WP/14/48

IMF Working Paper

Assessing Bias and Accuracy in the World Bank- IMF's Debt Sustainability Framework for Low-Income Countries

Andrew Berg, Enrico Berkes, Catherine Pattillo, Andrea F. Presbitero, Yorbol Yakhshilikov

INTERNATIONAL MONETARY FUND

WP/14/48

IMF Working Paper

Research Department and Strategy, Policy, and Review Department

Assessing Bias and Accuracy in the World Bank- IMF's Debt Sustainability Framework for Low-Income Countries*

Prepared by Andrew Berg, Enrico Berkes, Catherine Pattillo, Andrea F. Presbitero, Yorbol Yakhshilikov

Authorized for distribution by Andrew Berg and Catherine Pattillo

March 2014

This Working Paper should not be reported as representing the views of the IMF or DFID. The views expressed in this Working Paper are those of the authors and do not necessarily represent those of the IMF, IMF policy, or DFID. Working Papers describe research in progress by the authors and are published to elicit comments and to further debate.

Abstract

The World Bank and the IMF have adopted a debt sustainability framework (DSF) to evaluate the risk of debt distress in Low Income Countries (LICs). At the core of the DSF are empirically-based thresholds for each of five different measures of the debt burden (the "debt threshold approach" DTA). The DSF contains a rule for aggregating the information contained in these five different variables which we label the "worst-case aggregator" (WCA) in view of the fact that the DSF considers a breach of any one of the thresholds sufficient to indicate a high risk of debt distress. However, neither the DTA nor the WCA has heretofore been subject to empirical testing. We find that: (1) the DTA loses information relative to a simple proposed alternative; (2) the WCA is too conservative (predicting crises too often) in terms of the loss function used in the DSF; and (3) the WCA is less accurate than some simple proposed alternative aggregators as a predictor of debt distress.

JEL Classification Numbers: F33, F34, F35, O11

Keywords: Debt sustainability, DSF, IMF, Sovereign debt crises

Author's E-Mail Address: aberg@imf.org, enrico.berkes@u.northwestern.edu, cpattillo@imf.org, apresbitero@imf.org, yyakhshilikov@imf.org

[*] We would like to thank Saul Lizondo for initiating this project, Chris Geiregat, Andrea Fracasso, Andrew Jewell, Aart Kraay, Calvin McDonald, Sean Nolan, Juan Pradelli, Sam Ouliaris, and many IMF colleagues for useful comments and suggestions. This working paper is part of a research project on macroeconomic policy in low-income countries supported by the U.K.'s Department for International Development. This working paper should not be reported as representing the views of the IMF or of DFID. The views expressed in this Working Paper are those of the authors and do not necessarily represent those of the IMF, IMF policy, or of DFID. Working Papers describe research in progress by the authors and are published to elicit comments and to further debate.

Contents Page

1 Introduction

Low-income countries (LICs) are still recovering from the prolonged sovereign debt crisis of the 1980s and 1990s and then the grueling process of achieving debt relief under the Highly-Indebted Poor Country (HIPC) initiative and the Multilateral Debt Relief Initiative (MDRI). Debt levels are now low in most of these formerly highly-indebted countries. At the same time, many countries have improved their policy environments and economic performance. Low-income countries and their development partners want to seize on this opportunity to expand public investment in order to build scarce infrastructure and improve health and education. Aid flows have not accelerated as some had hoped, but substantial non-concessional financing is available from both private and official sources.

How should LICs and donors balance the competing objectives of financing development spending and avoiding a new round of debt crises? One answer is the World Bank and the International Monetary Fund (IMF)'s Debt Sustainability Framework (DSF). Introduced in 2005, the DSF is a standardized framework for conducting debt sustainability analysis (DSA) in LICs. Such a DSA consists of an analysis of the country's projected debt burden over the next 20 years and its vulnerability to shocks and an assessment of the risk of debt distress.

The LIC DSF is thus a central tool for the Fund's analysis of macroeconomic stability and contributes to determine access to IMF financing. The DSF ratings also determine the share of grants and loans in World Bank assistance to LICs, with countries at high risk of debt distress receiving all their World Bank support as grants. In addition, the DSF is an important input to the World Bank Non-Concessional Borrowing Policy and to the IMF Debt Limits Policy, designed to assist LICs in implementing sound borrowing policies and limiting the access to non-concessional lending in order to preserve long-term debt sustainability. Clearly it is important to ensure that the DSF strikes the right balance.

The DSF is a variant of standard methods that use probit regressions to predict crises (section 2 discusses related literature). The details of the DSF turn out to be important for the properties of the DSF. Thus, in section 3 we find ourselves descending into the weeds of the methodology, before emerging to make some ultimately simple points about how to correct some of the shortcomings.

The DSF is based on five separate policy-dependent debt thresholds. Each of these thresholds is derived from the estimation of a simple probit model. In each model, the dependent variable is the probability that a country will experience debt distress. Each of the five models includes a different measure of debt burden (public external debt as a share of GDP, debt-service as a share of exports, etc.), along with a common set of controls. The DSF defines a loss function in terms of missed crises and false alarms and from this loss function deduces an optimal risk of debt distress risk for each debt-burden measure. These probability cutoffs imply a threshold value for each debt-burden measure such that, if the measure breaches the threshold value, then the probability of debt distress is above the tolerable level. Under the DSF, countries are assigned a risk rating by comparing the debt-burden measures projected out for 20 years to the corresponding thresholds. In what follows we call this method the "debt threshold approach" (DTA).

The question arises as to how to aggregate the information from the five different debt-burden measures, because in any particular case they may imply different debt risk ratings. As described in section 4, in the DSF a country is classified at high risk of debt distress if *any one* of the debt measures breaches its threshold, irrespective of the evolution of the other four. We refer to this feature of the framework as the *DSF worst-case aggregator* (WCA). The IMF and the World Bank (2004) justify this rule partly on the grounds that collinearity among the debt-burden measures makes it difficult to combine them into one probit regression. However, the WCA "avoids" the multicollinearity problem only insofar as it does not seek to establish the parameters of the aggregator (e.g. the fact that each indicator is equally important) from the data.

So far, we have described the mechanical application of the DSF. In fact, there is room for judgment. However, the thresholds play a fundamental role, and deviations from a mechanical application require specific justification.[1] The most recent IMF-World Bank guidance note on the DSF emphasizes the use of judgment in cases of small and temporary breaches (and of near breaches) of the thresholds, though in principle broader considerations can be considered.[2] It seems that in practice the mechanical application usually predicts the actual rating, and that most of the exceptions reflect cases in which minor or near-breaches by a single variable are overruled, not rejections of the thresholds themselves or the WCA approach to aggregation.[3] In what follows, we therefore assess the DSF in its mechanical application. We return to the role of judgment in the conclusion.

In section 5, we investigate the merits of an alternative—and more standard—approach to assessing debt risks not by comparing debt measures to thresholds (as in the DTA) but rather by comparing the probability of debt distress with probability thresholds. We label this method the "probability threshold approach" (PTA).[4] The PTA turns out to have a number of advantages. As we quantify in this section, it is more accurate in part because it makes more effective use of the information contained in other predictors of debt distress such as policy and growth.

The PTA also facilitates the aggregation of the information contained in the various debt indicators. In section 6 we propose and analyze a variety of alternative aggregators, all based on some type of weighted average of the five debt measures in the DSF. We compare the performance of the WCA with that of single debt indicators and with the alternative aggregators. Despite the collinearity problem, we find some stark results. First, we show that the WCA is *biased*, in that it produces too many false alarms for each missed crisis, when evaluated in terms of the loss function that is used in the DSF to justify the individual debt burden thresholds. We also demonstrate

[1] Perhaps this reflects its roots in the HIPC initiative, where it was necessary to arrive at a common level of debt burden to which all countries would be reduced, which suggested a relatively simple and mechanical rule.

[2] As explained in (IMF and World Bank, 2013), "Although the indicative thresholds play a fundamental role in the determination of the risk rating, they should not be interpreted mechanistically. The assessment of risk needs to strike a balance between paying due attention to debt levels rising toward or above thresholds and using judgment. Thus, a marginal or temporary breach of a threshold may not necessarily imply a significant vulnerability. Conversely, a near breach should not be dismissed without careful consideration."

[3] Of 60 recent DSAs analyzed, the mechanical and actual ratings corresponded in 42 cases; in 11 out of the 18 other cases the deviation was grounded in the fact that breaches were small and temporary, with one debt variable within 3 percent of the threshold. Thus in about 90 percent of cases the thresholds and the WCA seem to have played the determinative role.

[4] Of course, most applications of probit early-warning models use the PTA without belaboring the point; we find the label useful in this paper to clarify the distinction with the DTA.

that the WCA is *inaccurate* relative to our proposed alternatives. For any given weights on false alarms and missed crises in the loss function, the WCA performs significantly less well than our proposed alternatives in identifying high-risk countries.

The bias of the WCA is a matter of construction, as we will see. However, the accuracy of the WCA is an empirical question. An analogy may be helpful. In assessing the risk of death, a doctor may look at multiple indicators, e.g. an indicator for cancer and another for heart disease. It may be that the most accurate way to assess risk is to look at the worst of the two indicators: a strong heart may not protect against a cancer. On the other hand, it maybe be more informative to average these two indicators somehow, rather than focusing on the worst one: a low risk of heart disease reduces overall risk even in the face of an ominous cancer risk. A comparison of the various ways of aggregating the two indicators with actual life expectancy would clarify which way produces the most accurate forecasts. Turning to the application at hand, it may be argued that a very high debt service-to-export ratio should be very alarming, irrespective of debt stock measures such as debt-to-GDP ratio. Alternatively, a good score on one measure may mitigate the risk associated with the other. For example, uncorrelated measurement error in the different debt burden measures would be exacerbated by the WCA and mitigated by averaging. Rather than try to resolve these issues a priori, our approach here follows the philosophy of the LIC DSF and tries to let the data speak.

Having established the statistical superiority of some of the alternative estimators, we attempt to make some recommendations about *which* alternative estimator has the most appealing properties. This turns out to be a difficult question, but we make and justify some concrete suggestions.

A loss function in terms of missed crises and false alarms greatly facilitates all these model comparisons. Fortunately, the DSF presents and uses such a loss function, in which 'equal weight' is placed on both types of events. This metric allows us to talk clearly about bias and about the relative accuracy of different models. For this purpose, we generally take the parameters of this loss function as given. However, we close in section 7 by revisiting the parameters of this loss function. Because of the bias in the WCA described above and some definitional subtleties discussed in section 7, the DSF effectively places more than 20 times the weight on each missed crisis episode than on false alarms.

The purpose of this paper is limited. We analyze some specific and, as it turns out, important shortcomings and discuss corrections. However, these corrections would have their own implications. Most importantly, correcting the bias in the DSF would tend to result in lower risk ratings for many countries, on average. Would this be good or bad? We have little to say specifically about this critical issue based on the analysis in this paper, because we restrict ourselves here to judging the DSF on its own terms. It may be that the DSF as it stands gives the "right answer" (low debt thresholds) if for the "wrong reasons" (biased aggregation methods). This must be because other flaws in the DSF (e.g. excessively optimistic growth forecasts) bias ratings in the other direction. If that is the case, then adopting the corrections in this paper would also require making other changes to reduce these other shortcomings in the DSF. More generally, we believe that a transparent discussion of the various flaws is better — i.e. promotes better analysis and, we hope, decision-making — than hoping they cancel out without looking hard at them. We return to some of these issues in the conclusion.

2 Related Literature

This paper takes as given the main outlines of the debt-risk regressions used in the IMF's current DSF: the definition of debt distress episodes, the approach of estimating probit regressions on these episodes, the sample, and even the set of potential determinants.[5] Thus a general exploration of the determinants of sovereign debt crises is beyond the scope of this paper.[6] Similarly, we abstract from the broader discussion on how to assess debt sustainability, sticking here to the general approach of the LIC DSF.[7]

The WB-IMF LIC DSF looks at multiple debt burden measures in assessing debt distress risk, thus raising the issue of how to combine the information from these measures. Almost all academic work follows general econometric practice and combines multiple candidate determinants in a single regression, generally emphasizing the results with those variables that win this "horse race" of statistical significance.[8] In some applications, there are many more candidate variables than degrees of freedom, such that a fully general specification is not feasible (as in cross-country growth regressions). In this case, Bayesian model averaging is a systematic approach to picking the variables with the most explanatory power, taking into account correlations with other variables (see, for instance, Bandiera *et al.*, 2010, for an application to the determinants of sovereign debt defaults).

