

# **Joint validation of credit rating PDs under default correlation<sup>\*</sup>**

**Ricardo Schechtman**

**Research Department, Central Bank of Brazil<sup>\*\*</sup>**

**January 2007**

***Preliminary incomplete version  
Do not circulate without permission***

## **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. A greater focus is placed on validating calibration although validation of mappings between rating scales and of discriminatory power are also considered. The new tests recognize the existence of default correlation and, differently to previous literature, deal jointly with the default behaviour of all the ratings and control the error of validating incorrect models. 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.

---

<sup>\*</sup> 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 its members. Comments and suggestions are mostly welcome.

<sup>\*\*</sup> [ricardo.schechtman@bcb.gov.br](mailto:ricardo.schechtman@bcb.gov.br) 55-21-21895384.

## 1. Introduction

The purpose of this paper is to discuss validation issues for credit rating models (CRMs). In this paper, 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 credit agencies such as Moody's and S&P's and bank's internal credit rating models.

CRMs have had their relevance highly increased recently as the new Basel II accord (BCBS(2004)) allows the PDs of the internal ratings to enter as inputs for the computation of banks' regulatory levels of capital<sup>1</sup>. 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.<sup>2</sup> However, the great challenge for Basel II, in terms of implementation, lies on the validation of CRMs, in particular the validation of the bank estimated rating PDs<sup>3</sup>.

In fact, validation 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 a few observations available for back testing.<sup>4</sup> This means, for instance, that if yearly default rates are to be compared with *ex-ante* yearly PD estimates then the risk analyst will, in most practical situations, have to judge the model based solely on 5 to 10 observations<sup>5</sup>. Secondly, as borrowers are usually sensitive to a common set of factors in the economy (e.g. industry, region), variation of macro-conditions over the time horizon induces correlation among defaults. Default correlation, in turn, results in larger uncertainty in the process of estimating the true credit quality of borrowers. Both these factors contribute to decreasing the power of quantitative methods of validation.

In light of that picture, BCBS(2005b) perceives validation of credit rating models 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. To the extent that the aforementioned difficulties are unavoidable, because they reflect the real world in which credit risk assessment is undertaken and validated, this paper addresses which general statistical tests can be proposed to examine the issue of validation in a scientifically appropriate manner. The study is not aimed at a final formula but at discussing the many trade-offs, strategies and limitations involved in the validation task from a statistical perspective. Further, another important characteristic of this study's approach involves taking the rating models as "black boxes". In other words, the tests discussed here examine the appropriateness of the model forecast (i.e. whether *ex-post* default rates are close to *ex-ante* PD estimates) rather than the model fit (i.e. whether the underlying model behind the PD estimation has a good fit). This avoids making the statistical tests model dependent and allows the discussions of this paper to assume a general nature.

---

<sup>1</sup> The higher the PD, the higher is the regulatory capital.

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

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

<sup>4</sup> Notice that this problem is not present in validating market risk, where the time horizon is typically in the order of days.

<sup>5</sup> For statistical standards a small sample.

The performance of credit rating models can be generally judged by calibration, discriminatory power and mapping. 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. Finally, the performance of a CRM could also be assured through a mapping established with another CRM already recognized as correctly specified.

As Basel II is explicit about the demand for banks' internal models to possess good calibration, testing calibration is the main focus of this paper.<sup>6</sup> According to BCBS(2005b), 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) and most of them make the unrealistic assumption of cross-sectional independency. Further, they do not control for the error of accepting a miscalibrated model<sup>7</sup>. This paper presents a framework in which joint PD testing in a default correlated context is possible. The approach is close in spirit to Balthazar (2004), although here the testing problem is stated, since the beginning, in an important distinguished way.

Good discriminatory power is also a desirable property of rating models 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 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 rating models. 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 between borrowers. This fact potentially represents an undesired property of the 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<sup>8</sup>. 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.

Finally, this paper also briefly discusses tests for mappings between two different rating scales. Mappings are usually established between a bank and a rating agency scale and make each rating of the bank correspond to a rating of the agency. They are useful at the early stages of development of a bank's internal model when data on default rate time series is scarce so that the bank can hardly validate its PD assignments<sup>9</sup>. In this case, BCBS (2005b) views the validation task as comprised of two steps: the validation of the rating agency model and the validation of the mapping itself. As

---

<sup>6</sup> According to BCBS (2004), PDs should resemble long-run default rate averages for each rating.

