard for the subsequent step of the MIP, in which the gravity of macroeconomic imbalances is specified. The results of our study help to identify ways to improve the current set-up and ultimately to deliver more transparent and effective policy advice.

Key words: Macroeconomic Imbalance Procedure, early warning indicators, signals approach, financial crises.
JEL classification: E02, E61, C25.

*The paper is based on a project conducted for the Federal Ministry for Economic Affairs and Energy, Germany. The views expressed in this paper are those of the authors and do not necessarily represent those of the ministry. We thank Klaus-Jürgen Gern and Stefan Kooths for helpful comments and suggestions.

†Kiel Institute for the World Economy, Kiellinie 66, 24105 Kiel. Correspondence: *jens.hogrefe@ifw-kiel.de*, *nils.jannsen@ifw-kiel.de* and *martin.ploedt@ifw-kiel.de*.

‡Swiss National Bank, Börsenstrasse 15, 8022 Zürich. Correspondence: *Tim.Schwarzmueller@snb.ch*. The views expressed in this paper are those of the authors and do not necessarily represent those of the Swiss National Bank.

1 Introduction

Numerous international institutions such as the IMF, the European Commission, and ASEAN have engaged in macroeconomic surveillance, and, in the light of the recent crisis experiences, often call for a further strengthening of macroeconomic surveillance in the future. The Macroeconomic Imbalance Procedure (MIP) of the European Union represents one of the most prominent macroeconomic surveillance mechanisms. The MIP was implemented at the end of 2011, against the backdrop of the global financial crisis and the beginning of the European sovereign debt crisis with an aim to prevent and correct macroeconomic imbalances within the EU.[1] It is embedded in the European Semester of the EU and is carried out for all member states on an annual basis.

The MIP is based on the Scoreboard that consists of eleven macroeconomic indicators and corresponding thresholds. It serves as a signaling device for potentially harmful macroeconomic developments (European Parliament (2011)). An analysis of Scoreboard signals in the context of the Alert Mechanism Report constitutes the starting point of the MIP. Based on the eleven indicators (and on an additional set of auxiliary indicators), the Alert Mechanism Report evaluates whether there are potential macroeconomic imbalances in member states that make further analysis in an In-Depth Review warranted. The In-Depth Review consists of a detailed analysis of country-specific circumstances, and it classifies the degree of macroeconomic imbalances into six categories, from 'no imbalance' to 'excessive imbalances, which require decisive policy action and the activation of the Excessive Imbalance Procedure'.

Despite the high relevance of the Scoreboard and the MIP for macroeconomic policy in the EU, little is known about whether the Scoreboard is indeed useful for identifying macroeconomic imbalances at a reasonably early stage and how the warning signals of the Scoreboard are related to the outcomes of the MIP. This study aims at filling this gap by means of a detailed empirical evaluation of the MIP. We analyze two of the most important aspects of indicator-based macroeconomic surveillance mechanisms. In the first part of our study, we evaluate the early warning properties of the Scoreboard. In the second part of our study, we analyze the relevance of the Scoreboard for the final outcome of the MIP, that is, whether an In-Depth Review is prepared and which category a macroeconomic imbalance is eventually classified.

[1] The macroeconomic imbalances the MIP aims to prevent or correct are very generally defined as imbalances that are 'adversely affecting, or have the potential to adversely affect, the proper functioning of the economy of a Member State or of the Economic and Monetary Union, or of the Union as a whole' (European Parliament (2011)).

The empirical evaluation of the early warning properties of the Scoreboard indicators is crucial for assessing the reliability of the MIP. On one hand, the Scoreboard should identify all potentially harmful macroeconomic imbalances and, on the other hand, it should not deliver too many false warning signals. False warning signals may possibly provoke misguided policy measures, which could have high economic costs. Furthermore, the official thresholds were chosen in a rather ad-hoc manner and are mainly based on the statistical distributions of the indicators (for instance, by using the lower and/or upper quartiles of the distributions), and not on rigorous empirical evaluations or theoretical considerations.[2] Our empirical evaluation of the early warning properties, therefore, also contributes to the identification of ways to improve the current set-up.

The evaluation of the Scoreboard is complicated by the fact that the Scoreboard is not meant to provide warning signals for specific events, but for broadly defined imbalances. For an empirical evaluation, however, a specific target variable is needed, which the early warning mechanism is supposed to predict. In this study, we evaluate the early warning properties of the Scoreboard indicators with regard to financial crises, that is, banking crises, currency crises, and sovereign debt crises. We do so for three reasons. First, financial crises seem to be the most obvious economic events that are 'adversely affecting, or have the potential to adversely affect, the proper functioning of the economy of a Member State or of the Economic and Monetary Union, or of the Union as a whole' (European Parliament (2011)). Second, the detailed description of the European Commission for the choice of indicators and thresholds frequently refers to several seminal papers that are concerned with early warning mechanisms for financial crises (for instance, Frankel and Saravelos (2012)), indicating that such crises are of particular interest for the MIP (European Commission (2012)). Third, while the economic consequences of financial crises are well understood and there exist well-defined dating schemes for financial crises, the identification and the consequences of other macroeconomic imbalances are not understood as well. Additionally, it is not always obvious how other such imbalances are linked to macroeconomic risks that are relevant for the MIP.

We apply the signals approach to evaluate the early warning properties of the Scoreboard indicators. This approach corresponds to that used in the Scoreboard and it has also been widely used in the literature to evaluate early warning mechanisms or indicators. We choose a forecast horizon of 2 to 5 years for the evaluation and thus deviate from

[2]See again European Parliament (2011) for details. The authors mention that the official thresholds are generally consistent with thresholds found in the empirical literature. Such a general assessment, however, cannot substitute for rigorous empirical evaluation.

most of the literature that usually considers a forecast horizon of 0 to 3 years (see, for instance, Borio and Drehmann (2009), Duca and Peltonen (2013), and Knedlik (2014)).[3] Using a forecast horizon of 2 to 5 years allows us to account for the specific institutional framework of the MIP. This refers to publication lags, which imply that for a MIP in year t, data for the year $t-2$ are taken into consideration. From a policy perspective, it is also crucial to receive warning signals as early as possible in order to have sufficient time for putting appropriate policy measures into effect.

While early warning mechanisms or indicators for financial crises have been intensively analyzed in the literature, early warning properties of the Scoreboard indicators have only been analyzed by Knedlik (2014) so far, who put particular emphasis on the preferences of policymakers when choosing the thresholds. Moreover, Knedlik (2014) only considers sovereign debt crises, which might be insufficient for a comprehensive assessment of the early warning properties of the Scoreboard. Other crises, in particular banking crises, have proven to have severe economic consequences (Cerra and Saxena (2008)) that have the potential to adversely affect the proper functioning of the economic and monetary union. In addition, it is well-understood that sovereign debt crises are frequently preceded by banking crises, making it much more important for policy makers to receive early warning signals for banking crises than for sovereign debt crises (Reinhart and Rogoff (2011)).

In the second part of our study, we investigate how warning signals from the Scoreboard indicators are related to the subsequent steps of the MIP. In doing so, we strive for a better understanding of how the Scoreboard is used by the European Commission within the MIP.[4] We first investigate the link between warning signals from the Scoreboard and the decision regarding whether an In-Depth Review is prepared. Second, we investigate the link between warning signals from the Scoreboard and the category in which an imbalance is classified within the In-Depth Review. To investigate these links, we employ probit as well as ordered probit models comprising a single indicator or multiple indicators. In addition, we not only consider the pure (binominal) warning signals, but also the deviation of the Scoreboard indicators from their corresponding official thresholds and analyze if these deviations contain additional information for the final imbalance classification.

Our main findings are as follows. Overall, the early warning properties of the Score-

[3] One notable exception is Drehmann and Juselius (2014) who evaluate early warning indicators with a special focus on the preferences of policy makers and use a forecast horizon of 5 years.

[4] The European Commission interprets the Scoreboard indicators neither as policy targets nor as policy instruments. If, however, the Scoreboard indicators were uninformative for the final classification of macroeconomic imbalances, this would cast doubts on the consistency and the credibility of the MIP.

board indicators with regard to financial crises are modest. The indicators provide early warning signals for only a limited number of financial crises (on average, they signal less than 50 percent of all crises) and provide several incorrect signals (on average, more than 30 percent of all signals are wrong). If the indicators provide correct signals for financial crises, they provide these signals sufficiently early to allow for counteractive policy measures (three to four years). The best indicators are found to be those that are also characterized as being particularly useful in the literature on early warning mechanisms for financial crises, namely, house prices, private sector debt, and private sector credit flow. Other indicators, such as the unemployment rate or nominal unit labor costs, do not provide useful information for predicting financial crises. Deriving optimal thresholds based on a standard weighting scheme regarding the trade-off between correct signals for crisis periods and incorrect signals for non-crisis periods, we find that the early warning properties of the indicators could be considerably improved. While the largest gains can be achieved for those indicators that originally exhibited the lowest usefulness, the most useful indicators based on the official thresholds remain the most useful indicators based on the optimized thresholds (with the indicator total financial liabilities now becoming about as useful as the three above-mentioned indicators). Given the institutional framework of the MIP, it seems appropriate to put a relatively high weight on correct signals for financial crises and a relatively low weight on incorrect signals for non-crisis periods when deriving optimal thresholds. Based on this preference, our results show that the early warning properties of most Scoreboard indicators could be considerably improved by choosing somewhat more restrictive (e.g., lower positive) thresholds.