An important exception to the general practice is the indicator-by-indicator approach of Kaminsky *et al.* (1997), which has been particularly influential in policy circles, e.g. in the Fund's own vulnerability exercise for emerging markets and advanced countries (IMF, 2012). Here, a large number of potential predictive variables are examined statistically one by one. Each variable is assumed to have a nonlinear effect on crisis risk, such that it is considered to signal a crisis when it is above a critical threshold derived to maximize the signal-to-noise ratio. The resulting indicator variables are combined using weights that reflect the univariate explanatory power of each indicator. This approach has the disadvantage that it does not take into account the correlations among the various indicators when arriving at the overall risk. Moreover, the assumption that each variable matters nonlinearly is generally not subject to statistical testing.[9] However, it has important attractions for policymakers. First, it can be applied even when the number of right-hand-side

[5] All these are discussed at length in IMF and World Bank (2012), which itself draws heavily on IMF and World Bank (2004) and the influential paper by Kraay and Nehru (2006). The latest DSF guidance note provides a comprehensive guide to the use of the DSF, explicitly intended for readers without an extensive prior knowledge of the framework (IMF and World Bank, 2013).

[6] For an updated and comprehensive review of the determinants of debt crises, focused on developing countries, see Pradelli (2012).

[7] See Buffie *et al.* (2012) for an alternative country-specific and scenario-based approach to assessing debt sustainability.

[8] The regression tree approach used in Manasse and Roubini (2009) uses a different technique that aims to find which combinations of indicator variables best sort observations into high-risk and low-risk pools. This may have promise as an alternative to the LIC DSF, but it represents a completely different technique and hence does not lend itself to the agenda of this paper, which is to examine the current approach and suggest modifications.

[9] Though this can be tested; see for example Berg and Pattillo (1999).

variables is large relative to the number of observations, and when collinearity among potential predictive variables is extreme. Second, it allows the user to identify readily which particular variables are above the threshold and thus are contributing to risk. The resulting risk factor "heat map" facilitates the integration of statistical analyses and judgment, in some contrast with the fundamental role of the mechanical thresholds in the LIC DSF.[10]

Finally, we have not been able to find any example in the literature of the use of the WCA as a tool for combining indicators to predict risk (even, after an admittedly brief search, in medicine/epidemiology), that is of an approach that focuses on the most alarming indicator for any particular observation.

3 The Debt Sustainability Framework

In the IMF-WB LIC DSF the probability of a country experiencing debt distress is estimated with a set of simple probit models on a sample of developing countries. Separate probits are run for each of five different debt burden measures, and from each such probit a threshold for the corresponding debt burden is derived. The DSF assigns a debt risk rating depending on whether there is a projected breach of these thresholds under baseline and stress-test scenarios. We call this approach the "debt threshold approach" (DTA).

More precisely, in a first step, the likelihood of experiencing a distress episode is estimated on a sample of low- and middle-income countries, observed between 1970 and 2007, using a parsimonious probit model. For each debt variable j, we follow the DSF and estimate the probit:

$$Prob_j(y_{it} = 1) = \Phi(\beta_{Debt_j} Debt_j + \beta_{MIC} Debt_j \times MIC + \beta_{CPIA} CPIA + \beta_{Growth} Growth) \qquad (1)$$

where $Debt_j$ is the debt variable, with j indexing the five alternatives: 1) the present value of external debt over GDP ($DGDP$), 2) the present value of external debt over exports ($DExp$), 3) the present value of external debt over revenues ($DRev$), 4) debt service over exports ($DsExp$), and 5) debt service over revenues ($DsRev$).[11] Thus, the five stand-alone debt measures are included separately into five different regressions. Their interaction with a dummy for middle-income countries (MIC) controls for a possible heterogeneous effect of external debt across different levels of development. Each regression also includes a measure of policies and institutional quality (the Country Policy and Institutional Assessment—CPIA—score produced by the World Bank) and GDP growth as proxies for governance and economic shocks.[12]

[10] "The unique nature of crises inherently limits the ability of formal statistical tools to extract information that may be useful for identifying the next crisis. 'Preparing to fight the last war' is an obvious pitfall. The [Fund's] EWE thus complements empirical analysis with more heuristic methods, including wide-ranging consultations, as well as judgment informed by economic expertise. Both approaches are complementary: quantitative methods provide a systematic basis for the identification and analysis of vulnerabilities and a useful cross-check on judgment; qualitative analysis helps identify new sources of vulnerabilities and assess consonance among the conclusions stemming from empirical work" (IMF, 2012, p. 15).

[11] A debt distress episode is defined as a period lasting three or more years in which at least one of the following signals of distress is observed: (i) the accumulation of arrears on public guaranteed (PPG) external debt in excess of five percent of the outstanding PPG external debt stock; (ii) a rescheduling of obligations due to Paris Club creditors; or (iii) the disbursement by the IMF of GRA resources exceeding 50 percent of IMF quota.

[12] All explanatory variables are lagged one period to attenuate endogeneity issues. Previous estimates also included

The first five columns of Table 1 show the results of the estimation of equation (1) using the five debt indicators separately. As in IMF and World Bank (2012, Table A3, p. 51), each debt variable is individually highly significant with a positive coefficient, so that more debt is associated with a higher likelihood of debt distress, while the CPIA score and the growth rate are negatively correlated with the probability of a debt distress event.

Again following the DTA, we then search through all values of candidate probability cutoffs $\overline{P_j}$ to find the best one. More specifically, for each candidate $\overline{P_j}$, we calculate the resulting false alarms and missed crises according to Table 2. We also calculate the value of the loss function, defined as a weighted average of false alarms (Type 1 errors, occurring when the model mistakenly predicts a debt distress episode) and missed crises (Type 2 errors, occurring when the model fails to predict a debt distress episode):

$$L = \alpha \times \frac{MC}{A + MC} + (1 - \alpha) \times \frac{FA}{B + FA} \tag{2}$$

where MC is the number of missed debt crises, A is the number of crises that are correctly called, FA is the number of false alarms, and B is the number of tranquil (i.e. non-crisis) periods correctly called. Following the IMF and the World Bank (2012), we set equal weights to Type 1 and Type 2 errors (i.e. we set $\alpha = \frac{1}{2}$).[13] The chosen probability cutoff ($\overline{P_j^*}$) is the value of $\overline{P_j}$ that minimizes the loss function.[14]

So far, we have done nothing more than find the optimal probability cutoff given the probit model, the data, and the loss function. Now, though, the approach taken is quite distinctive. The DSF calculates the associated *debt* threshold for LICs $\overline{D_j^{DTA}}$, by inverting equation (1):

$$\overline{D_j^{DTA}} = \frac{\Phi^{-1}(\overline{P_j^*} - \hat{\beta}_{CPIA} \times CPIA^G - \hat{\beta}_{Growth} \times \overline{Growth})}{\hat{\beta}_{Debt_j}} \tag{3}$$

where $\overline{D_j^{DTA}}$ is the value of the threshold for the debt variable $Debt_j$.

a dummy for African countries and the (log of) GDP per capita and adopted slightly different definitions of the distress indicator and debt ratios. For additional details, see IMF and World Bank (2012, Annex 1 and Table A1), and the 2010 and 2013 DSF guidance notes (IMF and World Bank, 2010, 2013). The list of all the LICs in the debt sustainability analysis and several related documents are available at: http://www.imf.org/dsa.

[13] This loss function is the "preferred method" of finding optimal cutoffs in IMF and World Bank (2012), with $\alpha = \frac{1}{2}$. This formulation is standard in the literature on EWS ((Alessi and Detken, 2011; Lo Duca and Peltonen, 2013)). We return in Section 7 to the specification of the loss function.

[14] This description hides a fair amount of complexity which need not concern us here. In the 2012 revision of the DSF, the IMF and the World Bank (IMF and World Bank, 2012, p.19) derive thresholds using three different concepts of probability of debt distress: (1) the unconditional probability of debt distress; (2) the probability of debt distress corresponding to the median value of the relevant debt burden indicator immediately prior to an outbreak of debt distress; and (3) the probability of debt distress that minimizes the number of missed crises and false alarms. This last option is the preferred one (see IMF and the World Bank (2012, p. 42): this probability "simultaneously minimizes the number of missed crises and false alarms produced by the model, thus ensuring that the thresholds are neither too permissive nor unduly conservative." Our results do not depend on whether approach (1), (2), or (3) is used. However, we follow the third approach because it makes it much easier to demonstrate the internal inconsistency of the WCA approach, as we show below. A further detail is that the weights in the DSF are not strictly speaking equal. Rather, the DSF calibrates the thresholds using the average probability minimizing Type 1 and 2 errors over the different weights, with the relative weight of Type 2 errors being "gradually increased" from one to almost three times the weight of Type 1 errors. The points we make in this paper are also robust to this detail.

Importantly, the threshold $\overline{D_j^{DTA}}$ depends on the values of the other determinants of debt distress in (3). In order to be able to produce a reasonably small number of thresholds, the DSF assigns each CPIA score to one of three categories (low, medium, and high), and these groups are assigned a value for $CPIA^G$ of 3.25, 3.5, and 3.75 respectively. Meanwhile, the country-and-time-specific $Growth$ variable in (1) is replaced by the historic average growth rate for all LICs ($\overline{Growth}$). In this way, there are three values for the debt burden thresholds $\overline{D_j^{DTA}}$ for each debt burden measure D_j (one for each category of CPIA score). The resulting set of debt thresholds is reported in Table 3.[15]

On the basis of the estimated policy-dependent thresholds and on the projected evolution of external public debt stock and flows over the 20 years (the "baseline scenario") and some standardized "stress tests"[16], countries are assigned one of these four possible risk of debt distress ratings:

- Low risk: All the debt burden indicators are well below the thresholds.
- Moderate risk: Debt burden indicators are below the thresholds in the baseline scenario, but stress tests indicate that *at least one threshold* [emphasis added] would be breached if there are external shocks or abrupt changes in macroeconomic policies.
- High risk: *One or more debt burden indicators* [emphasis added] breach the thresholds on a protracted basis under the baseline scenario.
- In debt distress: The country is already experiencing difficulties in servicing its debt, as evidenced, for example, by the existence of arrears.

3.1 Goodness-of-fit

To set a benchmark for comparisons with other models, we can consider standard metrics of the goodness-of-fit (GOF) of this DSF approach. To recap the DTA, we estimate the probits in (1), calculate the associated debt thresholds in (3), and compare debt levels to these thresholds period by period through the estimation sample. When the actual debt indicator is above the thresholds, we "call" a crisis for the next period. We then compare these calls with the actual incidence of crises.[17]

Figure 1 shows the so-called ROC (Receiver Operating Characteristic) curve for each of the five single-variable probits. The ROC represents the effectiveness of a given probit model at correctly classifying crisis and tranquil periods. For any given probability cutoff, the model will produce a certain number of false alarms and missed crises. The ROC graphs this point for all values of

[15] The official values of the cutoffs, as described in IMF and World Bank (2012) actually follow method [2] in footnote 14. These differ slightly from those in IMF and World Bank (2004) and from those shown in Table 3. However, all the resulting probability cutoffs and debt thresholds are about the same, as IMF and World Bank (2012) emphasizes. The last DSF guidance note revises the threshold of debt service over revenue, and incorporates remittances in the denominators of the debt ratios for countries which receive large remittances inflows (IMF and World Bank, 2013, Tables 2 and 5); however, in the paper we continue to refer to the data and methodology outlined in IMF and World Bank (2012), because this document presents in more detail the methodology behind the DSF.

[16] To improve the flexibility of the DSF, the assessment of the risk of debt distress may involve also the use of customized scenarios, if it captures an important vulnerability of the country which is overlooked by the standardized stress tests (IMF and World Bank, 2013). And as discussed in the introduction, there is a role for judgment in the application of the thresholds and the WCA.

