

Joint validation of credit rating PDs under default correlation *

Ricardo Schechtman

Research Department, Central Bank of Brazil**

July 2007

Abstract

The Basel Committee on Banking Supervision recognizes that one of the greatest technical challenges to the implementation of the new Basel II Accord lies on the validation of the banks' internal credit rating models (CRMs). This study investigates new proposals of statistical tests for validating the PDs (probabilities of default) of CRMs. It distinguishes between proposals aimed at checking calibration and those focused at discriminatory power. The proposed tests recognize the existence of default correlation, deal jointly with the default behaviour of all the ratings and, differently to previous literature, control the error of validating incorrect CRMs. Power sensitivity analysis and strategies for power improvement are discussed, providing insights on the trade-offs and limitations pertained to the calibration tests. An alternative goal is proposed for the tests of discriminatory power and results of power dominance are shown for them with direct practical consequences. Finally, as the proposed tests are asymptotic, Monte-Carlo simulations investigate the small sample bias for varying scenarios of parameters.

* The author would like to thank Axel Munk, Dirk Tasche, Getulio Borges da Silveira and Kostas Tsatsaronis for helpful conversations along the project. The author also thanks the Bank for International Settlements for its hospitality during his fellowship there. The views expressed herein are those of the author and do not necessarily reflect those of the Central Bank of Brazil, the Bank for International Settlements, or their members. Comments and suggestions are welcome.

** ricardo.schechtman@bcb.gov.br 55-21-21895384.

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 S&P's and banks' internal credit rating models.

CRMs have had their relevance increased recently as the new Basel II accord (BCBS(2004)) allows the PDs of the internal ratings to function as inputs in the computation of banks' regulatory levels of capital¹. Its goal is 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.² However, the great challenge for Basel II, in terms of implementation, lies on the validation of CRMs, particularly on the validation of bank estimated rating PDs³.

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*.⁴ This means, for instance, that the bank/supervisor will, in most practical situations, have to judge the CRM based solely on 5 to 10 (independent) observations available at the database⁵. 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.

In light of that picture, BCBS(2005b) perceives validation of CRMs as necessarily comprising a whole set of quantitative and qualitative tools rather than a single instrument. This study focuses solely, however, on a particular set of quantitative tools, namely the statistical tests. Having in mind the aforementioned unavoidable difficulties, this paper scientifically examines the validation of CRMs by means of general statistical tests, not dependent on the particular technique used in their development⁶. Furthermore, framework to be developed does not aim at a final

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

² On top of that, the transparency requirements contained in Basel II can also be seen as an important element aimed at enhancing financial stability.

³ According to BCBS (2005b) validation is above all a bank task, whereas the supervisor's role should be to certificate this validation.

⁴ Notice that this problem is not present in the validation of market risk, where the time horizon is typically in the order of days.

⁵ For statistical standards a small sample.

⁶ This allows the discussions of this paper to assume a general nature.

prescription but at discussing the trade-offs, strategies and limitations involved in the validation task from a statistical perspective.

Even restricting to general statistical tests, the judgment of the performance of a CRM is a multifaceted issue. It involves mainly 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(2004) is explicit about the demand for banks' internal models to possess good calibration, testing calibration is the starting point of this paper.⁷ According to BCBS(2005b), quantitative techniques for testing calibration are still on 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.* (2003)). These techniques have all the disadvantage of being univariate (i.e. designed to test a single rating PD per time) or to make the unrealistic assumption of cross-sectional default independency⁸. Further, they do not control for the error of accepting a miscalibrated CRM⁹. This paper presents an asymptotic framework to jointly test several PDs under the assumption of default correlation and controlling the previous error. The approach is close in spirit to Balthazar (2004), although here the testing problem formulation is remarkably distinct.

Good discriminatory power is also a desirable property of CRMs as it allows rating based yes/no decisions (e.g. credit granting) to be taken 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 ROC curve (Engelmann *et. al.* (2003)), the Accuracy Ratio and the Kolmogorov-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 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 the appendix A for a graphical example. Second, although not remarked in the literature, usual measures of discriminatory power are function of the cross-sectional dependency

⁷According to BCBS (2004), PDs should resemble long-run average default rates for all ratings.

⁸ Most of them suffer from both problems.

⁹ They control for the error of rejecting correct CRMs.

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 CRMs¹⁰. The framework of this paper leads to theoretical tests of “discrimination power” that 1) can be seen as a necessary requisite to perfect calibration and 2) are not a function of the default dependency structure.

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 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 small 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 an extension of the Basel II underlying model of capital requirement. In fact, this paper generalizes the idea of Balthazar(2004), of using the Basel II model for validation, to a multi-rating setting^{11,12}. The applied extension is based on Demey *et. al.* (2004)¹³ 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 (c.f. Gordy (2003)), for validation purposes a richer structure is necessary to allow for non-singular variance matrix among the ratings, as it becomes clearer ahead in the 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 Basel II model, z_{in} is expressed as:

$$z_{in} = \rho_B^{1/2} x + (\rho_W - \rho_B)^{1/2} x_i + (1 - \rho_W)^{1/2} \varepsilon_{in} \text{ for each rating } i=1 \dots I \text{ and each borrower } n=1 \dots N.$$

¹⁰ It is not solely a portfolio characteristic because default correlation among the ratings potentially depends on the design of the CRM too.

¹¹ This paper’s approach also differs from Balthazar(2004) in reversing the role of the hypothesis, as section 3 explains.

¹² The reader is referred to BCBS(2005a) for a detailed presentation of the Basel II underlying model.