In the second part of our study, we find that only some of the indicators that have proven to be useful in providing early warning signals for financial crises actually seem to play an important role for the final imbalance classification of the MIP. Interestingly, we find the export market share to be the single most important indicator in the MIP; it is informative for both the decision on whether an In-Depth Review is prepared and for the final imbalance classification. This result is robust to changing from a single-indicator set-up to a multi-indicator set-up and to considering deviations from the threshold instead of signals only. Other important indicators are private sector debt and public debt and, to a lesser extent, the net international investment position of a country. Overall, the Scoreboard indicators do have informative content for the final outcome of the MIP. It seems, however, that in the first four years of the MIP, the European Commission has made very selective use of the Scoreboard indicators. This potentially selective use does not concur with the findings regarding the early warning properties. Indicators that exhibit relatively good early warning properties are not part of the group of indicators

that are found to be relevant for the final outcome of the MIP. In the initial years of the MIP, the European Commission has targeted imbalances other than those directly linked to emerging financial crises. This is probably because the MIP was introduced after the financial crises of the years 2008 and 2009 and, hence, was mainly concerned with addressing the macroeconomic consequences of these crises. However, since these other imbalances are not understood as well as those related to financial crises and the focus on other imbalances is not communicated, the MIP lacks transparency. Such a lack of transparency and an unclear definition of imbalances reduce the ownership of member states in the MIP. This potentially explains the very low implementation rate of country specific recommendations that resulted from the MIP (European Parliament (2014), European Parliament (2015)).

The remainder of the paper is structured as follows. In Section 2, we assess the usefulness of the Scoreboard, that is, we empirically evaluate the early warning properties of the indicators based on the official and optimal thresholds. We also conduct several robustness checks regarding the forecast horizon and the evaluation criterion. In Section 3, we assess how the Scoreboard outcomes are used within the MIP, that is, we empirically evaluate the importance of the Scoreboard indicators for the likelihood that an In-Depth Review is prepared and the final classification of a macroeconomic imbalance. Section 4 concludes.

2 Assessing the usefulness of the Scoreboard

In this Section, we assess the usefulness of the Scoreboard for providing early warning signals for macroeconomic imbalances. Our assessment is based on the signals approach and puts particular emphasis on the choice of an appropriate forecast horizon (Section 2.1). Since the Scoreboard aims at providing early warning signals for only roughly defined macroeconomic imbalances, we choose to employ a financial crisis dummy as a specific target variable for the empirical evaluation. This target variable is discussed in Section 2.2. We then evaluate the Scoreboard indicators with their official thresholds (Section 2.3) and we derive optimal thresholds based on standard assumptions about the preferences of policy makers with regard to the trade-off between correct signals for crisis periods and incorrect signals for non-crisis periods (Section 2.4). Finally, we present some robustness checks in Section 2.5.

For the empirical evaluation, we use annual data for the eleven Scoreboard indicators for the 28 Member States of the European Union that are available on the Scoreboard

data platform.[5] The eleven indicators are: current account balance, net international investment position, real effective exchange rate, export market share, nominal unit labor costs, real house prices, private sector debt, private sector credit flow, public debt, unemployment rate, and total financial liabilities. The available vintages for these indicators are unbalanced across countries and across time. The longest available vintages run from 1970 to 2012. The shortest available vintages start at 2009 (real house prices for some of the countries). The number of available observations per indicator ranges from less than 300 (real house prices) to more than 600 (current account balance).

2.1 The signals approach and the forecast horizon

We use the signals approach for our empirical evaluation of the Scoreboard indicators since it corresponds more closely to the set-up of the Scoreboard than alternative approaches such as discrete choice models. Moreover, the signals approach has been widely used in the empirical literature on early warning mechanisms for financial crises (Reinhart and Kaminsky (1999)).

The idea of the signals approach is that an indicator signals the beginning of a financial crisis in a predetermined forecast horizon whenever this indicator exceeds (or falls below) a certain threshold. Overall, there are four cases depending on whether a crisis signal or the absence of a crisis signal was correct or incorrect (Table 1). A denotes the

Table 1: Evaluation matrix for signals approach.

	Crisis occurs	No crisis occurs
Signal	A	B
No signal	C	D

number of years in which the indicator (with its corresponding threshold) correctly signals the beginning of a financial crisis (indicator signals a crisis and a crisis begins within the forecast horizon). B denotes the number of years in which the indicator provides a signal but no financial crisis begins. Correspondingly, C denotes the number of years in which the indicator provides no signal but a financial crisis begins, and D denotes the number of years in which the indicator correctly provides no signal. Setting a threshold value for an indicator variable always entails a trade-off between the number of crisis signals that eventually turn out to be correct and the number of crisis signals that do not. The lower the threshold, the more crises are correctly predicted (A), but at the same time

[5] See Table 11 in Appendix A for an overview of all indicators and corresponding thresholds. All Scoreboard data can be downloaded from the Scoreboard data platform of the European Commission. We downloaded the data in spring 2014.

the number of incorrect signals (B), and thus the number of type II errors increase.[6] In the same vein, a higher threshold increases the number of cases in which one correctly predicts that no crisis occurs (D), but it also increases the number of crises without any preceding signal (C), and thus the number of type I errors.

When evaluating indicators and the corresponding thresholds, therefore, the number of correct and incorrect signals has to be weighted according to the preferences of the policy maker. A commonly used criterion to assess the usefulness of early warning indicators is the Noise-to-Signal Ratio (NSR). The NSR is calculated as the share of incorrectly signaled crises in number of years in which no financial crisis begins, in relation to the share of correctly signaled crises in the number of years in which a financial crisis begins:

$$NSR = \frac{\frac{B}{B+D}}{\frac{A}{A+C}}. \tag{1}$$

The better is an early warning indicator, the lower the NSR. An indicator is useful in providing early warning signals if NSR < 1. If NSR $= 1$ the indicator does not provide better early warning signals than a purely random signal. However, a NSR of below one is only a very general criterion for the usefulness of an indicator. It does not necessarily imply that policy makers consider an indicator (or the corresponding threshold) useful, for instance, because the resulting number of type I and type II errors may be unacceptable to them. In general, minimizing the NSR when looking for optimal thresholds for early warning indicators usually leads to a choice of threshold values for which the number of correctly predicted crises is relatively low and for which the number of correct predictions that no crisis occurs is relatively high. However, given that financial crises can be very costly, policy makers may be more interested in early warning indicators that correctly predict a high share of financial crises, tolerating at the same time a relatively high share of wrong signals in years in which no financial crisis occurs. This may be particularly true for the MIP because its institutional setting allows for a further examination of emerging warning signals in the Alert Mechanism Reports and In-Depth Reviews (and possibly for a conclusion that no macroeconomic imbalances exist). In contrast, no further action can be taken within the MIP in case the Scoreboard does not provide any warning signals.

For these reasons, empirical evaluations of early warning indicators are frequently based on a usefulness function U, that in turn is based on a loss function L (see, e.g., Alessi and Detken (2011)). In the loss function certain weights θ and $(1-\theta)$ are explicitly

[6] This applies to indicators that provide a signal when a certain threshold value is exceeded. The opposite is true for indicators that send a signal when falling below a certain threshold.

assigned to the share of type I errors ($C/(A+C)$) and type II errors ($B/(B+D)$). The weights refer to the preferences (or the risk aversion either to type I or to type II errors) of a policy maker:

$$L = \theta \frac{C}{A+C} + (1-\theta)\frac{B}{B+D}. \tag{2}$$

The usefulness of an indicator is then given by:

$$U = \min[\theta; 1-\theta] - L. \tag{3}$$

The greater the usefulness of an early warning indicator, the higher is the value of U; an early warning indicator is generally considered useful if $U > 0$. This is because a policy maker can always realize a loss of $\min[\theta; 1-\theta]$ by disregarding the early warning indicator (that is, either by always assuming that a crisis occurs, or by assuming that a crisis never occurs). We use a value of 0.5 of θ for our empirical investigation. This value indicates a balanced risk aversion of a policy maker with regard to the share of type I and type II errors and is commonly used in the literature. In Section 2.5 we check the robustness of our results by using a higher value of θ, that is, by assuming a higher risk aversion of a policy maker for missing a financial crisis.

