Joint Validation of Credit Rating PDs under Default Correlation*

Ricardo Schechtman
Central Bank of Brazil

This study investigates new proposals of statistical tests for validating the PDs (probabilities of default) of credit rating models (CRMs). The proposed tests recognize the existence of default correlation, deal jointly with the default behavior of all the ratings, and, in contrast to previous literature, control the error of validating incorrect CRMs. Power-sensitivity analysis and strategies for power improvement are discussed for the calibration tests, whereas a non-typical goal is proposed for the tests of discriminatory power, leading to results of power dominance. Finally, Monte Carlo simulations investigate the finite sample bias for varying scenarios of parameters.

JEL Codes: C12, G21, G28.

1. Introduction

This paper studies issues of validation for credit rating models (CRMs). In this article, CRMs are defined as a set of risk buckets (ratings) to which borrowers are assigned and which indicate the likelihood of default (usually through a measure of probability of default, PD) over a fixed time horizon (usually one year). Examples include rating models of external credit agencies such as Moody's and Standard & Poor's and banks' internal credit rating models.

*The author would like to thank Axel Munk, Dirk Tasche, Getulio Borges da Silveira, and Kostas Tsatsaronis for helpful conversations along the project as well as seminar participants at the Bank for International Settlements (BIS), the 17th FDIC Annual Derivatives Securities and Risk Management Conference, the C.R.E.D.I.T Conference, and Central Bank of Brazil research workshops. The author also acknowledges the hospitality at the BIS during his fellowship there, when most of the project was conducted. Views expressed in this paper do not necessarily reflect those of the Central Bank of Brazil or the BIS. Author contact: Research Department, Central Bank of Brazil, Av. Presidente Vargas 730, Rio de Janeiro, Brazil, 20071-900. E-mail: ricardo.schechtman@bcb.gov.br.

CRMs are key tools to credit risk management and have had their relevance increased, as the Basel II Accord (Basel Committee on Banking Supervision 2006a) allows the PDs of the internal ratings to function as inputs in the computation of banks' regulatory levels of capital.[1] Its goal was not only to make regulatory capital more risk sensitive, and therefore to diminish the problems of regulatory arbitrage, but also to strengthen stability in financial systems through better assessment of borrowers' credit quality. However, the great challenge for Basel II, in terms of implementation, lies still in the validation of CRMs, particularly the validation of bank-estimated rating PDs.[2] Besides satisfying regulatory demands, PD validation is also crucial for banks not to be left in competitive disadvantage towards their peers. However, the recent financial crisis has also promoted doubts about the efficacy with which banks and rating agencies had been validating their CRMs. Regulators are currently examining whether to place limits on the use of models to prevent banks from attempting to understate the riskiness of their portfolios[3] (Watt 2013).

In fact, validation of CRMs has been considered a difficult job due to two main factors. Firstly, the typically long credit time horizon of one year or so results in few observations available for *backtesting*. This means, for instance, that the bank/supervisor will, in most practical situations, have to judge the CRM based solely on five to ten observations available at the database. Secondly, as borrowers are usually sensitive to a common set of factors in the economy (e.g., industry, geographical region), variation of macro conditions over the forecasting time horizon induces correlation among defaults. Both these factors contribute to decreasing the power of quantitative methods of validation. This paper does not aim at a prescription to surpass the aforementioned unavoidable difficulties but instead at discussing the trade-offs and limitations involved in the validation task from a statistical perspective.

[1] The higher the PD, the higher is the regulatory capital.

[2] According to BCBS (2005b), validation is above all a bank task, whereas the supervisor's role should be to certificate this validation.

[3] That would represent a further enhancement of the Basel Accords, after the recent Basel III Accord (BCBS 2011). Basel III didn't bring any major modifications on the use of CRMs for regulatory purposes.

The judgment of the performance of a CRM is generally a twofold issue. It involves the aspects of calibration and discriminatory power. Calibration is the ability to forecast accurately the ex post *(long-run)* default rate of each rating (e.g., through an ex ante estimated PD). Discriminatory power is the ability to ex ante discriminate, based on the rating, between defaulting borrowers and non-defaulting borrowers.

As BCBS (2006a) is explicit about the demand for banks' internal models to possess good calibration, testing calibration is the starting point of this paper.[4] According to BCBS (2005b), quantitative techniques for testing calibration are still in the early stages of development. BCBS (2005b) reviews some simple tests, namely, the binomial test, the Hosmer-Lemeshow test, a normal test, and the traffic lights approach (Blochwitz et al. 2004). These techniques all have the disadvantage of being univariate (i.e., designed to test a single rating PD per time) and/or making the unrealistic assumption of cross-sectional default independency. An approach that tests each rating PD per time may translate into a joint procedure with rather higher error rates than those of the employed univariate test (e.g., Hochberg and Tamhane 1987). Similarly, a false assumption of default independency generally produces substantially higher error rates and can also lead to similar probabilities of rejecting correctly and incorrectly specified CRMs (e.g., BCBS 2005b, Bluemke 2013).

More recent proposed approaches have similar or new limitations. Balthazar (2004) proposes using the same Basel II model of capital requirement, which recognizes default dependency, for PD validation, but restricts the analysis to the univariate case. Miu and Ozdemir (2008) explore deeper the idea of Balthazar (2004), but still their analysis remains restricted to the univariate case. Blochlinger (2012) proposes a multivariate method that also recognizes default correlation but, on the other hand, is inconsistent with the functional form of the Basel II model (see the discussion in Gordy 2000). Inconsistency with the Basel II model cannot ensure that the validated PDs would be appropriate inputs to the Basel II capital requirement formula. Bluemke (2013) addresses the multivariate case in a Basel II-like model, but his approach does not provide a closed formula for the critical region or the power of the test. Therefore, apart from

[4] According to BCBS (2006a), PDs should resemble long-run average default rates for all ratings.

simulations, the author cannot discuss trade-offs and strategies for power improvement involved in the proposed validation test.[5] Additionally, a crucial concern common to all approaches in the literature is that they do not control for the error of accepting a miscalibrated CRM. Instead, they control for the error of rejecting correct CRMs, which, from a prudential viewpoint, is of secondary importance.

This paper reverses the roles of the hypothesis used throughout the CRM validation literature, in order to control for the error of validating incorrectly specified CRMs. Furthermore, this paper presents an asymptotic analytical framework to jointly test several PDs under the assumption of default correlation. The approach generalizes the Basel II model in a similar fashion to Demey et al. (2004) but with a new configuration oriented towards validation purposes. The results include a new simple one-sided CRM calibration test and the discussion of the relative roles played by the distinct elements that affect the power of the proposed test (e.g., differences between consecutive PD ratings, indifference regions of validation, asset correlations, ratings driving the power). Under the new formulation of the hypothesis, power is the probability of accepting a correctly specified CRM and, therefore, should achieve minimum levels for the test to be practical. Strategies for power improvement are also analyzed. The paper also discusses, to a considerable extent, the greater particular difficulties and conceptual problems related to two-sided CRM calibration testing.

Good discriminatory power is also a desirable property of CRMs, as it allows rating-based yes/no decisions (e.g., credit granting) to be made with less error and therefore less cost by the bank (see Blochlinger and Leippold 2006, for instance). BCBS (2005b) comprehensively reviews some well-established techniques for examining discriminatory power, including the area under the receiver operating characteristic (ROC) curve (Engelmann, Hayden, and Tasche 2003), the accuracy ratio, and the Kolgomorov-Smirnov statistic.

Although the use of the above-mentioned techniques of discriminatory power is widespread in banking industry, two constraining points should be noted. First, the pursuit of perfect discrimination is inconsistent with the pursuit of perfect calibration in realistic

[5] Additionally, Miu and Ozdemir (2008) and Blochlinger (2012) examine only very briefly power considerations.

CRMs. The reason is that to increase discrimination, one would be interested in having, over the long run, the ex post rating distributions of the default and non-default groups of borrowers as separate as possible, and this involves having default rates as low as possible for good-quality ratings (in particular, lower than the PDs of these ratings) and as high as possible for bad-quality ratings (in particular, higher than the PDs of these ratings). See appendix 1 for a graphical example. Second, although scarcely remarked in the literature (e.g., Blochlinger 2012), usual measures of discriminatory power are a function of the cross-sectional dependency between borrowers. This fact potentially represents an undesired property of traditional measures to the extent that the level and structure of default correlation is mainly a portfolio characteristic rather than a property intrinsic to the performance of CRMs.[6] Using the same framework employed in calibration testing, this paper proposes and discusses tests of “rating” discriminatory power that (i) can be seen as a necessary requisite to perfect calibration and (ii) are not a function of the default dependency structure. Power of these tests is also discussed, including results of power dominance between distinct proposed tests.

This text is organized as follows. Section 2 develops a default rate asymptotic probabilistic model (DRAPM) upon which validation will be discussed. The model leads to a unified theoretical framework for checking calibration and discriminatory power. Section 3 discusses briefly the formulation of the testing problem for CRM validation. The discussion of calibration testing, both one-sided and two-sided, is contained in section 4. Theoretical aspects of discriminatory power testing are investigated in section 5. Section 6 contains a Monte Carlo analysis of the finite sample properties of DRAPM and their consequences for calibration testing. Section 7 concludes.

2. The Default Rate Asymptotic Probabilistic Model (DRAPM)

The model of this section provides a default rate probability distribution upon which statistical testing is possible. It is based on

[6] It is not solely a portfolio characteristic because default correlation among the ratings potentially depends on the design of the CRM too.

an extension of the Basel II underlying model of capital requirement. In fact, this paper generalizes the idea first proposed by Balthazar (2004), of using the Basel II model for validation, to a multi-rating setting.[7] The applied extension is close to Demey et al. (2004)[8] and refers to including an additional systemic factor for each rating. While in Basel II the reliance on a single factor is crucial to the derivation of portfolio-invariant capital requirements (cf. Gordy 2003), for validation purposes a richer structure is necessary to allow for non-singular variance matrix among the ratings, as becomes clearer ahead in this section.

The formulation of DRAPM starts with a decomposition of z_{in}, the normalized return on assets of a borrower n with rating i. Close in spirit to the Basel II model, z_{in} is expressed as

$$z_{in} = \rho_B^{1/2} x + (\rho_W - \rho_B)^{1/2} x_i + (1 - \rho_W)^{1/2} \varepsilon_{in}, \tag{1}$$

for each rating $i = 1 \ldots I$ and each borrower $n = 1 \ldots N$, where x, x_i, ε_{ij} ($i = 1 \ldots I$, $j = 1 \ldots N$) are independent and standard normal distributed.

Above, x represents a common systemic factor affecting the asset return of all borrowers, x_i a systemic factor affecting solely the asset return of borrowers with rating i, and ε_{in} an idiosyncratic shock. The parameters ρ_B and ρ_W lie in the interval [0 1]. Note that $\mathrm{Cov}(z_{in}, z_{jm})$ is equal to ρ_W if $i = j$ and to ρ_B otherwise, so that ρ_W represents the "within-rating" asset correlation and ρ_B the "between-rating" asset correlation. The Basel II model (and the validation approaches that are based on it such as Balthazar 2004) is a particular case of DRAPM when $\rho_W = \rho_B$. In other words, there is no systemic factor associated with rating i in the latter.

The model description continues with the statement that a borrower n with rating i defaults at the end of the forecasting time horizon if $z_{in} < \Phi^{-1}(PD_i)$ at that time, where Φ denotes the standard

[7] This idea is also adopted in Miu and Ozdemir (2008) and Bluemke (2013), among others. The reader is referred to BCBS (2005a) for a detailed presentation of the Basel II underlying model.

[8] The purpose of Demey et al. (2004) is to estimate correlations, while the focus here is on developing a minimal non-degenerate multivariate structure useful for testing.

normal cumulative distribution function.[9] Consequently, the conditional probability of default $\mathrm{PD_i}(\mathbf{x})$, where $\mathbf{x} = (\mathrm{x},\mathrm{x_1},\ldots,\mathrm{x_i})$' denotes the vector of systemic factors, can be expressed by

$$\mathrm{PD_i}(\mathbf{x}) \equiv \mathrm{Prob}(\mathrm{z_{in}} < \Phi^{-1}(\mathrm{PD_i})|\mathbf{x}) = \Phi((\Phi^{-1}(\mathrm{PD_i}) - \rho_B^{1/2}\mathrm{x} - (\rho_W - \rho_B)^{1/2}\mathrm{x_i})/(1 - \rho_W)^{1/2}). \quad (2)$$

Let $\mathrm{DR_{iN}}$ denote the default rate computed using a sample of N borrowers with rating i at the start of the forecasting horizon. It is easy to see, as in Gordy (2003), that

$$\Phi^{-1}(\mathbf{DR_N}) - \Phi^{-1}(\mathbf{PD}(\mathbf{x})) \rightarrow\ 0 \text{ a.s. when } \mathrm{N} \rightarrow \infty, \quad (3)$$

where $\mathbf{DR_N} = (\mathrm{DR_{1N}}, \mathrm{DR_{2N}},\ldots,\mathrm{DR_{IN}})$' and $\mathbf{PD}(\mathbf{x}) = (\mathrm{PD_1}(\mathbf{x}), \mathrm{PD_2}(\mathbf{x}),\ldots, \mathrm{PD_i}(\mathbf{x}))$'.

More concretely, the limiting default rate joint distribution is

$$\Phi^{-1}(\mathbf{DR}) \approx \mathrm{N}(\boldsymbol{\mu}, \textstyle\sum), \quad (4)$$

where $\mu_i = \Phi^{-1}(\mathrm{PD_i})/(1 - \rho_W)^{1/2}, \sum_{ij} = \rho_W/(1 - \rho_W)$ if i = j, and $\sum_{ij} = \rho_B/(1 - \rho_W)$ otherwise.