¹³ 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.

where x, x_i, ε_{ij} ($i=1\dots I, j=1\dots N$) are independent and standard normal distributed. Here, 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 $\text{Cov}(z_{in}, z_{jm})$ 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 model description continues with the statement that a borrower j with rating i defaults at the end of the forecasting time horizon if $z_{in} < \Phi^{-1}(\text{PD}_i)$ at that time, where Φ denotes the standard normal cumulative distribution function. Note that the probability of this event is therefore, by construction, PD_i ¹⁴. Consequently, the conditional probability of default $\text{PD}_i(\mathbf{x})$, where $\mathbf{x}=(x, x_1, \dots, x_I)'$ denotes the vector of systemic factors, can be expressed by:

$$\text{PD}_i(\mathbf{x}) \equiv \text{Prob}(z_{in} < \Phi^{-1}(\text{PD}_i) | \mathbf{x}) = \Phi\left(\frac{\Phi^{-1}(\text{PD}_i) - \rho_B^{1/2} x - (\rho_W - \rho_B)^{1/2} x_i}{(1 - \rho_W)^{1/2}}\right).$$

Let's focus now on the asymptotic behaviour of the observable default rates. Let 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:

$$\text{DR}_{iN} - E(\text{DR}_{iN} | \mathbf{x}) \equiv \text{DR}_{iN} - \text{PD}_i(\mathbf{x}) \rightarrow 0 \text{ a.s. when } N \rightarrow \infty$$
¹⁵

Therefore, as Φ^{-1} is continuous, it is also true that

$$\Phi^{-1}(\text{DR}_{iN}) - \Phi^{-1}(\text{PD}_i(\mathbf{x})) \rightarrow 0 \text{ a.s. when } N \rightarrow \infty$$

so that in DRAPM the Φ^{-1} transformed default rates have asymptotically the same distribution as the Φ^{-1} transformed conditional probabilities, which are normal distributed^{16,17}.

More concretely, the limiting default rate joint distribution is:

$$\Phi^{-1}(\mathbf{DR}) \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\mathbf{DR} = (\text{DR}_1, \text{DR}_2, \dots, \text{DR}_I)'$, $\mu_i = \Phi^{-1}(\text{PD}_i)/(1 - \rho_W)^{1/2}$, $\Sigma_{ij} = \rho_W/(1 - \rho_W)$ if $i=j$ and $\Sigma_{ij} = \rho_B/(1 - \rho_W)$ otherwise.

This 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

¹⁴ Without generalization loss, PD_i is assumed to increase in i .

¹⁵ a.s. stands for almost sure convergence.

¹⁶ See the expression for $\text{PD}(\mathbf{x})$.

¹⁷ Although the choice of the normal distribution for the systemic factors may seem arbitrary in Basel II, for the testing purposes of this paper it is a pragmatic choice.

asymptotic, so that the discussions and results of this paper are not suitable for CRMs with a small number of borrowers per rating, such 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 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 Σ are warranted¹⁸. To the extent that borrowers of each rating present similar distributions of economic and geographic sectors of activity, that define the default dependency, ρ_B is likely to be very close to ρ_W , as this situation resembles the one factor case. By its turn, this paper assumes $0 < \rho_B < \rho_W$, in opposition to $\rho_B = \rho_W$, in order to leave open the possibility of some degree of association between PDs and borrowers' sectors of activity and with the technical purpose of obtaining a non-singular matrix Σ ^{19,20}. As a result, borrowers in the same rating behave more dependently than borrowers in different ratings, possibly because the profile of borrowers' sectors of activity is more homogeneous within than between ratings. Indeed, a more realistic modelling 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.

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 Basel II accord to extract some information on correlations²¹. By matching the variances of the non-idiosyncratic parts of the asset returns in Basel II and DRAPM models, ρ_W can be seen as the asset correlation parameter present in the Basel II formula²². For corporate borrowers, for example, Basel II accord chooses $\rho_W \in [0.12 \ 0.24]$ ²³. Sensitivity analysis of the power of the tests on the choices of these 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.

¹⁸ Note that the structure of Σ defines DRAPM more concretely than the chosen decomposition of the normalized asset return, because the decomposition is not unique given Σ .

¹⁹ To the best of the author's knowledge, the empirical literature lacks studies on that association.

²⁰ Even if the bank or the supervisor is convinced of the appropriateness of $\rho_B = \rho_W$, the approach of this paper is still defensible, provided, for instance, the default rates of different ratings are computed based on distinct sectors of activity.

²¹ An important distinction to the Basel II model, 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.

²² Note that Basel II can also be seen as the particular case of DRAPM when the coefficient of x_i is null, i.e. when $\rho_B = \rho_W$.

²³ On the other hand, Basel II accord doesn't provide information on ρ_B because it is based on a single systemic factor.

Finally, it is assumed serial independency 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 Σ/Y in place of Σ , where Y is the number of years available to *backtest*. According to BCBS(2005b), serial independency is less inadmissible than cross-sectional independency.

3. The formulation of the testing problem

Any configuration of a statistical test should start with the definitions of the null hypothesis H_0 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?²⁴ 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 BCBS (2005b) or in Balthazar (2004)²⁵. 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_0 . 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 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 the best of the author’s knowledge this paper is the first to feature the CRM validation problem in this way.

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 set²⁶. This implies that the statement that “the CRM is correctly specified” needs to be translated into some statement about the parameters PD_i s 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_i s 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_i s. 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 more the reality that not necessarily all the borrowers

²⁴ For this general discussion, one can think of “correctly specified” as meaning either correct calibration or good discriminatory power.

²⁵ Although they do not remark the consequences of this choice.

²⁶ H_0 and $H_0 \cup H_1$ need to be closed sets in order to guarantee that the maximum of the likelihood function is attained.

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.²⁷

4. Calibration testing

This section distinguishes between one-sided and two-sided tests for calibration. One-sided tests (which are only concerned about 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 excess of regulated capital is not desired by banks but also BCBS(2004) states that the PD estimates should ideally be consistent with the banks' managerial activities such as credit granting and credit pricing²⁸. To accomplish these goals, PD estimates must undistortly reflect the likelihood of default of every rating, something to be verified more effectively by two-sided tests (which are concerned about 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 the section. The analysis of one-sided calibration testing starts the section.

Based on the arguments of the previous section about the proper roles of H_0 and 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 H_1 .

H_0 : $PD_i \geq u_i$ for some $i=1 \dots I$

H_1 : $PD_i < u_i$ for every $i=1 \dots I$

where $PD_i \equiv \Phi^{-1}(PD_i)$, $u_i \equiv \Phi^{-1}(u_i)$. (This convention of representing Φ^{-1} transformed figures in italic is followed throughout the rest of the text)²⁹.

Here u_i is a fixed known number that defines an indifference acceptable region for PD_i . Its value should ideally be slightly larger than the value postulated for PD_i so that the latter is within the indifference region. Besides, u_i should preferably be

²⁷ 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 of information needed for validation.

²⁸ 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 for validation purposes.