In addition to the choice of appropriate evaluation criteria, the choice of the forecast horizon is of particular importance for the empirical evaluation. The forecast horizon determines the period in which a financial crisis must begin (or must not begin) to classify a signal as being correct or incorrect. Given a signal in year t, empirical studies have typically considered a forecast horizon of 0 to 3 years for the occurrence or non-occurrence of a crisis; see, for instance, Borio and Drehmann (2009), Duca and Peltonen (2013), and Knedlik (2014). We deviate from this literature and set the beginning of the forecast horizon to $(t+2)$. We do so for two reasons. First, from the perspective of a policy maker, it is crucial that an early warning mechanism signals crises sufficiently early to allow for implementing appropriate policy measures (Drehmann and Juselius (2014)). Second, due to the institutional framework of the MIP and the publications lags, for the MIP in year t the Scoreboard comprises data only up to the year $(t-2)$. With regard to the end of the forecast horizon, we follow Drehmann and Juselius (2014) and choose the year $(t+5)$ since early warning mechanisms should not signal crises too early given that policy measures to respond to the signals can also be associated with significant costs, and forecast uncertainty usually increases for longer horizons. Therefore, we use a forecast horizon of $[t+2, t+5]$ in our baseline specification. Accordingly, an indicator correctly predicts a crisis when it crosses its threshold in year t, followed by a crisis dummy that has a value of one in one of the years $t+2$ to $t+5$ (see Figure 1). We also report for each

indicator the average number of periods between a crisis signal and a subsequent crisis in order to determine which indicators tend to provide the earliest warnings. Finally, to ensure that the potentially very specific behavior of the indicators in the course of a crisis does not distort our results, we follow Detragiache and Spilimbergo (2001) and others and refrain from evaluating signals in a period in which a crisis occurs, as well as in the period thereafter.

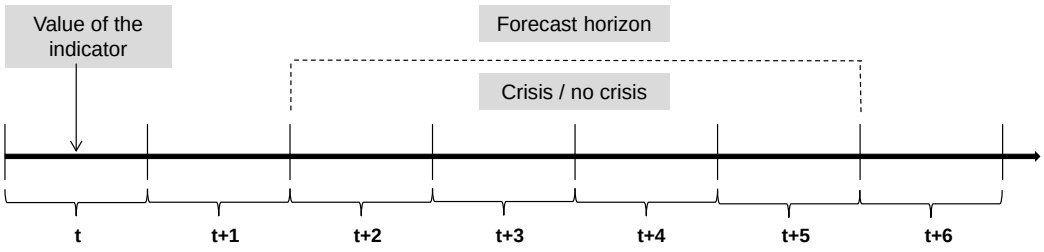

Figure 1: Forecast horizon.

2.2 Defining an economic crisis

There are various approaches in the literature for identifying and dating economic crises. In our empirical analysis, we use the comprehensive database of Laeven and Valencia (2013) that is based on the narrative approach. It comprises three types of crises during the period 1970-2012: systemic banking crises, sovereign debt crises, and currency crises. Laeven and Valencia (2013) define a systemic banking crisis based on two criteria, both of which have to be met. On the one hand, there have to be significant signs of financial distress in a banking system (losses in the banking system, bank runs, and/or bank liquidations). On the other hand, a country has to take significant policy measures in response to distress in its banking system (for instance, significant bank guarantees or liquidity support). A sovereign debt crisis is defined based on information on sovereign defaults to private creditors and debt rescheduling. A currency crisis is defined as an event of 'a nominal depreciation of the currency vis-à-vis the U.S. dollar of at least 30 percent that is also at least 10 percentage points higher than the rate of depreciation in the year before' (Laeven and Valencia (2013), p.250). In the database, the beginning and the end of the crises are defined on an annual basis. For our empirical investigation, we only use the dates for the beginning of a crisis.

Based on this approach we observe 31 banking crises, 11 currency crises, and 4 debt

crises for the EU-28 countries during the period 1970-2012 (see Figure 2 in Appendix B).[7] We merge these sets of dummies and construct a new dummy variable (labeled as 'financial crisis dummy'), which equals one if at least one of these three types of crises occurs. We use this financial crisis dummy for the empirical analysis. We focus on the financial crisis dummy instead of discriminating between different types of crises because the Scoreboard is intended to provide early warning signals for a broad set of potentially harmful crises or imbalances. Moreover, the low number of currency and particularly debt crises in our sample would not allow for a meaningful empirical evaluation of the Scoreboard indicators.

The availability of the Scoreboard data for the 11 indicators significantly varies across indicators and countries; especially for the 1970s, the Scoreboard data platform provides data only for a few indicators and countries. Consequently, not all crises can be used for the evaluation of all indicators. Overall, we can use a total of 31 financial crises for our evaluation (see Figure 3 in Appendix B). In three countries, no financial crisis took place in our evaluation period (Estonia, Malta, and Poland).

2.3 Empirical results 1: official thresholds

We first evaluate the forecasting performance of the Scoreboard indicators with their official thresholds.[8] We particularly focus on the usefulness U of an indicator (choosing θ=0.5) in our evaluation. The criterion for an indicator to be useful is $U > 0$, because usefulness $U = 0$ can always be achieved by signaling a crisis in each period or by never signaling a crisis.[9] However, we also report other relevant criteria as the percentage of correctly predicted crises and the NSR.

Overall, seven out of eleven indicators are useful in predicting financial crises according to our criterion (Table 2): current account balance, real effective exchange rate, real house prices, private sector debt, private sector credit flow, public debt, and total financial liabilities. In general, indicators that have been found in the literature to be particularly useful in forecasting financial crises exhibit relatively high values for their usefulness, namely, real house prices, private sector debt, and private sector credit flow. The real house prices indicator exhibits the highest usefulness with a value of 0.17, that is, the

[7] We additionally include dummy variables for Malta and Cyprus, which are not covered by Laeven and Valencia (2013). Crises dummies for these countries are based on the quarterly data set of Babecký et al. (2012) for the period 1970-2010. In addition, to update the database, we define a banking crisis for Cyprus in 2012.

[8] See Table 11 in Appendix A for an overview of all indicators and corresponding thresholds.

[9] Note that the choice of θ=0.5 implies that NSR < 1 whenever $U > 0$. Given θ=0.5, the maximal value of U is 0.5.

indicator is able to reduce the number of false predictions by about 17 percentage points compared to a situation in which one simply predicts in each period the occurrence (or non-occurrence) of a crisis.[10] However, in general, the usefulness of the indicators is relatively low. This is mainly due to low ability of many indicators to forecast financial crises correctly (or due to the high share of type I errors). Real house prices and private sector debt are the only indicators that provide correct early warning signals for more than 50 percent of the financial crises. In contrast, export market shares and the unemployment rate correctly signal a crisis in only 12 and 15 percent of the cases, respectively.

For all indicators the share of type II errors is significantly lower than the share of type I errors. The share of overall correct forecasts for the indicators is between 57 percent (unemployment rate and nominal unit labor costs) and about 73 percent (export market share, real house prices). The average lead of the indicators, which is for all indicators (except the export market share) between three and four years, is reasonably high.

Table 2: Evaluation based on official Scoreboard thresholds.

Indicator	NSR	Usefulness $\theta = 0.5$	type I error (in %)	type II error (in %)	correct forecasts (in %)	correct crisis signals (in %)	identified crisis (in %)	average lead (years)
CA	0.77	0.05	54.22	35.22	61.88	45.78	51.85	3.16
$NIIP$	1.16	-0.02	75.29	28.54	63.76	24.71	25.93	3.20
$REER$	0.78	0.05	53.42	36.39	60.41	46.58	61.90	4.64
EMS	1.16	-0.01	88.00	13.95	72.79	12.00	29.17	2.00
$NULC$	1.08	-0.01	65.38	37.22	57.11	34.62	52.17	3.88
HP	0.38	0.17	44.44	21.38	72.36	55.56	61.11	3.83
PSD	0.70	0.08	46.67	37.36	60.99	53.33	52.00	4.00
$PSCF$	0.60	0.09	56.16	26.45	68.35	43.84	52.00	3.44
PD	0.81	0.04	58.67	33.53	62.00	41.33	41.67	4.40
UR	2.23	-0.09	84.93	33.55	56.74	15.07	20.00	3.33
TFL	0.73	0.04	67.14	24.00	68.35	32.86	50.00	3.46

Notes: NSR = B/(B+D)/(A/(A+C)); Usefulness = min[θ; (1-θ)]-Loss; Loss = θC/(A+C)+(1-θ)B/(B+D); Type I error = C/(A+C); Type II error = B/(B+D); Share of correct forecasts = (A+D)/(A+B+C+D); Share of correct crisis signals = A/(A+C).
CA = current account balance, NIIP = net international investment position, REER = real effective exchange rate, EMS = export market share, NULC = nominal unit labor costs, HP = real house prices, PSD = private sector debt, PSCF = private sector credit flow, PD = public debt, UR = unemployment rate, TFL = total financial liabilities.

