WORKING PAPERS

109

Mária Širaňová – Karol Zeleňák

OUTLIERS DO MATTER – COMPARING THE DATABASES OF FINANCIAL CRISIS EVENTS

ISSN 1337-5598 (electronic version)

The WORKING PAPER SERIES is intended to convey preliminary, partial results of ongoing research achieved by fellows or research groups of the Institute of Economic Research which can be prepared for later publications.

AUTHORS

Ing. Mária Širaňová, MA., PhD., Institute of Economic Research SAS, Slovakia
Ing. Karol Zeleňák, PhD., National Bank of Slovakia; Comenius University in Bratislava, Slovakia

REVIEWERS

Ing. Tomáš Domonkos, PhD., Institute of Economic Research SAS, Slovakia
PhDr. Štefan Rychtárik, PhD., National Bank of Slovakia, Slovakia

We thank Dávid Hojdan, Boris Fišera and Tiffany Kajurová for excellent research assistance. We thank Marek Radvanský for sharing his database. We thank Filip Ostrihoň for helpful comments and suggestions. Authors acknowledge that this paper is the partial result of the project APVV under Grant No APVV-17-0551. Širaňová acknowledges that this paper is the partial result of the project VEGA under Grant No 2/0101/21.

ABSTRACT

Outliers do Matter – Comparing the Databases of Financial Crisis Events

In this paper we examine the consistency in the timing of crisis events of the most prominent databases of banking and fiscal crises. In order to do so we calculate Cohen's kappa indicator measuring level of commonality across databases. Additionally, we employ panel logit models with random effects to investigate predictive properties of early warning indicators selected from the Macroeconomic Imbalances Procedure scoreboard. We also identify the most influential crisis observations unique to each database using Pregibon's delta-beta influence statistics. Our results confirm that the degree of commonality across databases is indeed relatively high, especially if introducing a one-year lag due to possible beginning- and end-of-the-year discrepancies. However, there is still a significant role played by a few influential observations that determine several heterogeneous findings for statistically significant EWIs. This problem is more pronounced in the banking crisis literature. Based on the empirical findings, we propose several suggestions that should be discussed and potentially adopted by research community in order to address the existing concerns.

KEYWORDS: financial crisis, fiscal crisis, banking crisis, dating of crises, crisis prediction

ABSTRAKT

Na extrémnych hodnotách záleží – porovnanie databáz finančných kríz

V tomto príspevku skúmame súlad v identifikácii krízových udalostí najvýznamnejších databáz bankových a fiškálnych kríz. V rámci vykonanej analýzy vypočítavame Cohenov ukazovateľ kappa merajúci mieru prekrývania databáz krízových udalostí. Okrem toho využívame panelové logitové modely s náhodnými efektmi na skúmanie prediktívnych vlastností indikátorov včasného varovania (EWI) vybraných z hodnotiacej tabuľky Procedúry makroekonomických nerovnováh. Taktiež identifikujeme najvplyvnejšie pozorovania v rámci databáz krízových udalostí unikátne pre každú individuálnu databázu pomocou Pregibonovej delta-beta štatistiky. Naše výsledky potvrdzujú, že stupeň zhody naprieč databázami je skutočne relatívne vysoký, najmä ak zavádzame jednoročné oneskorenie kvôli možným nezrovnalostiam pri identifikácii krízových udalostí vznikajúcich na začiatku alebo na konci kalendárneho roka. Na druhú stranu však významnú úlohu zohráva niekoľko vplyvných pozorovaní, ktoré determinujú heterogénne výsledky pre vybrané štatisticky významné EWI. Tento problém je výraznejší pri databázach bankových kríz. Na základe týchto empirických zistení formulujeme niekoľko návrhov, ktoré by mali byť diskutované v rámci širšej vedeckej komunity, aby bolo možné adresovať otvorené problematické aspekty identifikácie krízových udalostí.

KĽÚČOVÉ SLOVÁ: finančná kríza, fiškálna kríza, banková kríza, identifikácia krízových udalostí, predikcia krízových udalostí

JEL CLASSIFICATION: D31, E21, E52, E58

The views expressed in the WP and the language revision is those of the authors.
Layout by: Ing. Mária Širaňová, MA., PhD.

Institute of Economic Research SAS/Ekonomický ústav SAV, Šancová 56, 811 05 Bratislava, www.ekonom.sav.sk
CONTACT: maria.siranova@savba.sk

Contents

1 Introduction

Over the last decade, the literature addressing various aspects of financial crises has expanded in an unprecedented fashion in the wake of the Great Recession. According to the IDEAS/RePEc database, the number of articles submitted to the database mentioning the term 'financial crisis' was more than 8 times higher in 2009 (2344) than in 2007 (278). In the last ten years, more than 32,000 research papers in the database include the term 'financial crisis' at some point. In addition to other topics, researchers and policy makers have primarily concerned themselves with analysing the causes and consequences of such disruptive events, as well as discussing the most appropriate response from a policy perspective (Claessens and Kose, 2013). Banking and fiscal crises have played a central role in this burgeoning discussion, not least due to their involvement in the most recent global financial meltdown.

Among the influx of research papers, several have enjoyed considerable success (Reinhart and Rogoff, 2009; Laeven and Valencia, 2008) by becoming the source of the much needed empirical evidence for the follow-up stream of quantitative and qualitative studies. Given their widespread use, it has been taken for granted (Kauko, 2014) that significant overlap among a few prominent crisis databases should warrant robust and mutually comparable findings.

Few studies contradict this generally accepted narrative (Frydl, 1999; Chaudron and de Haan, 2014; Boyd et al., 2009). Using the data from the top four banking crisis databases, Boyd et al. (2009) show how the predictive properties of selected early warning indicators (EWIs henceforth) differ substantially given the database used. The presence of even relatively small, yet influential, differences (Claessens and Kose, 2013) can thus affect analyses.

To contest the generally accepted view of robust crisis identification strategies, in this paper, we analyse the consistency in the timing of crises of the most prominent databases of banking and fiscal crises. First, we identify the most 'popular' databases in each category according to our algorithm measuring the expected number of absolute and relative citations of respective publications. Then, we combine the approaches of Chaudron and de Haan (2014) and Boyd et al. (2009) to calculate the degree of commonality across databases using Cohen's κ indicator and to investigate how the use of a different crisis event identification can alter the results of logit-based predictive models. Last, we identify the most influential crisis observations not shared by other databases according to Pregibon's delta-beta influence statistics. The set of EWIs is taken from the official list of EWI-like indicators used by the European Commission as part of the Macroeconomic Imbalances Procedure assessment. As the majority of the EWIs used in our analysis have already been adopted by the relevant policy makers, challenging their predictive power may have serious consequences for the conduct of a macroeconomic stabilization policy.

Our results suggest that the degree of commonality is relatively high in both the fiscal and banking literature after one accounts for possible discrepancies in the dating of a crisis event by introducing a one-year lag. Second, in most cases, the list of weak EWIs, i.e., variables with low or no predictive power, is consistent regardless of the identification strategy employed. On the other hand, the presence of a small number of influential observations differently identified across databases often results in highly heterogeneous findings for statistically significant EWIs. This problem is less urgent in the case of fiscal crisis events but more pronounced in the banking crisis field. Our findings thus corroborate the discussion in Claessens and Kose (2013), who argue that while (external) sovereign crises are relatively easily identified, banking crises often pose a larger challenge.

Since the degree of heterogeneity within existing crisis identification strategies may alter empirical findings, we discuss possible avenues that the research community could take to mitigate these adverse consequences. Among other recommendations, we strongly advocate the use of meta-analysis in the field of financial crisis research to identify the effect of the identification strategy on reported empirical results (e.g., Hamdaoui (2017)).

The remainder of the paper is organized as follows. Section 2 outlines the relevant literature, while Section 3 describes the empirical methodology and introduces our data. The results are reported in Section 4, and Section 5 concludes the paper.

2 Literature Review

Authors often acknowledge that the definition of a crisis event differs across studies (Boyd et al., 2009; Babecky et al., 2014; Frydl, 1999; Baldacci et al., 2011). In van den Berg et al. (2008), the sensitivity of the results to the definition of a crisis is listed among the four major issues in the crisis literature. However, usually no further rigorous analysis is conducted to address this concern. As an example, Kauko (2014) delivers a comprehensive overview of the banking crisis literature and discusses various approaches to crisis event specification. However, he argues that most of the literature on early warning banking crisis indicators delivers highly comparable results, as these papers analyse almost identical sets of cases.

To the best of our knowledge, very few papers address the consequences of database selection in a more quantitative way.

Frydl (1999) compares the dating, length and resolution costs of five important studies on banking crisis events published in the 1990s. They report considerable discrepancies among studies and caution that if similar disagreement were to be found in dating the economic recessions, the concept of a recession would be seriously impaired when used in empirical analysis. Similarly, Chaudron and de Haan (2014) qualitatively analyse the overlap among the three most authoritative sources of systemic banking crises in recent literature (Laeven and Valencia, 2008; Caprio and Klingebiel, 2002; Reinhart and Rogoff, 2009). In accordance with Frydl (1999), they conclude that databases differ significantly in terms of dating the beginning and length of a crisis even if referring to the same type of event (systemic banking crisis). Then, they proceed to quantitative assessment of four crisis events identified by all three databases based on a measure of bank failures and conclude that the database developed by Laeven and Valencia (2008) provides the most accurate specification of a systemic banking crisis. For fiscal crises, Baldacci et al. (2011) compare the outcomes of their crisis identification strategy with four additional papers and conclude that their approach produces more events due to the use of a more comprehensive definition by including IMF-supported programs and government bond yields.

Boyd et al. (2009) provide an exploratory analysis of differences across four major banking crisis databases (Demirgüç-Kunt and Detragiache, 2005; Caprio and Klingebiel, 2002; Reinhart and Rogoff, 2009; Laeven and Valencia, 2008). They conclude that for many crisis episodes, the dating classifications differ considerably across databases in terms of both the starting date and the duration. This finding casts serious doubt about either the robustness or the comparability of results obtained in a large empirical literature. They estimate and compare differences in predictive properties of selected variables given the choice of a database by logit model, and theirs is the only paper to date to conduct such an analysis of these databases.

In the majority of cases, as the focus is generally on the right-hand side of the equation, i.e., investigating the predictive properties of a single indicator or a group of indicators, one of the available databases is usually chosen without any further discussion. As an alternative, authors present their own definition of a crisis event that is consequently used (von Hagen and Ho, 2007; Boyd et al., 2009). In these instances, only a few of the studies compare their findings to results achieved when switching to another, alternative database (e.g., Boyd et al. (2009)).

This is rather surprising given how many different approaches one can adopt when addressing such a comprehensive topic as financial crises. Taking inspiration from the comprehensive overviews by Kauko (2014) and Claessens and Kose (2013), one can produce a rough taxonomy of databases at hand. The first dimension assesses each database based on the definition of a crisis event. In this case, the authors define the timing of a crisis based on a pre-determined quantitative criterion (e.g., the number of bank bankruptcies for banking crises as in Chaudron and de Haan (2014)), on expert judgment (Caprio and Klingebiel, 2002) or a combination of both (Baldacci et al., 2011).

The second dimension differentiates according to the time coverage of an identified event. Some authors only focus on the proper specification of the beginning of a crisis (e.g., Schularick and Taylor (2012)), but others also assess the length of a crisis event (Laeven and Valencia, 2008; Reinhart and Rogoff, 2009) or even provide estimates of fiscal or real costs associated with a crisis event (Laeven and Valencia, 2008; Duca et al., 2017).[1]

Financial crisis databases can also be characterized by three additional features: i) their scope in terms of time and country coverage (world or regional, short or long time span), ii) frequency (annual or higher frequency data), and iii) type of event (e.g., fiscal, banking, financial, currency, or external sector crises or few of them together). Contrary to the first two dimensions, these three characteristics should not result in significant differences in the timing of the length of a crisis event, as they only reduce the time-country space of possible crisis observations. As such, this heterogeneity should therefore not be translated into different predictive power of explanatory variables when using a common dataset.[2]

Finally, some authors comment on the systematic nature of a crisis event (Reinhart and Rogoff, 2009) or distinguish whether the event was caused by internal or external forces (Duca et al., 2017).

3 Comparing the Databases

3.1 Measuring the Popularity of a Database

To illustrate the consequences of significant differences in crisis event specifications across databases, we select the most prominent databases. This has the advantage that the potential bias due to the differences in the timing of a crisis event is expected to be significantly higher due to their relative popularity in the

[1]Kauko (2014) also distinguishes between a dichotomous and a continuous crisis event specification, with the latter having been introduced in the third generation of the banking crisis literature. In this paper, we consider only a dichotomous specification, which is, however, widely used in the field.

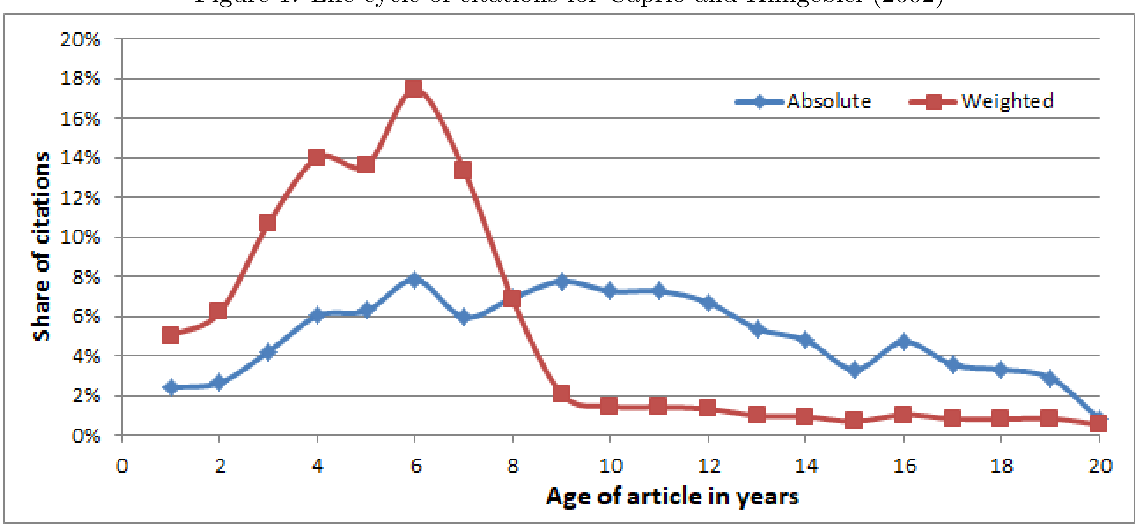
[2]The date of database publication may also matter, in particular in the case of expert judgment assessment. The use of 'insider information' in characterizing an event that occurred recently might introduce a certain level of subjectivity into the identification strategy. On the other hand, when evaluating the properties of some long-forgotten events, an important piece of information that may have remained hidden might also cloud the final judgment. Additionally, the role of data revisions that might update previously incorrectly dated events should also be considered. However, we do not further comment on these issues, as we were not able to find relevant sources that pay sufficient attention to them in our context.

scientific community, as well as among policy makers. We also do not further investigate whether the citing documents use the database for further empirical work or only as part of their literature review. Even if mentioned only as a reference, the inclusion of a database already signifies its importance (and that of its findings) for a research community.

To measure the 'popularity' of a database, we use publicly available data from two scientific databases, namely, Google Scholar and IDEAS/RePEc. The simple 'popularity' measure is given by the number of citations for each database. To count the number of citations, we choose the Google Scholar platform over the two official scientific databases (Web of Science, Scopus) because we would like to also measure the 'popularity' of a database among policy makers as well as wider audiences. Both scientific databases include only articles that were published in the form of a scientific document, hence neglecting the broader impact of a paper on the conduct of economic policy (captured by official documents released by national or international institutions). Second, many scholars from emerging and developing economies do not regularly publish in high-quality scientific outlets but are still able to follow the standard literature and hence are expected to be influenced by it. As argued by Martin-Martin et al. (2017), Google Scholar is able to efficiently identify highly cited documents.