²⁹ As Φ^{-1} is strictly increasing, statements about italic figures imply equivalent statements about non-italic figures.

smaller than the value postulated for PD_{i+1} so that at least the rejection of H_0 could conclude that $PD_i < \text{postulated } PD_{i+1}$.^{30,31}

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.³²

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

$$\overline{DR}_i \leq u_i / (1 - \rho_w)^{1/2} - z_\alpha (\rho_w / (Y(1 - \rho_w)))^{1/2} \text{ for every } i = 1 \dots I$$

where $\overline{DR}_i = \frac{\sum_{y=1}^Y DR_{iy}}{Y}$ is the (transformed) average annual default rate of rating i and $z_\alpha = \Phi(1-\alpha)$ is the $1-\alpha$ percentile of the standard normal distribution.³³

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 test³⁴. This means that the CRM is validated at size α if each 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 (1989)), which gives a solid theoretical foundation for the procedure since monotonicity is generally a desired property.³⁵ 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 *et al.* (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 the CRM does not depend on the parameter ρ_B , which is useful for applications since

³⁰ As banks have the capital incentive to postulate lower PDs one could argue that $PD_i < \text{postulated } PD_{i+1}$ also leads to $PD_i < \text{true } PD_{i+1}$.

³¹ Specific configurations of u_i are discussed later in the section.

³² Size of a test is the maximum probability of rejecting H_0 when it is true.

³³ This definition of \overline{DR}_i is used throughout the paper.

³⁴ 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.

³⁵ 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.

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 expression below.

$$\text{Power} = \Phi_I(-z_\alpha + (u_I - PD_I)/(\rho_W/Y)^{1/2}, \dots, -z_\alpha + (u_i - PD_i)/(\rho_W/Y)^{1/2}, \dots, -z_\alpha + (u_I - PD_I)/(\rho_W/Y)^{1/2}, \rho_B/\rho_W)$$

where $\Phi_I(\dots, \rho_B/\rho_W)$ is the cumulative distribution function of a I^{th} -variate normal of mean 0, variances equal to 1 and covariances equal to ρ_B/ρ_W .

Berger (1989) remarks that if the ratio ρ_B/ρ_W is small then the power of this test can be quite low for the PD_i s only slightly smaller than u_i s and/or a large number of ratings I . This is intuitive as a low ratio ρ_B/ρ_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 ρ_B/ρ_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 PD_i s decrease, u_i s increase, Y increases, I decreases, ρ_B increases or ρ_W decreases³⁶. In fact, it is worth examining the trade-off between the configuration of the indifference region in the form of the u_i s and the attained power. If high precision is demanded (u_i s close to postulated PD_i s) then power must be sacrificed; if high power is demanded then precision must be sacrificed (u_i s far from postulated PD_i s). 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:

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

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:

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

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

³⁶ Obviously the power also increases when the level α increases.

Table 1: u_i X 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

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 PD_i ^{37,38}. 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 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 favour of that design choice. Notice that this feature of increasing PD differences is directly related to the non-linearity 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 $\mathbf{u}=(u_1, u_2, \dots, u_I)'$ 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_i s depicted in table 2. CRMs of table 2 can have PD_i s 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_i s figures of table 2 are chosen with the purpose that the resulting sets of ratings of each CRM cover equal ranges in the PD scale. More specifically, this goal is interpreted here as all CRMs having equal u_0 and u_I ^{39,40}.

Table 2: PD_i s (%) chosen according to u_i specification and CRM design

	PD_i s follow arithmetic progression		PD_i s follow geometric progression	
	$u_i = PD_{i+1}$	$u_i = (PD_{i+1} + PD_i)/2$	$u_i = PD_{i+1}$	$u_i = (PD_{i+1} + PD_i)/2$
$I=3$	1.22, 11.82, 22.42	6.52, 17.17, 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

³⁷ Approximate because the computation was based on $I=1$. In fact the true attained power in a multi rating setup is smaller.

³⁸ The discussion of this paragraph assumes true \mathbf{PD} = postulated \mathbf{PD} .

³⁹ u_0 corresponds to the fictitious PD_0 . At table 2, PD_0 can be easily figured out from the constructional logic of the PD_i progression.

⁴⁰ 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 to 2 for $I=4$.

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 favourable one characterized by 10 years of data and a low within-rating correlation of 0.12, a unfavourable one characterized by the minimum number of 5 years prescribed by Basel II (c.f. Basel (2004)) and a high ρ_w at 0.18 and an in-between scenario⁴¹.

Table 3: Power comparison among CRM designs and u_i choices, I=3

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

	PD _i s follow arithmetic progression		PD _i s follow geometric progression	
	$u_i = PD_{i+1}$	$u_i = (PD_{i+1} + PD_i)/2$	$u_i = PD_{i+1}$	$u_i = (PD_{i+1} + PD_i)/2$
$\rho_w = 0.12, Y=10$	0.97	0.57	0.99	0.95
In-between	0.85	0.42	0.97	0.81
$\rho_w = 0.18, Y=5$	0.72	0.33	0.91	0.67

Table 4: Power comparison among CRM designs and u_i choices, I=4

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

	PD _i s follow arithmetic progression		PD _i s 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

Table 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 solely focusing on the former, more demanding requirements for u_i (c.f. $u_i = (PD_{i+1} + PD_i)/2$) may produce overly conservative tests, with for example power on the level of only 37%. Therefore liberal strategies for u_i (c.f. $u_i = PD_{i+1}$) seem to be necessary for realistic validation attempts and attention is focused on these strategies to 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

⁴¹ 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 as characterized by adjusting both Y and ρ_w or just one of them. At tables 3 and 4 it is given by $(\rho_w/Y)^{1/2} = 0.15$.

increase more than 80% from the worst to the best scenario (c.f. 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.

Table 5: Effect of ρ_B when **PDs 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

Table 6: Effect of ρ_B when **PDs 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

Tables 7 and 8 below 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 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) **PDs** drive the power of the test. This is partly intuitive because the highest (lowest) **PDs**

correspond to the smallest differences $(u_i - PD_i)$ in the CRMs of table 7 (table 8) and because distinct PD_i s contribute to the power differently just to the degree their differences $(u_i - PD_i)$ vary⁴². The surprising part of the result refers to the degree of relative low importance of the dropped PD_i s: the variation of power between $I=1$ and $I=4$ can be merely around 10%. This 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 .

Table 7: Influence of distinct PD_i s on power

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

PD_i s	$\alpha=5\%$	$\alpha=10\%$	$\alpha=15\%$
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

Table 8: Influence of distinct PD_i s on power

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

PD_i s	$\alpha=5\%$	$\alpha=10\%$	$\alpha=15\%$
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

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% at the postulated PD). Another related problem refers to the test not being similar on the boundary between the hypotheses and therefore biased (if $I>1$)⁴³. 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 Dermott and Wang (2002). The new

⁴² It is easy to see that for the CRMs with equally spaced PD_i s, $(u_i - PD_i)$ 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.