In line with our results, Knedlik (2014) concludes that the predictive power of the Scoreboard is low. However, the results of Knedlik (2014) for individual indicators differ significantly from our results. While he finds a positive usefulness of the unemployment

[10] When interpreting these results it is important to note that for the real house prices indicator, the lowest number of observations are available, which may be advantageous in the empirical evaluation.

rate and a poor performance of real house prices and private sector credit flow, we find the latter two variables to be among the most useful ones, and the unemployment rate to be a poor performer. The discrepancy in results seems to be primarily driven by differing crisis definitions. The definition chosen by Knedlik (2014) is based on long-term government bond spreads and includes only a small number of crisis events, which occurred mainly in Central and Eastern Europe. In contrast, the differing forecast horizon does not seem to play a significant role. As discussed later in section 2.5, the relative usefulness of the indicators is only mildly affected by choosing, for instance, an alternative forecast horizon of $[t, t+3]$, which includes contemporaneous signals and is more similar to Knedlik (2014).

One important reason for the relatively poor predictive power of the Scoreboard indicators is that the indicative thresholds are based on the statistical distributions of the indicators, but are not the results of a specific empirical evaluation. In the next section, we explore the extent to which the predictive power of the Scoreboard can be improved by choosing alternative thresholds that are based on an empirical optimization.

2.4 Empirical results 2: optimized thresholds

We define an optimal threshold as the threshold that maximizes the usefulness of an indicator or minimizes the NSR of an indicator. We derive the optimal threshold for each indicator by a grid search over the interval between the 5th and the 95th percentile of the distribution of all observations of this indicator. For indicators with both, an upper and a lower threshold (current account balance and real effective exchange rate), we add the condition that one of the thresholds has to be negative while the other has to be positive.

Table 3 shows the optimized thresholds for each indicator that are obtained from an optimization based on the usefulness and an optimization based on the NSR. For the purpose of comparison, Table 3 also shows the official Scoreboard thresholds. Table 4 and Table 5 summarize all evaluation results using the optimized thresholds based on the usefulness and the NSR, respectively.

In general, using optimized instead of official thresholds leads to a noticeable improvement of the forecasting performance of many indicators. This holds especially true for the optimized thresholds based on the usefulness criterion. Comparing Table 4 with Table 2 reveals that using optimized thresholds in many cases markedly increases the share of correct crisis signals. With respect to specific indicators, some findings are worth noting. The four most useful indicators are real house prices, private sector debt, private sector credit flow, and total financial liabilities. These indicators have already been among the most useful based on the official Scoreboard thresholds, and are also often emphasized as meaningful early warning indicators in the theoretical literature (see, for instance, Frankel

Table 3: Optimized thresholds.

Indicator	SB thresholds		optimized thresholds			
			NSR		Usefulness	
	lower	upper	lower	upper	lower	upper
CA	-4.0	6.0	-10.7	6.6	-6.3	2.9
$NIIP$	-35.0		-77.3		-57.5	
$REER$	-5.0	5.0	-7.8	3.1	-7.8	0.3
EMS	-6.0		-2.6		6.6	
$NULC$		9.0		4.7		1.8
HP		6.0		8.9		5.3
PSD		133.0		185.7		84.3
$PSCF$		14.0		24.6		8.7
PD		60.0		86.6		58.5
UR		10.0		4.0		4.0
TFL		16.5		14.9		8.2

Notes: CA = current account balance, NIIP = net international investment position, REER = real effective exchange rate, EMS = export market share, NULC = nominal unit labor costs, HP = real house prices, PSD = private sector debt, PSCF = private sector credit flow, PD = public debt, UR = unemployment rate, TFL = total financial liabilities.

and Saravelos (2012) or Borio and Drehmann (2009)). While a further improvement of the usefulness of real house prices is only possible to a limited extent, the usefulness of private sector debt, private sector credit flow, and total financial liabilities can be significantly increased by choosing thresholds that are lower than the official thresholds. The real effective exchange rate and the export market share experience the biggest improvement in usefulness due to the optimization process. However, this improvement comes along with a strong increase in type II errors, and with thresholds that seem less plausible from a theoretical point of view (upper threshold of almost zero for the real effective exchange rate and a lower threshold of +6.6 percent for the export market share; see Table 3). The unemployment rate is the only indicator that is not useful as an early earning indicator for financial crises even with an optimized threshold.

As expected, the optimization based on the NSR results in upper thresholds that are always higher than the upper thresholds derived from the optimization based on the usefulness, and in lower thresholds that are always lower than the lower thresholds derived from the optimization based on the usefulness. Often, these upper (lower) thresholds are also higher (lower) than the official ones. Accordingly, the percentage rate of type II errors indeed decreases for more than half of the indicators but at the same time the percentage rate of type I errors increases. Thus, the percentage rate of correct crisis signals decreases.

As mentioned before, given that the Scoreboard's central task is to correctly predict at least a certain share of crises, thresholds derived from an optimization based on the usefulness appear more reasonable.

Table 4: Evaluation based on optimized thresholds: Usefulness

Indicator	NSR	Usefulness $\theta = 0.5$	type I error (in %)	type II error (in %)	correct forecasts (in %)	correct crisis signals (in %)	identified crisis (in %)	average lead (years)
CA	0.56	0.12	45.78	30.43	67.22	54.22	51.85	3.76
$NIIP$	0.80	0.02	83.53	13.23	75.19	16.47	18.52	3.29
$REER$	0.75	0.10	24.66	56.33	49.61	75.34	80.95	4.19
EMS	0.78	0.11	4.00	74.71	37.95	96.00	83.33	4.26
$NULC$	0.93	0.03	10.26	83.50	31.27	89.74	91.30	4.57
HP	0.40	0.18	40.74	23.45	71.86	59.26	66.67	3.85
PSD	0.71	0.13	12.00	62.64	46.34	88.00	80.00	4.45
$PSCF$	0.63	0.14	24.66	47.38	56.59	75.34	76.00	3.90
PD	0.78	0.05	52.00	37.28	60.10	48.00	41.67	4.55
UR	0.99	0.00	4.11	95.21	22.02	95.89	95.00	4.65
TFL	0.68	0.14	15.71	56.92	50.38	84.29	81.82	4.14

Notes: NSR = B/(B+D)/(A/(A+C)); Usefulness = min[θ; (1-θ)]-Loss; Loss = θC/(A+C)+(1-θ)B/(B+D); Type I error = C/(A+C); Type II error = B/(B+D); Share of correct forecasts = (A+D)/(A+B+C+D); Share of correct crisis signals = A/(A+C).
CA = current account balance, NIIP = net international investment position, REER = real effective exchange rate, EMS = export market share, NULC = nominal unit labor costs, HP = real house prices, PSD = private sector debt, PSCF = private sector credit flow, PD = public debt, UR = unemployment rate, TFL = total financial liabilities.

2.5 Robustness checks

In this section, we provide several robustness checks with regard to our baseline results. First, we change the forecast horizon to a period that is more frequently used in the literature. Second, we re-run our empirical evaluation assuming that the policy maker has a higher preference for correctly predicting financial crises than in our baseline specification (that is, we choose a higher weight for type I errors in the loss function).

When using a forecast horizon of $[t, t+3]$, which is frequently used in the literature, the results are similar to our baseline results (Table 12, Appendix C). In contrast to the baseline results, private sector credit flow now exhibits the highest usefulness of all indicators. Moreover, the real effective exchange rate now reveals a strongly negative value of U, which is driven by a markedly higher share of type I errors.

Table 13 shows the evaluation results when keeping the forecast horizon $[t+2, t+5]$ but assuming that the policy maker has a higher preference for correctly predicting financial crises (choosing a weight of $\theta = 0.7$ in the loss function). Interestingly, based on the official

Table 5: Evaluation based on optimized thresholds: NSR

Indicator	NSR	Usefulness $\theta = 0.5$	type I error (in %)	type II error (in %)	correct forecasts (in %)	correct crisis signals (in %)	identified crisis (in %)	average lead (years)
CA	0.41	0.05	81.93	7.39	81.22	18.07	29.63	2.45
$NIIP$	0.71	0.01	90.59	6.73	79.46	9.41	14.81	2.67
$REER$	0.73	0.08	38.36	45.25	56.04	61.64	76.19	4.32
EMS	0.75	0.05	57.33	31.98	63.48	42.67	62.50	3.11
$NULC$	0.92	0.03	24.36	69.90	39.28	75.64	82.61	4.18
HP	0.32	0.13	62.96	11.72	74.37	37.04	55.56	3.27
PSD	0.38	0.10	68.00	12.07	78.01	32.00	32.00	3.40
$PSCF$	0.27	0.06	83.56	4.36	81.77	16.44	20.00	2.50
PD	0.62	0.03	82.67	10.69	76.48	17.33	16.67	4.00
UR	0.99	0.00	4.11	95.21	22.02	95.89	95.00	4.65
TFL	0.64	0.08	55.71	28.31	66.84	44.29	68.18	3.39

Notes: NSR = B/(B+D)/(A/(A+C)); Usefulness = min[θ; (1-θ)]-Loss; Loss = θC/(A+C)+(1-θ)B/(B+D); Type I error = C/(A+C); Type II error = B/(B+D); Share of correct forecasts = (A+D)/(A+B+C+D); Share of correct crisis signals = A/(A+C).
CA = current account balance, NIIP = net international investment position, REER = real effective exchange rate, EMS = export market share, NULC = nominal unit labor costs, HP = real house prices, PSD = private sector debt, PSCF = private sector credit flow, PD = public debt, UR = unemployment rate, TFL = total financial liabilities.