[17] It is infeasible to test the full procedure of 20-year baseline forecasts and stress tests outlined above.

the cutoff from 100 to 0 percent. The y-axis measures the rate of crises correctly called or 'true positives' as a share of crisis observations (corresponding to A/(A+MC) in Table 2). The x-axis measures the rate of false alarms or 'false positives' (corresponding to FA/(B+FA) in the Table). Thus for example a cutoff of 0 will yield the point furthest to the north-east on the figure and will imply 100 percent false positives and true positives, while a cutoff of 100 percent (to the extreme south-west) implies zero false alarms and zero true positives, for any model. The further the ROC curves above the 45 degree line the better the model predicts both crisis and tranquil periods. Thus, the area under the ROC (or AUROC) is a measure of overall predictive accuracy of the model that is independent of the policy maker's cutoff. An uninformative model would have a value of 0.5; a perfect predictor would have a value of 1.[18]

The point on the ROC chosen by the DTA is indicated by the green circle in Figure 1, given a value of α (the weight on missed crises) of $\frac{1}{2}$. The associated AUROC is also reported in the Figure. Figure 2 shows the loss function (2) associated with each of the probit regressions. The vertical line correspond to the probability cutoff $\overline{P_j^*}$ chosen by the DTA.

The observant reader will note that the minimum of the loss function in Figure 2 does not correspond to the cutoff chosen in the DTA. To see why, we need to descend deeply into those weeds we mentioned in the introduction. The point that actually minimizes the loss function when using the WCA to call crises is indicated by the red 'x' in Figure 1. To see this distinction between the chosen and the optimal point, note that for $\alpha = \frac{1}{2}$, the loss falls as the point moves to the northwest on the ROC diagram, and equal-loss lines are 45-degree lines in this graph. The tangent between such a 45 degree line and the ROC is at the optimal point.

Why doesn't the chosen DTA cut-off $\overline{P_j^*}$ actually minimize the loss? Because $\overline{P_j^*}$ is chosen to minimize the loss function according to predictions based on the probit in equation (1), which uses country-specific information on the $CPIA$ and growth. However the DTA uses equation (3)—which uses grouped values for the $CPIA$ and sample-average growth—to call crises and calculate goodness-of-fit. Thus, the probability cutoff that would minimize losses according to the DTA is not in general the cutoff that is actually chosen by the procedure.

This difference between the chosen and the optimal cut-off probability for the DTA—the "DTA cut-off discrepancy", to coin a term—is evidently readily fixable by modifying the procedure for choosing $\overline{P_j^*}$.[19] But we need to keep track of it to see how much of the differences we will observe in performance between various methods is due to this discrepancy.

[18] The AUROC is asymptotically normally distributed. For a recent use of the AUROC in a similar context, see Schularick and Taylor (2012), Catão and Milesi-Ferretti (2013), and Drehmann and Juselius (2013). There is one tricky distinction between the text, which describes a standard ROC, and what we do here. Remember that in the DSF and the variants we examine, the threshold is chosen to minimize the loss function. Following this approach, we draw the ROCs in this paper not by calculating the goodness-of-fit for each possible probability cut-off value from 0 to 1, but rather by calculating the optimal cutoff probability as α (the weight on missed crises in the loss function) varies from 1 to 0, and then calculating the associated goodness-of-fit. (Note that $\alpha = 1$ implies full weight on missed crises and hence an optimal cut-off of 0.) In effect, the resulting "alpha" ROC is a convex version of the standard ROC. The two curves coincide for all those cutoffs probabilities that would be chosen for some value of alpha. For the WCA, only the "alpha" ROC can be calculated: it is not possible to map from a given cutoff to a point on the ROC for the WCA, because each of the five debt indicators that make up the WCA is associated with a different probability cutoff. We therefore always use these 'alpha' ROCs here, so we are comparing apples to apples.

[19] In particular, we would choose the $\overline{P}_j$ that actually minimizes the loss according to the DTA, rather than $\overline{P_j^*}$. This would be the minimum of the loss function shown in Figure 2.

3.2 Limitations of the DSF

We can now highlight two narrow but important shortcomings that we will address in this paper. First, the DSF allows little room for country-specific characteristics. Most concretely, the debt thresholds are calculated using a LIC-average GDP growth rate and with three categories for the CPIA score rather than the continuous variable that is in the original probit regression. This methodology ignores potentially useful information contained in the original continuous variables.[20]

This may be an appropriate place to step back and ask, why would such a method have been chosen in the first place? For example, why search for debt thresholds rather than say thresholds for the probability of debt distress itself? This is perhaps a result of the original purpose of deriving thresholds for the debt burden indicators. The determination of these thresholds was an imperative in the context of the original HIPC initiative, because the objective of that initiative was to decide on the amount of debt relief to be delivered to each eligible country. A threshold (or for the HIPC initiative a target) level of say the debt/export ratio provided an objective recipe for allocating the scarce resources available for debt relief across countries. Moreover, thresholds on debt measures are arguably more intuitive than thresholds on proabilities, even if the former are derived from the latter. In the current context, though, this feature implies that the inclusion of further explanatory variables, or even the use of actual CPIA values or growth rates rather than categories, is unfeasible because it would lead to a proliferation of debt thresholds. We deal with this issue in Section 5.

The second shortcoming we address in this paper is that the desire to produce debt burden thresholds has muddied the question of how to aggregate information when the different debt burden measures convey conflicting information. An obvious way to proceed is to run a multivariate probit regression instead of the five separate probit regressions in equation (1). The final column of Table 1 immediately reveals the problem with this approach: none of the debt variables is significant. Table 4 gives an indication of why this is the case, showing the correlation of the five debt burden measures in the estimation sample. They are high enough to be problematic, though perhaps not so high that efforts to extract further information seem futile.

The question naturally arises as how to aggregate the information from the five different debt-burden measures, because in any particular case they may imply different debt risk ratings. Of course, by construction the five-variable probit fits the data at least weakly better than any of the single-debt-measure regressions, as indicated for example by the higher likelihood reported in the table. Before we examine GOF of multivariate methods more carefully, however, we need to examine the DSF method of aggregation.

[20] This loss of information takes place even if the DTA cut-off discrepancy noted above is fixed in the calculation of $\overline{P_j^*}$. There may be some concern that measurement error—endemic to LICs—would reduce the value of the use of the country-specific variables. Of course, the analyst may still make a country-specific estimate based on cross-country average values, if it is the best that can be done in a particular case.

4 The DSF Method of Aggregation: the WCA

Under the DSF, a country in which any one indicator breaches the threshold under the baseline scenario is considered at a high risk of debt distress, even if the other four are safely below their thresholds. We refer to this feature of the framework as the *DSF worst-case aggregator* (WCA).[21] Thus, the WCA is a simple extension of the individual debt thresholds $\overline{D_j^{DTA}}$: a signal is issued (i.e. a crisis is called) if and only if $Debt_j \geq \overline{D_j^{DTA}}$ for any j. We can then calculate GOF following Table 2 and equation (3). Much ink has been spilled on the specification and properties of the individual probits such as those in (1), but none on the empirical features of the WCA.

4.1 Bias

It turns out that the WCA is biased and, in particular, too conservative. This is not a value judgment; rather, the forecasts of the WCA yield too many false alarms and too few missed crises given the loss function that is meant to justify these forecasts. This is because the WCA itself does not have a probability cutoff; rather, it draws on the probability cutoff and thresholds of the underlying variable-by-variable probit regressions in equations (1) and (3). The level of the probability cutoff for each variable is is calculated to minimize the loss function when using the single-debt-variable probit models. When these cut-off levels are applied to the variables in the WCA, where it is enough for any one variable to breach its cut-off, crisis calls are made too often, given the loss function.

We can see a sign of this by looking at the frequency with which the WCA predicts debt distress. In the sample, the individual debt measures call crises from 13 percent (for $DRev$) to 40 percent (for $DsRev$) of the time. The WCA, in contrast, calls crises 56 percent of the time (Table 5).

We can quantify the degree of bias in the WCA. To do so, consider the ROC for the WCA (Figure 3). The point indicated by the circle is that chosen by the DSF, given the loss function and a value of α (the weight on missed crises) of 0.5. The point that actually minimizes the loss function when using the WCA to call crises is indicated by the 'x' in the figure. Again, we can see this by noting that, for $\alpha = \frac{1}{2}$, the loss falls as the point moves to the northwest on the ROC diagram, and equal-loss lines are 45-degree lines in this graph. The tangent between such a 45 degree line and the ROC is at the optimal point. This point corresponds to variable-by-variable cutoffs ranging from 18 to 20 percent for the five debt variables (column 3 of Table 6). This compares to the variable-by-variable cutoffs of around 10-14 percent that minimizes the loss function when applied to each isolated debt measure (column 2).

This result should make sense: a higher probability cutoff for each individual debt measure offsets the feature of the WCA that it calls crises much more often. The difference between (about) 12 percent and 19 percent is a measure of the bias of the WCA: using the debt thresholds

[21] The application of the WCA—and the DSF more broadly—is not mechanical in practice. However, the use of judgment with respect to the use of the debt burden indicators is generally confined to "limited and temporary" breaches of the thresholds. Typically, a stable and significant breach of the threshold by one debt burden measure will imply a high risk rating. We return to this issue below.

associated with the variable-by-variable cutoffs of column 2, as does the DSF, result in more false alarms than is appropriate given the equal weights on missed crises and false alarms in the loss function. To put it another way, a cut-off of about 19 percent would minimize the loss function when using the WCA.[22]

We can translate this bias into implications for the debt thresholds themselves. Column 4 of Table 6 shows what the debt thresholds would be for the optimal cutoffs of about 19 percent (both for countries with medium policy $CPIA = 3.5$). The unbiased thresholds have to be much higher in order to call the correct ratio of missed crises to false alarms, given the properties of the WCA and the loss function. For example, the threshold on debt/GDP goes from 30 to 49 percent.

A final way to quantify the bias of the WCA is to ask what relative weight α on missed crises would justify using the lower debt thresholds employed in the DSF. Increasing α moves the optimal point to the northeast along the ROC curve. The optimal probability cutoff would nearly coincide with the chosen one when $\alpha = 0.67$. In other words, if the weight on missed crises is twice the weight on false alarms, then the thresholds used in the DSF are approximately optimal.

This bias problem is easily solved, of course. For example, the higher debt thresholds reported in Table 6 would yield a frequency of missed crises and false alarms consistent with the loss function with $\alpha = \frac{1}{2}$. Alternatively, an acknowledgment that in fact there should be a much higher weight on missed crises than false alarms in the loss function ($\alpha = \frac{2}{3}$) would justify the continued use of current debt thresholds. Absent some calculation of the opportunity cost of missed borrowing opportunities and of the cost of debt crises, the weight α is after all somewhat arbitrary. If the WCA reflects implicitly the beliefs of staff, it might make sense to reflect them in the loss function weights. Transparency would be the main benefit.

4.2 Accuracy

Bias is only part of the story. A distinct question is whether the WCA is an *accurate* way to predict crises. We can get a start on the question by comparing the goodness-of-fit of the WCA to that of the predictions based on the individual debt measures. This is a low bar, as one would hope that the use of additional information allows better predictions, at least in-sample. As we will see, this does not seem to be the case.

Figure 4 shows the ROC for the best individual debt measure (the ratio of debt service over exports) along with that for the WCA. We can see that the AUROC and the value of the loss for the WCA are very close to that of $DsExp$. Indeed the AUROC (at the points chosen by the DTA) is actually a bit lower than using just the single variable, suggesting that there are no gains in terms of additional accuracy from the greater information potentially available in the WCA.

Remember, however, from Table 1 that the multivariate probit had a substantially higher log-likelihood than did any of the single-debt-measure probits. This suggests that more sensible aggregators than the WCA may do better than any single debt measure. To see if this is true, we would like to examine the GOF of the probit (or other aggregators) and compare to the WCA.

[22] The DTA cut-off discrepancy we saw in section 3.1 looks similar to this difference, but there is an important distinction: that discrepancy did not represent a bias but rather an essentially random optimization error.