This asymptotic distribution is a multi-rating extension of the univariate limiting distribution presented in Gordy (2003) and also analyzed in Vasicek (2002). That is the distribution upon which all the tests of this paper will be derived. A limiting normal distribution is mathematically convenient to the derivation of likelihood ratio multivariate tests. The cost to be paid is that the approach is asymptotic, so that the discussions and results of this paper are not suitable for CRMs with a small number of borrowers per rating, such as, for example, rating models for large corporate exposures. Even for moderate numbers of borrowers, section 6 reveals that the departure from the asymptotic limit can be substantial, significantly

[9] Note that the probability of this event is therefore, by construction, $\mathrm{PD_i}$. Without generalization loss, $\mathrm{PD_i}$ is assumed to increase in i. The characterization of default as the event defined by the fall of the (normalized return of) assets below a certain threshold is motivated by the structural approach to credit risk modeling developed by Black and Scholes (1973) and Merton (1974). Other implications of structural models for default probabilities can be found, for example, in Leland (2004).

altering the theoretical size and power of the tests. Application of the tests of the next sections should then be extremely careful.

Some comments on the choice of the form of $\sum$ are warranted.[10] To the extent that borrowers of each rating present similar distributions of economic and geographic sectors of activity, which ultimately govern borrowers' asset correlations, ρ_B is likely to be close to ρ_W, as this situation resembles the single systemic factor case. Nevertheless, it is reasonable to assume $0 < \rho_B < \rho_W$, in opposition to $\rho_B = \rho_W$, the implicit assumption of previous Basel II-like approaches, in order to leave open the possibility of some degree of association between rating PDs and borrowers' sectors of activity. As a result, borrowers in the same rating are considered to behave more dependently than borrowers in different ratings, because the profile of borrowers' sectors of activity is likely more homogeneous within than between ratings. Indeed, more realistic modeling is likely to require a higher number of asset correlation parameters and a portfolio-dependent approach; therefore the choice of just a pair of correlation parameters is regarded here as a practical compromise for general testing purposes. Furthermore, notice that $\rho_B \neq \rho_W$ is crucial to guarantee a non-singular matrix $\sum$ and, therefore, a non-degenerate asymptotic distribution. A singular matrix in the context of equation (4) would mean that the default rates of all ratings are asymptotically transformations of one another, which is unrealistic for joint validation purposes.

This paper further assumes that the correlation parameters ρ_W and ρ_B are known. The typically small number of years that banks have at their disposal suggests that the inclusion of correlation estimation in the testing procedure is not feasible, as it would diminish considerably the power of the tests. Instead, this paper relies on the Basel II Accord to extract some information on correlations.[11] By matching the variances of the non-idiosyncratic parts of the asset returns in the Basel II and DRAPM models, ρ_W can be seen as the

[10] Note that the structure of $\sum$ defines DRAPM more concretely than the chosen decomposition of the normalized asset return, because the decomposition is not unique given $\sum$.

[11] An important distinction to the Basel II model or Balthazar (2004), however, is that this paper does not make correlations dependent on the rating. In fact, the empirical literature on asset correlation estimation contains ambiguous results on this sensitivity.

asset correlation parameter present in the Basel II formula and in studies that make use of a single systemic factor. For corporate borrowers, for example, the Basel II Accord chooses $\rho_W \in [0.12\ 0.24]$. On the other hand, as the configuration present in equation (1) is new, there is no available information on ρ_B. Sensitivity analysis of the power of the tests on the choices of both ρ_W and ρ_B parameters is carried out in section 4. It should be noted, however, that the supervisory authority may have a larger set of information to estimate correlations and/or may even desire to set their values publicly for testing purposes.

Finally, serial independency is assumed for the annual default rate time series. Therefore, the (Φ^{-1}-transformed) average annual default rate, used as the test statistic for the tests of the next sections, has the normal distribution above, with $\sum/Y$ in place of $\sum$, where Y is the number of years available to *backtest.* According to BCBS (2005b), serial independency is less inadmissible than cross-sectional independency. Furthermore, Blochinger (2012) argues that if the anticipated parts of the systemic factors are already factored into the allocations of borrowers to rating PDs, the resulting rating default rates are indeed serially independent.

3. The Formulation of the Testing Problem

Any configuration of a statistical test should start with the definitions of the null hypothesis H_o and the alternative one H_1. In testing a CRM, a crucial decision refers to where the hypothesis “the rating model is correctly specified” should be placed.[12] If the bank/supervisor only wishes to abandon this hypothesis if data strongly suggests it is false, then the “correctly specified” hypothesis should be placed under H_0, as in Balthazar (2004), BCBS (2005b), Miu and Ozdemir (2008), and Blochlinger (2012), among others. But if the bank/supervisor wants to know if the data provided enough evidence confirming the CRM is correctly specified, then this hypothesis should be placed in H_1 and its opposite in H_o. The reason is that the result of a statistical test is reliable knowledge only when the null hypothesis is rejected, usually at a low significance

[12] For this general discussion, one can think of “correctly specified” as meaning either correct calibration or good discriminatory power.

level. The latter option is pursued throughout this paper. Thus the probability of accepting an incorrect CRM will be the error to be controlled for at the significance level α.

To be precise, Bluemke (2013) also tries to control for the error of accepting an incorrect CRM but, in placing this hypothesis under the null, he is led to try to limit the type II error, which is generally not liable to uniform limitation. Indeed, Bluemke (2013) restricts the error II limitation to a single point of the alternative H_1. By reversing the role of the hypotheses found in previous validation approaches, this paper seems to be the first to uniformly control for the error of accepting an incorrect CRM.

Placing the "correctly specified" hypothesis under H_1 has immediate consequences. For a statistical test to make sense, H_0 usually needs to be defined by a closed set and H_1, therefore, by an open set.[13] This implies that the statement that "the CRM is correctly specified" needs to be translated into some statement about the parameters' PD_is lying in an *open* set—in particular, there shouldn't be equalities defining H_1 and the inequalities need to be strict. It is, for example, statistically inappropriate to try to conclude that the PD_is are equal to the bank-postulated values. In cases like that, the solution is to enlarge the desired conclusion by means of the concept of an indifference region. The configuration of the indifference region should convey the idea that the bank/regulator is satisfied with the eventual conclusion that the true **PD** vector lies there. In the previous case, the indifference region could be formed, for example, by open intervals around the postulated PD_is. The next sections make use of the concept to a great extent. At this point it is desirable only to remark that the feature of an indifference region shouldn't be seen as a disadvantage of the approach of this paper. Rather, it reflects better the fact that not necessarily all the borrowers in the same rating i have exactly the same theoretical PD_i and that it is, therefore, more realistic to see the ratings as defined by PD intervals.[14]

[13] H_0 and H_0 U H_1 need to be closed sets in order to guarantee that the maximum of the likelihood function is attained.

[14] However, in the context of Basel II, ratings need not be related to PD intervals but merely to single PD values. In light of this study's approach, this represents a gap in information needed for validation.

4. Calibration Testing

This section distinguishes between one-sided and two-sided tests for calibration. One-sided tests (which are only concerned about $\mathrm{PD_i}$s being sufficiently high) are useful to the supervisory authority by allowing to conclude that Basel II capital requirements derived by the approved PD estimates are sufficiently conservative in light of the banks' realized default rates. From a broader view, however, not only is excess of regulatory capital undesirable by banks, but also BCBS (2006b) states that the PD estimates should ideally be consistent with the banks' managerial activities such as credit granting and credit pricing.[15] To accomplish these goals, PD estimates must, without adding distortions, reflect the likelihood of default of every rating, something to be verified more effectively by two-sided tests (which are concerned about $\mathrm{PD_i}$s being within certain ranges). Unfortunately, the difficulties present in two-sided calibration testing are greater than in one-sided testing, as indicated ahead in this section. The analysis of one-sided calibration testing starts the section.

4.1 One-Sided Calibration Testing

Based on the arguments of the previous section about the proper roles of $\mathrm{H_o}$ and $\mathrm{H_1}$, the formulation of a one-sided calibration test is proposed below. Note that the desired conclusion, configured as an intersection of strict inequalities, is placed in $\mathrm{H_1}$.

$$\begin{aligned}\mathrm{H_o} &: PD_i \geq u_i \text{ for some i} = 1\ldots \mathrm{I}\\ \mathrm{H_1} &: PD_i < u_i \text{ for every i} = 1\ldots \mathrm{I},\end{aligned}$$

where $PD_i \equiv \Phi^{-1}(\mathrm{PD_i})$ and $u_i \equiv \Phi^{-1}(\mathrm{u_i})$. (This convention of representing Φ^{-1}-transformed figures in italic is followed throughout the rest of the text.)[16]

[15] More specifically, if the PDs used as inputs to the regulatory capital differ from the PDs used in managerial activities, at least some consistency must be verified between the two sets of values for validation purposes; cf. BCBS (2006b).

[16] As Φ^{-1} is strictly increasing, statements about italic figures imply equivalent statements about non-italic figures.

Here $\mathrm{u_i}$ is a fixed known number that defines an indifference acceptable region for $\mathrm{PD_i}$. Its value should ideally be slightly larger than the value postulated for $\mathrm{PD_i}$ so that the latter is within the indifference region. Besides, $\mathrm{u_i}$ should preferably be smaller than the value postulated for PD_{i+1} so that at least the rejection of $\mathrm{H_0}$ could conclude that $\mathrm{PD_i}$ < postulated $\mathrm{PD_{i+1}}$.[17] That is also an advantage of this paper's approach, since the monotonicity of PDs between individual rating grades is not always ensured in the methods proposed in the literature (e.g., Bluemke 2013).

According to DRAPM and based on the results of Sasabuchi (1980) and Berger (1989), which investigate the problem of testing homogeneous linear inequalities concerning normal means, a size-α critical region can be derived for the test.[18]

Reject $\mathrm{H_0}$ (i.e., validate the CRM) if

$$\overline{DR}_i \leq u_i/(1-\rho_{\mathrm{W}})^{1/2} - \mathrm{z}_\alpha(\rho_{\mathrm{W}}/(\mathrm{Y}(1-\rho_{\mathrm{W}})))^{1/2}$$
$$\text{for every i} = 1 \ldots \mathrm{I}, \tag{5}$$

where $\overline{DR}_i = \frac{\sum_{y=1}^{Y} \Phi^{-1}(\mathrm{DR}_{iy})}{Y}$ is the (transformed) average annual default rate of rating i, and $\mathrm{z}_\alpha = \Phi(1-\alpha)$ is the $1-\alpha$ percentile of the standard normal distribution.[19]

This test is a particular case of a min test, a general procedure that calls for the rejection of a union of individual hypotheses if each one of them is rejected at level α. In general, the size of a min test will be much smaller than α, but the results of Sasabuchi (1980) and Berger (1989) guarantee that the size is exactly α for the previous one-sided calibration test.[20] This means that the CRM is validated at size α if each $\mathrm{PD_i}$ is validated as such.

A min test has several good properties. First, it is uniformly more powerful (UMP) among monotone tests (Laska and Meisner

[17] As banks have the capital incentive to postulate lower PDs, one could argue that $\mathrm{PD_i}$ < postulated $\mathrm{PD_{i+1}}$ also leads to $\mathrm{PD_i}$ < true $\mathrm{PD_{i+1}}$. Specific configurations of $\mathrm{u_i}$ are discussed later in the section.

[18] Size of a test is the maximum probability of rejecting $\mathrm{H_0}$ when it is true.

[19] This definition of $\overline{DR}_i$ is used throughout the paper.

[20] More formally, this is the description of a union-intersection test, of which the min test is a particular case when all the individual critical regions are intervals not limited on the same side.

1989), which gives a solid theoretical foundation for the procedure since monotonicity is generally a desired property.[21] Second, as the transformed default rate variables are asymptotically normal in DRAPM, the min test is also asymptotically the likelihood-ratio test (LRT). Finally, the achievement of size α is robust to violation of the assumption of normal copula for the transformed default rates (Wang, Hwang, and Dasgupta 1999) so that, for size purposes, the requirement of *joint* normality for the systemic factors can be relaxed.

From a practical point of view, it should be noted that the decision to validate or not validate the CRM does not depend on the parameter ρ_{B}, which is useful for applications since ρ_{B} is not present in Basel II framework and so there is not much knowledge about its reasonable values. However, the power of the test—i.e., the probability of validating the CRM when it is correctly specified—does depend on ρ_{B}. The power is given by the following expression.

$$\begin{aligned}\text{Power} = \Phi_{\mathrm{I}}(&-\mathrm{z}_\alpha + (u_1 - PD_1)/(\rho_{\mathrm{W}}/\mathrm{Y})^{1/2}, \ldots,\\ &-\mathrm{z}_\alpha + (u_i - PD_i)/(\rho_{\mathrm{W}}/\mathrm{Y})^{1/2}, \ldots,\\ &-\mathrm{z}_\alpha + (u_I - PD_I)/(\rho_{\mathrm{W}}/\mathrm{Y})^{1/2}; \rho_{\mathrm{B}}/\rho_{\mathrm{W}}),\end{aligned} \tag{6}$$

where $\Phi_{\mathrm{I}}(\ldots; \rho_{\mathrm{B}}/\rho_{\mathrm{W}})$ is the cumulative distribution function of an I^{th}-variate normal of mean 0, variances equal to 1, and covariances equal to $\rho_{\mathrm{B}}/\rho_{\mathrm{W}}$.