⁴³ 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 & Monfort (1995))

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 them⁴⁴. 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 that 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 one of the PD_i s go to 0 while the remaining one is set fixed at u_i ^{45,46}. 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 PD_{i-1} , not 0, represents a lower bound for PD_i , for every rating i . A natural modification of the test consists then on replacing z_α by a smaller constant $c > 0$ to adjust to the removed unrealistic **PD** scenarios⁴⁷, with resulting enlargement of the critical region and achievement of a more powerful test⁴⁸. Hence, c is defined by the requirement that the size of the modified test (with c instead of z_α) 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 PD_i s equal to their lower bounds (the postulated PD_{i-1} s), and the remaining one, say PD_j , set at u_j , for j varying in $1 \dots I$. More formally,

$$\text{Max}_{1 \leq j \leq I} (\Phi_1(-c + (u_j - PD_{j-1}) / (\rho_W / Y))^{1/2}, \dots, -c, \dots, -c + (u_I - PD_{I-1}) / (\rho_W / Y))^{1/2}; \rho_B / \rho_W) = \alpha^{49,50},$$

from which the value of c can be derived.

⁴⁴ More specifically, the power is higher at every **PD** parameter in H_0 .

⁴⁵ This limiting **PD** vector is in H_0 and, therefore, should not be validated. It has a probability of validation equal to α .

⁴⁶ Note $PD_i \rightarrow 0 \Rightarrow PD_i \rightarrow -\infty$

⁴⁷ As the coordinates of the input to the power function cannot go to infinity as before, $-c > -z_\alpha$ for the size to be achieved.

⁴⁸ See the definition of the critical region in the beginning of the section.

⁴⁹ PD_0 is here just a lower bound to PD_1 . It could be $-\infty$ or defined subjectively based on accumulated practical experience.

⁵⁰ Note that the new critical region will now depend on ρ_B and that the calculation of c needs some computational effort.

However, produced results indicate the previous modification approach is of limited efficacy to power improvement⁵¹. On the other hand, one may also try to derive the LRT based on the restricted **PD** parameter space:

H_0 : $PD_i \geq u_i$ for some $i=1 \dots I$ and $PD_i \geq$ postulated PD_{i-1} for every $I=1 \dots I$ ⁵²

H_1 : $PD_i < u_i$ for every $i=1 \dots I$ and $PD_i \geq$ postulated PD_{i-1} for every $I=1 \dots I$ ⁵³

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 favour of H_1 , because, by assumption, they cannot be explained by very small PDs⁵⁴. 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, H_0 and H_1 do not share the same boundary in $|R^1$ 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.

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.

H_0 : $PD_i \geq u_i$ or $PD_i \leq l_i$ for some $i=1.. I$

H_1 : $l_i < PD_i < u_i$ for every $i=1 \dots I$

Now the acceptable indifference region is defined by two parameters u_i and l_i for each rating i , with ideally $l_i \geq$ postulated PD_{i-1} and $u_i \leq$ postulated 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 to demonstrate that drugs are equivalent.⁵⁵ Berger and Hsu (1996) comprehensively review the recent development of equivalence tests in

⁵¹ Produced results indicate that the power increase is relevant only in the region of small (probably unrealistic) ratio ρ_B/ρ_W or for ambitious choices of u_i (i.e. close to 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 (c.f. table 1 for example). Those results are consistent with the intuition derived from the analysis of tables 7 and 8.

⁵² Same observation about PD_0 applies here as well.

⁵³ H_1 need not be defined only by strict inequalities here since the union $H_0 \cup H_1$ does not span the full $|R^1$.

⁵⁴ Very small observed average default rates in the sense that $\Phi^{-1}(DR_i)/(1 - \rho_W)^{1/2} < \Phi^{-1}(\text{postulated } PD_{i-1})$.

⁵⁵ More specifically, those tests are referred as bioequivalent tests in the pharmaceutical industry.

the univariate case ($I=1$). The standard procedure to test univariate equivalence is the TOST test (two one-sided tests - called this way because the procedure is equivalent to performing two size- α one sided tests and concluding equivalence only if both reject). Wang *et.al.* (1999) discuss the extension of TOST to the multivariate case, making use of the intersection-union method⁵⁶. When applied to the DRAPM distribution, that extension results in the following critical region for the two-sided calibration test⁵⁷.

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

$$l_i / (1 - \rho_w)^{1/2} + z_\alpha (\rho_w / (Y(1 - \rho_w)))^{1/2} \leq \overline{DR}_i \leq u_i / (1 - \rho_w)^{1/2} - z_\alpha (\rho_w / (Y(1 - \rho_w)))^{1/2}$$

for every $i = 1 \dots I$

As the maximum power of the test occurs in the middle point of the cube $[l_i \ u_i]^I$, 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).

Similarly to the one-sided test, the two-sided version has similar problems of lack of power and bias⁵⁸. In this respect, the statistical literature contains some proposals for improving TOST (Berger and Hsu(1996), Brown *et. al.*(1998)), which are again subject to criticism from an intuitive point of view by Perlman and Wu (1999)⁵⁹. 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.^{60,61} That observation indicates the magnified difficulty in performing two-sided calibration testing.

Two yet different approaches to testing multivariate equivalence deserve comments. The first one is developed by Brown *et. al.*(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

⁵⁶ Wang *et. al.* (1999) also show that TOST is basically a LR test.

⁵⁷ 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.

⁵⁸ If $I > 1$, the test is not similar on the boundary between the hypotheses and therefore biased.

⁵⁹ 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).

⁶⁰ It can be shown that the degree of conservatism depends on ρ_B .

⁶¹ 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).

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 to validate the calibration. Brown *et.al.* (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*⁶². 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 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.^{63,64}

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 quadrant of the space $|\mathbb{R}^I$. 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.

$$H_0: \mathbf{e}'\mathbf{D}\mathbf{e} \geq \Delta$$

$$H_1: \mathbf{e}'\mathbf{D}\mathbf{e} < \Delta$$

where $\mathbf{e} = \mathbf{PD}$ – postulated **PD**, **D** is a positive definite matrix, that conceives a notion of distance in $|\mathbb{R}^I$, 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_i s, $i=1\dots I$). 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_i s). 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_i s. 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

⁶² The form of the set is not an ellipse, commonly found in multivariate analysis, but rather a figure known as the Limaçon of Pascal.