To do this, we need a way to calculate missed crises and false alarms for all these alternative aggregators on a comparable basis. The approach of inverting the univariate probits and calculating debt thresholds lends itself to the WCA but not to the multivariate probit or other aggregators, since the thresholds for each debt variable would depend on all the others. Any resulting system of debt thresholds would be hopelessly complex.

Fortunately, there is another approach that addresses two of the important weaknesses of the DSF discussed above: it allows the use of country-specific and continuously-measured variables such as the $CPIA$, and it facilitates aggregation. We describe this approach now.

5 The Probability Threshold Approach

A simple way to include country-specific information in the risk ratings, without substantially departing from the original framework, is to express the thresholds in terms of probabilities, rather than debt ratios.[23] This approach consists in estimating model (1) and, then, retrieving the estimated probabilities of debt distress on the basis of country-specific $CPIA$ scores, growth rates and debt projections. The tolerable threshold for the probability of debt distress (the one that was used in equation (3) to 'invert' the probit and infer a debt burden threshold) can instead itself become the threshold against which to compare projections of the debt-distress probability. Thus, if the projected probability of debt distress rises above this probability threshold in the baseline scenario, the country would be considered to be at high risk of debt distress.

Thus, we can again run the five probit models (1) with the debt measures inserted one at a time. For each of the debt measures indexed by j, we now calculate the resulting estimated probabilities of debt distress P_j. We can then iterate to find an optimal probability cutoff $\overline{P_j^*}$ such that, if we call crises if and only if $P_j \geq \overline{P_j^*}$, then we minimize the same loss function (2). We call this the "probability threshold approach" or PTA.

With the PTA, the key element of the debt sustainability assessment is the evolution of the five forecasted probabilities of debt distress (one for each debt indicator) that are compared with a sample-wide probability threshold. This probability approach is more transparent insofar as there is no need for the move from the probit to the debt burden thresholds.[24] Moreover, it is easy to use country-specific CPIA scores or growth rates, or for that matter to add additional explanatory variables.

Comparing the GOF of the DTA for one debt measure versus the PTA allows us to compare directly the benefits of using continuous covariates such as the $CPIA$. Figure 5 shows the ROC for the DTA for the best-performing single debt variable $DsExp$ as well as for the same variable using

[23] IMF and World Bank (2012) discusses the possibility of a *probability approach*, which is included in the last revision of the DSF (IMF and World Bank, 2013) as an alternative methodology to use for assessing the risk of debt distress in borderline cases (when the largest breach, or near breach, of a threshold falls within a 10-percent band around the threshold).

[24] Alternatively, some argue that the simple thresholds for individual debt burden measures are easier to understand than probabilities and thus a better tool to guide discussions on how to avoid the risk of debt distress. However, the probabilities underlie the debt thresholds, so any "understanding" of the debt thresholds that does not encompass an understanding of the underlying probabilities may be too shallow to form the basis for an ideal policy dialog.

the PTA.[25] The AUROC increases from 0.792 to 0.814 and the value of the loss function falls from 0.31 for the DTA to 0.25 with the PTA. This gain can be decomposed into two components. The first is the reduction in loss to 0.26 from choosing the optimum cutoff while still using the DTA (eliminating the DTA cut-off discrepancy explained in section 3.1). This is the move from the green circle to the red x on the DTA ROC in Figure 5. The further reduction to 0.25 reflects the gain from using continuous values of $CPIA$ and $Growth$ rather than following the DTA approach of grouping the $CPIA$ and using the LIC-average value of $Growth$. This is the move from the red x on the DTA ROC to the blue circle on the PTA ROC in the Figure.

The right-hand panel of Figure 5 illustrates the same points with the loss functions. Eliminating the DTA cut-off discrepancy explained in section 3.1 involves moving from the green circle (DTA chosen) to the red x (DTA optimal).[26] The second component of the increase in the fit of the model involves moving from the red x to the blue circle (PTA optimal). Here the gain is due to the use of country-specific rather than grouped values for the CPIA and growth.

We identified two problems with the DSF in Section (3): an inability to use country-specific covariates and difficulties in using standard multivariate methods. We have seen that the probability approach solves the first problem and in so doing substantially improves goodness-of-fit. Even more important, as we shall see, is that the probability approach facilitates the aggregation of the information contained in the different debt burden measures. It is now a simple extension to include more than one debt burden measure in the probit regression, still comparing the predicted probability to a tolerable threshold. Having established the superiority of the probability approach for a single debt measure, we now continue with this approach to take a closer look at alternative aggregators.

6 How to Aggregate the Debt Indicators

The evidence from the multivariate probit suggests that there are potential information gains from using multiple debt measures. But as the probit itself also shows, high multicollinearity among the various debt measures makes it difficult to see how to achieve these gains with confidence. In this section we examine several possible aggregators and compare their accuracy to that of the WCA. In particular, we will discuss several different composite indicators, each of them a specific linear combination of the five debt indicators:

$$CI_k = \sum_{j=1}^{5} w_j \times Debt_j \tag{4}$$

where k indexes the four methods and w_j are the weights associated to the single debt indicators $Debt_j$.

We judge the performance of each composite indicator by including it in equation 1 as a replace-

[25] Similar considerations hold when using the other four debt indicators, but figures are not shown for brevity.

[26] It is not a coincidence that the points chosen on the loss function for the PTA and the DTA line up exactly, given that the cutoff point chosen for the DTA (which minimizes the loss associated with using the probabilities, not the thresholds, to call crises) is none other than the one used in the PTA.

ment for the debt measures (alone and interacted with the middle-income-country dummy, and keeping both growth and the $CPIA$), and then analyzing the associated goodness of fit measures.

6.1 Equal Weights

The WCA has two remarkable features. First, it calls crises based on the most alarming debt measure for any particular observation. Second, and as a result, it makes no use of the data to determine the relative weight to be assigned to each of the alternative debt measures; all are equally important. We start by testing the first feature of the WCA. We construct an aggregator we dub the Equal-Weights Composite Indicator (CI_{EW}) that represents a simple average of the five debt measures in the WCA:

$$CI_{EW} = \sum_{j=1}^{5} 0.2 \times Debt_j \tag{5}$$

where j as usual indexes the five debt measures. By comparing the accuracy of this method to the WCA, we can answer the question of whether the worst-case approach or an averaging approach is more accurate, in both cases making no use of the data to determine relative weights.

We can see from the left-hand panel of Figure 6 that the CI_{EW} is much more accurate than the WCA. It has a larger AUROC and a lower value of the loss function. This difference in accuracy is statistically significant at the 1 percent level.[27] It is worth underscoring this important result: putting aside bias and the question of if or how to make use of the data to attach weights to the different debt measures, the WCA is inaccurate relative to a simple average.[28]

6.2 Multivariate Probit

We now turn to the question of whether the data—as collinear as they are—contain enough information to improve on the accuracy of the CI_{EW}. This is a harder question to answer. Any method that allows the data to estimate the parameters will by construction have smaller average errors over the estimation sample than a method that restricts the parameters. But this apparent increase in accuracy could be spurious: the application of the estimated parameters to a new set of data, even when generated by the same process that produced the estimation sample, could result in less accurate and more volatile predictions. This "overfitting" problem is aggravated by the collinearity of the debt measures, which means that the available data do not pin down the parameter values tightly.

[27] For this calculation we draw 1,000 bootstrap samples from our data and calculate the AUROC of the WCA and the CI_{EW} for each sample. In only 3 of the 1,000 samples is the AUROC of the WCA higher than that of the CI_{EW}. Note that, as discussed in section 3.1, the WCA is impaired partly by the fact that the cutoffs chosen for each individual debt measure are not optimal. Even using the optimal cutoffs discussed in section 3.1, the AUROC of the WCA is lower than that of the CI_{EW} with a p-value of 0.075.

[28] The relatively good performance of the CI_{EW} reassuringly echoes the results from macroeconomic forecasts in Stock and Watson (2004), who describe the "forecast combination puzzle" as the finding that simple combination forecasts, in particular, combinations that look a lot like unweighted averages, tend to do well in empirical applications. It may reflect simple debt-variable-specific measurement error — likely especially prevalent in LICs — which could make extreme values of any one debt variable particularly suspect and thus the WCA a particularly inaccurate aggregation method.

We draw on a standard method of addressing this problem, which is to use measures of goodness of fit that start by measuring the sum of squared errors of the estimation but then impose a penalty for each additional estimated parameter. In particular, we focus on the Bayesian Information Criterion (BIC), for which a lower number implies a better fit, accounting for the tendency of models with larger numbers of even spurious parameters.[29]

We define the Multivariate Probit (MP) Composite Indicator (CI_{MP}) as:

$$CI_{MP} = \sum_{j=1}^{5} \beta_j \times Debt_j \tag{7}$$

where β_j are the coefficients on the five debt measures (indexed as usual by j) in the multivariate probit as reported in the last column of Table 1.[30] We find the $DsRev$ has the largest weight, while the $DExp$ has a small weight and $DRev$ even a negative weight (Table 7). All these weights, however, are imprecisely estimated (see Tables 1 and 4).

The CI_{MP} is more accurate than the CI_{EW}, as the right hand-side panel of Figure 6 shows. The AUROC rises from 0.84 to 0.87 (Table 7). However, this improvement is likely the result of the multivariate probit estimating more parameters than the CI_{EW}. Indeed, the BIC increases from 333.0 to 372.4, a difference that is statistically significant.[31]

We may want to stop here and use the CI_{EW}. However, one may think that there may be a better way to aggregate the information than the CI_{EW}. In addition, some may feel that the inclusion of all five variables requires a statistical justification, so that it could be more efficient to drop some debt variable with a relatively poor predictive power. We thus, and after substantial exploration, we look at two families of restricted models that fall somewhere between the CI_{EW} and the CI_{MP} in how they trade off higher in-sample accuracy against more reliable coefficients and predictions.[32]

6.3 More Parsimonious Data-Based Models

The main concern with the CI_{MP} is that the weights are imprecisely estimated and may be extremely sensitive to small sample variations. Is there a more parsimonious model than the

[29] The Bayesian Information Criterion (BIC) is calculated as:

$$BIC = -2 \times ln(likelihood) + ln(N) \times k \tag{6}$$

where k is the number of parameters and N the number of observations. The BIC can be viewed as a measure that combines fit and complexity and tries to balance them. Fit is measured negatively by $-2 \times ln(likelihood)$ (the larger the value, the worse the fit), while complexity is measured positively, by $ln(N) \times k$. Given two models fit on the same data, the model with the smaller value of the information criterion is considered to be better.

[30] As with the other indices, we first standard the debt measures by dividing by their standard deviation, so that the coefficient values can be directly compared.

[31] To calculate statistical significance, we use the same bootstrap technique mentioned in footnote 27, except that here we calculate the BIC for each bootstrap sample, to adjust for the fact that the CI_{MP} estimates many more parameters than the CI_{EW}. In only 18 out of 10,000 of these samples is the BIC of the CI_{MP} lower than that of the CI_{EW}, for a p-value of 0.018.

[32] We also examined the use of univariate measures of predictive accuracy as weights for the five variables, such that each $w_j = AUROC_j / \sum_{k=1}^{5}(AUROC_k)$ in equation 4. The resulting composite index performed indistinguishably from the CI_{EW}, reflecting the similarity of the AUROCs in the single-debt-measure models.

CI_{MP} for which the parameters can be estimated with more confidence and which provides as good or better goodness-of-fit? It turns out the answer to this question is 'yes'. Unfortunately, though, we have an embarrassment of riches, in that there are several such models, and the data do not allow us to say clearly which one is best.

6.3.1 Step-wise models

An efficient way to see the challenge is to examine the BIC for every possible probit model with from one to all five of the debt variables (there are 31 such models), as presented in Table 8.[33] It turns out that the model with the lowest BIC has just one debt variable: *DsRev*. The BIC is nearly as low for the best two-variable model, which includes just *DsRev* and *DExp*. Either of these two models is significantly better (at the 10 percent level) than the full five-variable model CI_{MP} and many of the more parsimonious models. However, there is no strong statistical basis for picking among the best parsimonious models. Table 8 highlights in particular the five parsimonious models whose performance according to the BIC cannot be distinguished from the CI_{EW}. For these five models the hypothesis that the BIC is equal to that of the CI_{EW} cannot be rejected at the 10 percent level.[34] The other models perform significantly worse than the CI_{EW}.