Berger (1989) remarks that if the ratio $\rho_{\mathrm{B}}/\rho_{\mathrm{W}}$ is small, then the power of this test can be quite low for the $\mathrm{PD_i}$s only slightly smaller than $\mathrm{u_i}$s and/or a large number of ratings I. This is intuitive, as a low ratio $\rho_{\mathrm{B}}/\rho_{\mathrm{W}}$ indicates that ex post information about one rating does not contain much information about other ratings and so is less helpful to conclude for validation. On the other hand, as previously noted in section 2, DRAPM is more realistic when $\rho_{\mathrm{B}}/\rho_{\mathrm{W}}$ is close to 1 so that the referred theoretical problem becomes less relevant in the practical case.

More generally, it is easy to see that the power increases when $\mathrm{PD_i}$s decrease, $\mathrm{u_i}$s increase, Y increases, I decreases, ρ_B increases, or

[21] In the context of this paper, a test is monotone if the fact that average annual default rates are in the critical region implies that smaller average default rates are still in the critical region. Monotonicity is further discussed later in the paper.

Table 1. $\mathbf{u_i \times PD_i}$

$PD_i(\%)$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
u_i (%)	2	4	6	8	9	11	12	14	15	17	18	20	21	22	24	25	26	28	29	30

Note: PD_i validated if $H_0 : PD_i \geq u_i$ rejected in favor of $H_1 : PD_i < u_i$, considering the base scenario: $Y = 5$, $\rho_W = 0.15$, $\alpha = 15\%$, and $\beta = 80\%$, where Y = number of years, ρ_W = asset correlation, α = size of the test, and β = power at the true $PD \in H_1$.

ρ_W decreases.[22] In fact, it is worth examining the trade-off between the configuration of the indifference region in the form of the u_is and the attained power. If high precision is demanded (u_is close to postulated PD_is), then power must be sacrificed; if high power is demanded (u_is far from postulated PD_is), then precision must be sacrificed. Some numerical examples are analyzed below in order to provide further insights on this trade-off.

The case I = 1 represents an upper bound to the power expression above. In this case, for a desired power of β when the probability of default is exactly equal to the postulated PD, it is true that

$$u - PD = (z_\alpha - z_\beta) \times (\rho_W/Y)^{1/2}. \tag{7}$$

In a base-case scenario given by $Y = 5$, $\rho_W = 0.15$, $\alpha = 15\%$, and $\beta = 80\%$, the right-hand side of the previous equation is approximately equal to 0.32. This scenario is considered here sufficiently conservative, with a realistic balance between targets of power and size. In this case, it holds that

$$u_i = \Phi(0.32 + \Phi^{-1}(PD_i)). \tag{8}$$

Table 1 displays pairs of values of u_i and PD_i that conform to the equality above.

As, in a multi-rating context, any reasonable choice of u_i must satisfy $u_i \leq PD_{i+1}$, table 1 illustrates, for the numbers of the base-case scenario, an approximate lower bound for PD_{i+1} in terms of

[22] Obviously, the power also increases when the level α increases.

Table 2. PDs (%) Chosen According to u_i Specification and CRM Design

	PD$_i$s Follow Arithmetic Progression		**PD$_i$s Follow Geometric Progression**	
	$\mathbf{u_i = PD_{i+1}}$	$\mathbf{u_i = (PD_{i+1} + PD_i)/2}$	$\mathbf{u_i = PD_{i+1}}$	$\mathbf{u_i = (PD_{i+1} + PD_i)/2}$
I = 3	1.22, 11.82, 22.42	6.52, 17.12, 27.72	1.22, 3.66, 11	1.83, 5.5, 16.5
I = 4	2, 9.5, 17, 24.5	5.75, 13.25, 20.75, 28.25	2, 4, 8, 16	2.66, 5.33, 10.66, 21.33

Notes: CRMs validated if $H_0 : PD_i \geq u_i$ for some $i = 1 \ldots I$ rejected in favor of $H_1 : PD_i < u_i$ for every $i = 1 \ldots I$. Distinct CRMs have the same u_0 and u_1 for each I = 3,4, where I is the number of ratings.

PD_i.[23] More generally, table 1 provides examples of whole rating scales that conform to the restriction $PD_{i+1} \geq u_i$, e.g., $PD_1 = 1\%$, $PD_2 = 2\%$, $PD_3 = 4\%$, $PD_4 = 8\%$, $PD_5 = 14\%$, $PD_6 = 22\%$, $PD_7 = 36\%$ (note the shaded cells). Note that such conforming rating scales must possess increasing PD differences between consecutive ratings (i.e., $PD_{i+1} - PD_i$ increasing in i), a characteristic found indeed in the design of many real-world CRMs. Therefore, DRAPM suggests a validation argument in favor of that design choice. Notice that this feature of increasing PD differences is directly related to the nonlinearity of Φ, which in turn is a consequence of the asymmetry and kurtosis of the distribution of the untransformed default rate.

To further investigate the feature of increasing PD differences and choices of $\mathbf{u} = (u_1, u_2, \ldots, u_I)$' in the one-sided calibration test, the cases I = 3 and I = 4 are explicitly analyzed in the sequence. For each I, four CRMs are considered, with their PD_is depicted in table 2. CRMs of table 2 can have PD_is following either an arithmetic progression or a geometric progression. Besides, two strategies of configuration of the indifference region are considered: a liberal one with $u_i = PD_{i+1}$ and a more precise one with $u_i = (PD_{i+1} + PD_i)/2$. In order to allow for a fair comparison of power among distinct CRMs, PD_is figures of table 2 are chosen with the purpose that the resulting sets of ratings of each CRM cover equal ranges in the

[23] This is approximate because the computation was based on I = 1. (In fact, the true attained power in a multi-rating setup is smaller.) Also, the discussion of this paragraph assumes that true **PD** = postulated **PD**.

Table 3. Power Comparison among CRM Designs and u_i Choices, I = 3

$\rho_B/\rho_W = 0.8, \alpha = 0.15$

	PD_is Follow Arithmetic Progression		**PD_is Follow Geometric Progression**	
	$\mathbf{u_i = PD_{i+1}}$	$\mathbf{u_i = (PD_{i+1} + PD_i)/2}$	$\mathbf{u_i = PD_{i+1}}$	$\mathbf{u_i = (PD_{i+1} + PD_i)/2}$
$\rho_W = 0.12$, Y = 10	0.97	0.60	1.00	0.95
In-between	0.86	0.43	0.98	0.81
$\rho_W = 0.18$, Y = 5	0.72	0.34	0.91	0.67

Notes: Power of the test $H_0 : PD_i \geq u_i$ for some $i = 1 \ldots I$ against $H_1 : PD_i < u_i$ for every $i = 1 \ldots I$, computed at the postulated PDs of table 2 (case I = 3). ρ_W = within-rating asset correlation, ρ_B = between-rating asset correlation, in-between scenario characterized by $(\rho_W/Y) = 0.15^2$, Y = number of years, α = size of test.

PD scale. More specifically, this goal is interpreted here as all CRMs having equal u_0 and u_i.[24]

The power figures of the one-sided calibration test at the postulated **PDs** are shown in tables 3 and 4, according to values set to parameters ρ_W and Y. The values of these parameters are chosen considering three feasible scenarios: a favorable one characterized by ten years of data and a low within-rating correlation of 0.12, an unfavorable one characterized by the minimum number of five years prescribed by Basel II (cf. Basel 2006a) and a high ρ_W at 0.18, and an in-between scenario.[25]

Tables 3 and 4 show that CRMs with the feature of increasing $(PD_{i+1} - PD_i)$ usually achieve significantly higher levels of power than CRMs with equally spaced PD_is, confirming the intuition derived from table 1. The tables also reveal that, even when

[24] u_0 corresponds to the fictitious PD_o. In table 2, PD_o can be easily figured out from the constructional logic of the PD_i progression. For the construction of the CRMs of table 2, $u_0 = 1.22\%$ and $u_3 = 33\%$ for I = 3, and $u_0 = 2\%$ and $u_4 = 32\%$ for I = 4. Furthermore, the ratio of the PD_i geometric progression is set equal to 3 for I = 3 and 2 for I = 4.

[25] As ρ_B/ρ_W is fixed in tables 3 and 4, what matters for the power calculation is just the ratio (ρ_W/Y). Therefore, the in-between scenario can be thought of as characterized by adjusting both Y and ρ_W or just one of them. In tables 3 and 4 it is given by $(\rho_W/Y)^{1/2} = 0.15$.

Table 4. Power Comparison among CRM Designs and $\mathbf{u_i}$ Choices, I = 4

$\rho_B/\rho_W = 0.8, \alpha = 0.15$

	PD_is Follow Arithmetic Progression		**PD_is Follow Geometric Progression**	
	$u_i = PD_{i+1}$	**$u_i = (PD_{i+1} + PD_i)/2$**	**$u_i = PD_{i+1}$**	**$u_i = (PD_{i+1} + PD_i)/2$**
$\rho_W = 0.12$, Y = 10	0.82	0.39	0.95	0.68
In-between	0.62	0.28	0.81	0.48
$\rho_W = 0.18$, Y = 5	0.49	0.22	0.65	0.37

Notes: Power of the test $H_0 : PD_i \geq u_i$ for some $i = 1 \ldots I$ against $H_1 : PD_i < u_i$ for every $i = 1 \ldots I$, computed at the postulated PDs of table 2 (case I = 4). ρ_W = within-rating asset correlation, ρ_B = between-rating asset correlation, in-between scenario characterized by $(\rho_W/Y) = 0.15^{1/2}$, Y = number of years, α = size of test.

solely focusing on the former, more demanding requirements for u_i (cf. $u_i = (PD_{i+1}+ PD_i)/2$) may produce overly conservative tests, with, for example, power on the level of only 37 percent. Therefore liberal strategies for u_i (cf. $u_i = PD_{i+1}$) seem to be necessary for realistic validation attempts, and attention is focused on these strategies in the remainder of this section. Further from the tables, the power is found to be very sensitive to the within-rating correlation ρ_W and to the number of years Y. It can increase more than 80 percent from the worst to the best scenario (cf. last column of table 4).

While in previous tables the between-rating correlation parameter ρ_B is held fixed, tables 5 and 6 examine its effect, along a set of feasible values, on the power of the test. Power is computed at the postulated **PDs** of CRMs of table 2 with $u_i = PD_{i+1}$, I = 4 and for the in-between scenario of parameters of ρ_W and Y. The tables show just a minor effect of ρ_B, regardless of the size of the test and the CRM design. Therefore, narrowing down the uncertainty in the value of ρ_B value is not of great importance if just approximate levels of power are desired at postulated **PDs**. The elements that indeed drive the power of the test are unveiled in the next analysis.

Tables 7 and 8 provide insights on the relative role played by the different ratings on the power. Power is computed at postulated **PDs** for a sequence of four embedded CRMs, starting with the CRM with equally spaced PDs of the second line of table 7 (the CRM with

Table 5. Effect of ρ_B when $\mathrm{PD_i}$s Follow Arithmetic Progression

$u_i = PD_{i+1}$, $(\rho_W/Y)^{1/2} = 0.15$, $I = 4$

	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 15\%$
$\rho_B/\rho_W = 0.6$	0.32	0.47	0.58
$\rho_B/\rho_W = 0.7$	0.35	0.50	0.60
$\rho_B/\rho_W = 0.8$	0.38	0.52	0.62
$\rho_B/\rho_W = 0.9$	0.41	0.55	0.65

Notes: Power of the test $H_0 : PD_i \geq u_i$ for some $i = 1 \ldots I$ against $H_1 : PD_i < u_i$ for every $i = 1 \ldots I$, computed at the postulated PDs of table 2 (case $I = 4$, $u_i = PD_{i+1}$, $\mathrm{PD_i}$s follow arithmetic progression). ρ_W = within-rating asset correlation, ρ_B = between-rating asset correlation, Y = number of years, α = size of the test.

Table 6. Effect of ρ_B when $\mathrm{PD_i}$s Follow Geometric Progression

$u_i = PD_{i+1}$, $(\rho_W/Y)^{1/2} = 0.15$, $I = 4$

	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 15\%$
$\rho_B/\rho_W = 0.6$	0.54	0.69	0.78
$\rho_B/\rho_W = 0.7$	0.56	0.71	0.79
$\rho_B/\rho_W = 0.8$	0.60	0.73	0.81
$\rho_B/\rho_W = 0.9$	0.62	0.74	0.82

Notes: Power of the test $H_0 : PD_i \geq u_i$ for some $i = 1 \ldots I$ against $H_1 : PD_i < u_i$ for every $i = 1 \ldots I$, computed at the postulated PDs of table 2 (case $I = 4$, $u_i = PD_{i+1}$, $\mathrm{PD_i}$s follow geometric progression). ρ_W = within-rating asset correlation, ρ_B = between-rating asset correlation, Y = number of years, α = size of the test.

increasing PD differences of the second line of table 8). Each next CRM in table 7 (table 8) is built from its antecedent by dropping the less risky (riskiest) rating. Power is computed for the in-between scenario and $u_i = PD_{i+1}$. The tables reveal that as the number of ratings diminishes, the power increases just to a minor extent, provided the riskiest (less risky) ratings are always kept in the CRM. Thus it can be said that in table 7 (table 8) the highest (lowest) $\mathrm{PD_i}$s drive the power of the test. This is partly intuitive because the highest (lowest) $\mathrm{PD_i}$s correspond to the smallest differences $(u_i - PD_i)$ in

Table 7. Influence of Distinct PD_is on Power

PD_is follow arithmetic progression;
$\rho_B/\rho_W = 0.6$; $(\rho_W/Y)^{1/2} = 0.15$; $u_i = PD_{i+1}$

PD_is	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 15\%$
2%, 9.5%, 17%, 24.5%	0.32	0.47	0.58
9.5%, 17%, 24.5%	0.32	0.47	0.58
17%, 24.5%	0.34	0.49	0.59
24.5%	0.44	0.58	0.68

Notes: Power of the test $H_0 : PD_i \geq u_i$ for some i = 1...I against $H_1 : PD_i < u_i$ for every i = 1...I, for I = 4...1, computed at the PDs of the first column. ρ_W = within-rating asset correlation, ρ_B = between-rating asset correlation, Y = number of years, α = size of the test.