In the IDEAS/RePEc database, we count the number of instances per year when the term 'financial crisis' was mentioned in articles submitted to this database in that particular year. This serves two purposes: i) to measure the overall 'popularity' of the topic during the particular period of time and ii) to weight the number of citations by the overall trend of increasing the size of the publication market in the economics field. IDEAS/RePEc is the largest bibliographic database dedicated to economics and available freely on the Internet. According to the IDEAS/RePEc, it indexes over 3,300,000 items of research, including over 3,100,000 that can be downloaded in full text.

Figure 1: Life cycle of citations for Caprio and Klingebiel (2002)



Notes: The absolute measure reflects the share of citations over the total number of citations for the article as recorded in the Google Scholar database. The weighted measure represents the share of citations over the total number of citations for the article as recorded in the Google Scholar database weighted by the total number of instances of the term 'financial crisis' in the IDEAS/RePEc database.

As the relevant databases were published in different years, a simple count of the number of citations would be insufficient. The life cycle of citations per publication in the economics field was investigated in Anauati et al. (2016), who used EconLit and Google Scholar data to model the life cycle of citations across fields of economic research. Articles published in the field of applied research, which encompasses the financial crisis databases, peak in terms of the annual number of citations approximately 6-8 years after publication. To calculate the expected number of citations per database, we model the life cycle of citations according to the life cycle of citations for Caprio and Klingebiel (2002). As illustrated in Figure 1, the peak in citations for this article is achieved during the 6-9 year period (absolute) or more than 6 years after its publication (weighted), which is in line with Anauati et al. (2016).

The expected number of citations is given by the following formula:

$$cit_i^{abs} = \frac{\sum\limits_{t=1}^{T} cit_{it}}{\sum\limits_{t=1}^{T} share_t^{CK}} \tag{1}$$

where cit_i^{abs} stands for absolute expected number of citations of database i, cit_{it} denotes the number of citations of database i in year t, and $share_t^{CK}$ represents citations in year t over the total number of citations for Caprio and Klingebiel (2002). Year $t = 1$ is the year of publication i, with T being the number of years since publication.

To provide an example, we illustrate the calculation using the Reinhart and Rogoff (2009) publication. As of 2019, the cumulative number of citations recorded in Google Scholar for this article equals 8217. As of 2019, the total number of years (T) since publication in 2009 was 11. According to the benchmark life cycle derived from Caprio and Klingebiel (2002), 64.57 per cent of citations are recorded over the first eleven years after the paper is published. The expected number of citations for Reinhart and Rogoff (2009) is therefore set at 12,725 citations.

The weighted number of citations is given by the following formula:

$$cit_i^{rel} = \frac{\sum\limits_{t=1}^{T} \frac{cit_{it}}{count_t^{FC}}}{\sum\limits_{t=1}^{T} wshare_t^{CK}} \tag{2}$$

where cit_i^{rel} stands for the relative expected number of citations of database i, cit_{it} denotes the number of citations of database i in year t, and $count_t^{FC}$ represents the number of instances when the term 'financial crisis' was mentioned in year t in the IDEAS/RePEc database. The variable $wshare_t^{CK}$ measures the share of citations in year t over the total number of citations for Caprio and Klingebiel (2002), with citations being divided by the number of instances when term 'financial crisis' was mentioned in year t in the IDEAS/RePEc database. Year $t = 1$ is the year of publication i, with T being the number of years since publication.

We again illustrate the calculation of cit_i^{rel} using the Reinhart and Rogoff (2009) publication. As of 2019, the weighted cumulative number of citations recorded in Google Scholar for this article equals 2.779. As of 2019, the total number of years (T) since publication in 2009 was 11 years. According to the benchmark life cycle derived from Caprio and Klingebiel (2002), 92.13 per cent of weighted citations are recorded over the first eleven years after the paper is published. The relative expected number of citations for Reinhart and Rogoff (2009) is therefore set at 3.016.

3.2 Statistical Comparison of Databases

As part of our empirical strategy, we attempt to measure how strongly the databases match in recognizing crisis and non-crisis periods.

(Non-)identification of a crisis and the number of crisis years, as already mentioned, are highly dependent on the definition of how the respective authors specify the crisis period. Simply, a stricter and sophisticated definition leads to a lower number of crises. The very strict definition of a crisis (in the sense of the need to comply with a number of criteria) can, in general, lead to missing the crisis event (Type 1 error). A relatively relaxed definition (a few simple criteria or a relaxed specification when not all criteria need to be met) can result in Type 2 error when all potential events are classified as crisis events even though several of them may be considered borderline from the perspective of other relevant sources.

For the purpose of statistical assessment, we use several statistical indicators that are capable of describing the degree of overlap in individual databases.

The first indicator is the *absolute frequency*. This simple statistical indicator measures how many instances of a predefined event appear in a sample, i.e., how many events meet the predefined criteria. The second indicator is the indicator of *relative frequency*. This indicator is calculated as the ratio of the number of predefined events to the overall sample of all events that may appear. In other words, this indicator measures the probability with which an event can occur.

Expected frequency is the third indicator and is used to assess conditional probability when a two-way occurrence of an event may occur. In general, the expected frequency also relates to probability theory, since this statistic expresses the frequency (number) of events or occasions on which a certain event may theoretically occur on average in a given number of trials. In a two-dimensional table, the expected frequency may be calculated as follows:

$$E_{ij} = \frac{(T_i.T_j)}{N} \tag{3}$$

where E_{ij} stands for expected frequency of the $i-th$ row and $t-th$ column, T_i denotes the total sum in the $i-th$ row, T_j represents the total sum in the $j-th$ column, and N is the table total sum.

Using the indicators above, we calculate *Cohen´s κ coefficient*. This statistic introduced by Jacob Cohen (1960) represents the rate of agreement between the two samples (databases in our case). The difference between the frequency type of indicators and κ is that κ takes into account the agreement that occurred by chance. Therefore, the κ coefficient may be interpreted as the inter-class correlation coefficient.

Suppose that there are N events, which are assigned independently to k categories by the two separate databases. The results could be displayed in a kxk contingency table (Table 1), where each ij value represents the proportion that the database A had classified in a category i and a database B had classified in a category j. p_i and p_j denote the probabilities/frequencies of assignment into categories i and j in the respective databases A and B. The calculation of the κ coefficient is then based on the relative and expected frequencies of the diagonal of a square contingency table.

The observed proportional agreement (p_0) between databases A and B is calculated as:

Table 1: Square contingency table

	Database B				
Database A	**1**	**2**	**. . .**	**k**	**Total**
1	p11	p12	. . .	p1k	p1K
2	p21	p22	. . .	p2k	p2K
. . .	. . .	. . .	. . .	. . .	. . .
k	pk1	pk2	. . .	pkk	pkK
Total	pK1	pK2	. . .	pKk	pKK

$$p_0 = \sum_{i=1}^{k} p_{ii} \tag{4}$$

and the overall proportion of agreement expected by chance is calculated as:

$$p_e = \sum_{i=1}^{k} p_i . p_j \tag{5}$$

Cohen's κ statistic measuring the degree of database agreement is as follows:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{6}$$

κ is a non-dimensional statistic that usually acquires values between zero and one. A value close to one represents perfect agreement between the two reviewed databases, while a zero value represents a state of no significant agreement between the databases, e.g., no more agreement than would be expected by chance. κ can also obtain negative values, which represent the fact that there is no effective agreement between the two databases.

The standard error of estimated κ can be estimated as:

$$SE(\kappa) = \frac{SD(\kappa)}{\sqrt{N}} \tag{7}$$

The standard deviation can be calculated as:

$$SD(\kappa) = \sqrt{\frac{p_0(1-p_0)}{(1-p_e)^2}} \tag{8}$$

Fleiss et al. (1969), however, argue that such a formula is based on incorrect assumptions, since it uses results in conservative significance tests and confidence intervals, which leads to overestimation of the results. However, running a sufficient number of calculations on PASS software at the NCSS.com website confirms that Cohen´s approximation of the standard deviation is relatively close to the estimation of the standard deviation by Fleiss et al. (1969). Thus, Cohen´s proposed standard deviation is sufficient for the purpose of computing the standard deviation.

Then, the 100(1–α) confidence interval for κ can be computed as:

$$\kappa + -z_{\alpha/2} SD(\kappa) \tag{9}$$

Landis and Koch (1977) derived a table of κ value classification (Table 2). Although Landis and Koch (1977) provided no evidence to support their classification and the table is thus more result of their expert judgment, it serves as a good guide for the interpretation of results.

Table 2: Table of interpretation of Kappa coefficients introduced by Landis and Koch

	Interpretation
<0	No agreement
0.00-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Almost perfect agreement

3.3 Database Selection Bias and Predictive Properties of Early Warning Indicators

Rather than creating the best early warning system or selecting the best predictor, our aim is to demonstrate the change in predictive power of EWIs conditional on the specification of a crisis event, i.e., the choice of crisis database. Discrete choice models belong to the most commonly used approaches to assess the predictive properties of early warning systems (hereinafter EWSs) in the literature (Kaminski et al., 1997; Berg and Pattillo, 1999). In general, studies use either linear probability models (Davis et al., 2016; Kacer, 2013), probit (Berg and Pattillo, 1999; Frankel and Rose, 1996; Kacer, 2013; Mulder et al., 2016), logit (Ostrihon, 2020; Domonkos et al., 2017a; Arregui et al., 2013; Davis et al., 2016; Valinskytė and Rupeika, 2015; Davis and Karim, 2008), or multinomial logit (Ciarlone and Trebeschi, 2005; Caggiano et al., 2016) models with random effects. Given our objective, we opt for a simple bi-variate pooled logit model with random effects, as in Boyd et al. (2009).

Our model can be therefore specified as follows:

$$P(y = 1|X_{t-1}) = G(X_{t-1}\beta) \tag{10}$$

where y stands for binary dependent variable (1 representing the crisis event, 0 otherwise), X_{t-1} is the explanatory variable at time *t-1*, and $G(\cdot)$ is the logistic cumulative distribution function.

Among the space of possible EWIs drawn from the relevant literature, we select several relevant variables from the Macroeconomic Imbalance Procedure (MIP henceforth) scoreboard table (8 benchmarks, 1 auxiliary). The MIP was introduced in 2011 by the European Commission to serve as a robust monitoring mechanism that helps identify the accumulation of potential risks, correct existing imbalances and prevent them from re-emerging. Officially (EC, 2016), the MIP was not envisaged as a pure EWS system but has been treated as such by relevant academic literature (Knedlik, 2014; Domonkos et al., 2017a; Siranova and Radvansky, 2018; Sondermann and Zorell, 2019; Biegun and Karwowski, 2020). As a consequence, the list of potential EWIs in the MIP scoreboard table incorporates indicators that were already chosen by the relevant policy makers (the European Commission). Possible mis-specification and bias in the expected predictive properties of such indicators may therefore introduce real consequences for stabilizing economic policy in Europe. Additionally, these indicators have already passed the screening process for selection as the most suitable EWIs; thus, it is expected that their predictive properties are robust and a rather well-known quantity.

The final list of variables and their specification are provided in Table 3. These EWIs were chosen to ensure the broadest coverage in terms of countries and time span. The list of potential EWIs includes four benchmark variables reflecting the accumulation of external imbalances (current account balance, export market share, REER, NIIP), four variables assigned to a group focusing on the creation of internal imbalances (government debt, private sector debt, unemployment rate, house price index) and one variable (FDI inflow) associated with external imbalances and listed among the additional 40 auxiliary indicators.

Table 3: Early Warning Indicators Description

Indicator description	**Calculation**	**Source**
Current account balance as % of GDP	3 year average	World Bank
Export market share (% of world export)	5 year % change	World Bank
Real effective exchange rate (2010=100)	3 year % change	World Bank
Net international investment position	% of GDP	World Bank
General government sector debt	% of GDP	IMF Global Debt Database
Private debt, loans and debt securities	% of GDP	IMF Global Debt Database
House price index (real)	1 year % change	BIS
Foreign direct investment in the reporting economy (flows)	% of GDP	World Bank

3.4 Timing of a Crisis Event and Event Identification Strategy

When assessing the degree of consistency between the databases, we also have to take into account the characteristics of the respective database (Section 2), e.g., whether the databases map only the beginning year of a crisis or whether they comprise the duration of a crisis. From this perspective, in a simple comparison ("basic approach"), the databases that also comprise information about the duration of a crisis logically signal more crisis years for the same period and set of countries than the databases mapping only the beginning year of a crisis. To avoid these structural inconsistencies between the databases, we transform the "duration databases" to simple "beginning year databases". We perform this transformation by adding zero value (no crisis) to years after the beginning year when the crisis was indicated in a database. This approach is termed the "strict approach".

Alternatively, we also reflected the fact that "a criterion for a crisis identification matters". In other words, having a variety of criteria for crisis identification may result in the identification of the same crisis but in different years. For instance, if we opt for a relaxed criterion for crisis identification (e.g., a small number of reviewed indicators or lower signalling level of investigated indicators), in this way, we are able to detect the onset of a crisis at an earlier stage than a database that opted for stricter criteria that detects the same crisis at later stage (for instance, with a one-year lag). In this way, the two different databases may detect the same crisis but in different years. To address such a possibility, we opt for an approach in which we investigate whether the two individual databases have detected the onset of a crisis in three consecutive years. For instance, if in database A, the beginning of a crisis for a country X was detected in year T, we also investigate whether in database B, the onset of a crisis for a country X was detected not only during year T but also during years $T+1$ and $T-1$. If such a situation has occurred, we evaluate this situation as compliance, assuming that the two databases detect the same crisis, just in a different period. This approach is termed the "semi-strict approach".

As a consequence, the "semi-strict approach" is expected to result in a higher number of commonly detected crisis periods than the "strict approach" that abstracts from possible time lags of crisis identification stemming from the crisis detection criteria.

In the empirical part of the analysis, we ensure uniformity of a sample by estimating logit models on a set of countries and time spans, which is common across all selected databases. We exclude the last 15 years spanning the Great Recession period and end our sample in 2003, a termination year set by Caprio and Klingebiel (2002). Due to its global overreach and substantial spillovers across many countries, it is likely that this event will be (correctly) identified in most of the databases. As we are interested in analysing measures of commonality in identifying 'idiosyncratic' crisis events, the occurrence of one major global disturbance may introduce distortions into our analysis.

The literature has long recognized that the inclusion of years following the first crisis year introduces crisis duration bias (Bussiere and Fratzscher, 2006). In post-crisis periods, variables affected by a crisis experience an adjustment process before reaching a more sustainable level or growth path. From this reason they may produce distorted signals. As a consequence, empirical studies either exclude post-crisis periods from the sample or treat them as tranquil periods. In our approach, we opt for the second option for the following reasons. Most important, while in some studies, the crisis duration is specified (which would allow for the exclusion of all subsequent crisis years), this is not the case for all of our databases. Hence, decisions regarding the duration of a crisis would be highly arbitrary.[3] Additionally, by retaining the post-crisis years as tranquil periods, we introduce positive bias to the calculation of statistical commonality (the κ indicator) and predictive properties of EWIs (logit estimates). However, as we will show, even with this positive compensation, it will often not be enough to improve the level of commonality across selected databases, especially in the case of probability models.

4 Results

4.1 Ranking of Databases

We collect information on several papers published in the post-2008 period that either sought to investigate the causes or consequences of fiscal and banking crises or were analysing predictive properties of selected indicators linked to these types of crises. The paper by Caprio and Klingebiel (2002) serves as the reference point for modelling the life cycle of paper citation profiles. The consequent updates of the original paper by Laeven and Valencia (2008) are summed up and treated as a single contribution.[4]

[3]Boyd et al. (2009) replaces the missing information on crisis duration in some of the databases with data collected from other sources where the end of a crisis was indicated. However, it is not entirely clear how to adjust the database in the case of events that were identified by only one database.

[4]Some of the prominent databases, measured by the absolute and relative number of citations, were published in the pre-2008 period, such as Demirgüç-Kunt and Detragiache (2005). Replacement of Reinhart and Rogoff (2009) with Demirgüç-Kunt and Detragiache (2005) would restrict our list of the most 'popular' databases to the inclusion of only one paper published in the post-2008 period (Laeven and Valencia, 2008). As we prefer to analyse the impact of the newest contributions that have the potential to shape discussion in the future, we do not further comment on Demirgüç-Kunt and Detragiache (2005), although we acknowledge its importance.