We present the best two-variable model in Figure 6 (panel c) and Table 7 (column 3), dubbed the CI_{SW}.[35] The AUROC is almost identical to those of the CI_{MP} and CI_{EW}. The BIC is significantly lower (at 1 percent level) than that of the CI_{MP}, reflecting the greater parsimony of the model.[36]

Given the similar AUROCs and number of estimated parameters, it should not be surprising that the BIC criterion does not reveal a winner when comparing the CI_{SW} and the CI_{EW}. Thus, so far we have no basis for choosing between these two models. We return to this issue after analyzing one more family of models.

6.3.2 Equal-Weight-Prior

An alternative proposed parsimonious model starts from a different starting point: the CI_{EW} rather than the CI_{MP}. We may have a strong prior (i.e. not data-based) belief that all five debt variables are worth of attention in assessing the risk of debt distress, a belief implicit in the WCA

[33] All the models include the other regressors and the middle-income interactions as in equation (1)

[34] These statistical calculations follow the method outlined in footnote 31.

[35] There are many alternative ways of arriving at a similar model, and we have tried many of them. For example, one possible way to tackle the collinearity problem is to simplify the five-variable model by eliminating statistically insignificant variables until those that remain are significant (i.e. a stepwise 'general-to-specific' approach). This has superficial merit in our application, in that such a procedure that starts by dropping the least significant of the five debt variables identifies a single specification with two surviving variables, *DsRev* and *DExp*. However, this conclusion is path-dependent. Dropping a slightly less insignificant variable in the first step will tend to yield a different pair of variables in the final specification. Bootstrapping this process results in many possible two-variable end-points to the procedure, none of which is a clear winner at standard p-values. The most successful, the model with *DsRev* and *DExp*, wins in only 37 percent of the bootstrap samples. We pick the two-variable model with *DsRev* and *DExp* and name it as we do because it has (nearly) the lowest BIC and it is also the end-point of the step-wise procedure just described.

[36] Unlike the CI_{MP}, the coefficient values are tightly estimated and are statistically significant in the traditional sense (not shown). Of course this is only true conditional on the assumption that this particular two-variable model is the correct one, an assumption we cannot defend statistically.

approach. In this case, we may want to demand statistical evidence for a deviation from the CI_{EW}, rather than for the inclusion of the variable at all. We look for a good-fitting and parsimonious model that only deviates from equal weights when the data clearly demand it.

We start again by defining the CI_{EW} (equation 5) and then we estimate the following probit:

$$\begin{aligned} Prob(y_{it} = 1) = \Phi(&\beta_{CI_{EW}} CI_{EW} + \beta_{MIC} CI_{EW} \times MIC + \\ &+ \sum_{j=1}^{4} \beta_j \times Debt_j + \beta_{CPIA} CPIA + \beta_{Growth} Growth) \end{aligned} \tag{8}$$

in which we include four of the five individual debt measures plus the equal weights composite index (we must omit one variable to avoid perfect collinearity). We then examine the BIC for all of the models (there are 30 of them) in which we include from one to four of the five debt variables *along with* the CI_{EW}.

We find that the data reject the less parsimonious models. Indeed, only two models—both single-variable—are competitive with the CI_{EW}: those with an additional parameter to distinguish *DsRev* and *DExp*.[37] The best (lowest-BIC) model is again the one with *DsRev* (again, the best two-variable model is that with *DsRev* and *DExp*, but it does not seem to perform as well as the CI_{EW}). We thus define a new composite indicator we dub "Equal-Weight-Prior Composite Indicator" (CI_{EWP}). The weights are equal to 0.2 for all the four debt variables for which we cannot find clear evidence against equal weights and to 0.55 for *DsRev*.[38] Figure 6 (panel c) and Table 7 (column 4) show the ROC, AUROC, and loss function for this model.

The choice between the CI_{EW} and the CI_{EWP} is not obvious. The former is a restricted version of the latter and thus by construction has slightly worse fit, as indicated its slightly lower AUROC. The BIC of the CI_{EW} is lower, suggesting that the CI_{EWP} does not yield a big enough improvement in fit to justify the larger number of parameters required. Neither difference is statistically significant, however.[39]

6.4 Summary on model selection

We have examined five aggregators for accuracy: *WCA*, CI_{MP}, CI_{EW}, CI_{SW}, and CI_{EWP}. The *WCA* is a clear loser, in that it is less accurate (as well as biased) relative to the CI_{EW}. The CI_{MP} is also a loser: it has a statistically-significantly higher BIC than the final three models, implying that its slightly smaller errors do not justify the larger number of estimated parameters.

Unfortunately, there is no clear winner among the rest. The CI_{EW} has the lowest BIC and the CI_{EWP} has the highest, but this difference is not statistically significant. Moreover, the data do not say which of the many possible CI_{SW} and CI_{EWP} models is reliably the best within each family. It does seem clear, though, that any chosen model should have at most two variables singled out, and that the generally most promising candidates would seem to be *DsRev* and *DsExp*. But

[37] We again follow the same statistical approach as described in footnote 31.

[38] The weight in *DsRev* comes from: $0.55 = 0.47 + 0.2 \times \beta_{CI_{EW}}$, where $\hat{\beta}_{CI_{EW}} = 0.38$ and 0.47 is $\hat{\beta_j}$ of *DsRev* in equation 8.

[39] To test the significance of the difference in AUROCs, we bootstrap the AUROC for the CI_{EW} and note that the in-sample value of the AUROC for the CI_{EWP} is well within the resulting distribution.

to show how muddy the waters are, each of the five debt variables is included in at least one model whose performance cannot be statistically distinguished from the other competitive models (again, we can see this in Table 8).[40]

In our view, the choice among these three approaches depends on two considerations: the strength of prior beliefs that all five variables should be included in the model, and tolerance for sample-dependence with respect to specification (variables, parameter values) (see Table 9). Our own judgment falls in favor of the CI_{EW}: we accept the view embedded in the current DSF that none of the variables should be neglected, and we find unattractive the feature that the choice of which specific variables to include and the weights to be assigned should be strongly dependent on small sampling variations. However, tastes may differ. There is one blank cell in Table 9: weak priors about whether all five variables matter combine with low tolerance for sample-specific specification to yield no good model.[41]

7 Revisiting the loss function

So far, we have expressed the loss function as in equation (2), assuming equal weights on false alarms and missed crises, as is in the DSF (though see footnote 14).

We saw in Section 4 that the WCA actually implies that the weight on missed crises is implicitly twice as high as on false alarms, even when the notional weight (the α in (2)) is $^1/_2$. In this section we look at another, independent, reason why the current DSF is more conservative than implied by the notion of "equal weights on false alarms and missed crises."

To see this, we must look closely at the loss function (2). Following most of the literature, the weight on false alarms α is multiplied not by the *absolute number* of missed crises but by the number of missed crises *as a share of total crises*. Similarly, $(1-\alpha)$ multiplies the number of false alarms as a share of tranquil (i.e. non-crisis) periods. Thus, if there are many fewer crisis observations than tranquil periods in the sample (in our sample, only 12.1 percent of observations are crisis observations), the weight attached to each missed crisis observation is much higher than the weight attached to each false alarm.

To see this, and to retrieve the weights actually assigned to each false alarm and missed crisis episode, we can define $\alpha' \equiv \frac{\alpha}{A+MC}$ and $(1-\alpha') \equiv \frac{(1-\alpha)}{B+FA}$. Rearranging and substituting into (2) yields:

$$L = \alpha' \times MC + (1-\alpha') \times FA \tag{9}$$

Thus, the weights α' and $(1-\alpha')$ are the weights given to the actual number of missed crises and false alarms. Hence, the relative weight attached to false alarm observations relative to missed crisis *observations* is $\frac{(1-\alpha')}{\alpha'} = \frac{B+FA}{A+MC}$. In our sample there are 186 tranquil periods and 17 crisis

[40] Our expectation was that these different models might perform similarly on average but yield very different results observation by observation. This turns out not to be the case, as there is a great overlap observation-by-observation in the predictions of different aggregators. For instance, CI_{SW} and CI_{EW} predict, respectively, 161 and 167 debt distress episodes, with an overlap of 141 events.

[41] Of course, more major changes to the framework might yield more interesting results. For example, if the regressions distinguished between solvency and liquidity crises, then the different debt measures might play more distinct roles. This is beyond the scope of this paper.

periods, so that $\frac{B+FA}{A+MC} = 10.9$. This implies that the DSF "equal" weighting scheme used in the loss policy function (2) actually weights missed crises 10.9 times more than false alarms.

And this is in addition to the conservative bias in the WCA that doubles the relative weight on missed crises. Putting these two factors together, the DSF implicitly weighs each missed crisis as roughly 22 times more important than each false alarm.

What if we used truly equal weights? In other words, what if we set $\alpha' = (1 - \alpha')$? Figure 7 shows the implications for the optimal loss function, for the CI_{EW}. The optimal cutoff is extremely high (between 0.61 and 0.79), and the loss is only trivially lower if a cutoff of 1, which would mean no crises would be called. This happens because the combination of the very low incidence of actual crisis in the sample with the relatively low accuracy of even the best of these predictors makes calling a crisis a dubious proposition even when the predictors are taking on relatively alarming values. Figure 7 also shows that if a cutoff such as those used in earlier sections is used, but with this truly equal-weighted loss function, the losses would be much higher.

8 Discussion and conclusions

A large number of developing countries experienced over-borrowing, unsuccessful public investment programs, and many years of macroeconomic disarray and poor growth in the 1980s and 1990s, followed by the protracted resolution process in the form of the HIPC Initiative (1996), the enhanced HIPC (1999), and MDRI (2006). With renewed debt carrying capacity and market access, all involved are eager to avoid going through that again. At the same time, a return to growth, stronger borrowing prospects, still-great needs for public investment, and insufficient aid flows have left countries wanting to borrow more. There is thus a clear need for a framework to balance these competing imperatives.

The Debt Sustainability Framework (DSF) adopted by the World Bank and the IMF in low-income countries is a rule-based empirically-justified approach to assessing the risk of debt crisis. It is based on the estimation of five separate specifications of the risk of debt distress, separately including five stand-alone debt burden measures, and on the calibration of corresponding debt thresholds, conditioned on the quality of policy and institutions. This framework is intuitive, easily manageable, and provides plausible indications.

In this paper, we focus on the question of how the DSF uses and in particular aggregates the information contained in the five separate indicators. We draw three conclusions about the current approach.

First, we find that the DSF approach of using debt thresholds, which implies limited use of country-specific values for the covariates in the crisis prediction probits, results in a loss of information that reduces the accuracy of predictions.

Second, we find two distinct problems with the DSF method of resolving conflicting signals from the five individual debt measures. This method, which we dub the *worst-case aggregator* (WCA) and which has not been examined empirically until now, calls a crisis when any one of the five debt measures is above its threshold, whatever the value of the other four. It turns out that the WCA is implicitly *biased*, in that it is more conservative—it calls crises more often—than can be justified by the purported weights attached to missed crises and false alarms in the loss function. In addition, the WCA is statistically *inaccurate.* By this we mean that, for any loss function weights, it is possible to find simple alternative aggregators that have a better in-sample combination of missed crises and false alarms than the WCA.

Third, we highlight an additional source of implicit conservative bias in the DSF, related to the nature of the weights used in the loss function. The IMF and the World Bank (2012) suggest that a loss function with "equal weights" on false alarms and missed crises can justify the thresholds used in the DSF. In fact, though, this is so only insofar as—following the early-warning system literature as a whole—equal weights are applied to crises observations *as a share of total crises*, and false alarms *as as share of total tranquil periods.* Because there are many more tranquil periods than false alarms, the implicit weights in the DSF on each missed crisis observation is at least 10 times higher than on each false alarm. Taking into account in addition the bias in the WCA itself, the effective weight on missed crises is 22 times higher than on false alarms.

We also suggest some solutions. The bias problem is easy to fix. For example, the thresholds on each individual debt indicator could be raised so that when the WCA is used, the frequency of false alarms falls so that it does in fact minimize the loss function that gives equal weight to crisis observations as a share of total crises.