Table 8. Influence of Distinct PD_is on Power

PD_is follow geometric progression;
$\rho_B/\rho_W = 0.6$; $(\rho_W/Y)^{1/2} = 0.15$; $u_i = PD_{i+1}$

PD_is	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 15\%$
2%, 4%, 8%, 16%	0.54	0.69	0.78
2%, 4%, 8%	0.54	0.69	0.78
2%, 4%	0.56	0.71	0.79
2%	0.65	0.77	0.84

Notes: Power of the test $H_0 : PD_i \geq u_i$ for some i = 1...I against $H_1 : PD_i < u_i$ for every i = 1...I, for I = 4...1, computed at the PDs of the first column. ρ_W = within-rating asset correlation, ρ_B = between-rating asset correlation, Y = number of years, α = size of the test.

the CRMs of table 7 (table 8) and because distinct PD_is contribute to the power differently just to the degree their differences $(u_i - PD_i)$ vary.[26] The surprising part of the result refers to the degree of relative low importance of the dropped PD_is: the variation of power between I = 1 and I = 4 can be merely around 10 percent. This

[26] It is easy to see that for the CRMs with equally spaced PD_is, $(u_i - PD_i)$ is trivially constant in i but the Φ^{-1}-transformed difference $(u_i - PD_i)$ decreases in i. For the CRMs with increasing $(PD_{i+1} - PD_{i,})$, $(u_i - PD_i)$ trivially increases in i and the Φ^{-1}-transformed difference $(u_i - PD_i)$ increases in i too.

latter observation should be seen as a consequence of the functional form of DRAPM, particularly the choice of the normal copula for the (transformed) default rates and the form of Σ.[27]

A message embedded in the previous tables is that in some quite feasible cases (e.g., $Y = 5$ years available at the database, $\rho_W = 0.18$ reflecting the portfolio default volatility, $\alpha < 15\%$ desired), the one-sided calibration test can have substantially low power (e.g., lower than 50 percent at the postulated $\mathbf{PD}$). Another related problem refers to the test not being similar on the boundary between the hypotheses and therefore biased (if $I > 1$).[28] To cope with these *deficiencies*, the statistical literature contains some proposals of non-monotone uniformly more powerful tests for the same problem, such as in Liu and Berger (1995) and McDermott and Wang (2002). The new tests are constructed by carefully enlarging the rejection region in order to preserve the size α. The enlargement trivially implies power dominance. The new tests have two main disadvantages though. First, from a supervisory standpoint, non-monotone rejection regions are harder to defend on an intuitive basis because they imply that a bank could pass from a state of validated CRM to a state of non-validated CRM if default rates for some of the ratings *decrease*. Second, from a theoretical point of view, Perlman and Wu (1999) note that the new tests do not dominate the original test in the decision-theoretic sense because the probability of validation under H_0 (i.e., when the CRM is incorrect) is also higher for them. The authors conclude that UMP tests should not be pursued at any cost, particularly at the cost of intuition. This is the view adopted in this study, so the new tests are not explored further in this paper.

Yet, one may try to include some prior knowledge in the formulation of the one-sided calibration test as a strategy for power improvement. Notice, first, that the size α of the test is attained when all but

[27] Bluemke (2013) also shows situations in which the power of validation tests is driven by a single rating, but he does not address the influence of the CRM design in determining what this rating is.

[28] A test is α-similar on a set A if the probability of rejection is equal to α everywhere there. A test is unbiased at level α if the probability of rejection is smaller than α everywhere in H_0 and greater than α everywhere in H_1. Every unbiased test at level α with a continuous power function is α-similar in the boundary between H_0 and H_1 (Gourieroux and Monfort 1995).

one of the $\mathrm{PD_i}$s go to 0 while the remaining one is set fixed at $\mathrm{u_i}$.[29] This is probably a very unrealistic scenario against which the bank or the supervisor would like to be protected. The bank/supervisor may alternatively remove by assumption this unrealistic case from the space of **PD** possibilities and rather consider that part of the information to be tested is true. Notably, it can be assumed that the postulated $\mathrm{PD_{i-1}}$, not 0, represents a lower bound for $\mathrm{PD_i}$, for every rating i. A natural modification of the test consists then on replacing $\mathrm{z_\alpha}$ with a smaller constant $\mathrm{c} > 0$ to adjust to the removed unrealistic **PD** scenarios,[30] with resulting enlargement of the critical region and achievement of a more powerful test. (Recall the definition of the critical region in (5).) Hence, c is defined by the requirement that the size of the modified test (with c instead of $\mathrm{z_\alpha}$) in the reduced **PD** space is α. Similarly to Sasabuchi (1980), the determination of c needs the examination of only the **PD** vectors with all but one of their coordinates' $\mathrm{PD_i}$s equal to their lower bounds (the postulated $\mathrm{PD_{i-1}}$s), and the remaining one, say $\mathrm{PD_j}$, set at $\mathrm{u_j}$, for j varying in 1. . . I. More formally,

$$\begin{aligned}&\mathrm{Max}_{1\leq \mathrm{j}\leq \mathrm{I}}(\Phi_{\mathrm{I}}(-\mathrm{c} + (u_1 - PD_0)/(\rho_{\mathrm{W}}/\mathrm{Y})^{1/2}, \ldots, -\mathrm{c}, \ldots,\\ &\quad - \mathrm{c} + (u_I - PD_{\mathrm{I}-1})/(\rho_{\mathrm{W}}/\mathrm{Y})^{1/2}; \rho_{\mathrm{B}}/\rho_{\mathrm{W}}) = \alpha^{31},\end{aligned} \tag{9}$$

from which the value of c can be derived.

However, produced results indicate the previous modification approach is of limited efficacy to power improvement. More specifically, computed results indicate that the power increase is relevant only in the region of small (probably unrealistic) ratio $\rho_{\mathrm{B}}/\rho_{\mathrm{W}}$ or for ambitious choices of $\mathrm{u_i}$ (i.e., close to $\mathrm{PD_i}$). In the latter case, the increase is not sufficient, however, to the achievement of reasonable levels of power because the original levels are already too low (cf. table 1, for example). Those results are consistent with the intuition derived from the analysis of tables 7 and 8.

[29] Note $\mathrm{PD_i} \to 0 \Rightarrow PD_i \to -\infty$. The limiting **PD** vector is in $\mathrm{H_0}$ and, therefore, should not be validated. It has a probability of validation equal to α.

[30] As the coordinates of the input to the power function cannot go to infinity as before, $-\mathrm{c} > -\mathrm{z_\alpha}$ for the size to be achieved.

[31] PD_0 is here just a lower bound to PD_1. It could be $-\infty$ or defined subjectively based on accumulated practical experience. Note that the new critical region will now depend on ρ_{B} and that the calculation of c needs some computational effort.

On the other hand, one may also try to derive the LRT based on the restricted **PD** parameter space:

$$\begin{aligned} \mathrm{H_o} : & \ PD_{\mathrm{i}} \geq u_{\mathrm{i}} \text{ for some i} = 1 \ldots \mathrm{I} \text{ and} \\ & \ PD_{\mathrm{i}} \geq \text{postulated } PD_{\mathrm{i}-1} \text{ for every i} = 1 \ldots \mathrm{I} \\ \mathrm{H_1} : & \ PD_{\mathrm{i}} < u_{\mathrm{i}} \text{ for every i} = 1 \ldots \mathrm{I} \text{ and} \\ & \ PD_{\mathrm{i}} \geq \text{postulated } PD_{\mathrm{i}-1} \text{ for every i} = 1 \ldots \mathrm{I}.^{32} \end{aligned}$$

The LRT will differ from the modification approach with respect to the information contained in the observed default rates. The LRT will have very small observed average default rates, providing lower relative evidence in favor of $\mathrm{H_1}$ because, by assumption, they cannot be explained by very small PDs.[33] Accordingly, the null distribution of the likelihood-ratio (LR) statistic doesn't need to put mass on those unrealistic **PD** scenarios. Unfortunately, to the best of the author's knowledge, the derivation of the LRT critical region for such a problem is lacking in the statistical literature. Its complexity arises from the facts that, in contrast to the original one-sided calibration test, $\mathrm{H_0}$ and $\mathrm{H_1}$ do not share the same boundary in $\Re^{\mathrm{I}}$ and that the boundary indeed shared is a limited set. Thus, it is reasonable to conjecture that the null distribution of the LR statistic will be fairly complicated. And similarly to the previous strategy, if $u_i >>$ postulated PD_{i-1} for most ratings, the increase in power is likely to negligible again.[34]

4.2 Two-Sided Calibration Testing

The section now comments on two-sided calibration testing, mostly from a theoretical perspective. Similarly to the one-sided version, the hypotheses of a two-sided test can be stated as follows:

[32] $\mathrm{H_1}$ need not be defined only by strict inequalities here since the union $\mathrm{H_0}$ U $\mathrm{H_1}$ does not span the full $\Re^{\mathrm{I}}$ space.

[33] Very small observed average default rates in the sense that $\Phi^{-1}(\mathrm{DR_i})/(1-\rho_{\mathrm{W}})^{1/2} < \Phi^{-1}(\text{postulated PD}_{\mathrm{i}-1})$.

[34] It is important to remark that if I is large, strategies of power improvement will generally have more chances of *relative* success, although they depart from lower original levels of power.

$$\begin{aligned}&\mathrm{H_o} : PD_i \geq u_i \text{ or } PD_i \leq l_i \text{ for some } \mathrm{i} = 1 \ldots \mathrm{I}\\ &\mathrm{H_1} : l_i < PD_i < u_i \text{ for every } \mathrm{i} = 1 \ldots \mathrm{I}.\end{aligned}$$

Now the acceptable indifference region is defined by two parameters $\mathrm{u_i}$ and $\mathrm{l_i}$ for each rating i, with ideally $\mathrm{l_i} \geq$ postulated $\mathrm{PD_{i-1}}$ and $\mathrm{u_i} \leq$ postulated $\mathrm{PD_{i+1}}$. Under that formulation, the test belongs to the class of multivariate equivalence tests, which are tests designed to show similarity rather than difference and are widely employed in the pharmaceutical industry (under the denomination of bioequivalent tests) to demonstrate that drugs are equivalent. Berger and Hsu (1996) comprehensively review the recent development of equivalence tests in the univariate case ($\mathrm{I} = 1$). The standard procedure to test univariate equivalence is the TOST test (two one-sided tests—called this because the procedure is equivalent to performing two size-α one-sided tests and concluding equivalence only if both reject). Wang, Hwang, and Dasgupta (1999) discuss the extension of TOST to the multivariate case, making use of the intersection-union method.[35] When applied to the DRAPM distribution, that extension results in the following critical region for the two-sided calibration test.[36]

Reject $\mathrm{H_o}$ (i.e., validate the CRM) if

$$\begin{aligned}&l_i/(1-\rho_\mathrm{W})^{1/2} + \mathrm{z}_\alpha(\rho_\mathrm{W}/(\mathrm{Y}(1-\rho_\mathrm{W})))^{1/2}\\ &\quad \leq \overline{DR}_i \leq u_i/(1-\rho_\mathrm{W})^{1/2} - -\mathrm{z}_\alpha(\rho_\mathrm{W}/(\mathrm{Y}(1-\rho_\mathrm{W})))^{1/2}\end{aligned} \tag{10}$$

for every $i = 1 \ldots \mathrm{I}$.

As the maximum power of the test occurs in the middle point of the cube $[l_i u_i]^\mathrm{I}$, it is reasonable to make the cube symmetric around the postulated **PD** (in other words, to make $u_i - PD_i = PD_i - l_i$ for every i), so that the highest probability of validating the CRM occurs exactly at the postulated **PD**. Additional configurations of the indifference region may include, as in the one-sided test, choosing $u_i = PD_{i+1}$ or $l_i = PD_{i-1}$ (but not both).

[35] Wang, Hwang, and Dasgupta (1999) also show that TOST is basically an LR test.

[36] The standard TOST is formulated assuming unknown variance, while the proposed two-sided calibration test of this paper assumes known variance. Therefore the reference to the term TOST encompasses here some freedom of notation.

Similarly to the one-sided test, the two-sided version has problems of lack of power and bias.[37] In this respect, the statistical literature contains some proposals for improving TOST (Berger and Hsu 1996; Brown, Hwang, and Munk 1998), which are again subject to criticism from an intuitive point of view by Perlman and Wu (1999).[38] Furthermore, an additional drawback of the two-sided test, in contrast to the original TOST, is its excess of conservatism because the test is only level α, while its size may be much smaller.[39] That observation indicates the magnified difficulty in performing two-sided calibration testing.

Yet, two additional approaches to testing multivariate equivalence deserve comments. The first one is developed by Brown, Casella, and Hwang (1995). Applied to the problem of **PD** calibration testing, it consists of accepting an alternative hypothesis H_1 (i.e., validating the CRM) if the Brown confidence set for the **PD** vector is entirely contained in H_1. The approach would allow the bank or the supervisor to separate the execution of the test from the task of defining an indifference region because H_1 configuration could be discussed at a later stage, after the knowledge of the *form* of the set. In particular, the confidence set can be seen as the smallest indifference region that still permits validation of the calibration. Brown, Casella, and Hwang (1995) propose an optimal confidence set in the sense that if the true **PD** vector is equal to the postulated one, then the expected volume of that set is minimal, which means that, in average terms, maximal precision is achieved when calibration is *exactly* right. The cost of this optimality is larger set volumes for **PDs** different from the postulated one. Munk and Pfluger (1999) show in simulation exercises that the power of Brown's procedure can be substantially lower than those of more standard tests, like the TOST, for a wide range of **PDs** close to the postulated one. Therefore, in light of the view of this paper that ratings could more

[37] If $I > 1$, the test is not similar on the boundary between the hypotheses and is therefore biased.