<sup>7</sup> They control for the error of rejecting correct models.

<sup>8</sup> It is not solely a portfolio characteristic because default correlation among the ratings potentially depends on the design of the CRM too.

<sup>9</sup> Mappings can also arise naturally prompted by regulatory classification rules. For example, Brazilian regulation establishes a regulatory rating scale in which banks should classify their exposures for provisioning purposes. In this way a mapping between any two Brazilian banks is indirectly established.

literature is scarce on the latter, BCBS (2005b) stresses the need for developing mapping validation tools. The framework of this paper is shown to accommodate, from a theoretical point of view, mapping testing too.

This text is organized as follows. Section 2 develops a default rate asymptotic model (DRAM) upon which validation will be discussed. The model leads to a unified theoretical framework for checking calibration, mapping and discriminatory power. Section 3 discusses briefly the statement of the testing problem for CRM validation. The application of DRAM for calibration testing is discussed in section 4. Some theoretical aspects of the use of the model for mapping and discriminatory power testing are discussed in sections 5 and 6, respectively. Section 7 contains a Monte–Carlo analysis of the small sample properties of DRAM and their consequences for calibration testing. Section 8 concludes.

## 2. The default rate asymptotic model

The model of this section provides a default rate probability distribution upon which statistical testing will be conducted. 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<sup>10</sup>. The reader is referred to BCBS(2005a) for a detailed presentation of Basel II underlying model. The extension applied is based on the development of Demey *et. al.* (2004)<sup>11</sup> and refers to adding an additional systemic factor for each rating in order to allow joint PD testing. 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.

The DRAM 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, I express  $z_{in}$  as

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

where  $x$ ,  $x_i$ ,  $\varepsilon_{ij}$  are independent and jointly normal distributed with mean 0 and variance 1 for each  $i=1\dots I$  and  $j=1\dots N$  and  $\rho_B$  and  $\rho_W$  lie in the interval  $[0, 1]$ . Here,  $x$  represents the common systemic factor affecting the asset return of all borrowers,  $x_i$  the systemic factor affecting solely the asset return of borrowers with rating  $i$  and  $\varepsilon_{in}$  an idiosyncratic shock. Note that  $\text{Cov}(z_{in}, z_{jm})$  is equal to  $\rho_W$  if  $i=j$  and to  $\rho_B$  otherwise, so that  $\rho_W$  represents the “within-rating” asset correlation and  $\rho_B$  the “between-rating” asset correlation.

The model continues by stating that a borrower  $j$  with rating  $i$  is assumed to default in time  $T$  if  $z_{in} < \Phi^{-1}(\text{PD}_i)$  at that time, where  $\Phi$  denotes the standard normal cumulative distribution. Note that the probability of this event is, therefore, by construction  $\text{PD}_i$ <sup>12</sup>. As a consequence, 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:

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

<sup>10</sup> This paper’s approach also differs from Balthazar(2004) in reversing the role of the hypothesis, as section 3 explains.

<sup>11</sup> The purpose of Demey *et. al.* (2004) is however to estimate correlations while the focus here is on developing a minimal non-degenerate multivariate structure useful for testing.

<sup>12</sup> Without generalization loss,  $\text{PD}_i$  is assumed to increase in  $i$ .

Let's focus now on the asymptotic behaviour of the observable variable default rate. Let  $DR_{iN}$  denote the default rate computed based on a sample of  $N$  borrowers with rating  $i$  at the start of the period. It is easy to see, as in Gordy (2003), that

$$DR_{iN} - E(DR_{iN}|\mathbf{x}) \equiv DR_{iN} - PD_i(\mathbf{x}) \rightarrow 0 \text{ a.s. when } N \rightarrow \infty$$

Therefore, as  $\Phi^{-1}$  is continuous, it is also true that

$$\Phi^{-1}(DR_{iN}) - \Phi^{-1}(PD_i(\mathbf{x})) \rightarrow 0 \text{ a.s. when } N \rightarrow \infty$$

so that in DRAM the  $\Phi^{-1}$  transformed default rates have asymptotically the same distribution as the  $\Phi^{-1}$  transformed conditional probabilities, which are normal distributed<sup>13,14</sup>.