Alternatively, it might be argued that the conservative bias of the WCA, and the much higher implicit weights on missed crises than false alarms, give the right answer for the wrong reasons. There are two lines of argument here. First, it may be that other biases in the overall process of assessing debt sustainability, such as overly optimistic growth projections or assessments about the benefits of public investment, may lean in the other direction.[42]And second, it may be that the expected cost of a false alarm—and possible foregone borrowing—is in fact much lower than the cost of a missed crisis.

It is not clear to us that either of these arguments fully justifies the bias in the current approach. It seems plausible that there are other optimistic biases in the overall system. But it would be lucky if two wrongs made a right in this case. If fixing the bias in the DSF causes the system to be too optimistic, the solution is not to leave the problem buried but to have a serious discussion about growth forecasts and other possible sources of an optimistic bias. Similarly, it would be better to have an open and broad-based discussion about the relative costs of false alarms and missed crises than to obscure the question. These relative costs depend on some considerations well within the IMF's expertise and usual concerns, notably debt crises and macroeconomic disarray, and some well outside, such as the welfare payoffs to alternative spending trajectories.[43]

In contrast with bias, there would seem to be little likelihood of disagreement about the goal of improving the *accuracy* of the DSF. Here, we have proposed clear improvements. There is substantial multicollinearity among the five debt burden indicators in the DSF, but there is nonetheless enough information in the different measures that there are alternative simple aggregators that perform significantly better than the WCA. In particularly, simply averaging the five debt indi-

[42] Arguably, the bias created by overly optimistic growth projections is already flagged by the use of a "historical values" stress test in the DSF itself.

[43] On this point see Goldsbrough (2007).

cators into a single index, which we dub the CI_{EW}, produces forecasts of debt distress that are more accurate than the WCA, i.e. they will tend to call some combination of fewer missed crises and fewer false alarms.

There is some suggestion from the analysis that it may be possible to use the information in the data to do somewhat better than simply assigning equal weight to each of the five variables, but the evidence is not decisive on this point. In particular, we analyze some specific alternatives to the CI_{EW} that do somewhat better in predicting crises in-sample, but it is simply not clear whether this improvement is real or reflects the vagaries of the particular data sample.

The narrow perspective in this paper—taking most of the DSF framework as given—has allowed us to focus in on some important weaknesses and suggest solutions. But it also means that the implementation of these solutions may call for a broader rethink of the DSF than we conduct in this paper. For example, under the "right-answer-for-the-wrong-reasons" view of the current DSF, fixing the bias associated with the WCA would "upgrade" many countries in a way that a fuller analysis, taking into account other possible sources of bias, might not. Of course, as we have emphasized, hoping that these various weaknesses happen to cancel seems not the best way forward.

Stepping back, it should be clear that we are asking a lot of scarce data and simple techniques to predict the complex phenomenon of debt crises in low-income countries. We have uniformly used in-sample prediction as the benchmark. But we can be confident that the next round of crises will be at least a bit different from the last, at least in the details. More broadly, no simple empirical model can hope to capture the complexities of each case. In particular, there is perhaps a false precision in suggesting that any multivariate indicator can get the projections right. In this context, paying attention to the broader context in which debt sustainability is assessed would seem wise, combining judgment with the statistical models.

We do not believe there should be a complete swing towards judgment, however. Contexts such as the prediction of debt distress in LICs are ones in which such judgment may not work particularly well. Kahneman (2011) emphasizes that, in general, judgment works well when applied in situations where the judge receives frequent and rapid feedback as to how she is doing.[44] This is not the situation of the LIC DSF, where crises are few and far between. In this case, heuristics and biases may generate large systematic errors in subjective judgments, and a simple rule is often proved to be superior to experts judgments (Gilovich *et al.*, 2002; Ashenfelter, 2008). In this spirit, good multivariate predictors should, we believe, serve as useful tools.[45] In sum, then, our results suggest that a multivariate indicator like the CI_{EW} could usefully replace or at least complement the WCA in the DSF.

In this paper, we have kept very close to the current DSF approach of finding thresholds based on a very simple panel regression using a small number of variables. Looking forward to the possible evolution of the DSF, we are not confident that further large gains can be had from fine-tuning

[44] "If subjective confidence is not to be trusted, how can we evaluate the probable validity of an intuitive judgment? [. . .] The answer comes from the two basic conditions for acquiring a skill: an environment that is sufficiently regular to be predictable, and an opportunity to learn these regularities through prolonged practice. When both these conditions are satisfied, intuitions are likely to be skilled" (Kahneman, 2011).

[45] For a related argument in the context of early warning systems of currency crisis, see Bussiere (2013)

these panel regressions. Thus, entirely different approaches may play a role. For example, it may be useful to emphasize the role of the fiscal reaction function in determining sustainability, as in Mauro *et al.* (2013) and Ghosh *et al.* (2013). The country-specific framework in Buffie *et al.* (2012) has a similar emphasis, along with the possibly more LIC-specific focus on public investment.

References

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

Ashenfelter, O. (2008). Predicting the quality and prices of bordeaux wine. *Economic Journal*, **118** (529), F174–F184.

Bandiera, L., Cuaresma, J. C. and Vincelette, G. A. (2010). Unpleasant surprises: Determinants and risks of sovereign default. In C. A. Primo Braga and G. A. Vincelette (eds.), *Sovereign Debt and the Financial Crisis: Will This Time Be Different?*, Washington, DC: The World Bank.

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

Buffie, E. F., Portillo, R., Zanna, L.-F., Pattillo, C. A. and Berg, A. (2012). *Public Investment, Growth, and Debt Sustainability: Putting Together the Pieces*. IMF Working Papers 12/144, International Monetary Fund.

Bussiere, M. (2013). *In Defense of Early Warning Signals*. Working papers 420, Banque de France.

Catão, L. and Milesi-Ferretti, G.-M. (2013). *External Liabilities and Crises*. IMF Working Papers 13/113, International Monetary Fund.

Drehmann, M. and Juselius, M. (2013). *Evaluating early warning indicators of banking crises: Satisfying policy requirements*. BIS Working Papers 421, Bank for International Settlements.

Ghosh, A. R., Kim, J. I., Mendoza, E. G., Ostry, J. D. and Qureshi, M. S. (2013). Fiscal fatigue, fiscal space and debt sustainability in advanced economies. *Economic Journal*, **123** (566), F4–F30.

Gilovich, T., Griffin, D. and Kahneman, D. (eds.) (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Goldsbrough, D. (2007). *Does the IMF Constrain Health Spending in Poor Countries?* Report, Center for Global Development.

IMF (2012). *The IMF-FSB Early Warning Exercise: Design and Methodological Toolkit*. Occasional Paper 274, International Monetary Fund, Washington DC.

IMF and World Bank (2004). *Debt Sustainability in Low-Income Countries - Proposal for an Operational Framework and Policy Implications*. Tech. rep., IMF and The World Bank, Washington, DC.

— and World Bank (2010). *Staff Guidance Note on the Application of the Joint Bank-Fund Debt Sustainability Framework for Low-Income Countries.* Tech. rep., IMF and The World Bank.

— and World Bank (2012). *Revisiting the Debt Sustainability Framework for Low-Income Countries.* Tech. rep., IMF and The World Bank, Washington, DC.

— and World Bank (2013). *Staff Guidance Note on the Application of the Joint Bank-Fund Debt Sustainability Framework for Low-Income Countries.* Tech. rep., IMF and The World Bank, Washington, DC.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Kaminsky, G., Lizondo, S. and Reinhart, C. M. (1997). *Leading indicators of currency crises.* Policy Research Working Paper Series 1852, The World Bank.

Kraay, A. and Nehru, V. (2006). When is external debt sustainable? *World Bank Economic Review*, **20** (3), 341–365.

Lo Duca, M. and Peltonen, T. A. (2013). Assessing systemic risks and predicting systemic events. *Journal of Banking & Finance*, **37** (7), 2183–2195.

Manasse, P. and Roubini, N. (2009). "rules of thumb" for sovereign debt crises. *Journal of International Economics*, **78** (2), 192–205.

Mauro, P., Romeu, R., Binder, A. and Zaman, A. (2013). *A Modern History of Fiscal Prudence and Profligacy.* IMF Working Papers 13/5, International Monetary Fund.

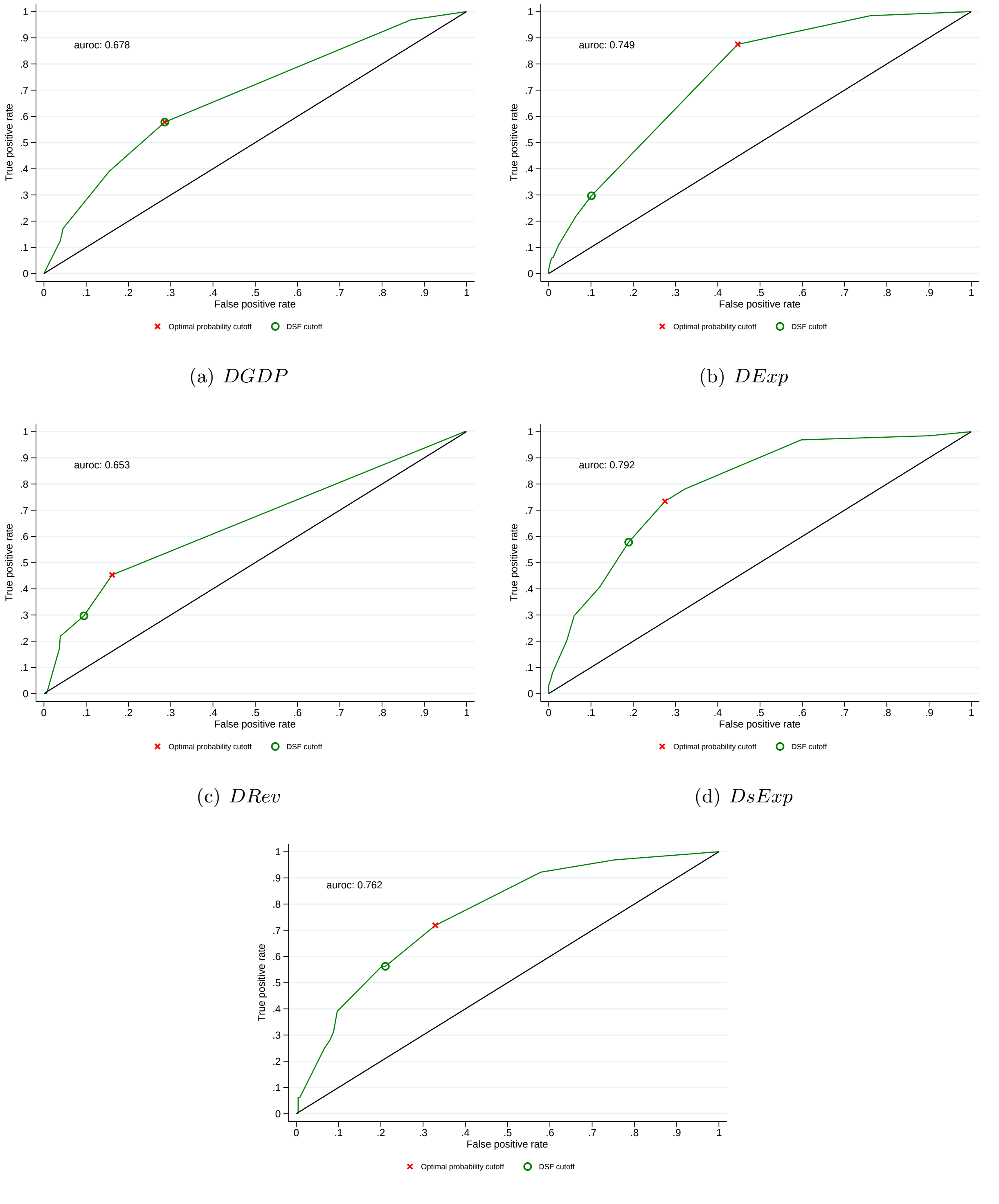
Pradelli, J. (2012). On external debt sustainability: Default probabilities and debt thresholds to monitor risk of distress, The World Bank.