[38] However, in the case of calibration testing with known variance, the bias is not as pronounced as in the standard TOST with unknown variance, due to the impossibility of making the variance go to 0 as in Berger and Hsu (1996).

[39] It can be shown that the degree of conservatism depends on ρ_B. The reason for the discrepancy with the standard TOST relates again to the impossibility of making the variance go to 0 as in Berger and Hsu (1996).

realistically be seen as PD intervals, the benefit of the optimality at a single point is doubtful at a minimum. Consequently, Brown's approach is regarded here as of more theoretical than practical value to calibration testing.[40]

The second different approach to testing multivariate equivalence is developed by Munk and Pfluger (1999). So far, this paper has just considered rectangular sets in the H_1 statements of the calibration tests. The goal has been to show that the true **PD** lies in a rectangle or in a quadrant of the space $\Re^{\mathrm{I}}$. The referred authors analyze instead the use of ellipsoidal alternatives for the multivariate equivalence problem, which, for purposes of calibration testing, can be exemplified as follows:

$$\mathrm{H}_o : \boldsymbol{e}^t \mathbf{D} \boldsymbol{e} \geq \Delta$$

$$\mathrm{H}_1 : \boldsymbol{e}^t \mathbf{D} \boldsymbol{e} < \Delta,$$

where $\boldsymbol{e} = \boldsymbol{PD}$ – postulated $\boldsymbol{PD}$, **D** is a positive definite matrix, which conceives a notion of distance in $\Re^{\mathrm{I}}$, and Δ denotes a fixed tolerance bound. **D** and Δ define an indifference region for **PD**.

Munk and Pfluger (1999) advocate this formulation to allow the notion of equivalence to be interpreted as a combined measure of several parameters (e.g., a combination of the PD_is, $\mathrm{i} = 1 \ldots \mathrm{I}$). As a consequence, this implies that very good *marginal* equivalence (e.g., the true PD_1 is very close to the postulated PD_1) should allow larger indifference regions for the other parameters (e.g., the other PD_is). Conceptually, though, this point is hard to justify in the validation of CRMs unless miscalibration were necessarily derived from a systematic erroneous estimation of all the PD_is. Nevertheless, the view of this paper is that miscalibration could be rather rating specific. Furthermore, note that the rectangular alternatives already permit a lot of flexibility in allowing different indifference interval lengths for different ratings. Consequently, for purposes of calibration testing, ellipsoidal alternatives are regarded here more as a practical complication.[41]

[40] Other confidence set approaches to calibration testing are also possible. Some of them are, however, dominated by the multivariate TOST (Munk and Pfluger 1999).

[41] However, for purposes of power improvement, it still might be useful to investigate ellipsoidal alternatives inscribed or approximating rectangular alternatives. This investigation is not addressed in this paper.

5. Tests of Rating Discriminatory Power

One of the most traditional measures of discriminatory power is the area under the ROC curve (AUROC).[42] Let n and m be two distinct random borrowers with probabilities of default $\mathrm{PD_n}$ and $\mathrm{PD_m}$, respectively. Following Bamber (1975), AUROC is defined as

$$\begin{aligned}\mathrm{AUROC} = {} & \mathrm{Prob}(\mathrm{PD_n} > \mathrm{PD_m} |\ \text{n defaults and m doesn't}) \\ & + {}^{1}\!/_{2} \cdot \mathrm{Prob}(\mathrm{PD_n} = \mathrm{PD_m} |\ \text{n defaults and m doesn't}).\end{aligned} \tag{11}$$

High values of AUROC (close to 1) are typically interpreted as evidence of good CRM discriminatory performance. However, the definition of AUROC as the probability of an event concerning the realizations of two (random) borrowers makes it a function not only of the **PD** vector but also of the default correlation structure.[43] To the extent that the CRM should not be held accountable for the effect of default dependency between borrowers, the AUROC measure of discrimination becomes distorted.[44] Blochlinger (2012) shows this distortion by means of a numerical example. The proposition below shows formally the dependency of AUROC on the asset correlation parameters.

Proposition. Consider an extension of DRAPM in which (ρ_{ij}) is the matrix of asset correlations between borrowers of ratings i and j, i,j = 1…I. Let P(i,j) denote the probability of two random borrowers having ratings i and j and P(i) the probability of one random borrower having rating i. Then

$$AUROC = \frac{\sum_{i>j} \Phi_2\left(PD_i, -PD_j, -\rho_{ij}\right)\mathrm{P(i,j)} + \frac{1}{2}\sum_i \Phi_2\left(PD_i, -PD_i, -\rho_{ii}\right)\mathrm{P(i)}}{\sum_{i,j} \Phi_2\left(PD_i, -PD_j, -\rho_{ij}\right)\mathrm{P(i,j)}} \tag{12}$$

[42] ROC = receiver operating characteristic curve (cf. Bamber 1975). $0 \leq \mathrm{AUROC} \leq 1$.

[43] It is a function of the distribution of borrowers across the ratings too.

[44] Note that, in contrast, the definition of good calibration is always *purely* linked to the good quality of the **PD** vector, although the way to *empirically* conclude that will typically depend on the default correlation values, as shown in section 4.

Proof. See appendix 2.

Blochlinger (2012) proposes a measure of discriminatory power that is not a function of default dependency. However, in contrast to AUROC, it is a function of the portfolio-wide true (unknown) PD and, besides, his asymptotic testing results are based on a multiplicative form for the conditional PDs not obeyed by the Basel II model or DRAPM (equation (2) is not of multiplicative form). This section describes alternatives for tests of *rating* discriminatory power built upon the DRAPM distribution. The qualifying term *rating* is added purposefully to the traditional expression "discriminatory power" to emphasize that the property desired to be concluded/measured here is different from that embedded in traditional measures of discriminatory power. Rather than verifying that the ex post rating distributions of the default and non-default groups of borrowers are as separate as possible, the proposed tests of *rating* discriminatory power aim at showing that $\mathrm{PD_i}$ is a strictly increasing function of i. In other words, the discriminatory power should be present *at the rating level* or, more concretely, low-quality ratings should have larger $\mathrm{PD_i}$s. Note that this is a less stringent requirement than correct two-sided calibration and the alternative hypothesis here will, therefore, strictly contain the $\mathrm{H_1}$ of the two-sided calibration test.[45] In this sense, the fulfillment of good rating discriminatory power is consistent with the pursuit of correct calibration. Furthermore, as the proposed tests are based on hypotheses involving solely the **PD** vector, they are not functions of default correlations; consequently, they address the two pitfalls of traditional measures of discriminatory power that were discussed in the introduction. Finally, showing PD monotonicity along the rating dimension is also useful to corroborate the assumptions of some methods of PD inference on low default portfolios (e.g., Pluto and Tasche 2005).

This section distinguishes between a test of *general* rating discriminatory power and a test of *focal* rating discriminatory power. The former addresses a situation where the bank or supervisor is

[45] Provided $\mathrm{u_i} < \mathrm{l_{i+1}}$ for $\mathrm{i} = 1 \ldots \mathrm{I} - 1$, as expected in practical applications. On its turn, Blochlinger (2012) investigates a less stringent requirement than correct two-sided calibration but a more demanding one than PD monotonicity, namely that PD ratios between ratings equal specific constants.

uncertain about the increasing PD behavior along the whole rating scale, whereas the latter focuses on a pair of consecutive ratings. The formulation of the general test is proposed below.

$$\mathrm{H_o} : PD_i \geq PD_{i+1} \text{ for some i} = 1 \ldots \mathrm{I} - 1$$
$$\mathrm{H_1} : PD_i < PD_{i+1} \text{ for every i} = 1 \ldots \mathrm{I} - 1.$$

By viewing $PD_{i+1} - PD_{\mathrm{i}}$ as the unknown parameter to be estimated (up to a constant) by $DR_{i+1} - DR_i$ for every rating i, the previous test involves testing strict homogeneous inequalities about normal means. (The key observable variables are now default rate differences between consecutive ratings, rather than the default rates themselves, as in the one-sided calibration test.) So, similarly to the one-sided calibration test, a size-α likelihood-ratio critical region can be derived.

Reject $\mathrm{H_0}$ (i.e., validate the CRM) if

$$\overline{DR}_{i+1} - \overline{DR}_i > \mathrm{z}_\alpha(2(\rho_\mathrm{W} - \rho_\mathrm{B})/(\mathrm{Y}(1 - \rho_\mathrm{W})))^{1/2} \quad \text{for every i} = 1 \ldots \mathrm{I} - 1. \qquad (13)$$

It is worth noting above that, differently from the calibration tests, there is no need for the configuration of an indifference region, as the desired $\mathrm{H_1}$ conclusion is already defined by strict inequalities. On the other hand, now the critical region and—therefore, the decision itself to validate the CRM—depends on the unknown parameter ρ_B. The Basel II case ($\rho_\mathrm{B} = \rho_\mathrm{W}$) represents the extreme liberal situation where just an observed increasing behavior of the average annual default rates along the rating dimension is sufficient to validate the CRM (regardless of the confidence level α), whereas the case $\rho_\mathrm{B} = 0$ places the strongest requirement in the incremental increase of the default rate averages along the rating scale.[46] In practical situations, the bank or the supervisor may want to determine the highest value of ρ_B such that the general test still validates the CRM and then check how this value conforms to its beliefs about reality.

When theoretically compared with the power of the one-sided calibration test, the power of the general test is notably affected by

[46] This is again intuitive, as low values of ρ_B mean that ex post information about one rating does not contain much information about other ratings.

a trade-off of three factors.[47] First, the fact that now the underlying normal variables are likely to have smaller variances ($\mathrm{Var}(DR_{\mathrm{i}+1} - DR_i) = 2(\rho_{\mathrm{W}} - \rho_{\mathrm{B}})/(1 - \rho_{\mathrm{W}}) < \mathrm{Var}(DR_i) = \rho_{\mathrm{W}}/(1 - \rho_{\mathrm{W}})$, provided $\rho_{\mathrm{B}}/\rho_{\mathrm{W}} > 1/2$) contributes to an increase in power. On the other hand, the now not positive underlying correlations ($\mathrm{Corr}(DR_{i+1} - DR_i, DR_j - DR_{j-1}) = -1/2$ if $\mathrm{i} = \mathrm{j}$ and 0 otherwise, compared with $\mathrm{Corr}(DR_i,DR_j) = \rho_{\mathrm{B}}/\rho_{\mathrm{W}} > 0$ for $\mathrm{i} \neq j$) contributes to a decrease in power.[48] Finally, the presence of I – 1 statements in H_1, instead of I, implies a slight increase in power too. In general, the resulting dominating force is to be determined by the particular choices of ρ_{B}, ρ_{W}, and I. However, computed results indicate that discrimination test power will usually be larger than calibration power for CRM designs including both arithmetic and geometric progressions for the $\mathrm{PD_i}$s and reasonable specifications for the testing parameters.[49] Finally, as with calibration testing, similar comments on possible strategies for power improvement and their limitations apply here as well.

It is also worthwhile to discuss the situation where the bank or the supervisor is satisfied by the "general level" of rating discrimination except for a particular pair of consecutive ratings. Suppose the bank/supervisor wants to find evidence that two consecutive ratings (say ratings 1 and 2, without loss of generality) indeed distinguish the borrowers in terms of their creditworthiness. From a supervisory standpoint, a suspicion of regulatory arbitrage may, for instance, motivate the concern.[50] To examine this issue, this section formulates a test of focal rating discriminatory power, whose hypotheses are stated as follows.[51]

[47] Similarly to the calibration case, the power expression can be easily derived.

[48] Therefore, not necessarily validating rating discriminatory power is easier than validating (one-sided) calibration.

[49] Also, computed results in line with previous calibration findings indicate that CRMs whose $\mathrm{PD_i}$s follow geometric progression will generally achieve higher levels of power than when $\mathrm{PD_i}$s follow arithmetic progression and their power is basically driven by the first pairs of consecutive ratings, in the high-credit-quality part of the scale.

[50] Suspicion of regulatory arbitrage may derive from a situation where large credit risk exposures are apparently rated with slightly better ratings so that the resulting capital charge of Basel II is diminished.

[51] The discussion of this section is easily generalized to the situation where more than one pair of consecutive ratings are to have their rating discriminatory power verified.

$$\mathrm{H_o} : PD_1 = PD_2 \leq PD_3 \leq \ldots \leq PD_I$$
$$\mathrm{H_1} : PD_1 < PD_2 \leq PD_3 \leq \ldots \leq PD_I.$$

From a mathematical point of view, the development of the likelihood-ratio test for such a problem is more complex than the majority of the tests considered so far in this paper, because now the union of the null and the alternative hypotheses do not span the full $\Re^{\mathrm{I}}$, nor do the hypotheses share a common boundary. But, in contrast to the section 4 one-sided calibration LRT under **PD** restriction, now both $\mathrm{H_0}$ and $\mathrm{H_1}$ are convex cones. This implies that the null distribution of the LR will depend on the structure of the cone $\mathrm{C} = \mathrm{H_o}$ U $\mathrm{H_1}$, whether obtuse or acute with respect to norm induced by $\sum^{-1}$.[52] In the first case, the LR statistic follows a $\chi 2$ *bar* distribution under $\mathrm{H_0}$ (Menendez, Rueda, and Salvador 1992b).[53] In the second case, the distribution of the LR statistic is intractable, but the test is dominated in power by a *reduced* test comprised of testing just the *different parts* of the hypotheses $\mathrm{H_o}$ and $\mathrm{H_1}$ (Menéndez and Salvador 1991; Menéndez, Rueda, and Salvador 1992a). It can be shown that the structure of $\sum$ adopted in this paper makes the cone C acute, so that the second case is the relevant one.[54] The reduced dominating test takes the form below.

$$\mathrm{H_o} : PD_1 = PD_2$$
$$\mathrm{H_1} : PD_1 < PD_2.$$

The test above is just a particular case of the general rating discriminatory power test with I = 2. Accordingly, its rejection rule is given as follows.