More concretely, the limiting joint default rate distribution is as follows:

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

where  $\mathbf{DR} = (DR_1, DR_2, \dots, DR_i)^T$ ,  $\mu_i = \Phi^{-1}(PD_i)/(1 - \rho_W)^{1/2}$ ,  $\Sigma_{ij} = \rho_W/(1 - \rho_W)$  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 convenient because it allows this paper's approach to build upon the literature of normal multivariate statistical testing. The cost is that the approach is asymptotic so that the discussions and results of this paper are not suitable for CRM 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, for some ranges of the parameters. Application of the tests of the next section should then be extremely careful.

Some comments on the choice of  $\boldsymbol{\Sigma}$  are warranted<sup>15</sup>. To the extent that borrowers of each rating present similar distributions of economic and geographical sectors of activity that define the default dependency,  $\rho_B$  is likely to be very close to  $\rho_W$ , 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 also technically to allow for a non-singular matrix<sup>16,17</sup>. 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 to be portfolio dependent; therefore the choice of just a pair of correlation parameters is regarded here as a practical compromise for testing purposes.

This paper further assumes that correlation parameters  $\rho_W$  and  $\rho_B$  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 correlations<sup>18</sup>. By matching the variances of the non-idiosyncratic parts of the asset returns in Basel II and in the extended DRAM,

---

<sup>13</sup> See the expression for  $PD(\mathbf{x})$ .

<sup>14</sup> 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 therefore a pragmatic choice.

<sup>15</sup> Note that the structure of  $\boldsymbol{\Sigma}$  defines DRAM more concretely than the chosen decomposition of the normalized asset return because, given  $\boldsymbol{\Sigma}$ , the latter is not unique.

<sup>16</sup> To the best of the author's knowledge, the empirical literature lacks studies on that association.

<sup>17</sup> Even if the bank or supervisor is convinced of the appropriateness of  $\rho_B = \rho_W$ , the tests of this paper are still defensible, provided the default rates of different ratings are computed based on distinct sectors for instance.

<sup>18</sup> 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 aspect.

$\rho_W$  can be seen as the asset correlation parameter present in the Basel II formula<sup>19</sup>. For corporate borrowers for example Basel II accord chooses  $\rho_W \in [0.12 \ 0.24]$ <sup>20</sup>. Sensitivity analysis of tests results on the choices of these parameters is pursued along the text. 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 it is assumed time independency for the default rates of each year. Therefore, the  $(\Phi^{-1}$  transformed) yearly average default rate, used as the test statistic for all the tests of the next sections, has the normal distribution above, just with  $\Sigma/Y$  in place of  $\Sigma$ , where  $Y$  is the number of years available to backtest. According to BCBS(2005b), time independency is less inadmissible than cross-sectional independency.

### 3. The statement of 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?<sup>21</sup> 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)<sup>22</sup>. But if the bank/supervisor wants to know if the data provided evidence confirming the CRM is correctly specified, then this hypothesis should be placed in  $H_1$  and the opposite of it 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 approach is pursued throughout this paper so that the probability of accepting an incorrect CRM will be the error to be controlled for at the significance level  $\alpha$ . To the best of the author’s knowledge this paper is first one to state 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 set<sup>23</sup>. This implies that the statement that “the model is correctly specified” needs to be translated into some statement about the parameters (PD<sub>i</sub>s) lying in an *open* set, in particular there shouldn’t be equalities defining  $H_1$  and the inequalities should be strict. It is, for example, statistically inappropriate to try to conclude that the PD<sub>i</sub>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<sub>i</sub>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 in the same rating  $i$  have exactly the same theoretical PD<sub>i</sub> and that it is therefore more realistic to see the ratings as defined by PD intervals.<sup>24</sup>

---

<sup>19</sup> Note that the Basel II case can also be seen as the particular case of DRAM when  $\rho_B = \rho_W$ .

<sup>20</sup> On the other hand Basel II accord doesn’t provide information on  $\rho_B$  because it is based on a single systemic factor.

<sup>21</sup> For this general discussion, one can think of “correctly specified” as meaning either correct calibration or good discriminatory power.