Schularick, M. and Taylor, A. M. (2012). Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870-2008. *American Economic Review*, **102** (2), 1029–61.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, **23** (6), 405–430.

FIGURES

Figure 1: The DTA: trade-off of false alarms and missed crises.



(a) *DGDP*

(b) *DExp*

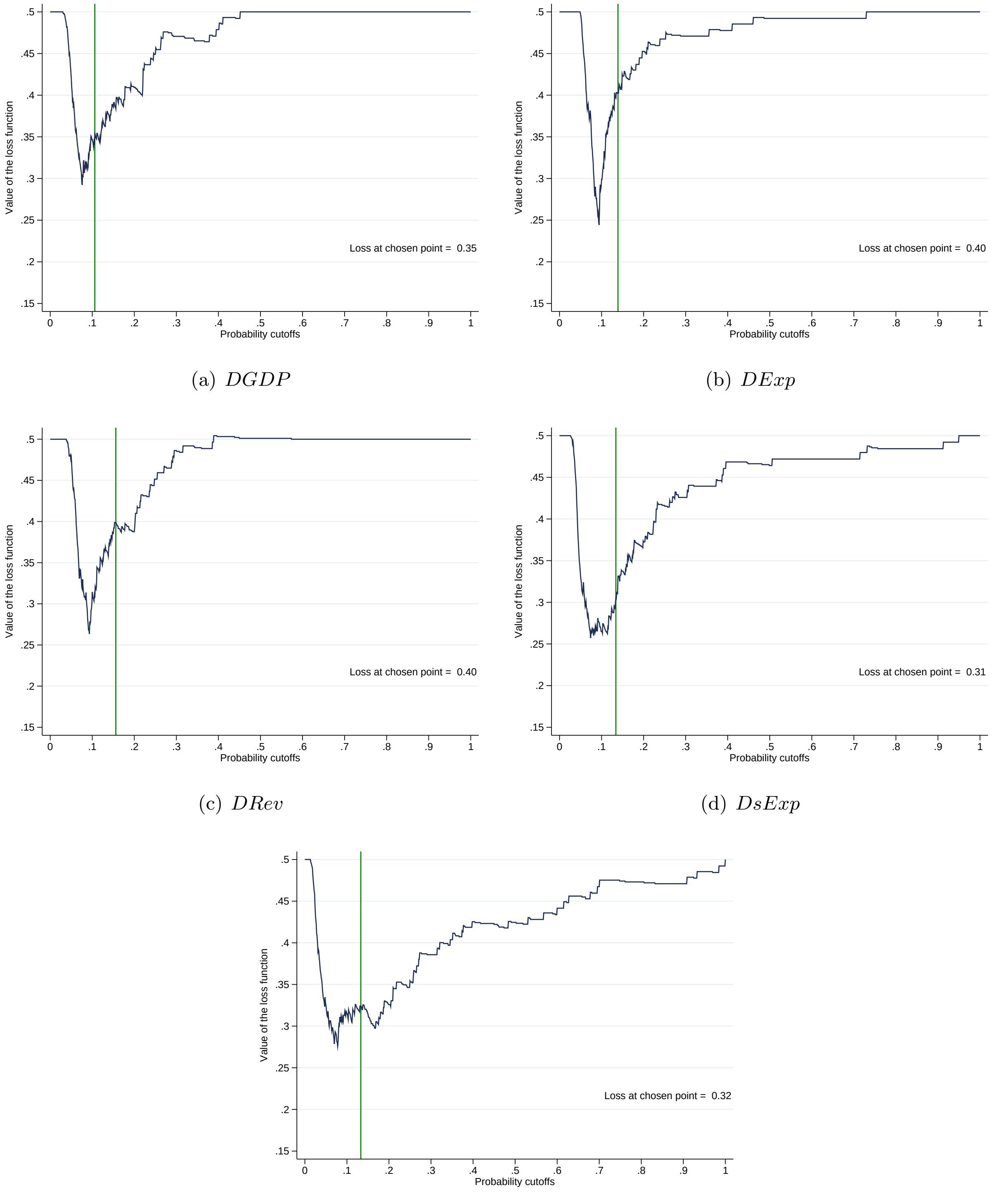
(c) *DRev*

(d) *DsExp*

(e) *DsRev*

Notes: the five figures are obtained from the estimates of the five different specifications of equation (1) as reported in Table 1, columns 1-5. The point on the ROC curves correspond to the probability cutoffs chosen by the DSF and the optimal ones, which minimize the loss function (2) for $\alpha = 0.5$.

Figure 2: Optimal cutoff and the loss function



(a) *DGDP*

(b) *DExp*

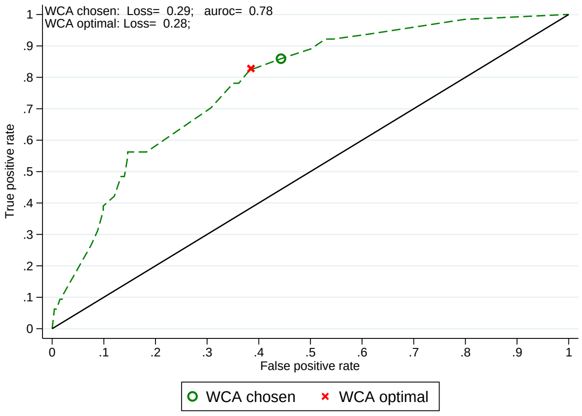
(c) *DRev*

(d) *DsExp*

(e) *DsRev*

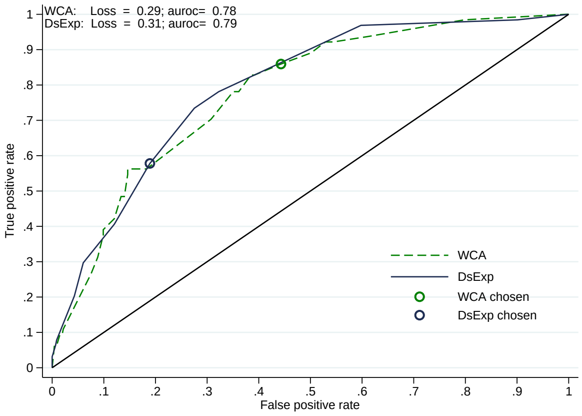
Notes: the five figures are obtained from the estimates of the five different specifications of equation (1) as reported in Table 1, columns 1-5. The vertical lines correspond to the values of the probability cutoffs chosen by the DTA.

Figure 3: Bias: the optimal and the chosen cutoffs



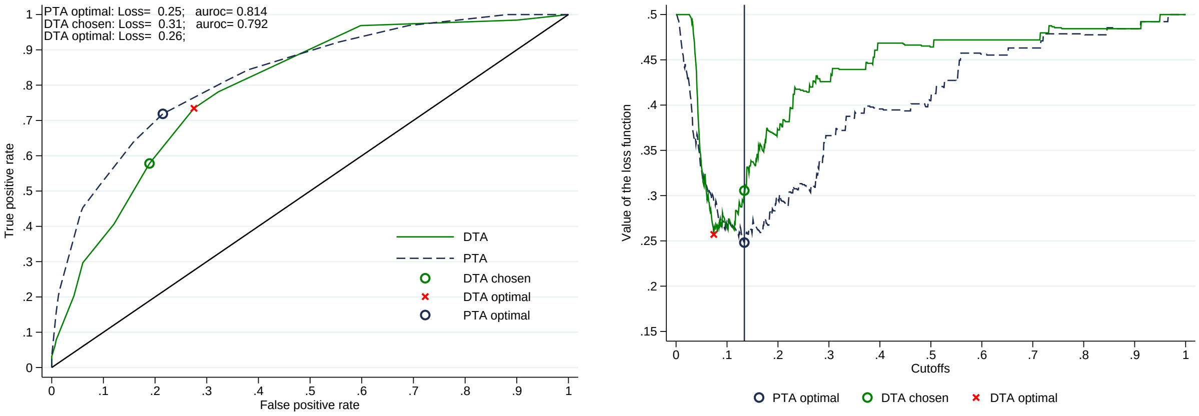
Notes: The ROC has been calculated using the WCA using the in-sample predictions.

Figure 4: Accuracy: WCA vs best single variable



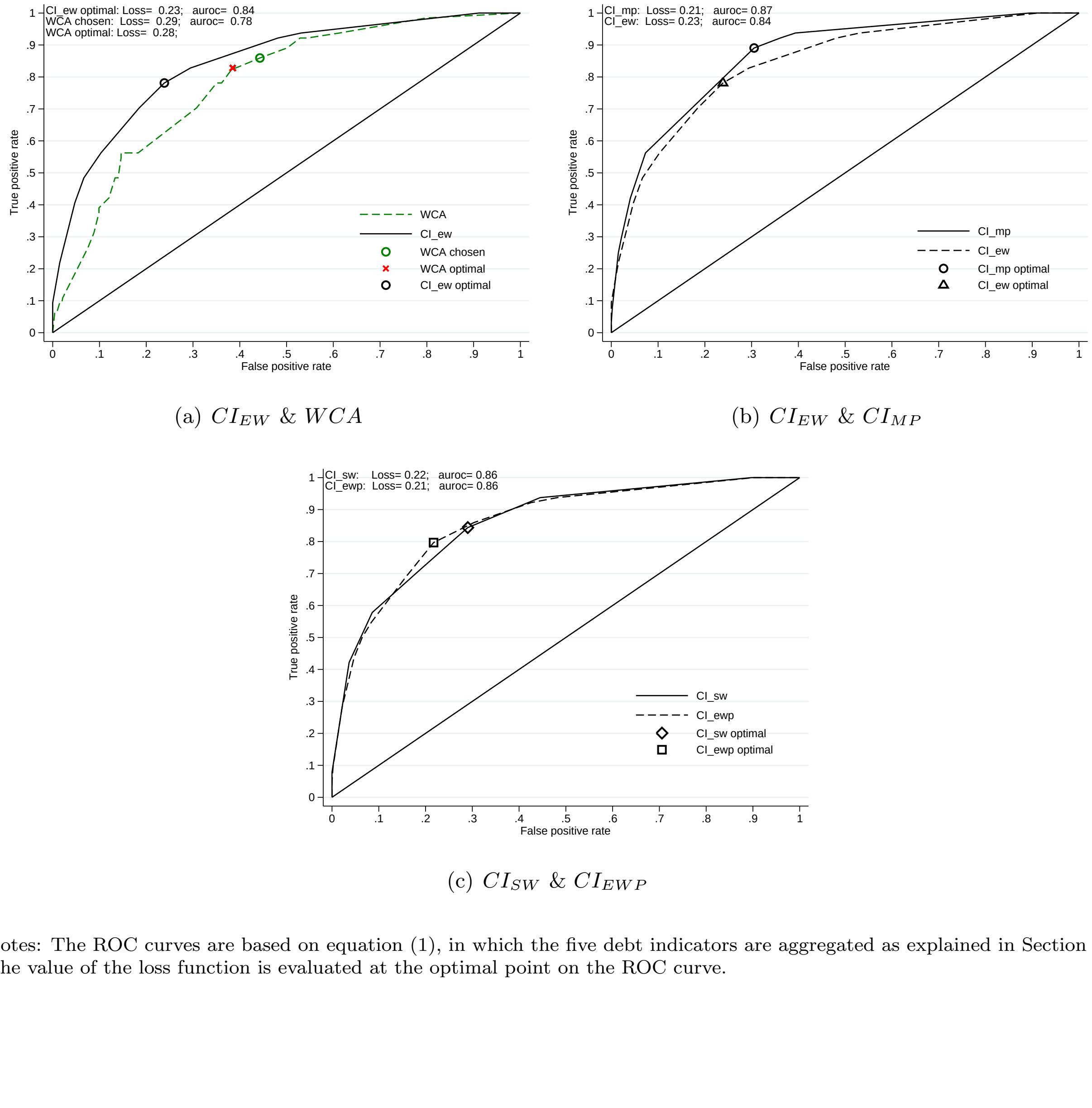
Notes: The ROC has been calculated using the WCA and the best individual debt measure ($DsExp$). using the in-sample predictions.