⁶³ Note also that DRAPM should be seen just an approximation to reality, so that, even if all borrowers in a rating have exactly the same PD, small deviations from the DRAPM assumptions may in practice distort the optimality at the true **PD** point.

⁶⁴ Other confidence set approaches to calibration testing are also possible. Some of them are, however, dominated by the multivariate TOST (Munk and Pflunger (1999)).

different ratings. Consequently, for purposes of calibration testing, ellipsoidal alternatives are regarded here more as a practical complication.⁶⁵

5. Tests of rating discriminatory power

One of the most traditional measures of discriminatory power is the area under the ROC curve (AUROC)⁶⁶. Let n and m be two distinct random borrowers with probabilities of default PD_n and PD_m , respectively. Following Bamber(1975), AUROC is defined as:

$$AUROC = \text{Prob}(PD_n > PD_m \mid n \text{ defaults and } m \text{ doesn't}) + \frac{1}{2} \text{Prob}(PD_n = PD_m \mid n \text{ defaults and } m \text{ doesn't})$$

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 makes it a function not only of the **PD** vector but also of the default correlation structure⁶⁷. 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.⁶⁸ The next proposition shows explicitly 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 , $i, j = 1 \dots I$. 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:

$$AUROC = \frac{\sum_{i>j} \Phi_2(\Phi^{-1}(PD_i), -\Phi^{-1}(PD_j), -\rho_{ij})P(i, j) + \frac{1}{2} \sum_i \Phi_2(\Phi^{-1}(PD_i), -\Phi^{-1}(PD_i), -\rho_{ii})P(i)}{\sum_{i,j} \Phi_2(\Phi^{-1}(PD_i), -\Phi^{-1}(PD_j), -\rho_{ij})P(i, j)}$$

Proof: Appendix B.

⁶⁵ However, for purposes of power improvement, it might be still useful to investigate ellipsoidal alternatives inscribed or approximating rectangular alternatives. This investigation is not addressed at this paper.

⁶⁶ ROC = Receiver Operating Characteristic curve (c.f. Bamber (1975)). $0 \leq AUROC \leq 1$.

⁶⁷ It is a function of the distribution of borrowers along the ratings too.

⁶⁸ Note that, in the 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.

The remainder of this section describes alternative proposals of 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 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 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 H_1 of the two-sided calibration test⁶⁹. In this sense, the fulfilment 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 function 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 & 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 uncertain about the increasing PD behaviour along the whole rating scale whereas the latter focuses on a pair of consecutive ratings.

The formulation of the general test is proposed below.

H_0 : $PD_i \geq PD_{i+1}$ for some $i = 1 \dots I-1$

H_1 : $PD_i < PD_{i+1}$ for every $i = 1 \dots I-1$

By viewing $PD_{i+1} - PD_i$ 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 means⁷⁰. So, similarly to the one-sided calibration test, a size- α likelihood-ratio critical region can be derived.

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

$$\overline{DR}_{i+1} - \overline{DR}_i > z_\alpha (2(\rho_W - \rho_B) / (Y(1 - \rho_W)))^{1/2} \text{ for every } i = 1 \dots I-1$$

⁶⁹ Provided $u_i < l_{i+1}$ for $i = 1 \dots I-1$, as expected in practical applications.

⁷⁰ 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.

It is worth noting above that, opposed to the calibration tests, there is no need to the configuration of an indifference region, as the desired 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 ($\rho_B = \rho_W$) represents the extreme liberal situation where just an observed increasing behaviour 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 $\rho_B = 0$ places the strongest requirement in the incremental increase of the default rate averages along the rating scale⁷¹. 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 compared to the power of the one-sided calibration test, the power of the general test is notably affected by a trade-off of three factors⁷². First, the fact that now the underlying normal variables are likely to have smaller variances ($\text{Var}(DR_{i+1}-DR_i) = 2(\rho_W - \rho_B)/(1 - \rho_W) < \text{Var}(DR_i) = \rho_W/(1 - \rho_W)$, provided $\rho_B/\rho_W > 1/2$) contributes to an increase in power. On the other hand, the now not positive underlying correlations ($\text{Corr}(DR_{i+1}-DR_i, DR_j-DR_{j-1}) = -1/2$ if $i=j$ and 0 otherwise, compared to $\text{Corr}(DR_i, DR_j) = \rho_B/\rho_W > 0$ for $i \neq j$) contributes to a decrease in power⁷³. Finally, the presence of I-1 statements in H_1 , instead of I, implies a slight increase in power too. The resulting dominating force is to be determined by the particular choices of ρ_B , ρ_W and I. In general, the same 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.⁷⁴ To examine this issue, this section formulates a test of focal rating discriminatory power, whose hypotheses are stated as follows.⁷⁵

$$H_0: PD_1 = PD_2 \leq PD_3 \leq \dots \leq PD_I$$

$$H_1: PD_1 < PD_2 \leq PD_3 \leq \dots \leq PD_I$$

⁷¹ 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.

⁷² Similarly to the calibration case, the power expression can be easily derived.

⁷³ Therefore, not necessarily validating rating discriminatory power is easier than validating (one-sided) calibration.

⁷⁴ 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.

⁷⁵ 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.

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 $|R|^1$ neither the hypotheses share a common boundary. But, in contrast to the section 4 one-sided calibration LRT under **PD** restriction, now both H_0 and H_1 are convex cones. This implies that the null distribution of the LR will depend on the structure of the cone $C = H_0 \cup H_1$, whether obtuse or acute with respect to norm induced by Σ^{-1} .^{76,77} In the first case, the LR statistic follows a χ^2 *bar* distribution under H_0 (Menendez *et. al.* (1992a)).⁷⁸ 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 H_0 and H_1 (Menendez and Salvador (1991), Menendez *et. al.* (1992b)). It can be shown that the structure of Σ adopted in this paper makes the cone C acute, so that the second case is the relevant one.⁷⁹ The reduced dominating test takes the form below.

$$H_0: PD_1 = PD_2$$

$$H_1: PD_1 < PD_2$$

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.