<sup>22</sup> Although they do not remark the consequences of their choices.

<sup>23</sup>  $H_0$  and  $H_0 \cup H_1$  need to be closed sets in order to guarantee that the maximum of the likelihood function is attained.

<sup>24</sup> 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.

## 4. Calibration testing

This section distinguishes between one-sided and two-sided tests for calibration. The one-sided test (which is only concerned about PD<sub>i</sub>s being greater than certain thresholds) is useful to the supervisor authority as it allows the latter 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 used in the banks' managerial activities such as credit granting and credit pricing. To accomplish these goals, PD estimates must undistortly reflect the likelihood of default of every rating, something to be verified more effectively only by the two-sided test (which is concerned to PD<sub>i</sub>s being within certain ranges). Unfortunately the difficulties of two-sided calibration testing are much greater than of one-sided testing, as the section reveals ahead. A discussion of the one-sided calibration test starts the analysis.

Based on the arguments of the previous section about the proper roles of H<sub>0</sub> and H<sub>1</sub>, the formulation of the one-sided calibration test is stated below. Note that the desired conclusion, configured as an intersection of strict inequalities, is placed in H<sub>1</sub>.

H<sub>0</sub>:  $PD_i \geq u_i$  for some  $i = 1 \dots I$

H<sub>1</sub>:  $PD_i < u_i$  for all  $i = 1 \dots I$

where  $PD_i \equiv \Phi^{-1}(PD_i)$ ,  $u_i \equiv \Phi^{-1}(u_i)$ . (This convention of representing  $\Phi^{-1}$  transformed figures in italic is followed throughout the rest of the text)<sup>25</sup>.

Here  $u_i$  is a fixed known number that defines an indifference acceptable region for PD<sub>i</sub>. Its value should ideally be slightly larger than the value postulated for PD<sub>i</sub> so that the latter is within the indifference region. Also  $u_i$  should ideally be smaller than the value postulated for PD<sub>i+1</sub> so that at least the rejection of H<sub>0</sub> could conclude that PD<sub>i</sub> < postulated PD<sub>i+1</sub>.<sup>26,27</sup>

According to DRAM and based on the results of Sasabuchi (1980) and Berger (1989), which investigate the problem of testing linear homogeneous inequalities concerning normal means, a size  $\alpha$  critical region can be derived for the test.<sup>28</sup>

Reject H<sub>0</sub> (i.e. validate the model) if

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

where  $\overline{DR}_i = \frac{\sum_{y=1}^Y DR_{iy}}{Y}$  is the yearly average (transformed) default rate of rating  $i$  and  $z_\alpha = \Phi^{-1}(1 - \alpha)$  is the  $1 - \alpha$  percentile of the standard normal distribution.<sup>29</sup>

<sup>25</sup> As  $\Phi^{-1}$  is strictly increasing, statements about the italic figures imply equivalent statements about the non-italic figures.

<sup>26</sup> As banks have the incentive to postulate lower PDs one could argue that  $PD_i < \text{postulated } PD_{i+1}$  also leads to  $PD_i < PD_{i+1}$ .

<sup>27</sup> Specific configurations of  $u_i$  are discussed later.

<sup>28</sup> Size of a test is the maximum probability of rejecting H<sub>0</sub> when it is true.

<sup>29</sup> This definition of  $\overline{DR}_i$  is used throughout the paper.

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  $\alpha$ . In general the size of a min test will be much smaller than  $\alpha$  but the results of Sasabuchi (1980) and Berger (1989) guarantee that the size is exactly  $\alpha$  for the one-sided calibration test<sup>30</sup>. This means that the CRM is validated at size  $\alpha$  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.<sup>31</sup> Second, as the transformed default rate variables are asymptotically normal in DRAM, the min test is also asymptotically the likelihood ratio test (LRT). Finally, the achievement of a size  $\alpha$  is robust to violation of the assumption of the normal copula for the transformed default rates (Wang *et. al.* (1999)) so that, for size purposes, the assumption of *joint* normal distribution of the systemic factors can be relaxed.