Figure 5: Accuracy and Bias: the DTA and the PTA for $DsExp$



Notes: the ROC curves (left panel) and the loss functions (right panel) have been calculated using the best individual debt measure ($DsExp$), using the in-sample predictions, adopting either the DTA or the PTA.

Figure 6: Comparing the composite indicators and the WCA



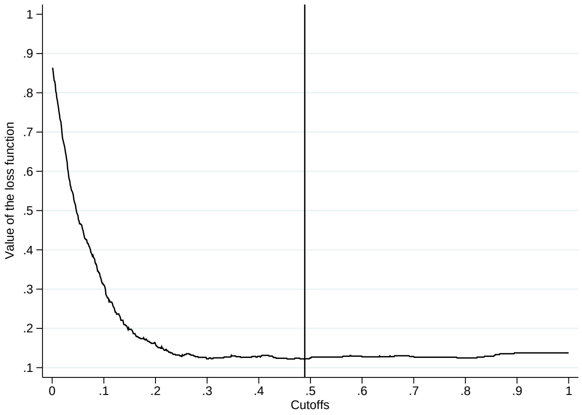
(a) CI_{EW} & WCA

(b) CI_{EW} & CI_{MP}

(c) CI_{SW} & CI_{EWP}

Notes: The ROC curves are based on equation (1), in which the five debt indicators are aggregated as explained in Section 6. The value of the loss function is evaluated at the optimal point on the ROC curve.

Figure 7: The loss function with equal weights, CI_{EW}



Notes: For a discussion of the (equal weights) loss function, see Section 7.

TABLES

Table 1: Baseline probit regressions

	(1)	(2)	(3)	(4)	(5)	(6)
$CPIA$	-0.454***	-0.393***	-0.410***	-0.353***	-0.399***	-0.402***
	(0.116)	(0.116)	(0.115)	(0.116)	(0.117)	(0.124)
$Growth$	-7.025***	-7.107***	-7.326***	-6.846***	-7.104***	-6.814***
	(1.753)	(1.792)	(1.755)	(1.814)	(1.786)	(1.898)
$DGDP$	0.333***					0.208
	(0.097)					(0.172)
$DGDP \times MIC$	0.013					-0.249
	(0.136)					(0.235)
$DExp$		0.227***				0.049
		(0.082)				(0.287)
$DExp \times MIC$		0.363**				0.564
		(0.151)				(0.561)
$DRev$			0.281***			-0.017
			(0.087)			(0.276)
$DRev \times MIC$			0.141			-0.254
			(0.134)			(0.518)
$DsExp$				0.450***		0.119
				(0.120)		(0.370)
$DsExp \times MIC$				-0.042		-0.205
				(0.149)		(0.508)
$DsRev$					0.697***	0.525*
					(0.149)	(0.286)
$DsRev \times MIC$					-0.363**	-0.090
					(0.168)	(0.459)
Observations	529	529	529	529	529	529
Pseudo-R^2	0.17	0.18	0.17	0.20	0.22	0.26
Log-Likelihood	-162.94	-160.04	-161.72	-155.49	-153.19	-145.42
BIC	357.24	351.43	354.79	342.34	337.73	372.36
AUROC	0.68	0.75	0.65	0.79	0.76	0.78
Loss (eq. 2)	0.34	0.40	0.40	0.31	0.32	0.29

Notes: The table reports the regression coefficients and, in brackets, the associated standard errors. * significant at 10%; ** significant at 5%; *** significant at 1%. All five debt variables have been standardized. All regressions are done on a sample of 529 observations. A constant is included. Variable definitions (see text for details):

$CPIA$: Country policy and institutional assessment
$Growth$: Real GDP growth
$DsRev$: The ratio of debt service to revenues
$DsExp$: The ratio of debt service to exports
$DRev$: The ratio of the NPV of debt to revenues
$DExp$: The ratio of the NPV of debt to exports
$DGDP$: The ratio of the NPV of debt to GDP
MIC: middle-income country dummy

Table 2: Calling a Crisis

	Debt distress episode	Tranquil period
Signal issued: $P_j \geq \overline{P_j}$	Correct signal (A)	False alarm (FA)
No signal issued: $P_j < \overline{P_j}$	Missed crisis (MC)	Correct signal (B)

Notes: P_j is the predicted probability of debt crisis from (1) for debt variable j; $\overline{P_j}$ is the probability threshold for calling crises for debt variable j.

Table 3: Debt thresholds

	PV of PPG external debt in percent of:			Debt service in percent of:	
	GDP	Exports	Revenue	Exports	Revenue
Estimated minimizing the loss function (2) on the sub-sample of 529 obs.					
Probability cutoffs	10.0%	10.0%	14.0%	14.0%	11.0%
Weak policy ($CPIA \leq 3.25$)	24	130	185	17	14
Medium policy ($3.25 < CPIA < 3.75$)	30	163	217	19	16
Strong policy ($CPIA \geq 3.75$)	36	196	250	21	18
IMF and WB 2012, Table 3, p. 20: minimizing the loss function on different sub-samples					
Probability cutoffs	14.0%	13.0%	15.0%	15.0%	14.0%
Weak policy ($CPIA \leq 3.25$)	28	131	184	18	17
Medium policy ($3.25 < CPIA < 3.75$)	36	179	217	20	20
Strong policy ($CPIA \geq 3.75$)	44	226	250	22	24
Under the current DSF					
Weak policy ($CPIA \leq 3.25$)	30	100	200	15	25
Medium policy ($3.25 < CPIA < 3.75$)	40	150	250	20	30
Strong policy ($CPIA \geq 3.75$)	50	200	300	25	35

The upper panel reports the thresholds that result choosing the probability cutoffs that minimize the loss function (2), setting $\alpha = 0.5$. They differ slightly from the corresponding thresholds in IMF and World Bank (2012, Table 3, p. 20), reported in the middle panel, because of slight differences in the loss function weights α, as described in footnote 14 as well as differences in the sample. The bottom panel reports the actual thresholds used in the current version of the DSF, as reported by IMF and World Bank (2012, Table 3, p. 20).

Table 4: The five debt indicators: pairwise correlations

	$DGDP$	$DExp$	$DRev$	$DsExp$	$DsRev$
$DGDP$	1				
$DExp$	0.5523*	1			
$DRev$	0.7872*	0.6792*	1		
$DsExp$	0.3892*	0.6689*	0.3692*	1	
$DsRev$	0.5355*	0.3003*	0.5721*	0.6630*	1

Notes: pairwise correlations based on the sample of 529 observations, non standardized debt variables. * significant at 1%.

Table 5: False Alarms, Missed Crises: single debt indicators and the WCA

Debt		$DGDP$		$DExp$		$DRev$		$DsExp$		$DsRev$		WCA	
distress		0	1	0	1	0	1	0	1	0	1	0	1
0	465	337	128	391	74	416	49	369	96	300	165	227	238
1	64	27	37	39	25	43	21	24	40	17	47	6	58
Total	529	364	165	430	99	459	70	393	136	317	212	233	296

Notes: in-sample frequencies.

Table 6: The bias of the DSF

Debt Measure	Optimal Cutoff	Debt thresholds	
		Using $\overline{P_j}$	Using $\overline{P} = 19\%$
	($\alpha = 0.5$)	from column 2	
$DGDP$	10.4%	30	49
$DExp$	10.4%	163	285
$DRev$	13.6%	217	284
$DsExp$	14.0%	19	24
$DsRev$	11.2%	16	23
WCA	19.0%	n.a.	n.a.

Notes: The second column reports the probabilities that minimize the loss function (2), when forecasting in-sample on the 529 observations used in this paper with $\alpha = (1 - \alpha)$. These are close, but not identical to those reported in Table A3 of IMF and World Bank (2012), for the reasons discussed in footnote 14. The last row reports the probability to be applied to each individual debt measure in order to minimize the same loss function when calling crises according to the WCA. The last two columns report the debt-specific thresholds calculated from equation 3, using the average LIC GDP growth rate, the CPIA score corresponding to medium policies ($CPIA = 3.5$), and using the probability cutoff from column 2 (for column 3) and 19% (for column 4).

Table 7: Weights of the Composite Indicators and goodness-of-fit measures

Debt indicator	CI_{EW}	CI_{MP}	CI_{SW}	CI_{EWP}	WCA
$DGDP$	0.200	0.208	0.000	0.200	
$DExp$	0.200	0.049	0.169	0.200	
$DRev$	0.200	-0.017	0.000	0.200	
$DsExp$	0.200	0.119	0.000	0.200	
$DsRev$	0.200	0.524	0.639	0.550	
Goodness-of-fit					
AUROC	0.84	0.87	0.86	0.86	0.78
BIC	333.02	372.36	337.74	338.44	
LOSS	0.23	0.21	0.22	0.21	0.29

Notes: all the weights (w_j) refer to standardized debt variables. The last three rows report the in-sample values of the AUROC, BIC and the loss function (eq. 2) for each of the aggregating rule.

Table 8: Accuracy of Various Models According to the BIC – Lower numbers are better[1]

CI_{EW}	CI_{SW}		CI_{EWP}	
333.0	**DsRev**	**337.7**	**DsRev**	**338.4**
	DExp DsRev	**337.7**	**DExp**	**340.6**
	DsExp_DsRev	**341.8**	DsExp	343.3
	DsExp	**342.3**	DRev	343.4
	DRev DsExp	345.2	DGDP	343.8
	DGDP DsRev	**345.5**	DExp DsRev	348.5
	DGDP DsExp	346.1	DGDP DsRev	348.9
	DRev DsRev	347.0	DsExp DsRev	349.5
	DGDP DExp DsRev	348.3	DExp DsExp	349.9
	DExp DRev DsRev	349.2	DExp DRev	349.9
	DExp DsExp DsRev	349.8	DRev DsRev	350.0
	DGDP DsExp DsRev	350.1	DGDP DRev	354.4
	DRev DsExp DsRev	350.4	DRev DsExp	355.0
	DExp	351.4	DGDP DsExp	355.1
	DExp DsExp	352.9	DGDP DRev DsExp	356.2
	DExp DRev DsExp	353.3	DGDP DExp DsExp	356.3
	DRev	354.8	DGDP DExp	357.5
	DGDP DRev DsExp	356.2	DExp DRev DsRev	359.9
	DGDP DExp DsExp	356.3	DExp DsExp DsRev	360.4
	DGDP	357.2	DGDP DRev DsRev	360.6
	DGDP DExp	357.5	DGDP DExp DsRev	360.8
	DGDP DRev DsRev	357.9	DGDP DsExp DsRev	360.8
	DExp DRev	358.9	DRev DsExp DsRev	361.6
	DGDP DExp DRev DsRev	360.0	DExp DRev DsExp	362.1
	DGDP DExp DsExp DsRev	360.2	DGDP DExp DRev	369.1
	DExp DRev DsExp DsRev	361.3	DGDP DExp DRev DsRev[3]	372.4
	DGDP DRev DsExp DsRev	361.5		
	DGDP DRev	364.6		
	DGDP DExp DRev DsExp	364.8		
	DGDP DExp DRev	369.1		
	DGDP DExp DRev DsExp DsRev[2]	372.4		

Notes:

[1] Lower numbers imply better fit, including a penalty for the number of estimated parameters. The **red bold** models do about as well as the CI_{EW}. (More technically, the hypothesis that the models indicated in **red bold** text perform as well as the CI_{EW} cannot be rejected at the 10 percent significance level, based on a boostrap analysis; see text for details).

[2] The CI_{EW} model with five variables is identical to the unrestricted five-variable probit.

[3] All CI_{EWP} models with four variables are identical to the unrestricted five-variable probit.

Variable definitions (see text for details):

DsRev: The ratio of debt service to revenues
DsExp: The ratio of debt service to exports
DRev: The ratio of the NPV of debt to revenues
DExp: The ratio of the NPV of debt to exports
DGDP: The ratio of the NPV of debt to GDP

Table 9: Choosing the aggregating rule

Tolerance for sample dependence Prior that all five variables matter	High	Low
Strong	CI_{EWP} (CI_{MP})	CI_{EW} (WCA)
Weak	CI_{SW}	

Notes: The CI_{MP} and the WCA are printed in gray since they are statistically less accurate than the CI_{EW}.