Table 4: Ranking of Databases - Absolute Citations

	Ranking - Top 3		**Expected #**	**Citations per**	**Type of crisis**				**Crisis**
	Fiscal	**Banking**	**of citations**	**crisis type**	**Fiscal**	**Banking**	**Financial**	**Currency**	**specification**
Babecky, J. et al. (2014)			320	107	1	1	0	1	duration
Baldacci, E. et al. (2011)	3		262	262	1	0	0	0	duration
Baron, M. et al. (2018)*			65	65	0	1	0	0	beginning
Gerling, K. et al. (2017)			108	108	1	0	0	0	duration
Jorda, O. et al. (2017)			1 623	1 623	0	0	1	0	beginning
Laeven, L. and Valencia, F.	2	1	5 149	4 435					
i) Laeven, L. and Valencia, F. (2008)			59	59	0	1	0	0	duration
ii) Laeven, L. and Valencia, F. (2013)			4 019	4 019	0	1	0	0	duration
iii) Laeven, L. and Valencia, F. (2018)			1 070	357	1	1	0	1	duration/beginning
iv) Laeven, L. and Valencia, F. (2020)			0	0	1	1	0	1	duration/beginning
lo Duca et al. (2017)			268	89	1	1	0	1	duration
Reinhart, C. M. and Rogoff, K. S. (2009)	1		12 725	12 725	1	0	0	0	duration
Reinhart, C. M. and Rogoff, K. S. (2011)		4	2 937	979	0	1	1	1	duration
Schularick, M. and Taylor, A.M. (2012)		2	3 500	3 500	0	1	0	0	beginning
Benchmark									
Caprio, G. and Klingebiel, D. (2002)		3	1 201	1 201	0	1	0	0	duration

Table 5: Ranking of Databases - Relative Citations

	Ranking - Top 3		**Expected #**	**Citations per**	**Type of crisis**				**Crisis**
	Fiscal	**Banking**	**of citations**	**crisis type**	**Fiscal**	**Banking**	**Financial**	**Currency**	**specification**
Babecky, J. et al. (2014)			54	18	1	1	0	1	duration
Baldacci, E. et al. (2011)	3		51	51	1	0	0	0	duration
Baron, M. et al. (2018)*			11	11	0	1	0	0	beginning
Gerling, K. et al. (2017)			18	18	1	0	0	0	duration
Jorda, O. et al.			282	282	0	0	1	0	beginning
Laeven, L. and Valencia, F.	2	2	1 424	1 289					
i) Laeven, L. and Valencia, F. (2008)			17	17	0	1	0	0	duration
ii) Laeven, L. and Valencia, F. (2013)			1 204	1 204	0	1	0	0	duration
iii) Laeven, L. and Valencia, F. (2018)			203	68	1	1	0	1	duration/beginning
iv) Laeven, L. and Valencia, F. (2020)			0	0	1	1	0	1	duration/beginning
lo Duca et al. (2017)			48	16	1	1	0	1	duration
Reinhart, C. M. and Rogoff, K. S. (2009)	1		3 016	3 016	1	0	0	0	duration
Reinhart, C. M. and Rogoff, K. S. (2011)		4	645	215	0	1	1	1	duration
Schularick, M. and Taylor, A.M. (2012)		3	885	885	0	1	0	0	beginning
Benchmark									
Caprio, G. and Klingebiel, D. (2002)		1	1 937	1 937	0	1	0	0	duration

Because some of the databases collect information on different types of crises (banking, fiscal, financial, currency), we calculate the average number of citations per crisis type by dividing the total number of expected citations by the number of types of crises included. This is to account for the fact that wider coverage in terms of types of crisis typically results in a higher overall number of citations. Consequent empirical analysis will be based on the 'citations per crisis type' indicator.

As is apparent in Tables 4 and 5, the three most prominent databases in the case of fiscal crises include Reinhart and Rogoff (2009), Laeven and Valencia (2008) and Baldacci et al. (2011). For banking crises, Schularick and Taylor (2012) and Laeven and Valencia (2008) are accompanied by Caprio and Klingebiel (2002). The relative measure of database prominence (Table 5) benefits the position of Caprio and Klingebiel (2002) in the ranking of banking crises by attributing greater relevance to the relative importance of the paper given the popularity of the topic in the years after its publication. At its peak in 2006 (Table A1), Caprio and Klingebiel (2002) was cited 95 times, with only 281 papers having mentioned the term 'financial crisis', i.e., a share of 33.81 per cent. This exceeds the performance of the most-famous Reinhart and Rogoff (2009) paper in 2012 with 1008 citations per 3168 papers mentioning the 'financial crisis' term, i.e., a share of 31.82 per cent.

The gap between the top three prominent studies and the rest of the group in both categories is noticeable. As shown in Hamermesh (2018), the distribution of citations across papers in economics is highly skewed, with relatively few articles accounting for the overwhelming majority of citations. We observe a similar pattern among our group of papers, with the top three papers accounting for more than 98 (84) per cent of the total expected number of citations in the fiscal (banking) crisis category.

The most prominent databases present an interesting mixture of heterogeneous approaches based on the taxonomy discussed in Section 2. Two databases on banking crises (Laeven and Valencia, 2008; Caprio and Klingebiel, 2002) have global country coverage and medium time spans. In terms of crisis event specification, both employ a qualitative approach ranging from semi-qualitative (Laeven and Valencia, 2008) to highly qualitative (Caprio and Klingebiel, 2002).[5] The paper by Schularick and Taylor (2012) uses semi-qualitative event assessment but focuses on only a selective sample of advanced economies that are observed over a very long time span. Similar in time coverage but with an extensive number of countries covered is the database by Reinhart and Rogoff (2009). For the definition of sovereign crisis events, both Reinhart and Rogoff (2009) and Laeven and Valencia (2008) base their assessment on a qualitative definition of debt restructuring. Baldacci et al. (2011) work similarly with the global sample as Laeven and Valencia (2008) and combine quantitative with qualitative assessment of fiscal stress periods. Two databases identify only the beginning of a crisis event (Laeven and Valencia (2008) for sovereign and currency crises, Schularick and Taylor (2012)), and three databases also provide additional information on their duration (Baldacci et al., 2011; Reinhart and Rogoff, 2009; Caprio and Klingebiel, 2002).

For the purpose of the following empirical analysis, we decided to exclude paper by Schularick and Taylor (2012) and replace it with Reinhart and Rogoff (2009) despite its high prominence among the banking crisis databases. This is due to the very narrow country coverage that would preclude us from assessing the robustness of our findings conditional on the choice of heterogeneous subsamples. Smaller overlap with other databases might also render our findings less robust.

[5]There is no strict distinction between semi-qualitative and highly qualitative approaches. We comment on this categorization, ironically, using our expert judgment.

4.2 Stylized Facts on Crisis Databases

4.2.1 What Can We See from the Fiscal Databases?

When assessing the available databases, we can obtain several findings. Intuitively, we can assume that the number of identified crises in the literature correlates with the number of investigated years and countries in the sample. This partially holds for our vintage of the three papers under our review that contain fiscal crisis databases, where Reinhart and Rogoff (2009) has the highest number of identified fiscal crises (1516 identified crisis years). This paper compiled and reviewed larger samples than the other two contributions. However, the paper with the lowest number of identified crises (Laeven and Valencia, 2008) is surprisingly not the one with the smallest sample under investigation (Baldacci et al., 2011).

On the one hand, Laeven and Valencia (2008) consider a sovereign debt crisis to be an event when public debt payments were suspended or when public debt was restructured without a suspension of payments. A similar approach is also adopted by Reinhart and Rogoff (2009), but their database also includes both externally and internally fuelled sovereign defaults. On the other hand, Baldacci et al. (2011) relax the definition beyond the possibility of debt default or restructuring (implicit or explicit), as they also consider for fiscal event crises the recourse to exceptional official financing and fiscal distress episodes that are severe enough to alter the attainment of macroeconomic stability and growth but do not result in defaults or near defaults.[6]

Although the sovereign debt crisis definitions in Laeven and Valencia (2008) and Reinhart and Rogoff (2009) exhibit several similarities (e.g., measurement based on an episode of sovereign debt default and restructuring), the sources of information in the two papers differ. Laeven and Valencia (2008) primarily base their database on Beim and Calomiris (2001), Bank (2002), Sturzenegger and Zettelmeyer (2006) and IMF staff reports, while Reinhart and Rogoff (2009) depart from Sarkees and Schafer (2000) and Maddison (2003).

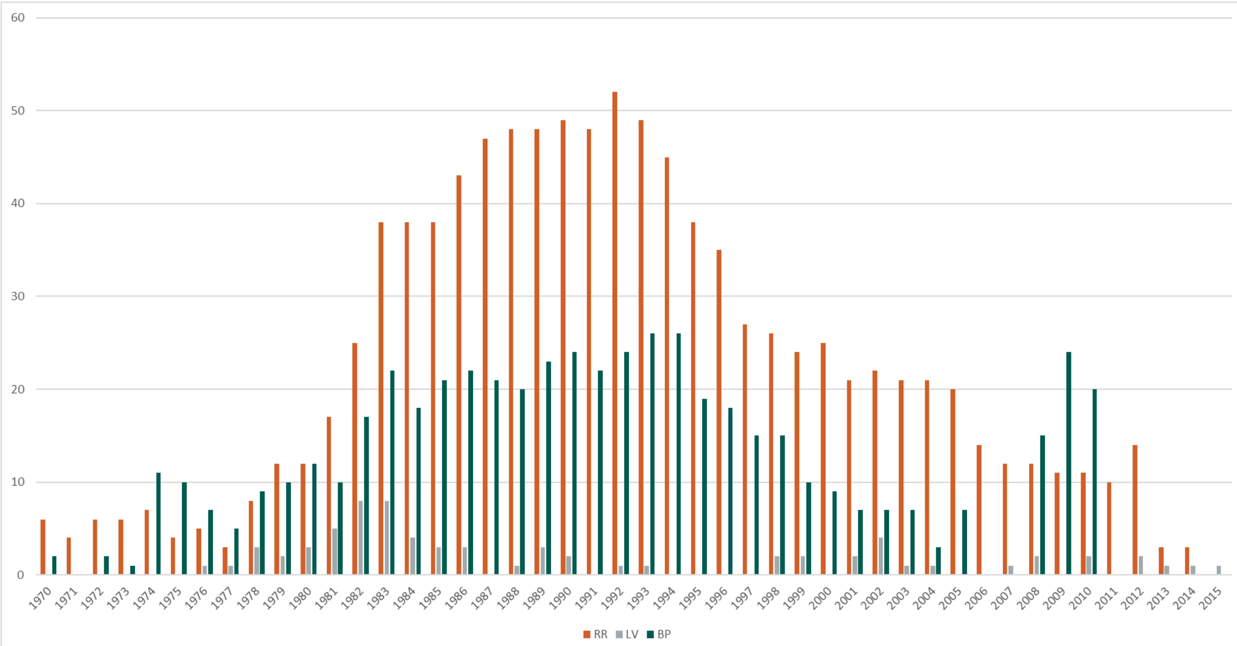
Reinhart and Rogoff (2009) report 1028 years of fiscal crises, Baldacci et al. (2011) identify 541 crisis years, and finally Laeven and Valencia (2008) report 71 years of fiscal crises. However, as noted above, Laeven and Valencia (2008) refer only to the starting year and not the duration of the crisis.

As visible in Figure 2, the 1980s and 1990s were periods with a high frequency of fiscal crisis occurrence. This period is famous for oil price shocks and the collapse of a number of centrally planned economies formed in the former socialist block of countries. Of course, these events do not explain the whole increase in the number of fiscal crises during this period but indisputably contributed to market stress and problems in the sovereign sector during that time. The papers Reinhart and Rogoff (2009) and (Baldacci et al., 2011) are quite consistent in their conclusion that the worst situation was during the years 1992-1994, which are the years with the most frequent occurrence of fiscal crises since 1970. Contrary, (Laeven and Valencia, 2008) finds the most crisis events at the beginning of the 1980s. After moderation of the situation at the beginning of the new millennium, when the number of fiscal crises decreased by more than half of their previous peak, the situation deteriorated after 2008 again due to the occurrence of world financial crises that took a toll on and had negative implications for the sovereign sector. The situation culminated in 2010 – 2012, when especially in Europe, the stress on sovereign debt financing increased in a number of countries due to doubts

[6] See Baldacci et al. (2011), p. 7.

about the sustainability of debt financing. However, the number of years with fiscal crises has not returned to its peak levels of the early 1990s in any paper.

Figure 2: Number of identified fiscal crises during the period of years 1970 − 2014 in the respective papers



Notes: RR denotes Reinhart and Rogoff (2009), LV denotes Laeven and Valencia (2008), BP denotes Baldacci et al. (2011)

When examining the fiscal crises that were identified in each of three reviewed papers (a crisis identified in the same year and country), the number of crises markedly shrinks to 24. The highest intersection of fiscal crises among all three papers is at the beginning of the 1980s, when four crises were identified in 1982 and five crises were identified in 1983. This fact is interesting since the two papers Reinhart and Rogoff (2009) and (Baldacci et al., 2011) find that the most crisis periods occurred in the early 1990s. However, this was not the case for (Laeven and Valencia, 2008), who also investigated the same countries where the first two papers identified fiscal crisis events but did not identify those fiscal events. The explanation for this probably stems from the fact, which will be explained later, that while Reinhart and Rogoff (2009) applied a quite arbitrary methodology for crisis detection based on incidence of certain events, such as bank runs,Laeven and Valencia (2008) imposed precise criteria with certain limits that were set in a strict way.

When we observe a low number of identified fiscal crises in common by all three studies under review, we understand that such a criterion is probably truly strict. Therefore, we turn to a softer rule and attempt to compare common crisis events that were identified commonly by at least two of the three papers under review.

After relaxing the rule to just two papers out of three that agreed on a crisis event identification, we see that the period with a higher occurrence of fiscal crises was during the 1980s and 1990s, while the new millennium brought relaxation of fiscal stress and crisis episodes in general. The peak period of fiscal crisis events was at the turn of 1980s and early 1990s. Surprisingly, an increase in the number of fiscal crises was not visible after 2008, when the global financial crisis arrived. This might be the result of an effort that

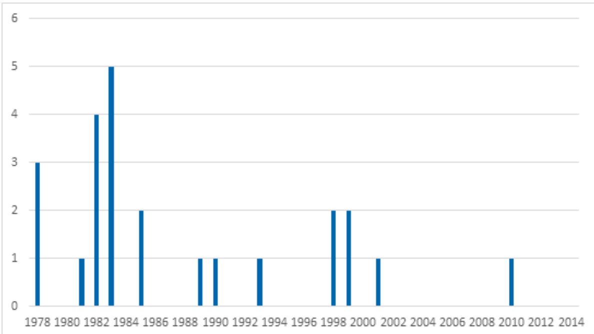


Figure 3: Number of fiscal crises identified in all three papers under review in respective years

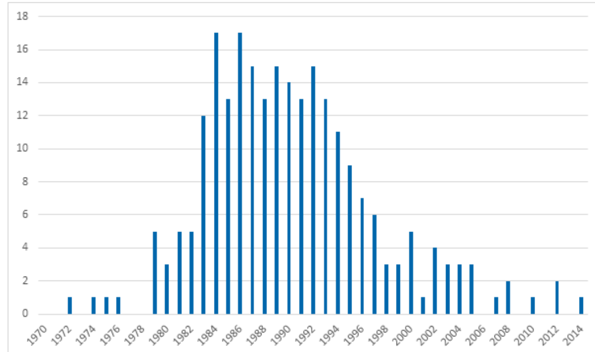


Figure 4: Number of fiscal crises identified in at least two papers under review in respective years

appeared in some parts of the world after the new millennium to implement more prudent fiscal policies (e.g., the Stability and growth pact in the EU), but it is definitely also the result of the implementation of non-standard monetary policy measures, such as quantitative easing, which undoubtedly helped some governments avoid bankruptcy.

4.2.2 What Can We See from the Banking Databases?

Among the banking crisis-related papers, we focus on Reinhart and Rogoff (2009), Laeven and Valencia (2008) and Caprio and Klingebiel (2002).