From a practical point of view it should be noted that the critical region does not depend on the parameter  $\rho_B$ , which is good in applications since  $\rho_B$  is not present in Basel II framework so that there is not much knowledge about its reasonable values. However, there is no free lunch: the power of the test, i.e. the probability of validating the CRM when it is correctly specified, does depend on  $\rho_B$ . The power is given by the expression below.

$$\text{Power} = \Phi_I(-z_\alpha + (u_1 - PD_1)/(\rho_W / Y)^{1/2}, \dots, -z_\alpha + (u_I - PD_I)/(\rho_W / Y)^{1/2}, \dots, -z_\alpha + (u_I - PD_I)/(\rho_W / Y)^{1/2}, \rho_B / \rho_W)$$

where  $\Phi_I(\dots, \rho_B / \rho_W)$  is the cumulative distribution of a I-th multivariate normal of mean 0, variances equal to 1 and covariances equal to  $\rho_B / \rho_W$ .

Berger (1989) remarks that if the ratio  $\rho_B / \rho_W$  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  $\rho_B / \rho_W$  indicates that *ex-post* information about one rating does not contain much information about other ratings and therefore is less helpful to conclude for validation. More generally, it is easy to see that the power increases when the  $PD_i$ s decrease, the  $u_i$ s decrease, Y increases, I decreases,  $\rho_B$  increases or  $\rho_W$  decreases<sup>32</sup>. 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). I now analyze some numerical examples in order to provide some further insights on this trade-off and on reasonable choices for  $u_i$ s.

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

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

---

<sup>30</sup> More formally the description just given is the one 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.

<sup>31</sup> In the context of this paper, a test is monotone if the fact that yearly average 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.

<sup>32</sup> Obviously the power also increases when the level  $\alpha$  increases.

In a base case scenario of  $Y=5$ ,  $\rho_W = 0.15$ ,  $\alpha = 15\%$  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 target balance between power and size. It is then true that:

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

Table 1 below displays values of  $u_i$  for varying values of  $PD_i$ .

**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 in the base case scenario, an approximate lower bound for  $PD_{i+1}$  in terms of  $PD_i$ <sup>33</sup>. 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  $PD_{i+1} - PD_i$ , a characteristic found indeed in the design of many real-world CRMs. Therefore DRAM provides a validation argument explaining that configuration choice. Notice that this feature of increasing  $PD_{i+1} - PD_i$  is directly related to the non-linearity of  $\Phi$ , which in turn is a consequence of the asymmetry and heavy tails of the distribution of the untransformed default rate.

To further investigate the feature of increasing PD differences I now analyse explicitly the case  $l=3$ . Two CRMs are considered: CRM 1 has equally spaced PDs and CRM 2 has increasing PD differences. 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  $u_i = (PD_{i+1} + PD_i)/2$ <sup>34</sup>. The choices for the values of  $\rho_W$  and  $Y$  are made considering three feasible scenarios: a favourable one characterized by 10 years of data and a low within-rating correlation of 0.15, a unfavourable one characterized by the minimum number of 5 years prescribed by Basel II (c.f. Basel (2004)) and a high  $\rho_W$  at 0.20 and an in-between scenario<sup>35</sup>. The power figures of the one-sided calibration test at the postulated PDs are shown in table 2.

**Table 2: Power comparison among CRMs and  $u_i$  choices**

CRM 1:  $PD_1=4\%$ ,  $PD_2=6\%$ ,  $PD_3 = 8\%$ ; CRM 2:  $PD_1=4\%$ ,  $PD_2=8\%$ ,  $PD_3 = 16\%$ ;  $\alpha = 15\%$ ;  $\rho_B / \rho_W = 2/3$

	CRM 1 with equally spaced PDs		CRM 2 with increasing $PD_{i+1} - PD_i$	
	$u_i = PD_{i+1}$	$u_i = (PD_{i+1} + PD_i)/2$	$u_i = PD_{i+1}$	$u_i = (PD_{i+1} + PD_i)/2$
$\rho_W = 0.15, Y=10$	0.38	0.18	0.96	0.65
In-between	0.27	0.14	0.84	0.46
$\rho_W = 0.20, Y=5$	0.20	0.11	0.70	0.35

The table shows that only model 2 achieves reasonable levels of power so that the feature of increasing  $PD_{i+1} - PD_i$  seems to be really necessary for realistic validation attempts. Therefore, the attention is focused on CRMs of this type to the remainder of this text. Table 2 also reveals that, even

<sup>33</sup> Approximate because the computation was based on  $l=1$ . In fact the true attained power in a multi rating setup is smaller.