[52] See Martín and Salvador (1988) and Menéndez, Rueda, and Salvador (1992b) for the definitions of those cone types. The norm induced by $\sum^{-1}$ is defined as $\|x\|_{\Sigma^{-1}} = x^T \mathbf{\Sigma}^{-1} x$.

[53] Although $\chi 2$ *bar* distributions are common in the theory of order-restricted inference (Robertson, Wright, and Dykstra 1988), application of the focal test in this circumstance is not very practical, as the determination of both the LRT statistic and the p-values are computationally intensive.

[54] This is true because $\mathbf{a_i'\Sigma a_j} \leq 0$, $\mathrm{i} \neq \mathrm{j}$, where the $\mathbf{a}_\mathrm{i}$'s ($\mathbf{a}_\mathrm{i} = (0, \ldots, -1, 1, \ldots, 0)$')) generate the linear restrictions defining the cone C. More specifically, it is true that $\mathbf{a_i'\Sigma a_j} = (\rho_\mathrm{B} - \rho_\mathrm{W})/(1 - \rho_\mathrm{W})$ if $|\mathrm{i} - \mathrm{j}| = 1$ or 0 if $|\mathrm{i} - \mathrm{j}| \geq 2$. See the mentioned references for further details. Whether more general but still realistic variance structures $\mathbf{\Sigma}$ might lead to a different conclusion is an interesting question not addressed in this paper.

Reject H_0 (i.e., validate the CRM) if

$$\overline{DR}_2 - \overline{DR}_1 > z_\alpha(2(\rho_W - \rho_B)/(Y(1-\rho_W)))^{1/2}. \tag{14}$$

The dominance of the focal test by a reduced test is a surprising result and was long considered an anomaly of the LR principle (e.g., Warrack and Robertson 1984). In the context of CRMs, this means that in order to judge the discriminatory performance of a particular pair of consecutive ratings, the bank or the supervisor would be in a better position if it simply disregards the prior knowledge of the performance of the other ratings. But how can less information be better? Only most recently Perlman and Wu (1999) showed that indeed the overall picture was not so much in favor of the "dominating" test, arguing that the latter presents controversial properties. For example, it rejects **PDs** *closer* to H_0 than to H_1.[55] Nevertheless, the practitioner does not have another choice besides using the power dominating test, because, as just observed, the null distribution of the LRT statistic for the focal test is unknown. Keeping that in mind, the analysis of this section provides the theoretical foundation to an easy-to-implement procedure that focuses solely on the supposedly problematic pair of ratings. More interestingly, however, a generalization of the results discussed in this section suggests a uniform procedure to check rating discriminatory power: select the ratings whose discriminatory capacity are at stake and apply the general test to them.

6. Finite-Sample Properties

All the tests discussed in this paper are based on the asymptotic distribution of DRAPM, which assumes an infinite number of borrowers for each rating. This section analyzes the implications to the performance of the one-sided calibration test of a finite but still large number of borrowers (N = 100 is chosen as the base case).[56] Due

[55] Perlman and Wu (1999) conclude once again that UMP size-α tests should not be pursued at any cost.

[56] The analysis is restricted to the one-sided calibration test not only because it is the main focus of this paper but also because the finite sample properties of discriminatory tests are more complex to analyze when distributions of default rate *differences* are involved. Also, as perceived later in the section, the issues of most concern related to the finite-sample properties of the two-sided calibration test derive from the analysis of the one-sided case.

to the strong reliance of the test on the asymptotic normality of the marginal distributions of DRAPM, it is important to verify how the real marginals compare to the asymptotic ones. The focus on a particular marginal allows then, for the sake of clarity, to direct the attention initially to the case I = 1. This section conducts Monte Carlo simulations of DRAPM, at the stage in which idiosyncratic risk is not yet diversified away, for N = 100 and Y = 5, unless stated otherwise. Based on a large set of simulated average annual default rates and for I = 1, the effective significance level is computed as a function of the nominal significance level α, for varying scenarios of the parameters true PD and ρ_W. In general, 200,000 simulations are run for each scenario.

$$\text{Effective confidence level} = \hat{Prob}\left(\frac{\sqrt{1-\rho_W}\,\overline{DR}_n - PD}{\sqrt{\rho_W/Y}} < -z_\alpha\right), \tag{15}$$

where the probability is estimated by the empirical frequency of the event and $\overline{DR}_n$ denotes a particular simulation result.

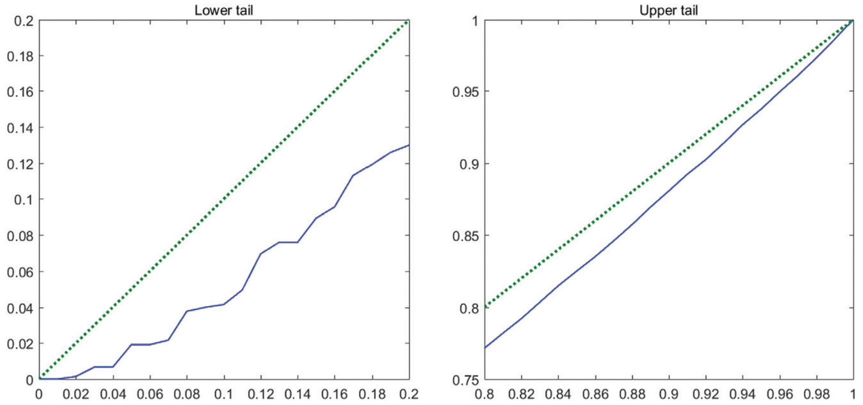
The effective level measures the real size of the asymptotic size-α one-sided test. Alternatively, since it is expressed in the form of a probability of rejection, the effective level can also be seen as the real power at the postulated PD, when the asymptotic power is equal to α, of an asymptotic size-δ one-sided test, with $\delta < \alpha$.[57] From both interpretations, the occurrence of effective levels lower than nominal levels means that the test is more conservative, with a smaller probability of validation in general than what is suggested by the analysis of section 4 based on DRAPM. Effective levels higher than nominal levels indicates the opposite: a finite-sample liberal *bias*.

A first general important finding derived from the performed simulations for the case I = 1 is that the convergence of the lower tails of the average (transformed) default rate distributions to their normal asymptotic limits is slower and less smooth than in the case of the upper tails, for realistic PD values. The situation is illustrated by the pair of graphs (figure 1) calculated based on the scenario PD = 3%, $\rho_W = 0.20$, N = 100, and Y = 5. The solid line represents

[57] More specifically, it is easy to see that $\delta = \Phi(-z_\alpha - (u - PD)/(\rho_W/Y)^{1/2})$.

Figure 1. Lower and Upper Tails

PD = 3%, $\rho_W = 0.20$, N = 100, Y = 5



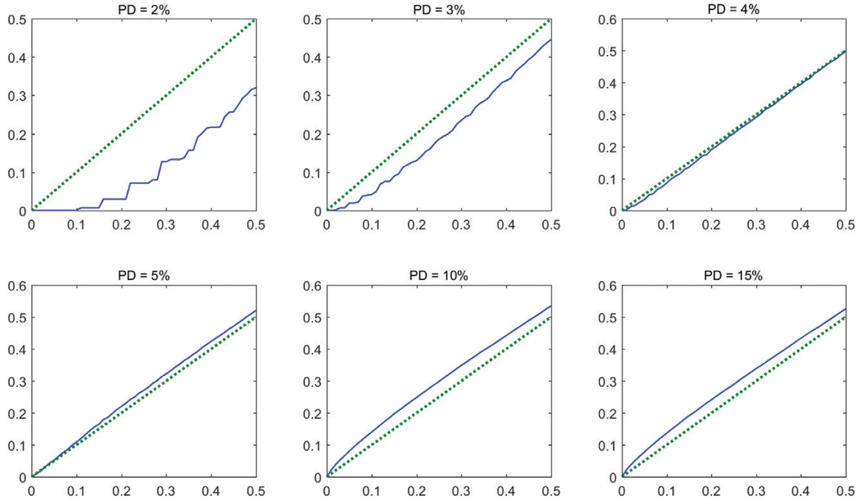
Notes: Solid line: Effective confidence level against the nominal size α of the asymptotic one-sided test H_0 : PD $\geq$ u against H_1 : PD $<$ u. Dotted straight line is the identity function to ease comparison. PD = true probability of default, ρ_W = asset correlation, N = number of borrowers, Y = number of years.

the effective confidence level for each nominal level depicted at the x-axes, while the dotted straight line is the identity function merely denoting the nominal level to facilitate comparison. Note that the effective level is much farther from the nominal value in the lower tail of the distribution (depicted in the right-hand graph) than in the upper tail (depicted in the left-hand graph). In particular, if the one-sided calibration test is employed at the nominal level of 10 percent, the test will be much more conservative in reality, as the effective size will be approximately only 4 percent.

Indeed, the fact that the lower tail is less well behaved is strongly relevant to this paper's one-sided calibration test. Under the approach of placing the undesired conclusion in H_0 (e.g., PD $\geq$ u), rejection of the null, or equivalently validation, is obtained if average default rates are small, so that the one-sided test is based in fact on the lower tail of the distribution. On the contrary, the upper tail would be the relevant part of the distribution had the approach of placing the "CRM correctly specified" hypothesis in H_0 been adopted, as in the rest of the literature. Since convergence of the upper tail is better behaved, the finite-sample departure from

Figure 2. Effect of PD

$\rho_W = 0.20$, N = 100, Y = 5



Notes: Solid line: Effective confidence level against the nominal size α of the asymptotic one-sided test H_0 : PD $\geq$ u against H_1 : PD < u. Dotted straight line is the identity function to ease comparison. PD = true probability of default, ρ_W = asset correlation, N = number of borrowers, Y = number of years.

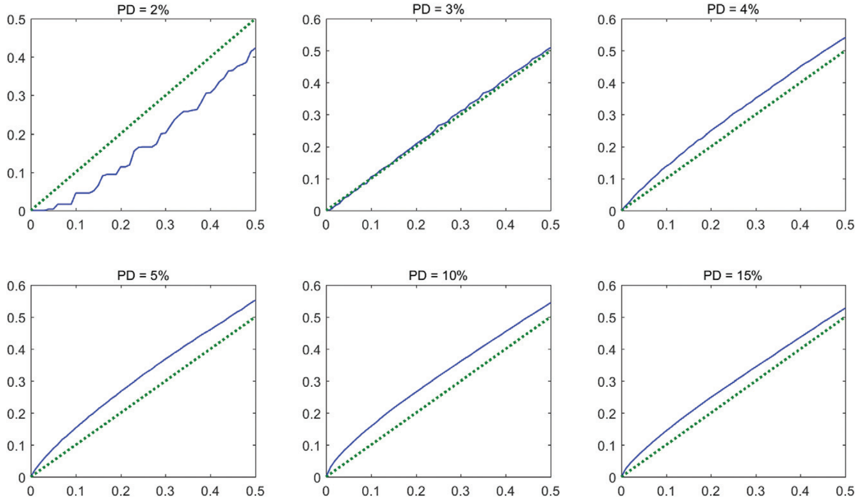
the normal limit would be smaller in this case. In the view of this paper, this would be, however, a misleading property of the latter approach because the worse relative behavior of the lower tail would not be revealed.

The main numerical findings regarding the finite sample power performance of the one-sided calibration test are described in the sequence, based on the analysis of the simulated lower tails.[58] The investigation starts with the effect of the true PD, when I = 1, on the effective confidence level. Figures 2 and 3 reveal that, in the region of 0% < PD < 10% and $0.15 < \rho_W < 0.20$, as PD increases, the test evolves from having a conservative bias (true power smaller than the

[58] Miu and Ozdemir (2008) also investigate finite-sample properties of similar validation tests, but their results are not comparable to those of this paper since they adopt H_0 : CRM correctly specified, they assume serial dependency of default rates, and they investigate different ranges of parameter values for PD, ρ_W, and Y.

Figure 3. Effect of PD

$\rho_W = 0.15$, N = 100, Y = 5



Notes: Solid line: Effective confidence level against the nominal size α of the asymptotic one-sided test H_0 : PD $\geq$ u against H_1 : PD < u. Dotted straight line is the identity function to ease comparison. PD = true probability of default, ρ_W = asset correlation, N = number of borrowers, Y = number of years.

asymptotic one) to having a liberal bias (true power larger than the asymptotic one). At PD = 4% for $\rho_W = 0.20$ or at PD = 3% for $\rho_W = 0.15$, the finite sample bias is approximately null as the test matches its theoretical limiting values. On the other hand, in the region of 10% < PD < 15% and $0.15 < \rho_W < 0.20$, as PD increases, the solid line comes back a bit closer to the dotted straight one, i.e., the test diminishes its liberal bias (but not sufficiently so as to turn conservative).

As the asymptotic one-sided test based on DRAPM already suffers from problems of lack of power, this section suggests, as a possible general recommendation, consideration of the real (unmodified) applications of the test solely in the cases where the finite-sample analysis indicates a non-conservative bias. Indeed, if instead an additional layer of conservatism is added to the already conservative asymptotic test, the resulting procedure may hardly validate at all. The restriction to the finite-sample liberal cases, when I = 1, suggests against, according to figures 2 and 3, attempts of validation of low PDs (e.g., PD $\leq 3\%$).