Reject H_0 (i.e. validate the CRM)

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

The dominance of the focal test by a reduced test is a surprising result and was long considered an anomaly of the LR principle (see 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 favour of the “dominating” test, arguing that the latter presents controversial properties. For example, it rejects **PDs** *closer* to H_0 than to H_1 .⁸⁰

⁷⁶ See (reference) for the definitions of those cone types.

⁷⁷ $\|x\|_{\Sigma^{-1}} = x^T \Sigma^{-1} x$

⁷⁸ Although χ^2 *bar* distributions are common in the theory of order-restricted inference (Robertson *et. al.* (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 computer intensive.

⁷⁹ This is true because $\mathbf{a}_i' \Sigma \mathbf{a}_j \leq 0$, $i \neq j$, where the \mathbf{a}_i 's ($\mathbf{a}_i = (0, \dots, -1, 1, \dots, 0)'$) generate the linear restrictions defining the cone C . More specifically, it is true that $\mathbf{a}_i' \Sigma \mathbf{a}_j = (\rho_B - \rho_W) / (1 - \rho_W)$ if $|i-j| = 1$ or 0 if $|i-j| \geq 2$. See the mentioned references for further details. May more general but still realistic variance structures Σ lead to a different conclusion is an interesting question not addressed in this paper.

⁸⁰ Perlman and Wu (1999) conclude once again that UMP size- α tests should not be pursued at any cost.

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. Having that in mind, the analysis of this section provides the theoretical foundation to an easy-to-implement and only procedure available: restrict the attention to the 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. Small 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 analyses 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)⁸¹. Due 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 ones⁸². The focus on a particular marginal allows then, for the sake of clarity, to restrict the attention to the case $I=1$ ⁸³. Hence this section conducts Monte-Carlo simulations of DRAPM, at the stage in which idiosyncratic risk is not yet diversified away⁸⁴ and for $I=1$, $N=100$ and $Y=5$, unless stated otherwise.⁸⁵ Based on a large set of simulated average annual default rates, 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 ⁸⁶.

$$\text{Effective confidence level} = \hat{\Pr} \text{ ob } \left(\frac{\sqrt{1-\rho_w} \overline{DR} - PD}{\sqrt{\rho_w / Y}} < -z_\alpha \right)$$

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

⁸¹ 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 small sample properties of discriminatory tests are more complex to analyse as distributions of default rate *differences* are involved. Also, as perceived later in the section, the issues of most concern related to the small-sample properties of the two-sided calibration test derive from the analysis of the one-sided case.

⁸² Review the form of the critical region in section 4.

⁸³ The issue of how the normal copula is distorted by the reality of a finite number of borrowers is not addressed in this version of the paper.

⁸⁴ In other words, before $N \rightarrow \infty$.

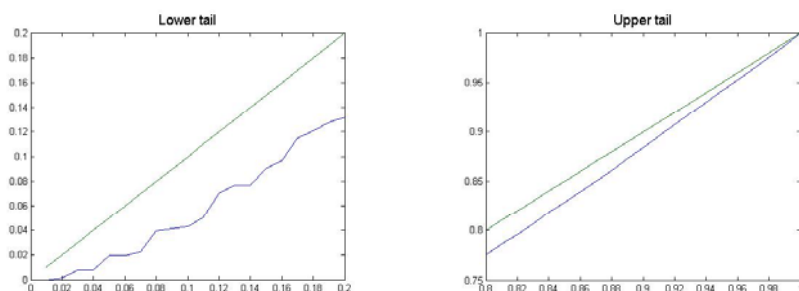
⁸⁵ Recently developed credit risk analytical methods to approximate distribution tails, such as the granularity adjustment, are not applicable here, as this paper deals with non-linear (Φ^1) transformed default rate distributions.

⁸⁶ In general 200000 simulations are run for each scenario.

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 $\delta < \alpha$ ⁸⁷. 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 small sample liberal *bias*.

A general important finding derived from the performed simulations is that the convergence of the lower tails of the (transformed) average 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 of⁸⁸. The situation is illustrated by the following pair of graphs calculated based on the scenario PD=3%, $\rho_w=0.20$, N=100 and Y=5. The blue line represents the effective confidence level for each nominal level depicted at the x-axes while the green 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 on the right-hand graph) than in the upper tail (depicted on the left-hand graph). In particular, if the one-sided calibration test is employed at the nominal level of 10%, the test will be much more conservative in reality, as the effective size will be approximately only 4%⁸⁹.

Graph 1: Lower and upper tails,
PD=3%, $\rho_w=0.20$ N=100, Y=5



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

⁸⁷ More specifically, it is easy to see that $\delta = \Phi(-z_\alpha - (u - PD)/(\rho_w/Y)^{1/2})$

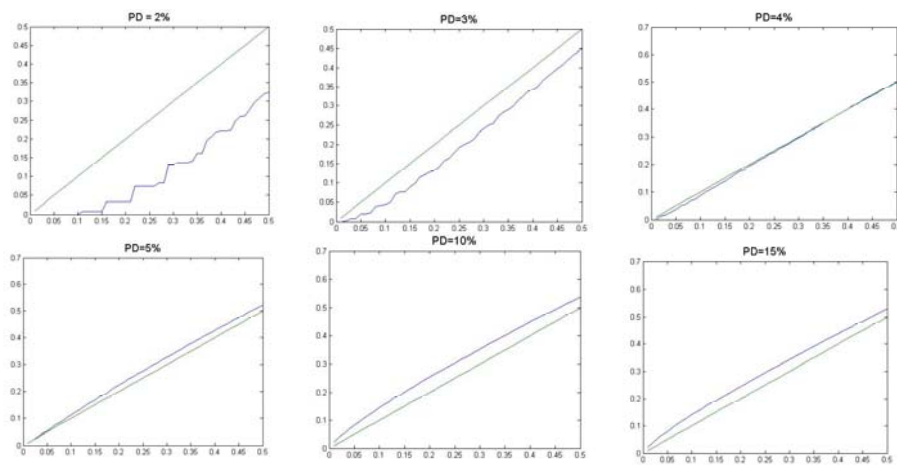
⁸⁸ The intuitive reason for this being that $\Phi^{-1}(PD) \rightarrow -\infty$ when $PD \rightarrow 0$.

⁸⁹ There is less mass in the simulated lower tail than in the respective tail of the DRAPM distribution.

relevant part of the distribution had the approach of placing the “CRM correctly specified” hypothesis in H_0 been adopted, as in BCBS(2005b). Since convergence of the upper tail is more well behaved, the small sample departure from 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 approach⁹⁰.