<sup>34</sup> The choice of an arithmetic mean makes sense as if the true PD cannot be shown to be smaller than the midpoint of the interval  $[PD_i, PD_{i+1}]$ , then validation should be denied.

<sup>35</sup> As  $\rho_B / \rho_W$  is fixed in table 2 what matters for the power calculation is just the ratio  $(\rho_W / Y)$ . Therefore the in-between scenario can be thought as characterized by adjusting both  $Y$  and  $\rho_W$  or just one of them. In table 2 it is given by  $\rho_W / Y = 0.0256$ .

when solely focusing on model 2, more demanding requirements for  $u_i$  (c.f. last column) may produce overly conservative tests, with power on the level of only 35%. Further, the power is found to be very sensitive to the within-rating correlation  $\rho_W$  and to the number of years  $Y$ . In most cases, it almost doubles from the worst to the best scenario.

While in table 2 the between-rating correlation parameter is hold fixed, table 3 examines its effect on the power of the test. Power is computed at postulated PDs of CRM 2 of table 2 for the worst scenario and  $u_i = PD_{i+1}$ . Table 3 shows just a minor effect of  $\rho_B$ , regardless of the size of the test. Results not shown reveal that a minor effect is also the case for the average PD configuration for  $u_i$ .

**Table 3: Effect of  $\rho_B$**

CRM:  $PD_1=4\%$ ,  $PD_2=8\%$ ,  $PD_3=16\%$ ;  $\rho_W = 0.15$ ,  $Y=10$ ,  $u_i = PD_{i+1}$

	$\alpha=5\%$	$\alpha=10\%$	$\alpha=15\%$
$\rho_B/\rho_W = 1/3$	0.37	0.54	0.65
$\rho_B/\rho_W = 1/2$	0.41	0.57	0.67
$\rho_B/\rho_W = 2/3$	0.44	0.60	0.70

Table 4 below gives insights on the relative role played by the different ratings on the power. Power is computed at postulated PDs for a sequence of three embedded CRMs starting with CRM 2. Each CRM is built from its antecedent by dropping the riskier rating. Power is computed at postulated PDs for the in-between scenario and  $u_i = PD_{i+1}$ . Table 4 reveals that, as the number of ratings diminishes, the power increases just to a minor extent, provided the less risky rating is always kept in the model. Therefore it can be said that the lower PD drives the power of the test. This is partly intuitive because the smallest  $PD_i$  corresponds to the smallest difference  $u_i - PD_i$  and because distinct PDs contribute to the power differently just to the degree their differences  $u_i - PD_i$  vary<sup>36</sup>. The surprising part of the result refers to the degree of relative low importance of the other PDs: the variation of power between  $l=1$  and  $l=3$  could be just 5%. It's important to remark that this result is strongly dependent on the feature of increasing  $PD_{i+1} - PD_i$ ; results not shown indicate that the largest  $PD_i$  drives the power for CRMs with equally spaced PDs.

**Table 4: PDs drivers of power**

$\rho_B/\rho_W = 1/3$ ;  $(\rho_W / Y)^{1/2} = 0.16$ ;  $u_i = PD_{i+1}$

PDs	$\alpha=5\%$	$\alpha=10\%$	$\alpha=15\%$
4%, 8%, 16%	0.59	0.74	0.82
4%, 8%	0.61	0.75	0.83
4%	0.70	0.81	0.87

An underlying message present in the analysis of the previous tables is that the one-sided calibration test can have substantially low power. Another related problem refers to the test not being similar on the boundary between the hypotheses and therefore biased (reference ?)<sup>37</sup>. To cope with these

<sup>36</sup> For the CRM of table 3 it's true, by construction, that  $u_i - PD_i$  increases in  $i$ . But it's also true that the greater the PDs, the smaller are equal differences in the  $\Phi^{-1}$  scale. It's not difficult to verify for table 3 that the net effect is of increasing  $u_i - PD_i$ .