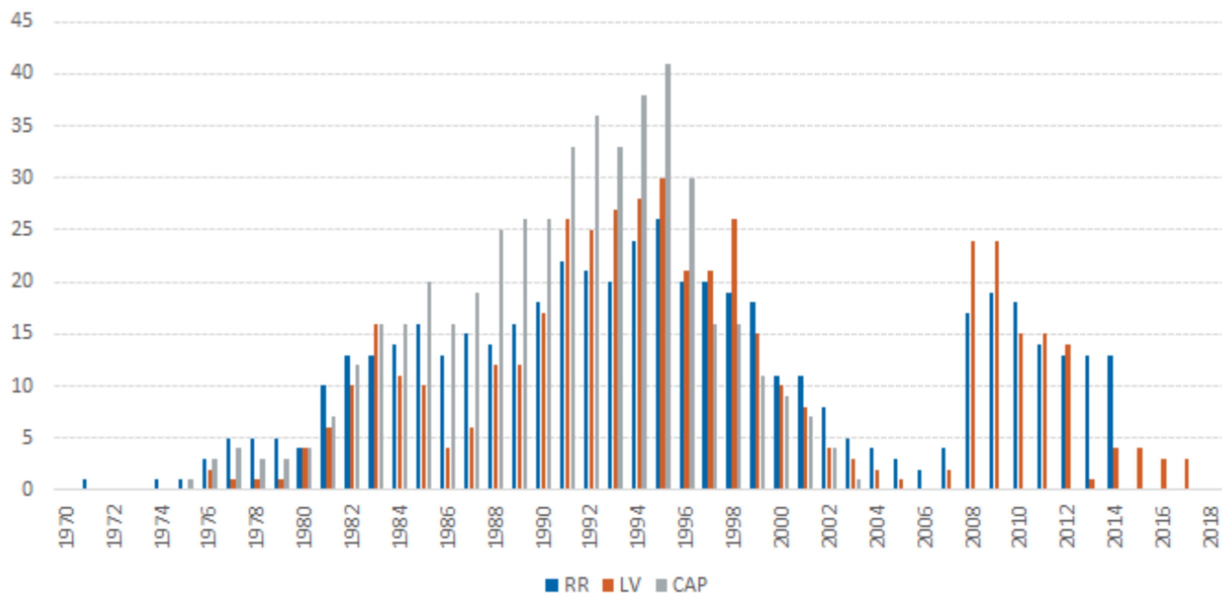
The picture of the frequency of banking crises is far more consistent across the databases under review than was the case for fiscal crises. This result is quite interesting since the definition of a crisis event is not entirely consistent across all three papers. Laeven and Valencia (2008) and Reinhart and Rogoff (2009) based their definition on the occurrence of bank runs, losses in the banking system and bank failures, as well as government assistance and financial injections into banking institutions. In contrast with these papers, Caprio and Klingebiel (2002) based their definition on net worth and as a banking crisis event consider an event when the net worth of the banking system is almost or entirely eliminated. Due to this more mechanical approach to crisis definition, Caprio and Klingebiel (2002) identifies that the majority of the years under review have a higher number of banking crises than the other two papers. However, drying up positive net worth likely triggers bad rumours and bank runs and leads to consequent losses and the need for government assistance, which probably results in coherent crisis event identification across all three papers at a higher rate, as one would expect.

Laeven and Valencia (2008) impose strict criteria with certain limits (bank restructuring fiscal costs of at least 3 per cent of GDP; a share of nonperforming loans above 20 per cent of total loans, etc.). However, by more closely analysing the imposed limits and caps, we can conclude that given limits were intentionally set at relatively high levels because the authors were primarily searching for systemic banking crises. A systemic banking crisis is characterized by a situation in which the banking system is under severe financial distress that requires significant policy intervention.

The same aim (i.e. identifying systemic banking crises), albeit from a different perspective, was also pursued by Caprio and Klingebiel (2002). They define the state of systemic banking crises as events when the net worth of the banking system is not positive, i.e., has been almost or entirely eliminated. Caprio and Klingebiel (2002) acknowledge that their approach is highly subjective and of a qualitative nature likely to cover only periods of severe disruption in the banking sector or, in some cases, borderline episodes of systemic banking crises.

Reinhart and Rogoff (2009) apply an arbitrary methodology for crisis detection based on the incidence of certain events (bank runs, closure, merger or takeover by the public sector, etc.). They consider bank runs on one or more financial institutions followed by policy intervention to define a systemic banking crisis. Additionally, a closure of one institution may spill over to the rest of the sector, which marks the start of a crisis period. In the database, major systemic crises are also accompanied by events of a less severe nature, thus increasing the total number of identified crisis periods. This database also extensively uses the Caprio and Klingebiel (2002) database as a source of the crisis data but enriches their database by adding an additional 112 crisis years (44 % of commonly identified years) in the reviewed sample of countries.

Figure 5: Number of identified banking crises during the period 1970 – 2018 in respective papers



Notes: RR denotes Reinhart and Rogoff (2009), LV denotes Laeven and Valencia (2008), CAP denotes Caprio and Klingebiel (2002)

The period with the most frequent occurrence of banking crises is, according to all three papers, the last decade of the 20th century. In particular, the first half of the 1990s was a markedly turbulent period with the highest number of identified banking crisis events in history. The number of banking crisis events during these years is even higher, as was the case during the global financial crisis (GFC, 2008-2010). The highest occurrence of banking crises at the beginning of the 1990s relates to the development of and increase in the interconnectedness of the financial system nearing the end of the 20th century. With the onset of digital technologies and globalization, the banking system has started to become far more interconnected and interdependent, as it had been decades before. However, with positive aspects of globalization and negative externalities, banking and financial market stress was far easier to transmit among countries. On the other hand, the prudent regulation of the national and international financial and banking system lagged far behind.

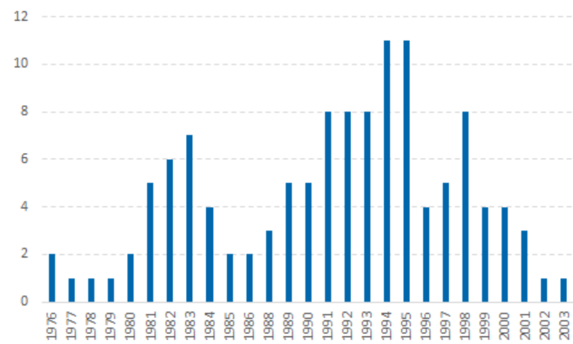


Figure 6: Number of banking crises identified in all three papers under review in respective years

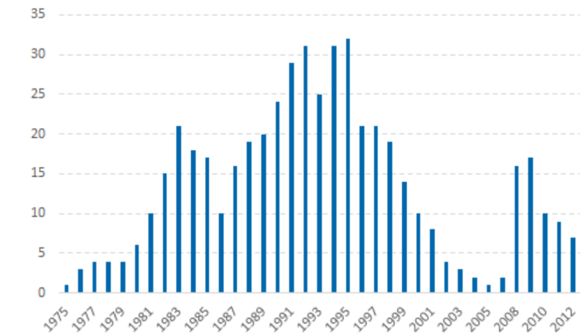


Figure 7: Number of banking crises identified in at least two papers under review in respective years

The first Basel Accord, known as Basel I, was issued only in 1988, and it merely imposed basic aspects of a prudent framework, predominantly targeting one aspect, the capital adequacy of financial institutions. The other aspects, such as liquidity, governance, disclosure policies, internal assessment processes and the financial stability of the banking system, were not directly considered. This issue arose with the subsequent adjustments of the Basel accords. Additionally, the number of countries implementing the Basel principles in their legislation was growing gradually. The extent of prudent legislative principles and the number of countries following the Basel principles is likely responsible for the fact that with the exception of the GFC, which was the largest crisis since the Great Depression and the largest one that hit the financial sector, the number of banking crises identified by all three papers under review was not as high as during the 1990s. This conclusion is consistent in both databases under review mapping this period.

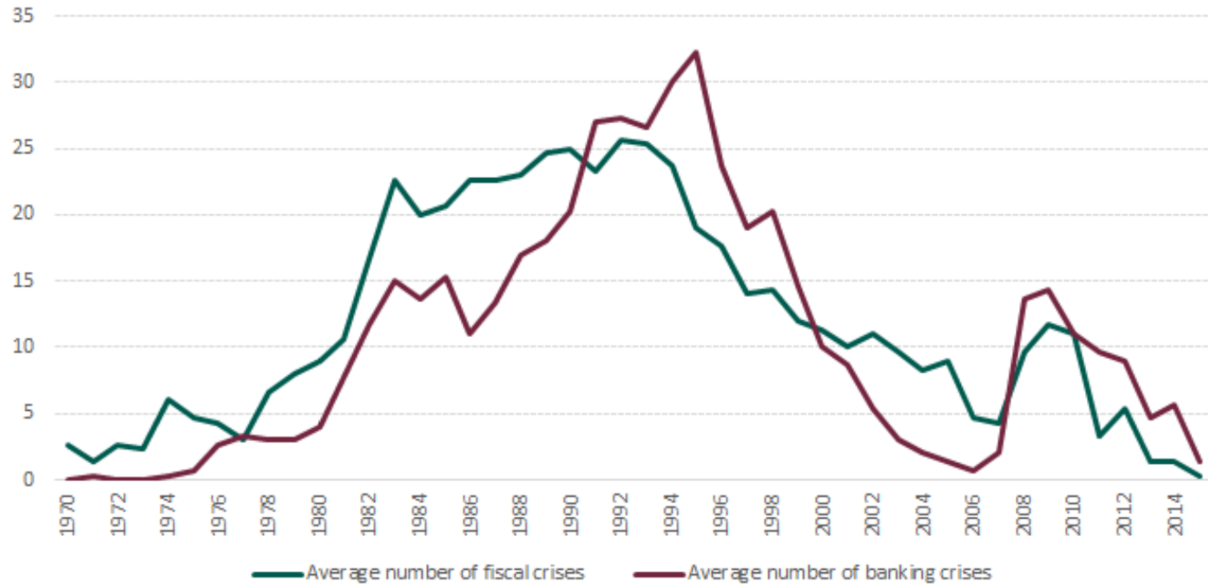
In contrast to fiscal crises, where Laeven and Valencia (2008) is only the beginning year type of the database, in banking crisis events, all three databases are duration-type databases. Therefore, the ratio of commonly identified banking crisis events by all three papers is higher than that for fiscal crises, reaching more than 25 % of crisis events identified by Laeven and Valencia (2008) and Caprio and Klingebiel (2002) and 20 % of crisis events identified by Reinhart and Rogoff (2009). Because Caprio and Klingebiel (2002) investigates the period ending in 2003, the wave of banking crises during the GFC was not captured. However, in contrast to chart 5, another wave of banking crisis episodes can be identified at the beginning of the 1980s by all three papers. This is something that we can also see in fiscal crisis episodes. While Reinhart and Rogoff (2009) and Caprio and Klingebiel (2002) did not record a significant decline in the number of crisis events during the 1980s, in Laeven and Valencia (2008), such a decline is quite visible. During the 1984-1986 period, Laeven and Valencia (2008) did not record 20 banking crisis events (altogether 37 crisis years) in the countries where the other two papers did. Most likely, the reason for this is arbitrariness in judgment regarding the consideration of some banking episodes, as Reinhart and Rogoff (2009) and Laeven and Valencia (2008) opted for quite similar definitions of a banking crisis episode.

Although the wave in the early 1980s is also visible when relaxing the rule and targeting the commonly identified crisis events by at least two out of three papers under review, its distinction from the wave of the 1990s is much less moderate. Additionally, the wave of banking crisis events connected to the GFC can be easily identified after 2008. However, due to the establishment of a more prudential regulatory framework, the number of banking crisis events was just half of that observed after 2008 at the beginning of the 1990s.

4.2.3 What is the Relation between Fiscal and Banking Crises?

According to data compiled from papers under review, there seems to be a strong correlation between the number of fiscal crises and banking crises (Figure 8). The correlation coefficient between the two datasets of fiscal and banking crisis events is equal to 0.83. This number is surprisingly high, close to an almost perfect positive correlation. This is in addition to the fact that not every banking crisis necessarily results in a sovereign debt crisis or a fiscal crisis and that not every fiscal crisis ends up in a systemic banking crisis.

Figure 8: Average number of fiscal and banking crises identified by papers under review



Notes: Number of crisis events calculated as the average of the databases under review.

The strong correlation between the banking and fiscal crisis datasets is surprising, since at the end of the 20th century and the beginning of the new millennium, various governments and public institutions across the world sought to develop strong independence between the banking and fiscal sectors. The reason for this was fear of the potential fiscal burden that sovereign sector would need to bear in the event of a banking crisis, when the need for fiscal injections or engagement of the sovereign sector could appear. The costs of a banking crisis would be in such a case borne by taxpayers, which was not politically acceptable. Various types of funds, such as deposit protection funds and retrieval facilities, were designed in the late 20th and early 21st centuries with the aim of placing the potential costs of banking crises outside the sovereign sector. As illustrated in the Figure 8 above, this policy was not fully successful. The sovereign sector has not remained undamaged by banking crises even in the new millennium. Additionally, the global financial crisis in 2008 and the subsequent onset of the sovereign debt crisis in 2011 in Europe showed that sovereign debt in countries where the banking sector was seriously affected by the global financial crisis was to a large extent exposed to costs of the banking sector crisis because the financial capacity of national protection funds and facilities was exhausted. As a consequence, new international facilities at the EU level were introduced with the aim of overcoming similar problems in the future.

4.3 Statistical Comparison

As part of our assessment of the databases, we analyse how consistent the most prominent databases from each sample (fiscal and banking crises) are in detecting crises. For this purpose, we construct pairwise consistency tables for each pair of databases summarizing the crisis and non-crisis periods detected by each of the databases. We also report relative and expected frequencies, as well as Cohen´s κ coefficient measuring the degree of agreement between the two databases. Table 6 summarizes aggregate pairwise results for the κ coefficients.

Table 6: Kappa coefficient statistics of comparison of the database in respective comparative approaches

Author	**Approach**	**Kappa**	**Author**	**Approach**	**Kappa**
fiscal crisis			*banking crisis*		
Reinhart and Rogoff (2009)	Simple	0.086	Reinhart and Rogoff (2011)	Simple	0.544
Laeven and Valencia (2018)	Strict	0.405	Laeven and Valencia (2018)	Strict	0.591
	Semi-strict	0.969		Semi-strict	0.987
Baldacci et al. (2011)	Simple	0.073	Caprio and Klingebiel (2002)	Simple	0.421
Laeven and Valencia (2018)	Strict	0.165	Laeven and Valencia (2018)	Strict	0.692
	Semi-strict	0.562		Semi-strict	0.893
Baldacci et al. (2011)	Simple	0.728	Caprio and Klingebiel (2002)	Simple	0.728
Reinhart and Rogoff (2009)	Strict	0.360	Reinhart and Rogoff (2011)	Strict	0.690
	Semi-strict	0.851		Semi-strict	0.808

4.3.1 Banking Crises

For Laeven and Valencia (2008) and Reinhart and Rogoff (2009), the comparison using the basic approach shows only moderate agreement between the two databases (with κ just equal to 0.544 and a standard deviation of 0.025, yielding a 95% confidence interval of 0.519 to 0.568). In the strict approach, we obtain almost substantial agreement (with κ just below 0.6 and a standard deviation of 0.0448), and in the semi-strict approach, we observe almost perfect agreement with κ close to 1.