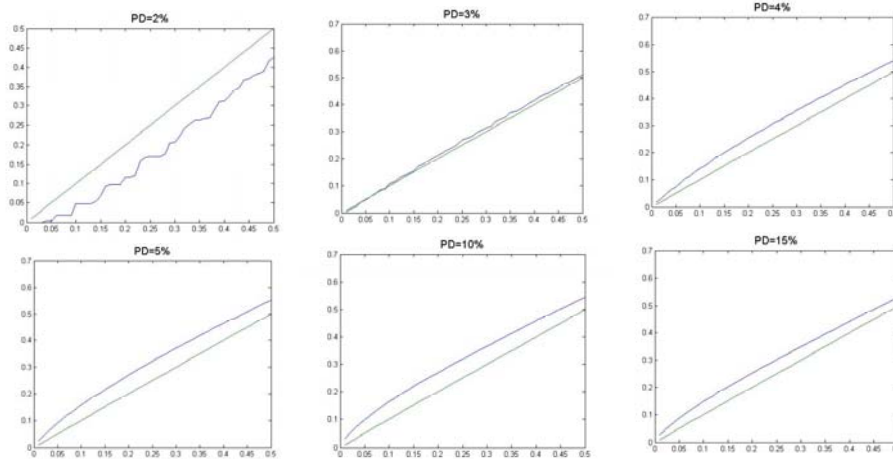
The main numerical findings regarding the small sample power performance of the one-sided calibration test are described in the sequence, based on the analysis of the simulated lower tails. The investigation starts with the effect of the true PD on the effective confidence level. Graphs 2 and 3 reveal 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 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 small 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 blue line comes back a bit closer to the green one, i.e. the test diminishes its liberal bias (but not sufficiently so as to turn conservative).

Graph 2: Effect of PD,
 $\rho_w=0.20$ $N=100$, $Y=5$



⁹⁰ Because the worse relative behaviour of the lower tail would not be revealed.

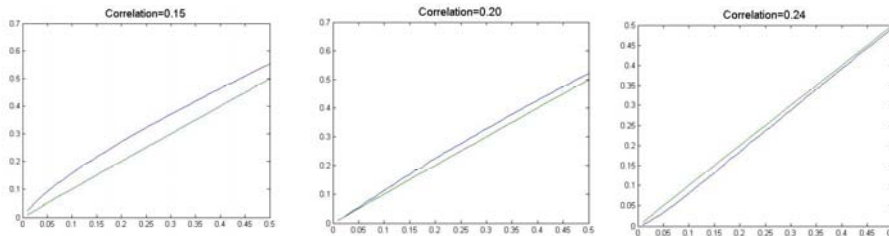
Graph 3: Effect of PD,
 $\rho_W=0.15, N=100, Y=5$



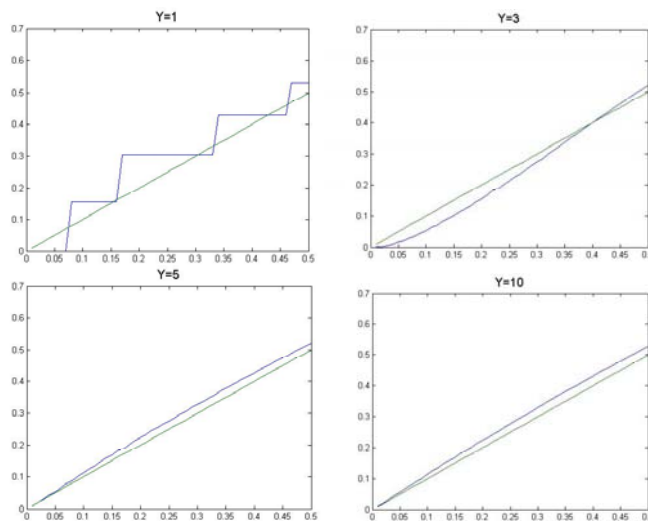
As the asymptotic one-sided test based on DRAPM already suffers from problems of lack of power, this section suggests, as possible general recommendation, to consider real (unmodified) applications of the test solely in the cases where the small sample analysis indicates a non-conservative bias. Indeed, if instead an additional layer of conservatism is added to the already conservative asymptotic test, the resulting procedure test may hardly validate at all. The restriction to the small sample liberal cases rules out, for example, according to graphs 2 and 3, validation of low PDs (e.g. $PD \leq 3\%$). Consequently, 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, a higher nominal level α could be applied to the low PDs.

The influence of correlation and the number of years under the base case of $N=100$ are analyzed in graphs 4 and 5. As the within-rating asset correlation ρ_W increases, the test evolves from a liberal bias to a small conservative one. Note that this represents a second channel, now through the small sample properties, by which ρ_W diminishes the power of the test. The effect of an increase in the number of years, in the region of 1 to 10 years, is to smooth considerably the distribution lower tail, although the *direction* of convergence is not clearly established. Results not shown also indicate that as N increases beyond 100, the blue and green lines come closer at every graph, as expected.

Graph 4: Effect of ρ_w
 PD=5%, Y=5, N=100



Graph 5: Effect of Y
 PD=5%, $\rho_w=0.20$ N=100



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 $Y=1$ (c.f. graph 6) or true PD=1%, for example, the lower tail of 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 $Y=1$ the first non-zero effective level is already approximately 15%; after that, the next value is approximately 30%. Therefore, validation at 5% or 10% significance level is not possible. Hence, Basel II prescription of a minimum of 5 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 small sample behaviour of the lower tail.

7. Conclusion

This study contributes to the CRM validation literature in 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 treatment of all ratings at once. 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 derive 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. A general message of the analysis is, however, 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.

The understanding of the implications of DRAPM to validation also includes an analysis of its small sample properties. As a matter of fact, DRAPM has the disadvantage of being an asymptotic model whose small 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 small number of years (e.g. $Y \leq 5$) in the one-sided calibration test. A possible recommendation is to rule out real (unmodified) applications of that test in those cases. On the other hand, when a liberal small sample bias is present, it may counterbalance the nominal conservatism, although some caution should always be exercised in the analysis.