<sup>37</sup> A test is  $\alpha$  similar on a set A if the probability of rejection is equal to  $\alpha$  everywhere there. A test is unbiased at level  $\alpha$  if the probability of rejection is smaller than  $\alpha$  everywhere in  $H_0$  and greater than  $\alpha$  everywhere in  $H_1$ . Every unbiased test at level  $\alpha$  with continuous power function is  $\alpha$ -similar in the boundary between  $H_0$  and  $H_1$ .

*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 tests are constructed by carefully enlarging the rejection region in order to preserve the size  $\alpha$ . 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 not-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 are not dominated in the decision theoretic sense because the probability of rejection under  $H_0$  (i.e. when the CRM is incorrect) is also higher for them<sup>38</sup>. The authors conclude that UMP tests should not be persecuted at any cost, particularly at the cost of intuition. This is the view adopted in this study so that I don't explore the new tests further.

I now examine yet a different approach to improve the power of the one-sided calibration test. Notice, first, that the size  $\alpha$  of the test is attained when all but one of the PD<sub>i</sub>s go to 0 while the remaining one is set fixed at  $u_i$ <sup>39,40</sup>. This is probably a very unrealistic scenario against which the bank or the supervisor would like to be protected. One may alternatively remove by assumption this unrealistic case from the space of **PD** possibilities and rather consider that part of the information postulated or believed by the bank is true. Notably, one can assume that  $PD_{i-1}$ , not 0, represents a lower bound for  $PD_i$ , for every rating  $i$ . Following Sasabuchi (1980) strategy, it's not difficult to show that, in that altered parameter space, the probability of rejection assumes a similar form to the original one, with the only difference being that now a constant  $c > 0$  plays the role of  $z_\alpha$ .

$$\text{Power} = \Phi_1(-c + (u_1 - PD_1)/(\rho_W / Y)^{1/2}, \dots, -c + (u_i - PD_i)/(\rho_W / Y)^{1/2}, \dots, -c + (u_l - PD_l)/(\rho_W / Y)^{1/2}; \rho_B / \rho_W)$$

The constant  $c$  is defined by the requirement that the maximum of the expression above over the intersection of  $H_0$  and the modified **PD** parameter space is  $\alpha$  (so that the size of the modified test is  $\alpha$ ). Similarly to Sasabuchi (1980), the determination of  $c$  needs the examination in  $H_0$  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  $1 \dots l$ . The  $j$  that maximizes power corresponds to the smaller difference  $u_j - PD_{j-1}$  and, therefore will, generally depend on the value postulated for **PD** and the choice of  $\mathbf{u}$ <sup>41</sup>. Suppose  $j=1$  for the sake of exposition. Then the requirement of size  $\alpha$  takes the form below.

$$\text{Power} = \Phi_1(-c, -c + (u_2 - PD_1)/(\rho_W / Y)^{1/2}, \dots, -c + (u_i - PD_{i-1})/(\rho_W / Y)^{1/2}, \dots, -c + (u_l - PD_{l-1})/(\rho_W / Y)^{1/2}; \rho_B / \rho_W) = \alpha$$

from which the value of  $c$  can be derived.

From that equation, it's possible to show that  $c < z_\alpha$ <sup>42</sup>. The replacement of  $z_\alpha$  by a smaller constant enlarges the critical region, when compared to the original test, and trivially produces, therefore, an aimed more powerful test<sup>43</sup>. From a methodological point of view, the drawbacks of the modified test

<sup>38</sup> More specifically, it is higher at every PD parameter in  $H_0$ .

<sup>39</sup> This limiting **PD** vector is in  $H_0$  and therefore, ideally, should not be validated. It has a probability of validation equal to  $\alpha$ .

<sup>40</sup> Note  $PD_i \rightarrow 0 \Rightarrow PD_i \rightarrow -\infty$

<sup>41</sup> If  $u_i$  is a non-decreasing function of  $PD_{i+1}$ , it's reasonable to assume that  $u_j - PD_{j-1}$  will increase in  $i$  in realistic CRMs because the low quality part of the rating scale tends to be sparser, as already mentioned. But it's also true that the greater the  $PD_i$ s, the smaller are equal differences in the  $\Phi^{-1}$  scale, reducing the effect on the transformed difference  $u_j - PD_{j-1}$ . Besides, there is the point that the first difference  $u_1 - PD_0 \equiv u_1$  may not be so small.