For the case $\mathrm{I} > 1$, it is easy to see that the effective level that measures the real size of the asymptotic size-α one-sided test is given by the maximum of the effective levels computed for each different rating assuming $\mathrm{I} = 1$.[59] Therefore, the effective confidence level takes graphically, for each nominal level α, the form of the maximum of the solid lines corresponding to the different rating PDs that constitute the CRM. As a result, the effective level may not be reduced in the presence of low PD ratings that introduce conservative bias on a marginal basis. Nevertheless, CRMs with low PD ratings will still be particularly hard to validate due to the lower true power derived from the corresponding real marginals. To better understand the multivariate case, this section computes the true effective power at the postulated $\mathrm{PD_i}$s of the CRMs of table 2. Effective power is estimated making use of the simulated average annual default rates, in a similar fashion to equation (15) but with $-z_\alpha + (u - PD)/(\rho_\mathrm{W}/\mathrm{Y})^{1/2}$ replacing $-z_\alpha$ and extending the formula to the multivariate case.[60]

The differences true powers at the postulated $\mathrm{PD_i}$s minus the asymptotic ones (shown in tables 3 and 4) are presented in tables 9 and 10.[61] Results indicate that the fall in power in the finite-sample setting is considerably more material in the CRM design where $\mathrm{PD_i}$s follow a geometric progression (unless the favorable scenario of low ρ_W and high Y is considered). This result is consistent with the finding of section 4 that the lowest $\mathrm{PD_i}$s drive the power of the multivariate one-sided test in this CRM design. Because the lowest PDs are exactly the ones that suffer most from a conservative finite-sample bias when $\mathrm{I} = 1$, as revealed in figures 2 and 3, this bias extends comprehensibly to the multivariate case. Consequently, when failing

[59] For $\mathrm{I} > 1$, the interpretation in terms of real power at the postulated PDs is also possible but a less direct one. For $\mathrm{I} > 1$, the effective confidence level can be seen as the maximum real power among I asymptotic size-δ_i one-sided tests, when the asymptotic power is equal to α, with $\delta_i < \alpha$ and the power being computed for test i at the postulated PD of rating i (while other ratings have PDs = 0). δ_i has a similar expression to the case $\mathrm{I} = 1$.

[60] Compare with equation (6), which determines the asymptotic power.

[61] The same (ρ_W,Y) scenarios of tables 3 and 4 are considered here. The in-between scenario is specified as $\rho_\mathrm{W} = 0.1575$ and $\mathrm{Y} = 7$ for the computation of the effective power. The assumption about $\rho_\mathrm{B}/\rho_\mathrm{W}$ is also kept here.

Table 9. Difference between Effective and Asymptotic Power for Several CRM Designs and u_i Choices, N = 100, I = 3

$\rho_B/\rho_W = 0.8, \alpha = 0.15$

	PD_is Follow Arithmetic Progression		**PD_is Follow Geometric Progression**	
	$u_i = PD_{i+1}$	**$u_i = (PD_{i+1} + PD_i)/2$**	**$u_i = PD_{i+1}$**	**$u_i = (PD_{i+1} + PD_i)/2$**
$\rho_W = 0.12$, Y = 10	–0.01	–0.01	0.00	–0.03
In-between	–0.01	0.00	–0.05	–0.09
$\rho_W = 0.18$, Y = 5	–0.01	–0.01	–0.11	–0.13

Notes: Effective and asymptotic power of the test $H_0 : PD_i \geq u_i$ for some $i = 1 \ldots I$ against $H_1 : PD_i < u_i$ for every $i = 1 \ldots I$, computed at the postulated PDs of table 2 (case I = 3). ρ_W = within-rating asset correlation, ρ_B = between-rating asset correlation, in-between scenario characterized by $(\rho_W/Y) = 0.15^2$, Y = number of years, α = asymptotic size of test. Effective power computed based on 200,000 Monte Carlo simulations of DRAPM.

Table 10. Difference between Effective and Asymptotic Power for Several CRM Designs and u_i Choices, N = 100, I = 4

$\rho_B/\rho_W = 0.8, \alpha = 0.15$

	PD_is Follow Arithmetic Progression		**PD_is Follow Geometric Progression**	
	$u_i = PD_{i+1}$	**$u_i = (PD_{i+1} + PD_i)/2$**	**$u_i = PD_{i+1}$**	**$u_i = (PD_{i+1} + PD_i)/2$**
$\rho_W = 0.12$, Y = 10	–0.02	–0.01	–0.03	–0.03
In-between	–0.02	–0.01	–0.08	–0.06
$\rho_W = 0.18$, Y = 5	–0.01	–0.01	–0.11	–0.08

Notes: Effective and asymptotic power of the test $H_0 : PD_i \geq u_i$ for some $i = 1 \ldots I$ against $H_1 : PD_i < u_i$ for every $i = 1 \ldots I$, computed at the postulated PDs of table 2 (case I = 4). ρ_W = within-rating asset correlation, ρ_B = between-rating asset correlation, in-between scenario characterized by $(\rho_W/Y) = 0.15^2$, Y = number of years, α = asymptotic size of test. Effective power computed based on 200,000 Monte Carlo simulations of DRAPM.

to validate CRMs with increasing PD differences because of large default rates of some low postulated $\mathrm{PD_i}$s, a possible practical advice is to apply the test only to the remainder of the postulated **PD** vector (e.g., ratings 3 to 7 in the example related to table 1). Alternatively, if suspicion of PD undercalibration is particularly placed on the low postulated $\mathrm{PD_i}$s, a higher nominal level α could be applied just to them.

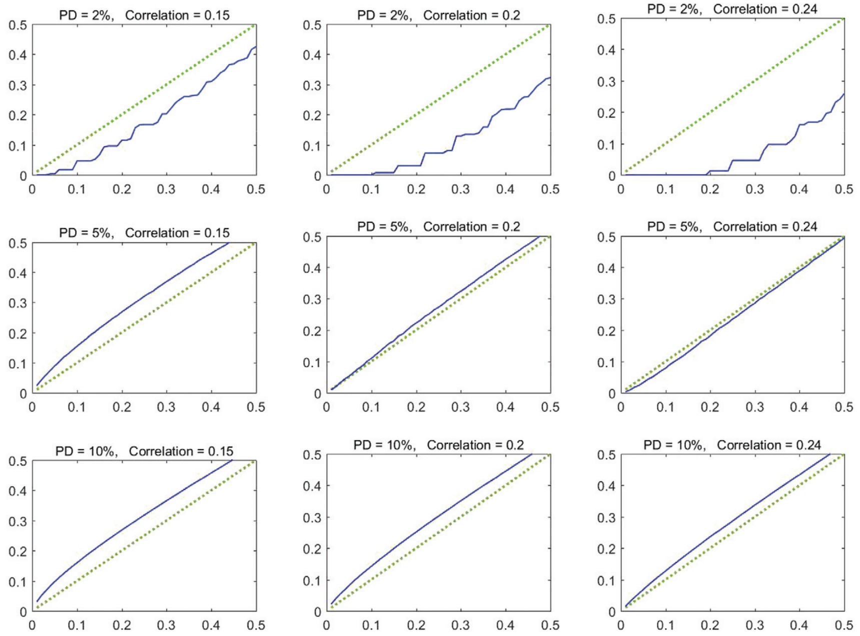
In the CRM design where $\mathrm{PD_i}$s follow an arithmetic progression, the highest $\mathrm{PD_i}$s drive the power of the test (see section 4), but they suffer instead from a small liberal finite-sample bias. Consequently, with equally spaced $\mathrm{PD_i}$s, the effective power of the multivariate test is not greatly affected, as noticed in tables 9 and 10. This is particularly useful since the asymptotic power in this CRM design is already low compared with CRMs with increasing PD differences (as discussed in section 4). Tables 9 and 10 finally reveal that the choice of the indifference region is generally of secondary importance to the magnitude of the finite-sample bias of the test in comparison with the choice of the rating $\mathrm{PD_i}$s, unless the favorable (ρ_W,Y) scenario is considered.

The influence of the within-rating asset correlation ρ_W under the base case of $N = 100$ is analyzed in figure 4. Considering initially the case $I = 1$, one notes that when ρ_W increases, the test evolves towards a more conservative bias (or towards a smaller liberal one), for every PD. Note that this represents a second channel, now through the finite-sample properties, by which ρ_W diminishes the power of the test. For the case $I > 1$, note first that the move towards a conservative bias is greater the lower the PD in figure 4. Because the lowest rating $\mathrm{PD_i}$s tend to drive the power of the test when CRMs possess increasing PD differences, it is expected that the finite-sample power-reducing effect of ρ_W will be stronger in precisely that type of CRM design. This is indeed the result found in tables 9 and 10 when one moves in the direction from $\rho_W = 0.12$ to $\rho_W = 0.18$ and $\mathrm{PD_i}$s follow a geometric progression.[62] Consequently,

[62] More precisely, in tables 9 and 10, Y also changes across the parameter scenarios. However, holding Y constant and just increasing ρ_W produces the same sort of result.

Figure 4. Effect of ρ_W

PD = 2%, 5%, or 10% depending on the row, Y = 5, N = 100



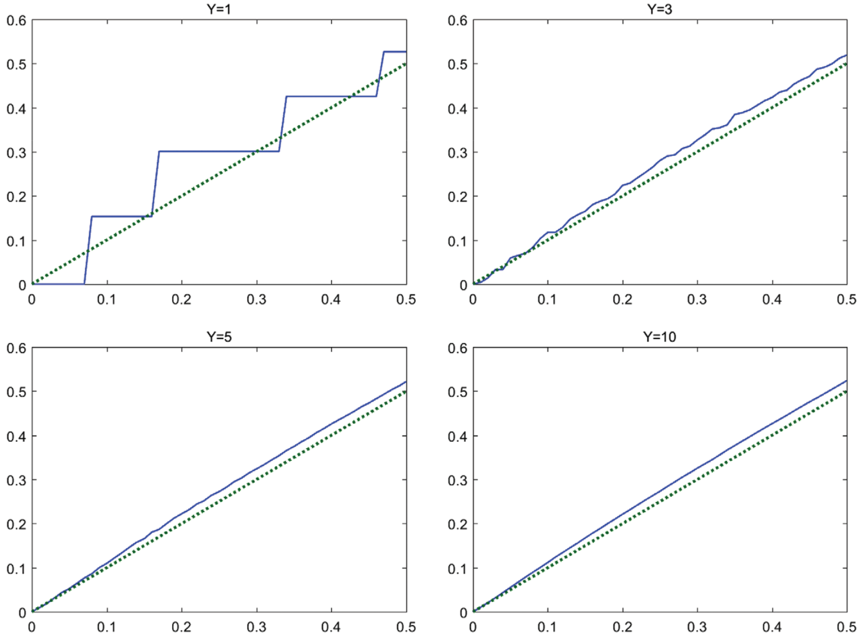
Notes: Solid line: Effective confidence level against the nominal size α of the asymptotic one-sided test H_0 : PD $\geq$ u against H_1 : PD $<$ u. Dotted straight line is the identity function to ease comparison. PD = true probability of default, ρ_W = asset correlation, N = number of borrowers, Y = number of years.

when a high value of ρ_W is an important consideration in the validation of CRMs with increasing PD differences, it may be advised to investigate separately the appropriateness of the calibration of postulated PD_is, in a similar fashion to what was suggested previously in this the section.

The influence of the number of years Y under the base case of N = 100 is analyzed in figure 5, considering again initially the case I = 1. The effect of an increase in the number of years, in the region of one to ten years, is to smooth considerably the distribution lower tail. Results not shown indicate that as N increases beyond 100,

Figure 5. Effect of Y

PD = 5%, $\rho_W = 0.20$, N = 100



Notes: Solid line: Effective confidence level against the nominal size α of the asymptotic one-sided test H_0 : PD $\geq$ u against H_1 : PD < u. Dotted straight line is the identity function to ease comparison. PD = true probability of default, ρ_W = asset correlation, N = number of borrowers, Y = number of years.

the solid and dotted lines come closer at every figure, as expected. Other produced results also indicate that, for the same Y, the lower-tail discontinuity is greater the lower the PDs. Consequently, based on the same reasoning underlying the previous analysis on the effect of ρ_W, the lower-tail discontinuity propagates more strongly to the multivariate case again when CRMs possess increasing PD differences.

Finally, it is important to observe that even if the one-sided test could be totally based on the simulated distributions of this section, there would still be some extreme cases where validation is virtually impossible at traditional low confidence levels. When I = 1 and Y = 1 (cf. figure 5) or true PD = 1%, for example, the lower tail of the distribution is quite discrete and presents significant probability of zero defaults. As a result, the effective confidence level jumps several times and assumes only a small finite number of values in the lower tail. When Y = 1 (and I = 1), the first non-zero effective level is already approximately 15 percent; after that, the next value is approximately 30 percent. Therefore, validation at the 5 percent or 10 percent significance level is not possible. Hence, the Basel II prescription of a minimum of five years of data is important not only to increase the asymptotic power of the test, according to section 4, but also to remove the quite problematic finite-sample behavior of the lower tail.

7. Conclusion

This study contributes to the CRM validation literature by introducing new ways to statistically address the validation of credit rating PDs. Firstly, it proposes new formulations for H_0 and H_1 in order to control the error of accepting an incorrect CRM. Secondly, it provides an integrated analytical treatment of all ratings at once, in a way that recognizes the effect of default correlation. Finally, it provides a unified framework for testing calibration and rating discriminatory power. All these aspects are interlinked with the development of a probabilistic asymptotic normal model for the average default rate vector that recognizes default correlation. Important empirical and practical consequences stem from these proposals, as outlined in the following paragraphs.