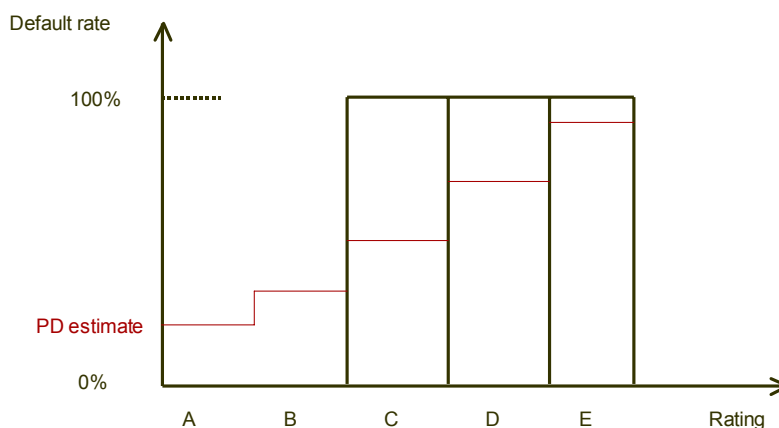
Above all, the bank or the regulator should not demand much from statistical testing of CRMs. Even under the simplifying assumptions of DRAPM, the power of the tests of this paper, as well as other tests discussed in the literature, is negatively affected by the unavoidable presence of default correlation 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 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.

8. Appendix

Appendix A

The figure below 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 behaviour of this CRM, knowing beforehand the rating of the obligor allows one to predict default or not default with certainty (perfect discriminatory power). The red line indicates the *ex-ante* PD estimate for each rating. Ratings A and B had 0% default rate, thus lower than the *ex-ante* prediction. Ratings C to E had 100% 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 don't 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.



Appendix B

Proof of proposition.

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

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

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

$$1/2 \text{ Prob}(PD_n = PD_m \mid n \text{ defaults and } m \text{ doesn't}) = \frac{\sum_{i=1}^I \Phi_2(\Phi^{-1}(PD_i), -\Phi^{-1}(PD_i), -\rho_{ii}) P(i)}{2 \sum_{i,j=1}^I \Phi_2(\Phi^{-1}(PD_i), -\Phi^{-1}(PD_j), -\rho_{ij}) P(i, j)}.$$

and the proposition is proved.

9. References

- Balthazar, L. (2004), "PD Estimates for Basel II", *Risk*, April 2004.
- Bamber, D. (1975), "The area above the ordinal dominance graph and the area below the receiver operating graph", *Journal of Mathematical Psychology*, 12, 387-415.
- Basel Committee on Banking Supervision (2004), "International Convergence of Capital Measurement and Capital Standards: A Revised Framework", *Bank for International Settlements*.
- Basel Committee on Banking Supervision (2005a), "An Explanatory Note on the Basel II IRB Risk Weight Functions", *Bank for International Settlements*.
- Basel Committee on Banking Supervision (2005b), "Studies on the Validation of Internal Rating Systems", *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*, Vol. 84, No. 405.
- Berger, R. L. and J. C. Hsu (1996), "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets", *Statistical Science*, Vol. 11, No. 4.
- Blochlinger, A. and M. Leippold (2006), "Economic Benefit of Powerful Credit Scoring", *Journal of Banking and Finance*, 30.
- Blochwitz, S., S. Hohl, D. Tasche and C. Wehn (2004), "Validating Default Probabilities on Short Time Series", *Working Paper*.
- Brown, L. D., G. Casella and G. Hwang (1995), "Optimal Confidence Sets, Bioequivalence and the Limaçon of Pascal", *Journal of the American Statistical Association*, Vol. 90 No. 431.
- Brown, L. D., G. Hwang and A. Munk (1998), "An Unbiased Test for the Bioequivalence Problem" *The Annals of Statistics*, Vol. 25.
- Demey P., J. F. Jouanin, C. Roget and T. Roncalli (2004), "Maximum Likelihood Estimate of Default Correlations", *Risk*, November 2004.
- Engelmann, B. E. Hayden and D. Tasche (2003), "Testing Rating Accuracy", *Risk*, January 2003.
- Gordy, M. B. (2000), "A Comparative Anatomy of Credit Risk Models", *Journal of Banking and Finance*, 24 (1-2), p.119-149.
- Gordy, M. B. (2003), "A Risk-Factor Model Foundation for Ratings-Based Bank Capital Rules", *Journal of Financial Intermediation*, Vol. 12, No. 3.
- Gourieroux, C. and A. Monfort (1995), "Statistics and Econometric Models", *Themes in Modern Econometrics*, Cambridge University Press.
- Laska, E. M. and M. J. Meisner (1989), "Testing Whether an Identified Treatment is Best", *Biometrics*, 45.
- Liu, H. and R. L. Berger (1995), "Uniformly More Powerful, One-Sided Tests for Hypotheses about Linear Inequalities", *The Annals of Statistics*, Vol. 23, No. 1.
- 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.

- Menéndez, J. A. and B. Salvador (1991), “Anomalies of the Likelihood Ratio Test for Testing Restricted Hypotheses”, *The Annals of Statistics*, Vol. 19, No. 2.
- Menéndez, J. A., C. Rueda and B. Salvador (1992a), “Testing Non-Oblique Hypotheses”, *Communications in Statistics - Theory and Methods*, 21(2).
- Menéndez, J. A., C. Rueda and B. Salvador (1992b), “Dominance of Likelihood Ratio Tests under Cone Constraints”, *The Annals of Statistics*, Vol. 20 No. 4.
- Munk, A. and R. Pfluger (1999), “ $1-\alpha$ Equivariant Confidence Rules for Convex Alternatives are $\alpha/2$ -level Tests – with Applications to the Multivariate Assessment of Bioequivalence”, *Journal of the American Statistical Association*, Vol. 94, No. 448.
- Perlman, M.D. and L. Wu (1999), “The Emperor’s New Test“, *Statistical Science*, Vol.14, No. 4.
- Pluto, K. and D. Tasche (2005), “Thinking Positively”, *Risk*, August 2005.
- 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.
- Shapiro, A. (1988), “Towards a Unified Theory of Inequality Constrained Testing in Multivariate Analysis”, *International Statistical Review*, 56, 1.
- Vasicek, O. (2002),”Loan Portfolio Value”, *Risk*, December 2002.
- Wang, W., J. T. G. Hwang and A. Dasgupta (1999), ”Statistical tests for multivariate bioequivalence”, *Biometrika*, 86, 2.
- Warrack, G. and T. Robertson (1984), “A Likelihood Ratio Test Regarding Two Nested but Oblique Order-Restricted Hypotheses”, *Journal of the American Statistical Association*, Vol. 79, No. 388.