Scoreboard thresholds, we now obtain a negative usefulness for all indicators. Hence, the current Scoreboard thresholds might not be in line with a notably stronger preference to avoid type I errors. Given the weight of $\theta = 0.7$, we additionally derive the thresholds that maximize the usefulness for each indicator (Table 14 and Table 15). In this case, the different trade-off between missing a crisis and too much noise becomes evident. Based on the optimized thresholds all indicators would now provide correct early warning signals for almost all financial crises. Accordingly, the share of type I errors is close to zero, whereas the share of type II errors increases tremendously; this results in a usefulness that is close to zero for many indicators and considerably lower as compared to our baseline results. House prices and private sector debt are still the most useful indicators.

3 Assessing the relevance of the Scoreboard for the final outcome of the MIP

In this section, we empirically assess the relevance of the Scoreboard indicators for the final outcome of the MIP. The question of how the European Commission deals with signals of the Scoreboard indicators is particularly relevant given the multitude of indicators and crisis signals. In fact, there was at least one signal from a Scoreboard indicator for each

country in all first four MIPs; on average there were about three signals from Scoreboard indicators per country and year. Directly linking the fact that one Scoreboard indicator provided a signal with a clear crisis indication for the respective country would therefore render the Scoreboard approach somewhat useless. Hence, the European Commission may weight signals from the Scoreboard in some specific way before taking any further actions like In-Depth Reviews.[11]

To further assess the use of the Scoreboard results in the MIP, we proceed in two steps. These two steps arise naturally from the way the European Commission proceeds; it first decides whether an In-Depth Review needs to be prepared and, if so, how the degree of the macroeconomic imbalance can be classified. Accordingly, we first empirically analyze the role of the Scoreboard results for the likelihood that an In-Depth Review is prepared (Section 3.1). Second, we analyze the role of the Scoreboard indicators for the final classification of imbalances in an In-Depth Review (Section 3.2). In particular, we analyze what Scoreboard indicators are most important for the final outcome of the MIP and whether deviations of the indicator values from their official thresholds ultimately affect the final outcome of the MIP.

It is worthwhile to keep in mind that in the current framework there is no direct link from the results of the Scoreboard to the final outcome of the MIP. However, the decision on whether or not an In-Depth-Review is prepared relies on an economic reading of the Scoreboard indicators. Thus, it would cast serious doubts on the consistency and the credibility of the MIP if its outcome were completely unrelated to the Scoreboard results. For the economic reading of the Scoreboard indicators and the final assessment of whether macroeconomic imbalances exist, the European Commission uses additional information besides that provided by the Scoreboard. There is little knowledge about the 'true model' of the behavior of the European Commission. We argue, however, that this additional information has 'news' characteristics as compared to the Scoreboard results. Accordingly, we regard the danger of omitted variable bias in our analysis as being low.

Our analysis covers the first four MIPs that have been conducted between 2011 and 2013. Unfortunately, the number of observations that we can use for our analysis is restricted because 'programme countries' are not covered by the MIP. In total, 98 observations are available, that is, 98 decisions were made on whether or not an In-Depth Review was warranted (Table 6). In 58 out of these 98 cases, an In-Depth Review was

[11] The European Commission itself mentions that at least unemployment represents an indicator that is less likely to provide satisfying early warning properties. Instead, it should 'be read in conjunction with forward-looking scoreboard indicators' (European Commission (2012), p. 23).

prepared. However, in 38 of the 58 cases the decision that an In-Depth Review was warranted was not the outcome of the Alert Mechanism Report, but was mandatory due to the fact that macroeconomic imbalances had been identified in the previous MIP. More precisely, the European Commission conducted 12 In-Depth Reviews in 2011/12, 13 In-Depth Reviews in 2012/13 (11 of them mandatory), 17 In-Depth Reviews in 2013/14 (13 of them mandatory), and 16 In-Depth Reviews in 2014/15 (14 of them mandatory). The total number of identified macroeconomic imbalances and the number of 'excessive imbalances' have increased over time. In 2014/2015, the European Commission identified macroeconomic imbalances in 16 countries; five of these imbalances were classified as 'excessive'. Interestingly, more than 50 percent of the countries that are monitored in the MIP ultimately exhibit some form of macroeconomic imbalance according to the MIP.

Table 6: Signals, In-Depth-Reviews, and Imbalances

year	# countries monitored	# signals	# in-depth reviews	# mandatory reviews	# (1)	# (2)	# (3)	# (4)
2011/12	24	77	12	0	6	4	2	0
2012/13	23	63	13	11	11	0	0	2
2013/14	25	76	17	13	11	0	0	3
2014/15	26	80	16	14	6	2	3	5
Σ	98	296	58	38	34	6	5	10

Notes: (1): 'imbalances which require monitoring and policy action', (2): 'imbalances which require decisive policy action and monitoring', (3): 'imbalances which require decisive policy action and specific monitoring', (4): 'excessive imbalances'.

As 38 In-Depth Reviews were mandatory, we can only use 60 of the 98 observations for our empirical analysis of the role of the Scoreboard indicators for the likelihood that an In-Depth Review is prepared (Section 3.1). We can use all 98 observations for our empirical analysis of the role of the Scoreboard indicators for the final classification of imbalances in the In-Depth Review (Section 3.2). When no In-Depth Review was prepared, we set the respective observation equal to an In-Depth Review that does not identify any macroeconomic imbalance (that is, an In-Depth Review with the outcome 'no imbalances').

Given the relatively low number of observations, our analysis should be regarded as a first assessment of the role of the Scoreboard results for the final outcome of the MIP. We believe that such a first assessment is tremendously important since the MIP is one of the most comprehensive macroeconomic surveillance mechanisms globally. Moreover, the outcome of the MIP can also lead to far-reaching policy recommendations. There-

fore, it is crucial to identify any inconsistencies and ways to improve the procedure as early as possible in order to avoid potentially large costs caused by inappropriate policy recommendations.

3.1 Signals and In-Depth Reviews

In this subsection, we analyze the role of the Scoreboard indicators for the likelihood that an In-Depth Review is prepared. We only focus on nine of the eleven indicators since an insufficient number of signals are available for private sector credit flow and total financial liabilities. Our analysis is based on a simple probit model. We assume that a latent variable Y_{it}^* is determined by

$$Y_{it}^* = \alpha + X_{it}\beta + e_{it}, \quad e_{it} \sim N(0,1), \tag{4}$$

and the observed variable Y_{it} takes on the value 1 (In-Depth Review) or 0 (no In-Depth Review) according to

$$Y_{it} = \begin{cases} 0, & \text{if } Y_{it}^* < 0, \\ 1, & \text{else.} \end{cases} \tag{5}$$

The regressors X_{it} are based on the Scoreboard results. We employ dummy variables that take a value of 1 if an indicator exceeds the official threshold ('signal') and 0 otherwise. In some specifications, we also include the absolute deviation from the threshold as regressor in order to analyze whether the strength of a signal is relevant.[12] We run regressions based on the information content of a single indicator (that is, we include either one signal dummy, the measured deviation of one indicator variable from its threshold, or both) as well as regressions based on the information content of all indicators jointly. For the latter case of 'multi-indicator' regressions, we also employ a general-to-specific approach to obtain a more parsimonious model. The general-to-specific approach rests either on the Akaike-Information-Criterion (AIC) or on the Bayesian-Information-Criterion (BIC). In each step of the general-to-specific approach, we drop the indicator with the highest p-value as long as all remaining indicators exhibit a p-value of at least 0.1.

The results for the 'single-indicator' regressions (including the signal and/or the deviation of one indicator) show that signals of only few indicators are frequently followed by In-Depth Reviews (Table 7). We reject the null hypothesis that the coefficient estimate of the signal dummy is zero for the export market share, private sector debt and public debt at common significance levels. The coefficient estimates of signals from the

[12] For those indicators that have two thresholds, we calculate the absolute deviation from the threshold that is closer to the actual value of the indicator.

net international investment position and the unemployment rate are also statistically distinguishable from zero. However, the coefficient estimates exhibit the wrong sign with regard to the intention of the Scoreboard, that is, the likelihood that an In-Depth Review is prepared is higher when these two indicators do not provide a signal. Next, we include the deviation from the threshold as a second explanatory variable for each indicator and test the null hypothesis that the coefficients for both explanatory variables are zero. We again reject the null hypothesis for the export market share, private sector debt, and public debt. Moreover, we now find evidence that the information content of the current account is also associated with the decision whether an In-Depth Review is prepared at the 10% level of significance. When only considering deviations (the last line of Table 7), we find that the coefficient estimate for private sector debt is not statistically distinguishable from zero while the coefficient for the current account is statistically distinguishable from zero even at the 5% level of significance. This also holds for the net international investment position, and its coefficient also exhibits the correct sign. It seems that the net international investment position might have some informative value for the MIP, but an emerging signal from this indicator per se might not.