Table 7: Statistics of comparison of the databases Laeven and Valencia (2018) and Reinhart and Rogoff(2011) in respective comparative approaches

			RR (2009)								
			Basic aproach			**Strict approach**			**Semi-strict approach**		
			Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total
LV (2018)	Crisis years	**Absolute freq.**	**229**	**27**	**256**	**62**	**19**	**81**	**79**	**2**	**81**
		Relative freq.	0.077	0.009	0.086	0.021	0.006	0.027	0.027	0.001	0.027
		Expected freq.	0.015	0.071	-	0.001	0.026	-	0.001	0.027	-
	Non-crisis years	**Absolute freq.**	**283**	**2,426**	**2,709**	**62**	**2,822**	**2,884**	**0**	**2,884**	**2,884**
		Relative freq.	0.095	0.818	0.914	0.021	0.952	0.973	0.000	0.973	0.973
		Expected freq.	0.158	0.756	-	0.041	0.932	-	0.026	0.947	-
	Total	**Absolute freq.**	**512**	**2,453**	**2,965**	**124**	**2,841**	**2,965**	**79**	**2,886**	**2,965**
		Relative freq.	0.173	0.827	1.000	0.042	0.958	1.000	0.027	0.973	1.000
		Expected freq.	-	-	-	-	-	-	-	-	-
		Kappa			**0.544**			**0.591**			**0.987**
		Std. dev.			**0.025**			**0.045**			**0.009**

When overcoming the differences in the types of databases and comparing just the first year of identified banking crises (strict and semi-strict approaches), in the strict approach, the Laeven and Valencia (2008) and Reinhart and Rogoff (2009) databases agree on only 43.4 % of crises detected, what is the result of the heterogeneous definition of a crisis period in the two databases. This results in a situation where Laeven and Valencia (2008) identifies only two-thirds of crisis events identified by Reinhart and Rogoff (2009). The events were detected at an earlier stage under the more relaxed crisis criteria applied by Reinhart and Rogoff (2009). This strong heterogeneity, however, diminishes when we apply the semi-strict criterion, where the rate of commonly identified crisis events increases to more than 97%.

Table 8: Statistics of comparison of the databases Caprio and Klingebiel (2002) and Laeven and Valencia (2008) in respective comparative approaches

			CK (2002)								
			Basic aproach			Strict approach			Semistrict approach		
			Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total
LV (2018)	Crisis years	**Absolute freq.**	**241**	**115**	**356**	**89**	**33**	**122**	**99**	**23**	**122**
		Relative freq.	0.057	0.027	0.085	0.021	0.008	0.029	0.024	0.005	0.029
		Expected freq.	0.013	0.072	-	0.001	0.028	-	0.001	0.028	-
	Non-crisis years	**Absolute freq.**	**399**	**3,455**	**3,854**	**43**	**4,045**	**4,088**	**0**	**4,088**	**4,088**
		Relative freq.	0.095	0.821	0.915	0.010	0.961	0.971	0.000	0.971	0.971
		Expected freq.	0.114	0.858	-	0.030	0.941	-	0.023	0.948	-
	Total	**Absolute freq.**	**640**	**3,570**	**4,210**	**132**	**4,078**	**4,210**	**99**	**4,111**	**4,210**
		Relative freq.	0.152	0.848	1.000	0.031	0.969	1.000	0.024	0.976	1.000
		Expected freq.	-	-	-	-	-	-	-	-	-
		Kappa			**0.421**			**0.691**			**0.893**
		Std. dev.			**0.024**			**0.035**			**0.022**

Despite the dissimilar definitions of systemic banking crises in Caprio and Klingebiel (2002) and Laeven and Valencia (2008), the results are not as heterogeneous as one would expect. In the strict approach, the rate of agreement of the databases is 53.9 % of crises detected in both databases. In this approach, the κ coefficient equals 0.6915 (with a standard deviation of 0.0351, yielding a 95%-confidence interval of 0.727 to 0.656), which is close to substantial agreement. However, in the semi-strict approach, the κ value increases to 0.8932 (with a standard deviation of 0.0222, yielding a 95% confidence interval of 0.915 to 0.871), which qualifies as in almost perfect agreement. In this respect, the two databases seem to be quite uniform in the metric of crisis detection, although they apply heterogeneous criteria for the task.

Table 9: Statistics of comparison of the databases Reinhart and Rogoff (2011) and Caprio and Klingebiel (2002) in respective comparative approaches

			RR (2011)								
			Basic aproach			Strict approach			Semi-strict approach		
			Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total
CK (2002)	Crisis years	**Absolute freq.**	**250**	**33**	**283**	**63**	**13**	**76**	**71**	**0**	**71**
		Relative freq.	0.136	0.018	0.154	0.033	0.007	0.040	0.037	0.000	0.037
		Expected freq.	0.031	0.124	-	0.002	0.038	-	0.002	0.035	-
	Non-crisis years	**Absolute freq.**	**112**	**1,437**	**1,549**	**40**	**1,796**	**1,836**	**32**	**1,809**	**1,841**
		Relative freq.	0.061	0.784	0.846	0.021	0.939	0.960	0.017	0.946	0.963
		Expected freq.	0.167	0.678	-	0.052	0.909	-	0.052	0.911	-
	Total	**Absolute freq.**	**362**	**1,470**	**1,832**	**103**	**1,809**	**1,912**	**103**	**1,809**	**1,912**
		Relative freq.	0.198	0.802	1.000	0.054	0.946	1.000	0.054	0.946	1.000
		Expected freq.	-	-	-	-	-	-	-	-	-
		Kappa			**0.728**			**0.690**			**0.808**
		Std. dev.			**0.022**			**0.042**			**0.034**

Interesting results also stem from the comparison of duration-type databases by Caprio and Klingebiel (2002) and Reinhart and Rogoff (2009). In the basic approach, the consistency between the two databases is quite high, with κ reaching the 0.728 level (and standard deviation of 0.0217, yielding a 95% confidence interval of 0.750 to 0.706), which represents substantial agreement between the two databases. This outcome demonstrates the significant rate of similarity in the classification of crisis years by both databases.

The two databases are consistent not only in crisis length but also in the identification of the beginning of crises, which illustrates the fact that the κ value in the strict approach is close to the κ value in the basic approach. This does not change even when applying the more relaxed definition imposed in the semi-strict approach, where the κ value increases only moderately to 0.808, which is on the border between substantial and almost perfect agreement. Despite that the two databases have some differences in their crisis definitions, they deliver quite consistent results.

4.3.2 Fiscal Crises

As noted in Section 4.1, the most prominent databases according to ranking in the fiscal field are Reinhart and Rogoff (2009), Laeven and Valencia (2008), and Baldacci et al. (2011), in addition to Schularick and Taylor (2012). The number of identified fiscal crisis events commonly detected by Reinhart and Rogoff (2009) and Laeven and Valencia (2008) is smaller than the number of commonly detected banking crisis events (41 fiscal events vs 62 banking events in the strict approach and 63 vs 79 in the semi-strict approach). The number of commonly identified fiscal crises is smaller, although both papers detected a higher number of fiscal than banking crisis events. The smaller coincidence in commonly identified fiscal crises is probably the result of an arbitrary rule for sovereign debt crisis detection in both papers (see the discussion in Section 4.2.1).

Table 10: Statistics of comparison of the databases Laeven and Valencia (2008) and Reinhart and Rogoff(2009) in respective comparative approaches

			RR (2009)								
			Basic aproach			**Strict approach**			**Semistrict approach**		
			Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total
LV (2018)	Crisis years	**Absolute freq.**	**58**	**922**	**980**	**41**	**89**	**130**	**63**	**0**	**63**
		Relative freq.	0.012	0.196	0.209	0.009	0.019	0.028	0.013	0.000	0.013
		Expected freq.	0.003	0.206	-	0.000	0.027	-	0.000	0.013	-
	Non-crisis years	**Absolute freq.**	**9**	**3,711**	**3,720**	**26**	**4,544**	**4,570**	**4**	**4,633**	**4,637**
		Relative freq.	0.011	0.780	0.791	0.006	0.967	0.972	0.001	0.986	0.987
		Expected freq.	0.168	0.805	-	0.014	0.958	-	0.014	0.973	-
	Total	**Absolute freq.**	**67**	**4,633**	**4,700**	**67**	**4,633**	**4,700**	**67**	**4,633**	**4,700**
		Relative freq.	0.014	0.986	1.000	0.014	0.986	1.000	0.014	0.986	1.000
		Expected freq.	-	-	-	-	-	-	-	-	-
		Kappa			**0.086**			**0.405**			**0.969**
		Std. dev.			**0.027**			**0.055**			**0.016**

After abstracting from the length of crises and assessing only the detection of beginning year of the fiscal crises (strict approach), the commonality in the detection of fiscal crisis events is relatively weak with a κ coefficient equal to 0.4051 (and a standard deviation of 0.0548, yielding a 95% confidence interval of 0.350 to 0.460) that is upper limit of just fair agreement. Only 26 % of crisis episodes detected in both papers were commonly identified. However, this is probably because the papers under review detect a fiscal crisis event at different stages.

Once we apply the semi-strict approach, the rate of agreement increases significantly to 94 % of commonly identified fiscal crises detected by both papers, and the κ coefficient equals 0.966 (with a standard deviation of 0.024, yielding a 95% confidence interval of 0.99 to 0.942). This rate of agreement is already very close to (or even higher than) those that we observe in the banking crisis event comparison. Thus, the two databases agree that a fiscal crisis has occurred (high κ value in semi-strict approach) but are quite heterogeneous in the detection of the first year of an event (relatively small κ value in the strict approach).

The fact that the overlap in fiscal crisis detection is much less pronounced than the overlap when using the first year of a fiscal crisis is even clearer when analysing databases by Laeven and Valencia (2008) and Baldacci et al. (2011). The two databases agree on just one-tenth of fiscal crisis events (18 out of 180 crises) under a strict approach. The Cohen´s κ value is very low and equals 0.165 (with a standard deviation of 0.064, yielding a 95% confidence interval of 0.101 to 0.228), which represents just slight agreement. However, when analysing overlap using the semi-strict approach, the rate of agreement between the two databases increases significantly to more than 40%, with κ equal to 0.562 (and a standard deviation of 0.081, yielding a 95% confidence interval of 0.481 to 0.642). This κ value falls into the moderate agreement range. However, we can see that the degree of agreement is significantly lower, as observed in the previous reviews.

Table 11: Statistics of comparison of the databases Laeven and Valencia (2008) and Baldacci et al. (2011) in respective comparative approaches

			BP (2011)								
			Basic aproach			**Strict approach**			**Semistrict approach**		
			Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total
LV (2018)	Crisis years	**Absolute freq.**	**27**	**4**	**31**	**18**	**13**	**31**	**19**	**12**	**31**
		Relative freq.	0.011	0.002	0.012	0.007	0.005	0.012	0.007	0.005	0.012
		Expected freq.	0.003	0.010	-	0.001	0.011	-	0.000	0.012	-
	Non-crisis years	**Absolute freq.**	**514**	**2,016**	**2,530**	**149**	**2,381**	**2,530**	**17**	**2,513**	**2,530**
		Relative freq.	0.201	0.787	0.988	0.058	0.930	0.988	0.007	0.981	0.988
		Expected freq.	0.209	0.779	-	0.064	0.923	-	0.014	0.974	-
	Total	**Absolute freq.**	**541**	**2,020**	**2,561**	**167**	**2,394**	**2,561**	**36**	**2,525**	**2,561**
		Relative freq.	0.211	0.789	1.000	0.065	0.935	1.000	0.014	0.986	1.000
		Expected freq.	-	-	-	-	-	-	-	-	-
		Kappa			**0.073**			**0.165**			**0.561**
		Std. dev.			**0.036**			**0.064**			**0.081**

A very similar conclusion can also be drawn from the comparison of the two duration types of databases collected by Reinhart and Rogoff (2009) and Baldacci et al. (2011). From the beginning, it is clear that both databases have a significant rate of agreement when the Cohen´s κ coefficient equals 0.728 under the simple approach (the standard deviation equals 0.036, yielding a 95 % confidence interval of 0.75 to 0.706). This falls into the range of substantial agreement, which is not surprising given the fact that the rate of commonly identified crisis years attains a level of 53 % (277 years out of 525 years). However, the two databases agree on crisis periods rather than on the beginning of crisis events.

Regarding the first year of crisis events, the rate of agreement declines significantly, with Cohen´s κ reaching 0.36, which is in the range of just fair agreement. This is because only 24 % of the initial years of fiscal events are identified by both databases in the same year (40 out of 168 fiscal events). However, when we extend the assessment criterion to the semi-strict approach, the rate of agreement between the two databases rises again to almost perfect agreement with a Cohen´s κ value of 0.851 (the standard deviation equals 0.036, yielding a 95% confidence interval of 0.881 to 0.814).

Generally, we can state that it is far more difficult to reach agreement on the initial year of a crisis event. However, if we relax the coincidence criterion (i.e., move to the semi-strict approach), the coincidence of crisis detection can increase significantly. What criterion is applied matters in the sense that various databases may not be as heterogeneous in their results as it appears at first glance. They can identify the same crisis event at different stages of the crisis due to the different detection criteria applied in the respective papers.

Table 12: Statistics of comparison of the databases Reinhart and Rogoff (2009) and Baldacci et al. (2011) in respective comparative approaches

			RR (2009)								
			Basic aproach			Strict approach			Semistrict approach		
			Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total	Crisis years	Non-crisis years	Total
BP (2011)	Crisis years	**Absolute freq.**	**277**	**37**	**314**	**40**	**23**	**63**	**47**	**16**	**63**
		Relative freq.	0.122	0.016	0.138	0.018	0.010	0.028	0.021	0.007	0.028
		Expected freq.	0.030	0.108	-	0.002	0.026	-	0.001	0.027	-
	Non-crisis years	**Absolute freq.**	**211**	**1,752**	**1,963**	**105**	**2,109**	**2,214**	**0**	**2,214**	**2,214**
		Relative freq.	0.093	0.769	0.862	0.046	0.926	0.972	0.000	0.972	0.972
		Expected freq.	0.185	0.677	-	0.062	0.910	-	0.020	0.952	-
	Total	**Absolute freq.**	**488**	**1,789**	**2,277**	**145**	**2,132**	**2,277**	**47**	**2,230**	**2,277**
		Relative freq.	0.214	0.786	1.000	0.064	0.936	1.000	0.021	0.979	1.000
		Expected freq.	-	-	-	-	-	-	-	-	-
		Kappa			**0.728**			**0.360**			**0.851**
		Std. dev.			**0.022**			**0.055**			**0.037**

4.4 Predictive Properties of EWIs

We investigate the predictive properties of individual EWIs using a balanced set of countries and years across selected databases. We include 59 countries from 1970-2003 and 67 countries from 1975-2003 for fiscal and banking crises, respectively.

For fiscal crisis events, the lowest number of identified events is found in Laeven and Valencia (2008), two times fewer than in Reinhart and Rogoff (2009) and four times fewer than in Baldacci et al. (2011). Interestingly, both databases (Reinhart and Rogoff, 2009; Baldacci et al., 2011) are relatively successful at identifying crisis events listed in Laeven and Valencia (2008); however, they add additional new crisis events.

Similarly, in the case of banking crises, the most selective database is by Laeven and Valencia (2008), followed by Caprio and Klingebiel (2002) and Reinhart and Rogoff (2009). Interestingly, even with this relatively vague definition based on expert judgment, Caprio and Klingebiel (2002) still identify more than 30% of the crisis events identified by Laeven and Valencia (2008).

A comparison of the average marginal effects from individual bi-variate logit models, as specified in equation 10, is provided in Table 13 for fiscal events and Table 14 for banking crisis events. As the results show, even the substantial synchronization across databases demonstrated by the κ coefficients does not result in perfect overlap in terms of predictive properties of individual EWIs. In contrast, there exist very few instances (current account balance and private sector debt in the fiscal databases) where a uniform positive agreement is formed across all three databases. For the negative confirmation, i.e., no effect of EWIs on predicted events, the agreement is reached more easily, with several EWIs being identified (NIIP, government debt, and HPI for fiscal crises and NIIP, government debt, private debt, and HPI for banking crises).

In general, no apparent pattern in commonality is observed, as no two databases are more similar in terms of the selection of statistically (in-)significant EWIs.