On calibration testing, the relative roles played by the distinct elements that affect the power are unveiled for the one-sided version. The feature of increasing PD differences between consecutive ratings, found in many real-world CRMs, and, particularly, the choice of liberal indifference regions are shown to be important to the achievement of reasonable levels of power. On the other hand, the correlation between the ratings, whose calibration is not present in Basel II, possesses only a minor effect on power. Also, appropriately restricting the set of PDs to be tested may do a job almost as good as the original test in terms of power, which may offer support, in many practical circumstances, to reduce joint validation of credit rating PDs to individual validation of a few rating PDs. Another important general message of the analysis is that the power of the one-sided calibration test is unavoidably and substantially low in some cases. Regarding this issue, strategies of power improvement are discussed, suggesting limited efficacy or inappropriateness. Additionally, the paper discusses the conceptual problems of applying modern ideas in multivariate equivalence to two-sided calibration testing.

As far as discrimination is concerned, a new goal of rating discriminatory power is established for CRMs. In contrast to traditional measures of discrimination, the new aimed property is less stringent than the requirement of perfect calibration and is not dependent on default correlation. Results of uniform power dominance provide a theoretical foundation for restricting the investigation of the desired property just to the pairs of consecutive ratings whose discriminatory capacity are at stake and, therefore, lead to an easy-to-implement procedure.

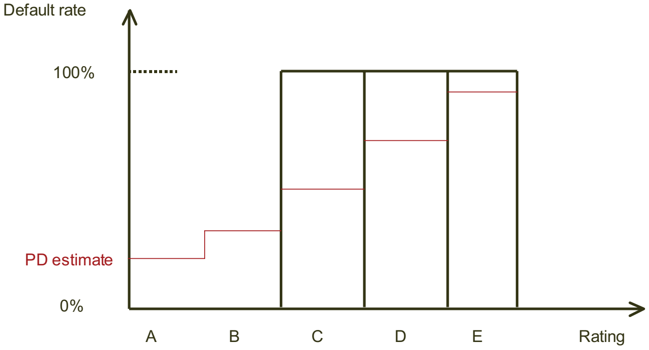
Understanding the implications of DRAPM to validation also includes an analysis of its properties when dealing with a finite-sample of borrowers. As a matter of fact, DRAPM has the disadvantage of being an asymptotic model whose finite-sample properties may introduce a significant additional layer of test conservatism besides the asymptotic one. Monte Carlo simulations show that this will likely be the case for small PDs (e.g., PD $\leq 3\%$) or a small number of years (e.g., $\mathrm{Y} < 5$) in the one-sided calibration test. A possible recommendation in the former case may be to apply the test just to the remaining ratings or to investigate the low postulated PDs at a higher nominal confidence level. On the other hand, when a liberal finite-sample bias is present, it may counterbalance the nominal conservatism, although some caution should always be exercised in the analysis. A general more robust procedure, however, would ideally try to incorporate the remaining non-systemic part of the credit risk into the validation process. Future research is warranted on this aspect.

Above all, the bank or the regulator should not demand much from statistical testing of CRMs. Even under the simplifying assumptions of DRAPM, the power of the tests of this paper, as well as other tests discussed in the literature, is negatively affected by the unavoidable presence of default correlation and by the small length of default rate time series available in banks' databases. Possibly due to this reason, BCBS (2005b) perceives validation as comprising not only quantitative but also somewhat qualitative tools. It is likely, for example, that the investigation of the continuous internal use of **PDs**/ratings by the bank may uncover further evidence, although subjective, supporting or not supporting the CRM validation. Nonetheless, this paper supports the view that the possibility of reliance on qualitative aspects opened by the Basel Committee should not dampen the incentives to extract as much quantitative feedback as possible from statistical testing, including a quantitative sense of its limitations.

Appendix 1

Figure 6 should be interpreted as a result over the long run and displays a rating model with perfect discrimination but not perfect calibration. The bars' heights represent the magnitude of the ex post default rate for each rating. All borrowers classified as C to E defaulted, whereas all borrowers classified as A to B survived. If this is the regular behavior of this CRM, knowing beforehand the rating of the obligor allows one to predict default or no default with certainty (perfect discriminatory power). The thin stepped line indicates the ex ante PD estimate for each rating. Ratings A and B had a 0 percent default rate, thus lower than the ex ante prediction. Ratings C to E had a 100 percent default rate, thus higher than the ex ante prediction. The CRM is therefore not correctly calibrated. Obviously, this example represents an extreme case (because realistic CRMs do not have perfect discriminatory power), but it is useful to illustrate that although both characteristics are desirable, they may well be inconsistent as they are aimed at their best.

Figure 6. Perfect Discrimination but Imperfect Calibration



Notes: The bold bar height of each rating represents the ex post long-run default rate of that rating, whereas the thin stepped line represents the ex ante PD estimates of the ratings.

Appendix 2

Proof of Proposition

The first parcel of the AUROC definition can be expressed as follows.

$$\begin{aligned}
&\mathrm{Prob}(\mathrm{PD_n} > \ \mathrm{PD_m}|\text{n defaults and m doesn't})\\
&= \frac{\mathrm{Prob}(\text{n defaults and m doesn't, } \mathrm{PD_n} > \ \mathrm{PD_m})}{\mathrm{Prob}(\text{n defaults and m doesn't})}\\
&= \frac{\sum_{i,j=1}^{I} \mathrm{Prob}(\text{n defaults and m doesn't, } \mathrm{PD_n} > \mathrm{PD_m}| \text{ n has rating i and m has rating j})\mathrm{P(i,j)}}{\sum_{i,j=1}^{I} \mathrm{Prob}(\text{n defaults and m doesn't}| \text{ n has rating i and m has rating j})\mathrm{P(i,j)}}\\
&= \frac{\sum_{i,j=1,\ i>j}^{I} \mathrm{Prob}(\text{n defaults and m doesn't } | \text{ n has rating i and m has rating j})\mathrm{P(i,j)}}{\sum_{i,j=1}^{I} \mathrm{Prob}(\text{n defaults and m doesn't}| \text{ n has rating i and m has rating j})\mathrm{P(i,j)}}\\
&= \frac{\sum_{i,j=1,\ i>j}^{I} \Phi_2\left(\Phi^{-1}\left(\mathrm{PD_i}\right), -\Phi^{-1}\left(\mathrm{PD_j}\right), -\rho_{\mathrm{ij}}\right)\ \mathrm{P(i,j)}}{\sum_{i,j=1}^{I}\ \Phi_2\left(\Phi^{-1}\left(\mathrm{PD_i}\right), -\Phi^{-1}\left(\mathrm{PD_j}\right), -\rho_{\mathrm{ij}}\right)\ \mathrm{P(i,j)}}.
\end{aligned}$$

where the last equality derives from the expression for a joint probability of default and non-default implicit in a DRAPM-style model (e.g., Gordy 2000). Similarly, the second parcel of the AUROC definition can be expressed as

$$1/2\ \mathrm{Prob}(\mathrm{PD}_{\mathrm{n}} = \mathrm{PD}_{\mathrm{m}} | \mathrm{n\ defaults\ and\ m\ doesn't}) \\ = \frac{\sum_{i=1}^{I} \Phi_2\left(\Phi^{-1}\left(\mathrm{PD_i}\right), -\Phi^{-1}\left(\mathrm{PD_i}\right), -\rho_{ii}\right)\mathrm{P(i)}}{2\sum_{i,j=1}^{I} \Phi_2\left(\Phi^{-1}\left(\mathrm{PD_i}\right), -\Phi^{-1}\left(\mathrm{PD_j}\right), -\rho_{\mathrm{ij}}\right)\mathrm{P(i,j)}}.$$

and the proposition is proved, observing the convention $PD_i \equiv \Phi^{-1}(\mathrm{PD_i})$.

References

Balthazar, L. 2004. “PD Estimates for Basel II.” *Risk* (April): 84–85.

Bamber, D. 1975. “The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph.” *Journal of Mathematical Psychology* 12 (4): 387–415.

Basel Committee on Banking Supervision. 2005a. “An Explanatory Note on the Basel II IRB Risk Weight Functions.” Bank for International Settlements.

———. 2005b. “Studies on the Validation of Internal Rating Systems.” Bank for International Settlements.

———. 2006a. “Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework — Comprehensive Version.” Bank for International Settlements.

———. 2006b. “The IRB Use Test: Background and Implementation.” Bank for International Settlements.

———. 2011. “Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems — Revised Version.” Bank for International Settlements.

Berger, R. L. 1989. “Uniformly More Powerful Tests for Hypotheses Concerning Linear Inequalities and Normal Means.” *Journal of the American Statistical Association* 84 (405): 192–99.

Berger, R. L., and J. C. Hsu. 1996. “Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets.” *Statistical Science* 11 (4): 283–319.

Black, F., and M. Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81 (3): 637–54.

Blochlinger, A. 2012. "Validation of Default Probabilities." *Journal of Financial and Quantitative Analysis* 47 (5): 1089–1123.

Blochlinger, A., and M. Leippold. 2006. "Economic Benefit of Powerful Credit Scoring." *Journal of Banking and Finance* 30 (3): 851–73.

Blochwitz, S., S. Hohl, D. Tasche, and C. When. 2004. "Validating Default Probabilities on Short Time Series." Working Paper.

Bluemke, O. 2013. "Probability of Default Validation: Introducing the Likelihood-Ratio Test and Power Considerations." *Journal of Risk Model Validation* 7 (2): 29–59.

Brown, L. D., G. Casella, and G. Hwang. 1995. "Optimal Confidence Sets, Bioequivalence and the Limaçon of Pascal." *Journal of the American Statistical Association* 90 (431): 880–89.

Brown, L. D., G. Hwang, and A. Munk. 1998. "An Unbiased Test for the Bioequivalence Problem." *Annals of Statistics* 25 (6): 2345–67.

Demey, P., J. F. Jouanin, C. Roget, and T. Roncalli. 2004. "Maximum Likelihood Estimate of Default Correlations." *Risk* (November): 104–8.

Engelmann, B., E. Hayden, and D. Tasche. 2003. "Testing Rating Accuracy." *Risk* (January).

Gordy, M. B. 2000. "A Comparative Anatomy of Credit Risk Models." *Journal of Banking and Finance* 24 (1–2): 119–49.

———. 2003. "A Risk-Factor Model Foundation for Ratings-Based Bank Capital Rules." *Journal of Financial Intermediation* 12 (3): 199–232.

Gourieroux, C., and A. Monfort. 1995. *Statistics and Econometric Models*, Vol. 2 (Themes in Modern Econometrics). Cambridge University Press.

Hochberg, Y., and A. C. Tamhane. 1987. *Multiple Comparison Procedures.* John Wiley & Sons, Inc.

Laska, E. M., and M. J. Meisner. 1989. "Testing Whether an Identified Treatment Is Best." *Biometrics* 45 (4): 1139–51.

Leland, H. E. 2004. "Predictions of Default Probabilities in Structural Models of Debt." *Journal of Investment Management* 2 (2): 5–20.

Liu, H., and R. L. Berger. 1995. “Uniformly More Powerful, One-Sided Tests for Hypotheses about Linear Inequalities.” *Annals of Statistics* 23 (1): 55–72.

Martín, M., and B. Salvador. 1988. “Validity of the ‘Pool-Adjacent-Violator’ Algorithm.” *Statistics and Probability Letters* 6 (3): 143–45.

McDermott, M. P., and Y. Wang. 2002. “Construction of Uniformly More Powerful Tests for Hypotheses about Linear Inequalities.” *Journal of Statistical Planning and Inference* 107 (1–2): 207–17.

Menéndez, J. A., C. Rueda, and B. Salvador. 1992a. “Dominance of Likelihood Ratio Tests under Cone Constraints.” *Annals of Statistics* 20 (4): 2087–99.

———. 1992b. “Testing Non-Oblique Hypotheses.” *Communications in Statistics — Theory and Method* 21 (2): 471–84.

Menéndez, J. A., and B. Salvador. 1991. “Anomalies of the Likelihood Ratio Test for Testing Restricted Hypotheses.” *Annals of Statistics* 19 (2): 889–98.

Merton, R. C. 1974. “On the Pricing of Corporate Debt: The Risk Structure of Interest Rates.” *Journal of Finance* 29 (2): 449–70.

Miu, P., and B. Ozdemir. 2008. “Estimating and Validating Long-Run Probability of Default with Respect to Basel II Requirements.” *Journal of Risk Model Validation* 2 (2): 3–41.

Munk, A., and R. Pfluger. 1999. “1-α Equivariant Confidence Rules for Convex Alternatives are α/2-Level Tests — With Applications to the Multivariate Assessment of Bioequivalence.” *Journal of the American Statistical Association* 94 (448): 1311–19.

Perlman, M. D., and L. Wu. 1999. “The Emperor’s New Tests.” *Statistical Science* 14 (4): 355–69.

Pluto, K., and D. Tasche. 2005. “Thinking Positively.” *Risk* (August): 72–78.

Robertson, T., F. T. Wright, and R. L. Dykstra. 1988. *Order Restricted Statistical Inference.* John Wiley & Sons.

Sasabuchi, S. 1980. “A Test of a Multivariate Normal Mean with Composite Hypotheses Determined by Linear Inequalities.” *Biometrika* 67 (2): 429–39.

Shapiro, A. 1988. “Towards a Unified Theory of Inequality Constrained Testing in Multivariate Analysis.” *International Statistical Review* 56 (1): 49–62.

Vasicek, O. 2002. “Loan Portfolio Value.” *Risk* (December).

Wang, W., J. T. G. Hwang, and A. Dasgupta. 1999. "Statistical Tests for Multivariate Bioequivalence." *Biometrika* 86 (2): 395–402.

Warrack, G., and T. Robertson. 1984. "A Likelihood Ratio Test Regarding Two Nested but Oblique Order-Restricted Hypotheses." *Journal of the American Statistical Association* 79 (388): 881–86.

Watt, M. 2013. "Mending the RWA Machine." *Risk* 26 (1, January): 17–20.