Overall, based on the 'single-indicator' regressions, the export market share of an economy exhibits the closest link to the realization of an In-Depth Review.

Table 7: Drivers of In-Depth-Reviews: Single-indicator models

	CA	*NIIP*	*REER*	*EMS*	*NULC*	*HP*	*PSD*	*PD*	*UR*
Number of signals	19	36	6	24	14	5	24	21	20
In-Depth-Reviews	8	9	3	14	3	2	13	12	4
Signal (p-value)	0.443	0.034	0.505	0.000	0.577	0.291	0.006	0.009	0.082
Wrong sign		yes			yes				yes
Signal and deviation (p-value)	0.055	0.003	0.222	0.000	0.181	0.927	0.012	0.004	0.176
Deviation (p-value)	0.027	0.003	0.137	0.039	0.139	0.807	0.289	0.075	0.330

Notes: Signal: Test of the null hypothesis that the coefficient for signal is zero (signal only regressor). Signal and deviation: Test of the null hypothesis that the coefficients for signal and deviation from threshold are zero. Deviation: Test of the null hypothesis that the coefficient for deviation from threshold is zero.

Results for the 'multi-indicator' probit regressions are shown in Table 8. In the first specification (I), we jointly include the signal dummies of those nine indicators for which the number of signals is sufficiently high. Interestingly, the export market share and the net international investment position (with the correct sign) exhibit by far the highest marginal effects, and these effects are statistically distinguishable from zero. When we

add deviations from thresholds as additional explanatory variables, the marginal effects of most signals are lower and none of the signals is statistically significant anymore (see specification II).[13] The only significant variable is the deviation from the threshold of the export market share variable.

Finally, we run a general-to-specific approach using the AIC or the BIC, respectively. When using the AIC (specification III) five variables remain in the model: the signal from the net international investment position and the deviations of the current account, export market share, private sector debt, and public debt. The deviation of the export market share has the highest marginal effect relative to its standard deviation. When using the BIC (specification IV) the signal from net international investment position drops out while the four deviations from specification III remain in the model.

The 'parsimonious' models resulting from the general-to-specific approach are only able to predict 12 (AIC) or 13 (BIC) of the 20 In-Depth Reviews that were not mandatory.[14] For both models, the number of correct predictions of decisions of the European Commission with 49 of 60 (non-mandatory) is reasonably high.

3.2 Signals and imbalance diagnosis

The In-Depth Reviews of the European Commission include a final classification of the macroeconomic imbalances. In this section, we analyze the role of the Scoreboard indicators for this final classification of imbalances. The European Commission differentiates between six categories: 'no imbalances' (0), 'imbalances which require monitoring and policy action' (1), 'imbalances which require decisive policy action and monitoring' (2), 'imbalances which require decisive policy action and specific monitoring' (3), 'excessive imbalances' (4), as well as 'excessive imbalances which require decisive policy action and the activation of the Excessive Imbalance Procedure' (5). We do not consider the last category in the subsequent analysis, for the simple reason that no imbalance so far has ever been classified into this category.

We use an ordered probit model to estimate the impact of signals and deviations from thresholds of the Scoreboard indicators on the final classification of imbalances because

[13] Note that we skipped the intensity of house prices here. Due to the low number of signals from this indicator, the regression would suffer from severe multicollinearity.

[14] These numbers are based on the notion that the models predict an In-Depth Review when the models signal a likelihood of above 50 percent.

Table 8: Drivers of In-Depth-Reviews: Multi-indicator models

	(I)		(II)		(III)		(IV)		std
	ME	p	ME	p	ME	p	ME	p	
CA	0.074	0.532	-0.055	0.715					0.444
NIIP	0.425	0.022	0.357	0.257	0.235	0.139			0.502
REER	0.086	0.624	-0.110	0.764					0.329
EMS	0.345	0.010	-0.034	0.859					0.497
NULC	-0.026	0.855	-0.059	0.682					0.389
HP	0.072	0.701							
PSD	0.235	0.047	0.085	0.663					0.503
PD	0.220	0.093	0.025	0.902					0.502
UR	-0.193	0.204	-0.449	0.315					0.467
dCA			0.090	0.472	0.089	0.099	0.102	0.056	1.236
dNIIP			0.001	0.870					23.422
dREER			0.078	0.422					1.466
dEMS			0.029	0.043	0.027	0.001	0.021	0.002	6.510
dNULC			-0.005	0.897					2.150
dPSD			0.002	0.139	0.002	0.063	0.001	0.243	51.834
dPD			0.010	0.337	0.013	0.047	0.009	0.016	18.292
dUR			0.038	0.633					
LL	-27.567		-20.035		-21.920		-23.129		
AIC	75.134		74.069		55.840		56.259		
BIC	96.077		109.673		68.406		66.730		

Notes: (I): Signals only. (II): Signals and deviations (w/o HP). (III): General-to-specific based on II (AIC). (IV): General-to-specific based on II (BIC). ME: Marginal effect. p: p-value for z-statistic. std: standard deviation of regressor.

the final classifications are ordered. The ordered probit model is given by

$$Y_{it} = \begin{cases} 0, & \text{if } Y_{it}^* < 0, \\ 1, & \text{if } 0 \leq Y_{it}^* < c_1, \\ 2, & \text{if } c_1 \leq Y_{it}^* < c_2, \\ 3, & \text{if } c_2 \leq Y_{it}^* < c_3, \\ 4, & \text{if } c_3 \leq Y_{it}^*, \end{cases} \tag{6}$$

where the latent part again reads as

$$Y_{it}^* = \alpha + X_{it}\beta + e_{it}, \quad e_{it} \sim N(0,1). \tag{7}$$

When no In-Depth Review was prepared, we set this case equal to an In-Depth Review with the classification 'no imbalances' (0). In sum, we have 98 observations: 43 times 'no imbalances', 34 times 'imbalances which require monitoring and policy action', 6 times 'imbalances which require decisive policy action and monitoring', 5 times 'imbalances which require decisive policy action and specific monitoring' and 10 times 'excessive im-

balances' (see also Table 6).

Table 9 shows the results for the 'single-indicator' ordered probit models. When including only the signal dummy as explanatory variable, signals from the export market share and public debt turn out to be informative for the final classification of the macroeconomic imbalances (Table 9, first line). When we also include the deviation from the threshold in our model, the results do not change much (third line). Again, the export market share and public debt appear to be informative for the final classification of the macroeconomic imbalances. However, when including deviations from thresholds as well, the net international investment position indicator is also informative. Overall, the deviation from the threshold seems to be of little relevance. Only the deviations of the current account and net international investment position are statistically significant at the 10% level of significance (fourth line).

Table 9: Drivers of imbalance classifications: Single-indicator models

	CA	*NIIP*	*REER*	*EMS*	*NULC*	*HP*	*PSD*	*PD*	*UR*
Signal (p-value)	0.461	0.776	0.940	0.000	0.176	0.264	0.117	0.000	0.642
Wrong sign	yes	yes			yes	yes			yes
Signal and deviation (p-value)	0.189	0.007	0.976	0.000	0.315	0.527	0.277	0.000	0.575
Deviation (p-value)	0.061	0.000	0.834	0.349	0.603	0.344	0.302	0.106	0.278

Notes: Signal: Test of the null hypothesis that the coefficient for signal is zero (signal only regressor). Signal and deviation: Test of the null hypothesis that the coefficients for signal and deviation from threshold are zero. Deviation: Test of the null hypothesis that the coefficient for deviation from threshold is zero.

Table 10 shows the results for the 'multi-indicator' ordered probit models. We first consider the results when including signal dummies only (specification I).[15] Altogether, the results are similar to the analysis of the drivers of In-Depth Reviews in the previous section. Signals arising from the export market share, public debt and net international investment position have an economically and statistically significant effect on the final classification of the macroeconomic imbalances. When we additionally include the deviations from official thresholds in the ordered probit model (specification II), the results are basically unchanged. However, for the public debt variable, the deviation from the threshold matters for the imbalance classification instead of the signal.

Finally, we employ a general-to-specific approach again. The parsimonious model based on the AIC (specification III) includes some additional variables, namely, the sig-

[15] Again, we have to omit the house prices indicator due to multicollinearity problems.

nal emerging from nominal unit labor costs and the deviation of the net international investment position, the real effective exchange rate, and the unemployment rate. In contrast, the parsimonious model based on the BIC (specification IV) includes exactly those variables that were found to be statistically significant in specification II: signals emerging from the export market share and net international investment position, as well as the deviation of public debt from its threshold.