Table 13: Predictive Features of EWIs for Fiscal Crises with Strict and Semi-strict Event Specification - Average Marginal Effects

	Baldacci et al. (2011)		Laeven and Valencia (2018)		Reinhart and Rogoff (2009)			
Variable	AME: Strict	AME: Semi-strict	AME: Strict	AME: Semi-strict	AME: Strict	AME: Semi-strict	Countries	# of obs.
CA	**-0.801*** **	**-0.811*** **	**-0.308** **	**-0.298** **	**-0.490*** **	**-0.480*** **	58	1,221
	(0.000)	**(0.000)**	**(0.011)**	**(0.014)**	**(0.003)**	**(0.003)**		
EMS	0.007	0.007	**0.008***	0.008	**0.017*** **	**0.017*** **	59	1,402
	(0.605)	(0.579)	**(0.097)**	(0.103)	**(0.004)**	**(0.005)**		
REER	-0.016	-0.016	0.012	0.012	-0.005	-0.004	41	804
	(0.668)	(0.668)	(0.122)	(0.122)	(0.688)	(0.716)		
NIIP	-0.023	-0.023	-0.005	-0.005	-0.006	-0.005	46	546
	(0.446)	(0.445)	(0.76)	(0.76)	(0.552)	(0.592)		
Gov. Debt	-0.029	-0.029	0.009	0.009	0.013	0.013	40	652
	(0.328)	(0.328)	(0.401)	(0.401)	(0.293)	(0.293)		
Pr. Debt	**-0.050*** **	**-0.051*** **	**-0.025** **	**-0.025** **	**-0.060*** **	**-0.060*** **	55	1,281
	(0.005)	**(0.004)**	**(0.039)**	**(0.040)**	**(0.005)**	**(0.007)**		
UR	-0.040	-0.040	-0.024	-0.024	0.047	0.043	59	707
	(0.761)	(0.762)	(0.722)	(0.722)	(0.463)	(0.457)		
HPI	-0.11	-0.11	-0.011	-0.011	0.000	0.000	22	480
	(0.35)	(0.35)	(0.709)	(0.709)	(0.757)	(0.757)		
FDI	-0.517	**-0.621***	-0.157	-0.154	**-0.524** **	**-0.506***	59	1,428
	(0.14)	**(0.083)**	(0.413)	(0.42)	**(0.049)**	**(0.056)**		

Notes: AME stands for estimated average marginal effect from individual bivarate panel logit models with random effects as specified in [10]. P-values are in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level. CA stands for current account balance, EMS for export market share, REER for real effective exchange rate, NIIP for net international investment position, Gov. Debt for general government debt to GDP ratio, Pr. Debt for private sector debt to GDP ratio, UR for unemployment rate, HPI for real house price index, FDI for foreign direct investments.

Table 14: Predictive Features of EWIs for Banking Crises with Strict and Semi-strict Event Specification - Average Marginal Effects

	Caprio and Klingebiel (2002)		Laeven and Valencia (2018)		Reinhart and Rogoff (2011)			
Variable	AME: Strict	AME: Semi-strict	AME: Strict	AME: Semi-strict	AME: Strict	AME: Semi-strict	Countries	# of obs.
CA	-0.071	-0.073	-0.150	**-0.156***	-0.116	-0.146	62	1,335
	(0.535)	(0.526)	(0.109)	**(0.094)**	(0.352)	(0.237)		
EMS	**0.023*****	**0.023*****	**0.020*****	**0.020*****	0.011	0.012	66	1,606
	(0.01)	**(0.01)**	**(0.006)**	**(0.006)**	(0.368)	(0.29)		
REER	**0.022****	**0.022****	0.008	0.008	**0.025****	**0.026****	49	1,016
	(0.035)	**(0.033)**	(0.511)	(0.550)	**(0.028)**	**(0.022)**		
NIIP	0.010	0.010	0.017	0.016	-0.016	-0.017	49	599
	(0.669)	(0.669)	(0.438)	(0.46)	(0.489)	(0.468)		
Gov. Debt	-0.003	-0.004	-0.011	-0.011	-0.001	-0.003	41	762
	(0.897)	(0.853)	(0.56)	(0.58)	(0.978)	(0.908)		
Pr. Debt	-0.004	-0.004	-0.012	-0.012	-0.017	-0.017	62	1502
	(0.733)	(0.714)	(0.214)	(0.219)	(0.149)	(0.153)		
UR	**-0.326***	**-0.358***	-0.174	-0.174	**-0.584****	**-0.508****	67	780
	(0.094)	**(0.076)**	(0.271)	(0.271)	**(0.017)**	**-0.03**		
HPI	-0.142	-0.142	-0.106	**-0.140***	-0.016	-0.029	27	548
	(0.179)	(0.179)	(0.194)	**(0.097)**	(0.886)	(0.786)		
FDI	**-0.690****	**-0.514***	**-0.471***	**-0.460***	-0.305	-0.321	66	1,635
	(0.019)	**(0.064)**	**(0.062)**	**(0.073)**	(0.246)	(0.227)		

Notes: AME stands for estimated average marginal effect from individual bivarate logit probit models with random effects as specified in [10]. P-values are in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level. CA stands for current account balance, EMS for export market share, REER for real effective exchange rate, NIIP for net international investment position, Gov. Debt for general government debt to GDP ratio, Pr. Debt for private sector debt to GDP ratio, UR for unemployment rate, HPI for real house price index, FDI for foreign direct investments.

In the case of fiscal crises, Baldacci et al. (2011) and Laeven and Valencia (2008) seem to share higher commonality, at first sight, with three statistically significant EWIs, while for Reinhart and Rogoff (2009), we report four statistically significant EWIs. However, by relaxing the limits for confidence intervals by a slight margin in the case of Laeven and Valencia (2008), one additional EWI can be added to the list (REER). In general, the results are highly comparable in terms of selecting EWIs with no predictive power (NIIP, government debt, unemployment rate, house price index) or strong predictive properties (current account balance, private debt). Differences across databases are mostly observed when comparing EWIs related to the external sector (export market share, REER) or international capital flows (FDI). When comparing the findings in the semi-strict specification, Baldacci et al. (2011) shares a commonality with Reinhart and Rogoff (2009) for the FDI indicator, but this is not the case for Laeven and Valencia (2008).

The situation is highly diverse for banking crisis databases. While one can at least confirm with a relatively high level of confidence that two EWIs do not have sufficient predictive power in any case (government debt, NIIP), this is not true for the rest of the indicators. Caprio and Klingebiel (2002) and Laeven and Valencia (2008) report high statistical significance for the export market share indicator and lower statistical significance for the FDI inflows indicator, which are both refuted by Reinhart and Rogoff (2009). Contrary to the other two studies, the statistical significance of the REER and unemployment rate indicators is rejected by Laeven and Valencia (2008), who is the only contribution to report the importance of the current account balance in a semi-strict specification. By relaxing the limits on statistical confidence intervals or by a slight alteration of a dataset, private sector debt (Reinhart and Rogoff, 2009) and the house price index (Laeven and Valencia, 2008) may obtain higher predictive power.

As argued by van den Berg et al. (2008), rather than simply pooling all available data, the presence of more homogeneous country clusters should be investigated before setting up the logit model. In our case, working with smaller subsamples constructed according to membership in an economic club (EU28, OECD) or level of income (developed versus developing economies) only accentuates the issues of a small sample with few influential observations. Tables A3 and A4 show average marginal effects of individual EWIs using different subsamples with the semi-strict specification. While some of the findings from the full samples are echoed here (current account balance for OECD group), it is almost impossible to find a common unifying element with robust findings in all selected databases. If anything, it is much easier to eliminate the least reliable EWIs than to find robust evidence of the high predictive power of a single indicator. Even after relaxing confidence interval limits for banking cries, in most of the cases, we find only two-database overlap (e.g., HPI and FDI for OECD countries; REER, UR, and EMS for developing countries). A similar picture is found in the case of fiscal crises, with practically no systematic commonality present.

Contrary to expectations, the use of the semi-strict specification does not substantially improve predictive properties or lead to better commonality across databases. As discussed previously, the number of identified crisis periods varies significantly even in cases when the event specification is sufficiently close, e.g., Laeven and Valencia (2008) and Reinhart and Rogoff (2009) for systemic banking crises. Thus, the observed heterogeneity is likely to be driven more by the country dimension than by a possible time lag in the event specification. As noted by van den Berg et al. (2008), aggregating countries might not only lead to a loss of information but could also severely affect the estimation and inferences. This is because many determinants of crisis events are highly idiosyncratic and thus do not satisfy the homogeneity of parameters assumption. As a result, the presence of 'switchers' among the list of EWIs might reflect differences in crisis event specifications to which these particular EWIs strongly respond.

4.4.1 Influential crisis observations

As discussed in Section 4.3, the level of commonality across databases is comparably high after one accounts for a possible one-year identification lag. However, findings from probability model estimates (Section 4.4) suggest the presence of heterogeneity that likely exists due to the limited number of influential observations. In this step, we identify the most influential observations per estimation using Pregibon's delta-beta influence statistic (Pregibon, 1981).[7] An observation is declared influential if Pregibon's delta-beta influence statistic is larger than the sum of one standard deviation and the empirical mean. We report only those influential observations for which the crisis event was (non-)identified by one particular database; thus, it differs from the specification applied by others.

Table 15 and Table 16 report the most influential observations per estimates with given variable and database used in case of fiscal and banking crisis events.

Overall, the number of influential observations reflects the strictness of the identification strategy in the respective databases. While Laeven and Valencia (2008) report only two and Reinhart and Rogoff (2009) seven crisis events not shared by two other databases, Baldacci et al. (2011) identifies 57 events that are not recognized by other sources. As acknowledged by Baldacci et al. (2011), a higher number of identified periods stems from the use of a more comprehensive definition and utilization of additional sources of data (e.g., IMF-supported programs). Nevertheless, even with this substantial number of individually identified events, the findings of probability models (Table 13) are highly consistent. The impact of these single events may be translated into more pronounced effects of individual indicators, i.e., higher average marginal effects. In the most apparent case, the AME associated with the current account balance in Baldacci et al. (2011) is twice and thrice the size of the effect reported by Reinhart and Rogoff (2009) and Laeven and Valencia (2008), respectively. Due to the nature of this indicator (i.e., reflecting the accumulation of external imbalances), it is reasonable to hypothesize that the substantial worsening of the current account balance results in crisis episodes that ultimately force a country to negotiate IMF funding.

Second, heterogeneity in the identification strategy may have been more apparent if using different sets of indicators, especially those able to predict greater pressure on government bond yields. As a result, the remaining MIP-based EWIs may not be informative about the materialization of bond market pressure in a decisive manner.

Among the banking crisis databases, both Caprio and Klingebiel (2002) and Laeven and Valencia (2008) report few event periods that are not identified by their peers. In the case of Reinhart and Rogoff (2009), the most influential observations include six instances of systemic and twelve instances of non-systemic banking crises. To see how the inclusion of borderline episodes (Caprio and Klingebiel, 2002) and non-systemic events (Reinhart and Rogoff, 2009) changes the comparability across databases, we calculate the κ indicator and re-estimate the logit model, as in [10], for Reinhart and Rogoff (2009) with a subset of systemic-only crisis events. The results are available upon request.

As expected, the number of commonly identified banking crisis events decreased in both cases (Laeven and Valencia (2008) and Caprio and Klingebiel (2002)). However, the κ values do not change significantly

[7]Instead of the panel random effect logit model, we use a pooled data structure. Differences between the panel model with random effects and the pooled data model are indistinguishable.

Table 15: The Most Influential Observations - Fiscal Crises with Semi-strict Specification

Country	Year	Type	CA	EMS	REER	NIIP	GD	PRD	UR	HPI	FDI
						Reinhart and Rogoff (2009)					
Jamaica	1987		1	0	0	0	0	1	0	0	1
Nigeria	1987		1	1	1	0	0	1	0	0	1
Peru	1980		1	1	0	0	0	1	0	0	1
Sri Lanka	1981		1	1	0	0	0	1	0	0	1
Tunisia	1979		1	1	1	0	0	0	0	0	1
Venezuela	1990		1	1	1	1	0	1	0	0	1
Venezuela	1995		1	0	1	1	0	1	1	0	1
						Laeven and Valencia (2018)					
Brazil	1993		1	1	1	0	0	0	1	0	1
Russia	1998		1	1	1	1	0	1	1	0	1
						Baldacci et al. (2011)					
Bulgaria	1997	Market - Infl.	1	1	1	0	0	1	1	0	1
Ukraine	1992	Market - Infl.	0	0	0	0	0	0	1	0	0
Australia	1986	Market - Yields	0	1	1	0	0	1	0	1	1
Australia	1989	Market - Yields	0	1	1	0	0	1	0	1	1
Canada	1990	Market - Yields	1	1	1	1	1	1	0	1	1
Denmark	1982	Market - Yields	1	1	1	0	1	1	0	1	1
Finland	1990	Market - Yields	1	1	1	1	1	1	0	1	1
Finland	1992	Market - Yields	1	1	1	1	1	1	1	1	1
Greece	1993	Market - Yields	1	1	1	0	1	1	0	0	1
Malaysia	1998	Market - Yields	1	1	1	0	1	1	1	1	1
New Zealand	1985	Market - Yields	0	1	1	0	0	1	0	1	1
Norway	1986	Market - Yields	1	1	1	1	1	1	0	1	1
Poland	2001	Market - Yields	1	1	1	1	1	1	1	0	1
Sweden	1990	Market - Yields	1	1	1	1	1	1	0	1	1
Albania	1998	Official - IMF	1	1	0	0	1	0	1	0	1
Argentina	1998	Official - IMF	1	1	0	1	0	1	1	0	1
Brazil	1998	Official - IMF	1	1	1	0	0	1	0	0	1
Brazil	2001	Official - IMF	1	1	1	0	0	1	0	0	1
Colombia	1999	Official - IMF	0	1	1	1	1	1	1	1	1
Colombia	2003	Official - IMF	0	1	1	1	1	1	1	1	1
Egypt	1978	Official - IMF	0	1	0	0	1	1	0	0	1
Guatemala	1983	Official - IMF	1	1	0	0	0	1	0	0	1
Hungary	1982	Official - IMF	0	0	0	0	0	1	0	0	0
Hungary	1991	Official - IMF	0	0	0	0	1	1	0	0	0
India	1981	Official - IMF	1	1	0	0	0	1	0	0	1
Indonesia	1997	Official - IMF	1	1	0	0	0	1	1	0	1
Jordan	1996	Official - IMF	1	1	0	0	0	1	1	0	1
Kenya	1975	Official - IMF	0	0	0	0	0	1	0	0	0
Kenya	1979	Official - IMF	1	1	0	0	0	1	0	0	1
Kenya	1982	Official - IMF	1	1	0	0	0	1	0	0	1
Kenya	1988	Official - IMF	1	1	0	0	0	1	0	0	1
Korea	1983	Official - IMF	1	1	0	0	0	1	0	1	1
Korea	1997	Official - IMF	1	1	0	1	1	1	1	1	1
Mexico	1977	Official - IMF	0	1	0	0	0	0	0	0	1
Mexico	1995	Official - IMF	1	1	1	0	0	1	1	0	1
Mexico	1999	Official - IMF	1	1	1	0	0	1	1	0	1
Morocco	1980	Official - IMF	1	1	0	0	0	1	0	0	1
Pakistan	1988	Official - IMF	1	1	1	0	0	1	0	0	1
Pakistan	1994	Official - IMF	1	1	1	0	0	1	1	0	1
Pakistan	2001	Official - IMF	1	1	1	0	0	1	1	0	1
Panama	1980	Official - IMF	1	0	0	0	1	0	0	0	1
Philippines	1976	Official - IMF	0	1	0	0	1	1	0	0	1
Philippines	1980	Official - IMF	1	1	1	0	1	1	0	0	1
Philippines	1998	Official - IMF	1	1	1	0	1	1	1	0	1
Portugal	1983	Official - IMF	1	1	1	0	1	1	0	0	1
Sri Lanka	1991	Official - IMF	1	1	0	0	0	1	0	0	1
Sri Lanka	1993	Official - IMF	1	1	0	0	0	1	1	0	1
Sri Lanka	2003	Official - IMF	1	1	0	0	0	1	0	0	1
Thailand	1981	Official - IMF	1	1	0	0	0	1	0	0	1
Thailand	1985	Official - IMF	1	1	0	0	0	1	0	0	1
Thailand	1997	Official - IMF	1	1	0	1	0	1	1	1	1
Tunisia	1988	Official - IMF	1	1	1	0	0	0	0	0	1
Turkey	1999	Official - IMF	1	1	0	1	0	1	1	0	1
Turkey	2002	Official - IMF	1	1	0	1	1	1	0	0	1
El Salvador	1981	Official - rest.	1	1	0	0	0	1	0	0	1
India	1989	Official - rest.	1	1	0	0	0	1	0	0	1
Nigeria	2001	Official - rest.	1	1	1	0	0	1	1	0	1

Notes: '1' indicates that the observation was ranked among the most influential observations in bivariate logit model. The most influential observations have their Pregibon's delta-beta influence statistic higher than sum of one standard deviation and empirical mean from all observations. The CA stands for current account balance, EMS for export market share, REER for real effective exchange rate, NIIP for net international investment position, GD for general government debt to GDP ratio, PRD for private sector debt to GDP ratio, UR for unemployment rate, HPI for real house price index, FDI for foreign direct investments.