Table 10: Drivers of imbalance classifications: Multi-indicator model

	(I)		(II)		(III)		(IV)	
	coeff	p	coeff	p	coeff	p	coeff	p
CA	-0.162	0.579	0.357	0.400				
NIIP	1.399	0.001	1.752	0.012	1.314	0.010	1.456	0.000
REER	-0.302	0.416	-0.186	0.766				
EMS	1.974	0.000	2.306	0.000	2.538	0.000	2.008	0.000
NULC	0.603	0.109	0.498	0.278	0.769	0.043		
PSD	0.238	0.396	0.613	0.154				
PD	0.761	0.012	-0.284	0.557				
UR	0.018	0.960	-0.069	0.893				
dCA			0.026	0.874				
dNIIP			0.012	0.302	0.017	0.071		
dREER			-0.124	0.367	-0.133	0.118		
dEMS			0.032	0.305				
dNULC			0.013	0.883				
dPSD			-0.004	0.309				
dPD			0.041	0.000	0.030	0.000	0.028	0.000
dUR			-0.084	0.261	-0.091	0.114		
LL	-101.569		-90.201		-92.507		-97.937	
AIC	227.137		220.402		207.015		209.874	
BIC	258.157		272.101		235.449		227.969	

Notes: (I): Signals only. (II): Signals and deviations (w/o HP). (III): General-to-specific based on II (AIC). (IV): General-to-specific based on II (BIC). coeff: estimated coefficient. p: p-value for z-statistic.

3.3 Interpretation of the results

Overall, the Scoreboard has some informative content for the final outcome of the MIP, as it determines whether an In-Depth Review is prepared and how the imbalances are finally classified in the In-Depth Reviews. However, our results suggest that the European Commission makes selective use of the Scoreboard indicators. The single most important indicator for the outcome of the MIP is the export market share followed by the public debt. This result is robust when using 'single-indicator' or 'multi-indicator' models and when the deviations of the indicators from their official thresholds are taken into account. The private sector debt indicator is only informative for the decision whether an In-Depth

Review is prepared and the net international investment position is only informative when other indicators are additionally taken into account ('multi-indicator' model). Other indicators appear informative only in very few specifications (the current account) or only in one specification (nominal unit labor costs, real effective exchange rate, and unemployment rate). The deviation from the official threshold seems particularly important for the public debt indicator.

Interestingly, Scoreboard indicator variables that are less relevant in predicting financial crises (cf. Section 2) are found to be relevant for the final outcome of the MIP. The export market share represents the most striking example, given that it is not entirely clear from an empirical or a theoretical perspective how a decline in the export market share in one country could adversely affect the proper functioning of the economy of other countries in the EU. One reason for this could be the specific historical situation in which the Scoreboard has been effective. As many European countries are currently in the aftermath of financial crises, predicting upcoming crises might be of less importance for the European Commission. In light of our results, we argue that the European Commission indeed uses the MIP not just as an early warning mechanism for financial crises, but rather targets other imbalances that are not understood as well as those related to financial crises.

4 Conclusions

The MIP of the European Commission is one of the most relevant indicator-based international macroeconomic surveillance mechanisms. Despite its relevance, relatively little is known about how efficiently and consistently the Scoreboard - the indicator-based signaling device of MIP - is linked to the final outcome of the MIP. In this paper, we shed light on these issues by empirically investigating the role of the Scoreboard in the MIP. In particular, we investigate two of the most relevant aspects of indicator-based surveillance mechanisms. First, we examine the usefulness of the Scoreboard as an early-warning mechanism. Second, we discuss the relevance of the results of the Scoreboard for the final outcome of the MIP. Addressing these questions is important because misguided policy recommendations based on wrong signals from the Scoreboard can have far-reaching consequences with large economic costs. Moreover, doubts in the transparency and consistency of the MIP with regard to how the results of the Scoreboard are linked to the final outcome of the MIP could lead to a loss of ownership for the policy recommendations at the country level. Ownership, however, is widely regarded to be the key for policy recommendations to be successfully implemented. Therefore, empirical evaluations of the MIP,

even at a relatively early stage, are helpful to improve the MIP and avoid large economic costs.

Our results indicate that there are several avenues along which the MIP could be improved. With regard to the early-warning properties of the Scoreboard, our results show that very few indicators have reasonably good early-warning properties for financial crises. This result holds even when we deviate from the official thresholds of the Scoreboard and derive optimal thresholds based on reasonable assumptions about the preferences of policy makers (relatively high preferences for correct early-warning signals for financial crises). We also show that, given these preferences, some of the thresholds of the indicators could be easily improved. In general, the Scoreboard does not aim at providing early warning signals exclusively for financial crises, but also for other imbalances at the national level that do not have significant spill-over effects and are not exactly specified. However, using an indicator-based signaling device for one set of imbalances (financial crises), for which it has at best moderate early-warning properties, and for another set of imbalances, which are not exactly specified and therefore cannot be evaluated, may cast doubts on the efficiency of the MIP and may eventually contribute to the extremely low implementation rate of the policy recommendations.

With regard to the role of the Scoreboard for the final outcome of the MIP (that is, for the decision whether an In-Depth Review is prepared and how a macroeconomic imbalance is classified), we find that only very few indicators (most importantly, export market share and public debt) are significantly related to the final outcome of the MIP. In general, the indicators that have good early-warning properties for financial crises are not relevant for the final outcome of the MIP. This result might be a consequence of the fact that the MIP was introduced after the financial crises of the years 2008 and 2009 and is now concerned with 'cleaning up the mess' rather than providing early-warning signals for future crisis events. However, our results indicate additional potential weaknesses of the MIP that may contribute to the low implementation rate of policy recommendations. First, only those indicators that are relevant for some types of (unspecified) imbalances but are irrelevant for future crisis events are significantly related to the final outcome of the MIP. Second, one of those indicators, public debt, is already a prominent indicator in the Stability and Growth Pact. This could create potential overlaps with other procedures of the European Semesters and might weaken the transparency of the European Semester as a whole. Third, the other important indicator, export market share, may exhibit some structural weaknesses. For example, to the degree that emerging and developing economies further integrate into world markets, losses of export market shares

of developed economies may be a typical consequence rather than an indicator for any macroeconomic imbalance.

Overall, our results suggest that the MIP could be improved by (i) adjusting the thresholds of the Scoreboard indicators to provide more accurate early warning signals for financial crises, (ii) being more explicit on what kind of macroeconomic imbalances besides those related to financial crises it seeks to address, and (iii) streamlining it to central and more specific aims.

References

Alessi, L. and C. Detken (2011). Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity. *European Journal of Political Economy 27*(3), 520–533.

Babecký, J., T. Havránek, J. Matějů, M. Rusnák, K. Šmídková, and B. Vašíček (2012). Banking, debt and currency crises: early warning indicators for developed countries. Working Paper Series 1485, European Central Bank.

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

Cerra, V. and S. C. Saxena (2008). Growth dynamics: The myth of economic recovery. *American Economic Review 98*(1), 439–57.

Detragiache, E. and A. Spilimbergo (2001). Crises and liquidity; Evidence and interpretation. IMF Working Papers 01/2, International Monetary Fund.

Drehmann, M. and M. Juselius (2014). Evaluating early warning indicators of banking crises: Satisfying policy requirements. *International Journal of Forecasting 30*(3), 759–780.

Duca, M. L. and T. Peltonen (2013). Macrofinancial vulnerabilities and future financial stress: assessing systemic risks and predicting systemic events. *Journal of Banking & Finance 37*, 2183–2195.

European Commission (2012). Scoreboard for the surveillance of macroeconomic imbalances. Occasional Papers 92, European Commission.

European Parliament (2011). On the prevention and correction of macroeconomic imbalances. Regulation (EU) of the European Parliament and of the council 1176/2011, European Parliament.

European Parliament (2014). Country Specific Recommendations (CSRs) for 2013 and 2014 - A comparison and an overview of implementation. Technical report, Directorate-General for Internal Policies - Economic Governance Support Unit (EGOV).

European Parliament (2015). Country Specific Recommendations (CSRs) for 2014 and 2015 - A comparison and an overview of implementation. Technical report, Directorate-General for Internal Policies - Economic Governance Support Unit (EGOV).

Frankel, J. and G. Saravelos (2012). Can leading indicators assess country vulnerability? Evidence from the 2008-09 global financial crisis. *Journal of International Economics 87*(2), 216–231.

Knedlik, T. (2014). The impact of preferences on early warning systems - The case of the European Commission's Scoreboard. *European Journal of Political Economy 34*(C), 157–166.

Laeven, L. and F. Valencia (2013). Systemic banking crises database. *IMF Economic Review 61*(2), 225–270.

Reinhart, C. M. and G. L. Kaminsky (1999). The twin crises: The causes of banking and balance-of-payments problems. *American Economic Review 89*(3), 473–500.

Reinhart, C. M. and K. S. Rogoff (2011). From financial crash to debt crisis. *American Economic Review 101*(5), 1676–1706.

A Scoreboard indicators

Table 11: Overview of indicators and indicative thresholds

Indicator	Transformation	Threshold
Current account balance (% of GDP)	3 year backward moving av.	+6% and -4%
Net int. investment position (% of GDP)		-35%
Real effective exchange rate	3 years percentage change	-/+5% (euro area) -/+11% (non-euro area)
Export market share	5 years percentage change	-6%
Nominal unit labor costs	3 years percentage change	+9% (euro area) +12% (non-euro area)
House prices (deflated)	year-on-year change	6%
Private sector debt (% of GDP)		133%
Private sector credit flow (% of GDP)		14%
Public (general govt.) debt (% of GDP)		60%
Unemployment rate	3 year backward moving av.	10%
Total financial sector liabilities	year-on-year change	16.5%

B Crisis events

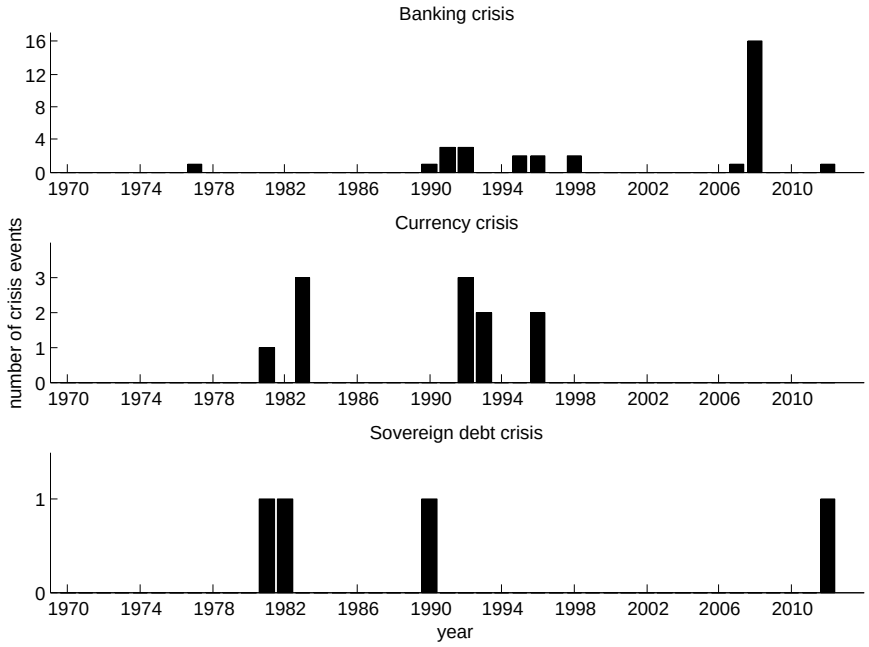

Figure 2: Different crisis events by years (EU-28) based on Laeven and Valencia (2013).

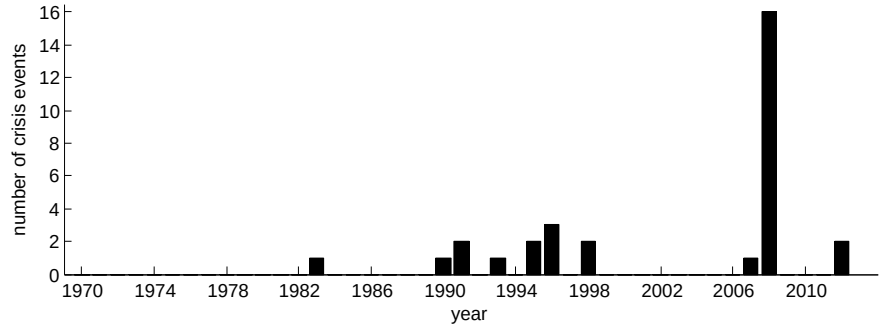

Figure 3: Crisis events by years (EU-28) that can be ultimately employed for the evaluation of the Scoreboard indicators.

C Robustness checks

Table 12: Evaluation based on official Scoreboard thresholds (Forecast horizon $[t, t+3]$, $\theta = 0.5$).

Indicator	NSR	Usefulness $\theta = 0.5$	type I error (in %)	type II error (in %)	correct forecasts (in %)	correct crisis signals (in %)	identified crisis (in %)	average lead (years)
CA	0.61	0.11	44.12	34.11	64.64	55.88	66.67	2.32
$NIIP$	1.04	-0.01	73.13	28.06	66.09	26.87	25.93	2.10
$REER$	2.15	-0.11	80.70	41.57	52.70	19.30	47.62	2.80
EMS	0.89	0.01	85.00	13.37	76.37	15.00	29.17	2.00
$NULC$	1.75	-0.08	77.59	39.21	55.04	22.41	30.43	1.92
HP	0.56	0.10	53.33	25.97	67.84	46.67	61.11	2.64
PSD	0.71	0.08	46.77	37.95	60.76	53.23	48.00	2.54
$PSCF$	0.44	0.16	43.33	24.93	72.42	56.67	60.00	2.40
PD	0.74	0.06	55.00	33.24	63.66	45.00	41.67	2.90
UR	2.29	-0.09	85.71	32.73	59.59	14.29	30.00	2.33
TFL	0.64	0.07	63.16	23.67	70.63	36.84	54.55	2.67

Notes: NSR = B/(B+D)/(A/(A+C)); Usefulness = min[θ; (1-θ)]-Loss; Loss = θC/(A+C)+(1-θ)B/(B+D); Type I error = C/(A+C); Type II error = B/(B+D); Share of correct forecasts = (A+D)/(A+B+C+D); Share of correct crisis signals = A/(A+C).
CA = current account balance, NIIP = net international investment position, REER = real effective exchange rate, EMS = export market share, NULC = nominal unit labor costs, HP = real house prices, PSD = private sector debt, PSCF = private sector credit flow, PD = public debt, UR = unemployment rate, TFL = total financial liabilities.

Table 13: Evaluation based on official Scoreboard thresholds (Forecast horizon $[t+2, t+5]$, $\theta = 0.7$).

Indicator	NSR	Usefulness $\theta = 0.7$	type I error (in %)	type II error (in %)	correct forecasts (in %)	correct crisis signals (in %)	identified crisis (in %)	average lead (years)
CA	0.77	-0.19	54.22	35.22	61.88	45.78	51.85	3.16
$NIIP$	1.16	-0.31	75.29	28.54	63.76	24.71	25.93	3.20
$REER$	0.78	-0.18	53.42	36.39	60.41	46.58	61.90	4.64
EMS	1.16	-0.36	88.00	13.95	72.79	12.00	29.17	2.00
$NULC$	1.08	-0.27	65.38	37.22	57.11	34.62	52.17	3.88
HP	0.38	-0.08	44.44	21.38	72.36	55.56	61.11	3.83
PSD	0.70	-0.14	46.67	37.36	60.99	53.33	52.00	4.00
$PSCF$	0.60	-0.17	56.16	26.45	68.35	43.84	52.00	3.44
PD	0.81	-0.21	58.67	33.53	62.00	41.33	41.67	4.40
UR	2.23	-0.40	84.93	33.55	56.74	15.07	20.00	3.33
TFL	0.73	-0.24	67.14	24.00	68.35	32.86	50.00	3.46

Notes: See Table 12.

Table 14: Evaluation based on optimized thresholds: Usefulness (Forecast horizon $[t+2, t+5]$, $\theta = 0.7$).

Indicator	NSR	Usefulness $\theta = 0.7$	type I error (in %)	type II error (in %)	correct forecasts (in %)	correct crisis signals (in %)	identified crisis (in %)	average lead (years)
CA	0.97	0.01	0.00	96.52	18.23	100.00	88.89	4.26
$NIIP$	1.02	-0.03	5.88	95.82	18.99	94.12	85.19	4.27
$REER$	0.98	0.01	0.00	98.10	20.31	100.00	90.48	4.62
EMS	0.78	0.05	4.00	74.71	37.95	96.00	83.33	4.26
$NULC$	0.96	0.00	2.56	93.53	24.81	97.44	91.30	4.57
HP	0.64	0.08	5.56	60.69	54.27	94.44	88.89	4.35
PSD	0.80	0.06	0.00	80.17	34.04	100.00	88.00	4.55
$PSCF$	0.79	0.03	6.85	73.84	37.89	93.15	80.00	4.18
PD	0.96	-0.01	5.33	91.33	23.99	94.67	79.17	3.92
UR	0.99	-0.01	4.11	95.21	22.02	95.89	95.00	4.65
TFL	0.83	0.04	2.86	80.62	33.16	97.14	86.36	4.27

Notes: See Table 12.

Table 15: Optimized thresholds (Forecast horizon $[t+2, t+5]$, $\theta = 0.7$).

Indicator	SB thresholds		optimized thresholds			
			NSR		Usefulness	
	lower	upper	lower	upper	lower	upper
CA	-4.0	6.0	-10.7	6.6	-0.2	0.1
$NIIP$	-35.0		-77.3		33.5	
$REER$	-5.0	5.0	-7.8	3.1	-0.1	0.3
EMS	-6.0		-2.6		6.6	
$NULC$		9.0		4.7		-0.6
HP		6.0		8.9		-1.9
PSD		133.0		185.7		62.2
$PSCF$		14.0		24.6		3.8
PD		60.0		86.6		10.6
UR		10.0		4.0		4.0
TFL		16.5		14.9		2.8

Notes: See Table 12.