Table 16: The Most Influential Observations - Banking Crises with Semi-strict Specification

Country	**Year**		**CA**	**EMS**	**REER**	**NIIP**	**GD**	**PRD**	**UR**	**HPI**	**FDI**
							Caprio and Klingebiel (2002)				
Iceland	1993		1	1	0	1	1	1	1	0	1
							Laeven and Valencia (2018)				
Ecuador	1998		1	1	1	1	0	1	1	0	1
Japan	1997		0	1	1	1	1	1	1	1	1
Nicaragua	1990		1	1	1	0	0	1	0	0	1
Romania	1998		1	1	1	0	1	1	1	0	1
							Reinhart and Rogoff (2011)				
Central Afr. Rep.	1988	Systemic	1	0	0	0	0	1	0	0	0
Myanmar	2002	Systemic	0	0	0	0	0	1	1	0	0
Norway	1987	Systemic	1	0	0	1	1	1	0	1	0
Peru	1999	Systemic	1	0	0	1	0	1	1	0	1
Philippines	1981	Systemic	1	0	0	0	1	1	0	0	1
Turkey	1991	Systemic	1	0	0	1	0	1	0	0	0
Bolivia	1999	Non-systemic	1	0	0	1	0	0	1	0	1
Brazil	1985	Non-systemic	1	0	0	0	0	0	0	0	0
Dominican Rep.	1996	Non-systemic	1	1	0	0	0	1	1	0	1
El Salvador	1998	Non-systemic	1	1	0	1	1	1	1	0	0
Guatemala	2001	Non-systemic	1	0	0	0	0	1	1	0	1
Honduras	1999	Non-systemic	1	1	0	0	1	1	1	0	0
Honduras	2001	Non-systemic	1	1	0	0	1	1	1	0	1
Indonesia	1992	Non-systemic	1	0	0	0	0	1	1	0	0
Korea	1985	Non-systemic	1	1	0	0	0	1	0	1	1
Paraguay	2002	Non-systemic	1	1	0	1	0	1	1	0	0
Thailand	1979	Non-systemic	1	0	0	0	0	1	0	0	0
United States	1984	Non-systemic	1	0	0	1	1	1	0	1	0

Notes: '1' indicates that the observation was ranked among the most influential observations in bivariate logit model. The most influential observations have their Pregibon's delta-beta influence statistic higher than sum of one standard deviation and empirical mean from all observations. The CA stands for current account balance, EMS for export market share, REER for real effective exchange rate, NIIP for net international investment position, GD for general government debt to GDP ratio, PRD for private sector debt to GDP ratio, UR for unemployment rate, HPI for real house price index, FDI for foreign direct investments.

and in most cases remain in the same range as in the full database. The greatest shifts in κ values can be seen in the strict approach, since in this approach, we take into account only the first years of crisis periods. Nevertheless, the shifts are not greater than 0.1 bp. which is likely because systemic crisis events included in the systemic crisis-only database are the events that mostly appear in the other two papers. This reflects the fact that Reinhart and Rogoff (2009) identify in the full database the highest number of banking crisis events among all three papers under review, notwithstanding the fact that their investigated sample is the smallest of the three papers.

Interpreting the results from the logit models, the commonality between Reinhart and Rogoff (2009) and Laeven and Valencia (2008) increases, with the two databases reporting two statistically significant (EMS, FDI) and four insignificant (REER, NIIP, UR and government debt) variables. However, differences exist in the case of current account, HPI and private sector debt indicators. Caprio and Klingebiel (2002) remain the only authors reporting the unemployment rate and REER among significant EWIs. As a result, the heterogeneity in the identification strategy based on a more or less precise definition of systemic events still negatively affects the level of commonality across selected databases.

4.5 Robustness checks

In the first robustness check, we address the critique by van den Berg et al. (2008), who argue that the logit model is preferred over pooled probit models if one's aim is to predict events that are in the tail of a distribution, i.e., if crisis events in the sample occur with a small frequency. Additionally, predicting a crisis event in a panel setup without country-specific fixed effects amplifies the underlying heterogeneity; however, it may result in an artificially small number of statistically significant EWIs whose effect is supposedly constant and homogeneous across the cross-sectional dimension. For these reasons, we re-estimate individual bi-variate regressions using a panel logit model with fixed effects. The results are available upon request.

For banking crisis events, the original findings remain unaffected in the case of export market share, REER, FDI and the unemployment rate, among statistically significant cases, and NIIP and government debt for insignificant EWIs. The current account balance gains statistical significance in the Caprio and Klingebiel (2002) data and loses its significance in the Laeven and Valencia (2008) data. HPI becomes significant in the Laeven and Valencia (2008) database. The major change is observed in the case of private sector debt, when the previously insignificant indicator becomes a highly statistically significant predictor in all three databases. Similarly, for fiscal crises, the performance of the majority of EWIs (current account balance, export market share, private sector debt, unemployment rate, HPI and NIIP) remains robust to the choice of modelling technique. In one case (FDI), statistical significance is lost in the model with fixed effects. REER becomes significant in Laeven and Valencia (2008), and government debt does so in Baldacci et al. (2011). Overall, the model with fixed effects reports three EWIs whose predictive power is sensitive to the choice of crisis event specification. Overall, the heterogeneity in the results is more pronounced in the case of banking than fiscal crisis events and remains a serious issue, especially in the case of external sector-related variables.

We also include all EWIs in a single multivariate logit model with random effects as part of the robustness check; however, the results from this exercise strongly suffer from the limited number of observations. In most of the cases, we are able to include only 16 countries for which we have data for all the control variables. In the case of Laeven and Valencia (2008), the model is not estimated, as this highly restricted sample does not include any crisis year. Not surprisingly, the model does not find any statistically significant indicator regardless of the database used.

4.6 Policy Implications

Our *first* important finding shows that the degree of commonality is indeed relatively high in both fiscal and banking databases. When accounting for possible discrepancies in the dating of a crisis event by introducing a one-year lag, the overlap is even more pronounced, with some of the combinations achieving almost perfect agreement. This alone could corroborate the argument advanced by Kauko (2014) or Claessens and Kose (2013) regarding the robustness of results in the empirical literature regardless of the database used.

However, the *second* finding highlights the significant role played by influential observations. The presence of a few observations differently identified across databases results in several heterogeneous findings in the case of statistically significant EWIs. This problem is less urgent in the case of fiscal crisis events but more pronounced in the banking crisis field. Our findings thus corroborate the discussion in Claessens and Kose (2013), who argue that while (external) sovereign crises are relatively easily identified, banking crises often pose a larger challenge.

Third, most of the variables with ambiguous statistical significance are to a larger extent associated with the external sector, be it the real (export market share, REER) or financial (FDI) side of an economy. This observation may therefore highlight the crucial role played by adverse external shocks. Hence, being able to distinguish whether the crisis event originated in domestic conditions or was triggered by an external shock may substantially improve commonality across databases. Recently, a database compiled by Duca et al. (2017) for EU countries has ventured exactly in this direction.[8]

Overall, our paper illustrates how important the choice of a crisis event database is, even when 'following the crowd' and choosing from the set of widely recognized papers. From the policy perspective, we consider several cases of heterogeneous results conditional on the choice of a database a highly disturbing finding, since the existence of an effective early warning system presupposes the inclusion of the best performing indicators with the highest predictive power.

As the need arises to address these issues, we advocate for a wider discussion in the community of researchers and policy makers along the following set of suggestions:

1) Because many databases differ in terms of event specification even if approaching a similar issue (e.g., systemic banking crisis), research should include a more detailed discussion on the reasons for selecting the particular databases and the possible implications for the empirical findings.

2) As part of the robustness checks, authors may consider estimating baseline regressions using another 'popular' database with the closest definition of a crisis event. This is especially pressing in the case of research that adopts its own individual approach to crisis event specification (e.g., the study by Boyd et al. (2009)).

3) Introducing a one-year time lag to crisis event specification (the semi-strict approach) may help eradicate one source of potential heterogeneity (already applied by Baldacci et al. (2011)). While this approach may be plausible in the case of macro-oriented, predominantly panel-based studies, for research

[8]The role of data revisions, especially in the case of externally driven adverse shocks, should also be taken into consideration. Domonkos et al. (2017b) show that in the case of the MIP scoreboard, the predictive power of several EWIs, predominantly from the external sector, is likely to be distorted due to data revisions.

focusing on higher frequency data (e.g., the reactions of financial markets and market-based indicators), one needs to adjust the beginning of a crisis given the preferred objective.

4) As the literature on financial crisis determinants has substantially increased in recent years (decades), it is the most appropriate time to apply a popular meta-analysis approach with crisis event specification acting as one of the important causal factors. Potentially, a newly designed taxonomy of crisis databases would provide an additional list of covariates explaining differences in the performance of individual EWIs across relevant studies. To the best of our knowledge, the only such study is by Hamdaoui (2017) and analyses the effect of financial liberalization on banking crisis incidence by employing meta-analysis standards. While it represents a good step in the desired direction, there is no conditioning factor included in the list of regressors that captures differences in the event identification strategy.

5) The creation of an open-access repository that integrates information from existing studies into one comprehensive dataset. This repository should be open to all authors willing to submit their 'home-made' databases on crisis events, as inclusion may increase their visibility in the research community. Additionally, a sufficient amount of information on the characteristic features of existing databases would naturally result in the creation of an established crisis database taxonomy enabling filtering along various dimensions (see Section 2). As a result, any follow-up research would benefit from the existence of such a repository by being able to discuss the sensitivity of its empirical findings and select the database that best fits its objective.

6) Distinguishing between internally and externally triggered crisis events. Newly compiled databases should strive to provide information regarding the source of the disturbance (e.g., Duca et al. (2017)). Alternatively, to control for external spillover effects, empirical studies using premade databases should always control for external sector (both real and financial) related variables.

7) The role of international organizations (such as the IMF) in establishing a benchmark definition of a crisis event should be discussed. In many cases, the existing 'popular' databases are the work of researchers affiliated with regional or global international institutions (International Monetary Fund in Baldacci et al. (2011), World Bank in Caprio and Klingebiel (2002), European Central Bank and European Systemic Risk Board in Duca et al. (2017)), which already implies some level of synchronization at the regional and international levels. It would be, however, advisable to strive to achieve a higher level of commonality due to persistent heterogeneity across compiled databases.

It is necessary to note that our suggestions are not intended to discourage additional work on a more precise yet diverse specification of crisis events. In contrast, we feel that adopting a widely accepted benchmark definition of a particular crisis event would make it substantially easier to mark out differences when proposing an adjustment to an already established identification strategy. As a side benefit, communication of research findings to a wider audience would become more transparent and easier to comprehend when using one commonly agreed-upon benchmark definition.

5 Conclusions

In this paper, we examine the consistency in the identification and timing of crisis events of the most prominent databases of banking and fiscal crises. To do so, we calculate the degree of commonality across databases using Cohen's κ indicator. To search for the differences in predictive power of selected early warning indicators, we use panel logit models with random effects on a common set of countries shared by all databases. Last, we identify the most influential crisis observations unique to each database using Pregibon's delta-beta influence statistics. A set of indicators is taken from the Macroeconomic Imbalances Procedure scoreboard that is utilized as a quasi-early warning system at the EU level. As the majority of the EWIs used in our analysis have already been adopted by the relevant policy makers, an increase in uncertainty in their estimated predictive power may bring about serious consequences when conducting macroeconomic stabilization policy.

Our results confirm that the degree of commonality across banking and fiscal databases is indeed relatively high. Once accounting for a one-year lag due to possible beginning- and end-of-the-year discrepancies, we achieve almost perfect agreement in most of the cases. However, there is still a significant role played by a few influential observations that result in several heterogeneous findings for statistically significant EWIs. This problem is more pronounced in the banking crisis literature. Most of the variables with ambiguous statistical significance are to a larger extent associated with the external sector, be it the real (export market share, REER) or financial (FDI) side of an economy.

Based on the empirical findings, we propose several suggestions that would help mitigate this issue. For example, we advocate the use of a more extensive list of robustness checks that would be enabled by the creation of an open-access repository integrating existing crisis databases. More important, the newly designed taxonomy of crisis databases would invite the preparation of studies that employ meta-analysis standards to investigate the effect of different crisis identification strategies on EWI performance.

References

Anauati, V., S. Galiani, and R. H. Galvez (2016). Quantifying the life cycle of scholarly articles across fields of economic research. *Economic Inquiry 54*(2), 1339–1355.

Arregui, N., J. Benes, I. Krznar, S. Mitra, and A. O. Santos (2013). Evaluating the net benefits of macro-prudential policy: A cookbook. IMF Working Paper WP/13/167, International Monetary Fund.

Babecky, J., T. Havranek, J. Mateju, M. Rusnak, K. Smidkova, and B. Vasicek (2014). Banking, debt, and currency crises in developed countries: Stylized facts and early warning indicators. *Journal of Financial Stability 15*(December), 1–17.

Baldacci, E., I. Petrova, N. Belhocine, G. Dobrescu, and S. Mazraani (2011). Assessing fiscal stress. IMF Working Paper WP/11/100, International Monetary Fund.

Bank, W. (2002). Global development finance. appendix on commercial debt restructuring. World Bank Publications ISBN-10 : 0821350854, World Bank.

Beim, D. O. and C. W. Calomiris (2001). *Emerging Financial Markets*. Boston : McGraw-Hill/Irwin: Boston : McGraw-Hill/Irwin.

Berg, A. and C. Pattillo (1999). Predicting currency crises: The indicators approach and an alternative. *Journal of International Money and Finance 18*(4), 561–586.

Biegun, K. and J. Karwowski (2020). Macroeconomic imbalance procedure (mip) scoreboard indicators and their predictive strength of "multidimensional crises". *Equilibrium - Quarterly Journal of Economics and Economic Policy 15*(1), 11–28.

Boyd, J., G. D. Nicolò, and E. Loukoianova (2009). Banking crises and crisis dating: Theory and evidence. Working Paper WP/09/141, International Monetary Fund.

Bussiere, M. and M. Fratzscher (2006). Towards a new early warning system of financial crises. *Journal of International Money and Finance 25*(6), 953–973.

Caggiano, G., P. Calice, L. Leonida, and G. Kapetanios (2016). Comparing logit-based early warning systems: Does the duration of systemic banking crises matter? *Journal of Empirical Finance 37*(June), 104–116.

Caprio, G. and D. Klingebiel (2002). Episodes of systemic and borderline banking crises. In D. Klingebiel and L. Laeven (Eds.), *Managing the real and fiscal effects of banking crises*, pp. 31–49. Washington, D.C.: World Bank.

Chaudron, R. and J. de Haan (2014). Identifying and dating systemic banking crises using incidence and size of bank failures. DNB Working Paper 406 / January 2016, De Nederlandsche Bank.

Ciarlone, A. and G. Trebeschi (2005). Designing an early warning system for debt crises. *Emerging Markets Review 6*(4), 376–395.

Claessens, S. and M. A. Kose (2013). Financial crises: Explanations, types, and implications. Working Paper WP/13/28, International Monetary Fund.

Davis, E. P. and D. Karim (2008). Comparing early warning systems for banking crises. *Journal of Financial Stability 4*(June), 89–120.

Davis, J. S., A. Mack, W. Phoa, and A. Vandenabeele (2016). Credit booms, banking crises, and the current account. *Journal of International Money and Finance 60*(February), 360–377.

Demirgüç-Kunt, A. and E. Detragiache (2005). Cross-country empirical studies of systemic bank distress: A survey. *National Institute of Economic Review 192*(1), 68–83.

Domonkos, T., F. Ostrihon, I. Sikulova, and M. Siranova (2017a). Analysing the relevance of the mip scoreboard's indicators. *National Institute Economic Review 239*(1), R32–R52.

Domonkos, T., F. Ostrihon, I. Sikulova, and M. Siranova (2017b). Working paper on the mip scoreboard (analysing the relevance and predictive power of the mip scoreboard's indicators). FIRSTRUN Deliverable 2.5, FIRSTRUN.

Duca, M. L., A. Koban, M. Basten, E. Bengtsson, B. Klaus, P. Kusmierczyk, J. H. Lang, C. Detken, and T. Peltonen (2017). A new database for financial crises in european countries. ecb/esrb eu crises database. Occasional Paper Series No 13/July 2017, European Central Bank.

EC (2016). The macroeconomic imbalance procedure. rationale, process, application: A compendium. Institutional Paper 039, European Commission.

Fleiss, J. L., J. Cohen, and B. S. Everitt (1969). Large sample standard errors of kappa and weighted kappa. Technical Report 5.

Frankel, J. A. and A. K. Rose (1996). Currency crashes in emerging markets: An empirical treatment. *Journal of International Economics 41*(3-4), 351–366.

Frydl, E. J. (1999). The length and cost of banking crises. Working Paper WP/99/30, International Monetary Fund.

Hamdaoui, M. (2017). Financial liberalization and systemic banking crises: A meta-analysis. *International Economics 152*(December), 26–54.

Hamermesh, D. S. (2018). Citations in economics: Measurement, uses, and impacts. *Journal of Economic Literature 56*(1), 115–156.

Kacer, M. (2013). Predikcie finančných kríz s využitím metód finančnej ekonometrie. *Biatec 21*(10), 23–26.

Kaminski, G., S. Lizondo, and C. R. Reinhart (1997). Leading indicators of currency crises. IMF Working Paper WP/97/79, International Monetary Fund.

Kauko, K. (2014). How to foresee banking crises? a survey of the empirical literature. *Economic Systems 38*(3), 289–308.

Knedlik, T. (2014). The impact of preferences on early warning system – the case of the european commission's scoreboard. *European Journal of Political Economy 34*(June), 157–166.

Laeven, L. and F. Valencia (2008). Systemic banking crises: A new database. IMF Working Paper WP/08/224, International Monetary Fund.

Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. Technical Report 1.

Maddison, A. (2003). *The World Economy: Historical Statistics*. OECD Publishing: OECD Publishing.

Martin-Martin, A., E. Orduna-Male, A.-W. Harzing, and E. D. López-Cózar (2017). Can we use google scholar to identify highly-cited documents? *Journal of Informetrics 11*(1), 152–163.

Mulder, C., R. Perrelli, and M. D. Rocha (2016). The role of bank and corporate balance sheets on early warning systems of currency crises: An empirical study. *Emerging Markets Finance Trade 52*(7), 1542–1561.

Ostrihon, F. (2020). Evaluating banking crisis predictions in eu and v4 countries. *Business Trends 10*(2), 64–72.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics 9*(4), 705–724.

Reinhart, C. M. and K. S. Rogoff (2009). *This Time Is Different: Eight Centuries of Financial Folly*. Princeton, New Jersey: Princeton University Press.

Sarkees, M. R. and P. Schafer (2000). Correlates of war data on war: An update to 1997. Technical Report 1.

Schularick, M. and A. M. Taylor (2012). Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870-2008. *American Economic Review 102*(2), 1029–1061.

Siranova, M. and M. Radvansky (2018). Performance of the macroeconomic imbalance procedure in light of historical experience in the cee region. *Journal of Economic Policy Reform 21*(4), 335–352.

Sondermann, D. and N. Zorell (2019). A macroeconomic vulnerability model for the euro area. Working Paper 2306, European Central Bank.

Sturzenegger, F. and J. Zettelmeyer (2006). *Debt Defaults and Lessons from a Decade of Crises*. The MIT Press: The MIT Press.

Valinskytė, N. and G. Rupeika (2015). Leading indicators for the countercyclical capital buffer in lithuania. Lietuvos Bankas Occasional Paper Series 4/2015, Lietuvos Bankas.

van den Berg, J., B. Candelon, and J.-P. Urbain (2008). A cautious note on the use of panel models to predict financial crises. *Economic Letters 101*(1), 80–83.

von Hagen, J. and T.-K. Ho (2007). Money market pressure and the determinants of banking crises. *Journal of Money, Credit and Banking 39*(5), 1037–1066.

Appendix

Table A1: Citation Profile

Database	**Publication Year**	**Citations**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Babecky, J. et al.	2014	94									
Baldacci, E. et al.	2011	131									
Baron, M. et al.*	2018	6									
Gerling, K. et al.	2017	10									
Jorda, O. et al.	2017	248									
Laeven, L. and Valencia, F.											
i) Laeven, L. and Valencia, F.	2008	42								1	1
ii) Laeven, L. and Valencia, F.	2013	3 080							9	31	181
iii) Laeven, L. and Valencia, F.	2018	99									
iv) Laeven, L. and Valencia, F.	2020	0									
lo Duca et al.	2017	41									
Reinhart, C. M. and Rogoff, K. S.	2009	8 217									55
Reinhart, C. M. and Rogoff, K. S.	2011	1 683									
Schularick, M. and Taylor, A.M.	2012	2 260									8
Benchmark											
Caprio and Klingebiel	2002	1 201	29	32	51	73	76	95	72	84	94
# of 'financial crisis' in REPEC			**299**	**266**	**246**	**269**	**288**	**281**	**278**	**636**	**2 344**

Database	**Publication Year**	**Citations**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Babecky, J. et al.	2014	94					2	10	17	19	25	21
Baldacci, E. et al.	2011	131		7	13	12	20	14	21	16	18	10
Baron, M. et al.*	2018	6								1	2	3
Gerling, K. et al.	2017	10								1	7	2
Jorda, O. et al.	2017	248							11	47	92	98
Laeven, L. and Valencia, F.												
i) Laeven, L. and Valencia, F.	2008	42	0	0	1	5	8	9	5	4	4	4
ii) Laeven, L. and Valencia, F.	2013	3 080	238	300	303	270	283	300	301	310	288	266
iii) Laeven, L. and Valencia, F.	2018	99								3	9	87
iv) Laeven, L. and Valencia, F.	2020	0										
lo Duca et al.	2017	51							1	6	13	21
Reinhart, C. M. and Rogoff, K. S.	2009	8 217	481	760	1 008	999	966	945	887	780	732	604
Reinhart, C. M. and Rogoff, K. S.	2011	1 683	32	87	158	184	186	219	242	197	185	193
Schularick, M. and Taylor, A.M.	2012	2 260	46	61	125	175	208	262	308	329	379	359
Benchmark												
Caprio, G. and Klingebiel, D.	2002	1 201	88	88	81	65	58	40	57	43	40	35
# of 'financial crisis' in REPEC			**3 150**	**3 237**	**3 168**	**3 468**	**3 248**	**3 033**	**2 894**	**2 725**	**2 585**	**2 184**

Table A2: Life-Cycle of Citation Profiles

Database	Publication Year	Citations	Year from publication								
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Babecky, J. et al.	2014	94	2	10	17	19	25	21			
Baldacci, E. et al.	2011	131	7	13	12	20	14	21	16	18	10
Baron, M. et al.*	2018	6	1	2	3						
Gerling, K. et al.	2017	10	1	7	2						
Jorda, O. et al.	2017	248	11	47	92	98					
Laeven, L. and Valencia, F.		0									
i) Laeven, L. and Valencia, F.	2008	42	1	1	0	0	1	5	8	9	5
ii) Laeven, L. and Valencia, F.	2013	3 080	9	31	181	238	300	303	270	283	300
iii) Laeven, L. and Valencia, F.	2018	99	3	9	87						
iv) Laeven, L. and Valencia, F.	2020	0									
lo Duca et al.	2017	41	1	6	13	21					
Reinhart, C. M. and Rogoff, K. S.	2009	8 217	55	481	760	1008	999	966	945	887	780
Reinhart, C. M. and Rogoff, K. S.	2011	1 683	32	87	158	184	186	219	242	197	185
Schularick, M. and Taylor, A.M.	2012	2 260	8	46	61	125	175	208	262	308	329
Benchmark											
Caprio and Klingebiel	2002	1 201	29	32	51	73	76	95	72	84	94
Caprio, G. and Klingebiel, D. (%)	**2002**	**100%**	**2.4%**	**2.7%**	**4.2%**	**6.1%**	**6.3%**	**7.9%**	**6.0%**	**7.0%**	**7.8%**

Database	Publication Year	Citations	Year from publication									
			10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Babecky, J. et al.	2014	94										
Baldacci, E. et al.	2011	131										
Baron, M. et al.*	2018	6										
Gerling, K. et al.	2017	10										
Jorda, O. et al.	2017	248										
Laeven, L. and Valencia, F.												
i) Laeven, L. and Valencia, F.	2008	42	4	4	4							
ii) Laeven, L. and Valencia, F.	2013	3 080	301	310	288	266						
iii) Laeven, L. and Valencia, F.	2018	99										
iv) Laeven, L. and Valencia, F.	2020	0										
lo Duca et al.	2017	41										
Reinhart, C. M. and Rogoff, K. S.	2009	8 217	732	604								
Reinhart, C. M. and Rogoff, K. S.	2011	1 683	193									
Schularick, M. and Taylor, A.M.	2012	2 260	379	359								
Benchmark												
Caprio, G. and Klingebiel, D.	2002	1 201	88	88	81	65	58	40	57	43	40	35
Caprio and Klingebiel (%)	**2002**	**100%**	**7.3%**	**7.3%**	**6.7%**	**5.4%**	**4.8%**	**3.3%**	**4.7%**	**3.6%**	**3.3%**	**2.9%**

Table A3: Average Marginal Effects of EWIs for Fiscal Crises with Semi-strict Event Specification (Sub-samples)

	EU28			Obs./	OECD			Obs./	Bottom half			Obs./
Variable	BL(2011)	LV(2018)	RR(2011)	Country	BL(2011)	LV(2018)	RR(2011)	Country	BL(2011)	LV(2018)	RR(2011)	Country
CA	**-0.662****	-0.041	-0.041	339/	**-0.949*****	-0.303	-0.212	471/	-0.365	-0.121	-0.283	467/
	(0.05)	(0.703)	(0.703)	17	**(0.007)**	(0.105)	(0.628)	23	(0.32)	(0.53)	(0.356)	21
EMS	0.002	0.002	0.002	335/	-0.029	-0.005	-0.004	585/	0.010	-0.001	**0.024****	529/
	(0.941)	(0.774)	(0.774)	17	(0.535)	(0.747)	(0.771)	24	(0.593)	(0.935)	**(0.026)**	21
REER	0.013	.	.	307/	0.050	.	.	469/	-0.028	0.017	0.008	178/
	(0.843)			16	(0.355)			21	(0.673)	(0.335)	(0.845)	9
NIIP	-0.023	.	.	237/	-0.035	.	.	367/	0.059	-0.007	0.059	87/
	(0.374)			16	(0.28)			24	(0.536)	(0.891)	(0.536)	11
Gov. Debt	-0.015	.	.	343/	-0.019	.	.	460/	-0.182	-0.007	0.008	109/
	(0.673)			17	(0.558)			22	(0.114)	(0.906)	(0.901)	9
Pr. Debt	0.007	.	.	378/	-0.026	-0.009	-0.009	591/	-0.055	0.024	0.018	408/
	(0.759)			17	(0.181)	(0.461)	(0.461)	24	(0.442)	(0.458)	(0.738)	19
UR	0.096	.	0.039	203/	-0.060	.	.	287/	-0.312	-0.119	-0.123	252/
	(0.718)		(0.737)	17	(0.804)			24	(0.214)	(0.437)	(0.498)	21
HPI	-0.117	.	.	232/	-0.058	.	.	418/	-0.233	-0.039	-0.233	33/
	(0.349)			8	(0.618)			15	(0.691)	(0.912)	(0.691)	2
FDI	-0.432	.	.	356/	-0.268	-0.129	-0.160	559/	**-1.793****	-0.503	**-1.335***	527/
	(0.425)			17	(0.539)	(0.598)	(0.546)	24	**(0.049)**	(0.298)	**(0.082)**	21

Notes: Table shows estimated average marginal effect from individual bivariate logit probit models with random effects as specified in [10]. Bottom Half indicates group of countries with income p.c. level below world average. BL(2011) stands for Baldacci et al. (2011), LV(2018) for Laeven and Valencia (2018), and RR(2011) for Reinhart and Rogoff (2011). P-values are in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level. CA stands for current account balance, EMS for export market share, REER for real effective exchange rate, NIIP for net international investment position, Gov. Debt for general government debt to GDP ratio, Pr. Debt for private sector debt to GDP ratio, UR for unemployment rate, HPI for real house price index, FDI for foreign direct investments.

Table A4: Average Marginal Effects of EWIs for Banking Crises with Semi-strict Event Specification (Sub-samples)

	EU28			Obs./	OECD			Obs./	Bottom half			Obs./
Variable	CK(2002)	LV(2018)	RR(2011)	Country	CK(2002)	LV(2018)	RR(2011)	Country	CK(2002)	LV(2018)	RR(2011)	Country
CA	-0.133	-0.268	-0.133	302/	**-0.513***	**-0.342***	-0.389	514/	0.298	-0.065	0.152	522/
	(0.638)	(0.225)	(0.638)	14	**(0.055)**	**(0.095)**	(0.147)	24	(0.174)	(0.66)	(0.494)	24
EMS	-0.037	-0.028	-0.037	363/	0.016	-0.017	-0.028	690/	0.017	**0.018***	0.002	603/
	(0.415)	(0.437)	(0.415)	16	(0.664)	(0.583)	(0.463)	27	(0.148)	**(0.06)**	(0.921)	25
REER	0.090	0.053	0.090	347/	0.731	0.030	**0.108***	572/	0.023	-0.001	0.027	275/
	(0.19)	(0.279)	(0.19)	16	(0.453)	(0.462)	**(0.063)**	25	(0.128)	(0.965)	(0.108)	14
NIIP	-0.008	-0.012	-0.008	235/	-0.010	0.008	-0.004	419/	0.039	-0.001	-0.012	98/
	(0.772)	(0.493)	(0.772)	14	(0.731)	(0.788)	(0.896)	26	(0.487)	(0.959)	(0.83)	12
Gov. Debt	0.010	-0.030	0.010	377/	-0.006	-0.016	-0.010	564/	-0.029	-0.010	-0.052	103/
	(0.703)	(0.298)	(0.703)	16	(0.819)	(0.474)	(0.695)	25	(0.671)	(0.867)	(0.568)	8
Pr. Debt	-0.002	-0.007	-0.002	399/	-0.001	0.004	-0.018	691/	0.059	0.014	**0.124****	600/
	(0.924)	(0.62)	(0.924)	16	(0.928)	(0.707)	(0.296)	27	(0.247)	(0.779)	**(0.032)**	23
UR	0.114	0.023	0.114	184/	-0.399	-0.248	-0.170	316/	-0.582	-0.222	**-1.011****	304/
	(0.508)	(0.901)	(0.508)	16	(0.216)	(0.362)	(0.515)	27	(0.115)	(0.334)	**(0.033)**	26
HPI	-0.089	-0.092	-0.089	267/	-0.167	**-0.177***	-0.036	484/	-0.025	.	-0.025	33/
	(0.492)	(0.331)	(0.492)	11	(0.131)	**(0.069)**	(0.745)	20	(0.961)		(0.961)	2
FDI	-1.888	-0.247	-1.888	350/	**-1.229***	-0.521	-0.960	669/	-0.165	-0.195	0.146	642/
	(0.182)	(0.574)	(0.182)	16	**(0.086)**	(0.283)	(0.135)	27	(0.675)	(0.615)	(0.592)	25

Notes: Table shows estimated average marginal effect from individual bivariate panel logit models with random effects as specified in [10]. Bottom Half indicates group of countries with income p.c. level below world average. CK(2002) stands for Caprio and Klingebiel (2002), LV(2018) for Laeven and Valencia (2018), and RR(2011) for Reinhart and Rogoff (2011). P-values are in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level. CA stands for current account balance, EMS for export market share, REER for real effective exchange rate, NIIP for net international investment position, Gov. Debt for general government debt to GDP ratio, Pr. Debt for private sector debt to GDP ratio, UR for unemployment rate, HPI for real house price index, FDI for foreign direct investments.