<sup>42</sup> As the components of the power function cannot go to infinity as before, the first component must increase for the size to be achieved.

<sup>43</sup> See the definition of the critical region in the beginning of the section.

lie, however, on the dependence of the new critical region on  $\rho_B$  and on the fact that the calculation of  $c$  needs some computational effort. From a performance perspective, preliminary produced results suggest that the power increase of the modified one-sided calibration test is relevant only in the region of small (possibly unrealistic) ratio  $\rho_B/\rho_W$  or for ambitious choices of  $u_i$  (i.e. close to  $PD_i$ ). In that 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).

### Tables to be included

I now comment on the two-sided calibration test. Similarly to the one-sided version, its hypotheses are stated as follows.

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

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

Now the indifference acceptable 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 becomes an example of the class of multivariate equivalence tests, which are tests designed to show similarity rather than difference and which widely employed in the pharmaceutical industry to demonstrate that drugs are equivalent.<sup>44</sup> Berger and Hsu (1996) comprehensively review the recent development of equivalence tests in the univariate case ( $I=1$ ). The standard procedure to test univariate equivalence is the TOST test (two one sided test - called this way because the procedure is equivalent to performing two size- $\alpha$  one sided tests and to conclude equivalence only if both reject). Wang *et.al.* (1999) discuss the extension of TOST to the multivariate case, making use of the intersection-union method. That extension, when applied to the DRAM distribution, results in the following critical region for the two-sided calibration test<sup>45</sup>.

Reject  $H_0$  (i.e. validate the model) if

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

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 all  $i$ ), 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 previous test has similar problems of lack of power and bias (the latter if  $I > 1$ )<sup>46</sup>. Indeed, the statistical literature contains some proposals of improvement for the TOST (Berger and Hsu(1996), Brown *et. al.*(1997)), which are again subject to criticism from an intuitive point of view by Perlman and Wu (1999)<sup>47</sup>. 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  $\alpha$  (Berger

<sup>44</sup> More specifically, these tests are referred as bioequivalent tests in the pharmaceutical industry.

<sup>45</sup> The standard TOST is framed assuming unknown variance while the two-sided calibration test of this paper assumes known variance. Therefore the reference to the term TOST represents here some freedom of notation.

<sup>46</sup> It is not similar on the boundary between the hypotheses and therefore biased.

<sup>47</sup> However, in the case of calibration testing with known variance, the bias is not as pronounced as in the case of TOST with unknown variance.

and Hsu (1996)) while its size may be much smaller.<sup>48,49</sup> These observations indicate the magnified difficulty in performing the two-sided calibration testing.

### Tables to be included

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, it consists of accepting an alternative hypothesis  $H_1$  (i.e. validating the model) if the Brown confidence set for the **PD** vector is entirely contained in  $H_1$ . The approach would allow the bank or supervisor to separate the execution of the test from the task of defining an indifference region. In fact, the confidence set can be seen as the smallest indifference region so that it is still possible to validate the calibration and, therefore,  $H_1$  configuration could be discussed at a later stage, after the knowledge of the *form* of the set. Brown *et.al.* (1995) propose an optimal confidence set in the sense that, if the true **PD** vector is the postulated one, then the expected volume of the set is minimal, which means that, in average terms, maximal precision is achieved when the calibration is *exactly* right<sup>50</sup>. 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 should 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 interest to calibration testing.<sup>51,52</sup>

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 alternative hypotheses 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  $\mathbb{R}^l$ . The referred authors analyze the use of ellipsoidal alternatives for the problem of multi-equivalence testing, which, for purposes of calibration testing, can be exemplified as follows.

$$H_0: \mathbf{e}^t \mathbf{D} \mathbf{e} \geq \Delta$$

$$H_1: \mathbf{e}^t \mathbf{D} \mathbf{e} < \Delta$$

where  $\mathbf{e} = \mathbf{PD} - \text{postulated } \mathbf{PD}$ ,  $\mathbf{D}$  is a positive definite matrix that conceives a notion of distance in the  $\mathbb{R}^l$  space and  $\Delta$  denotes a fixed tolerance bound.  $\mathbf{D}$  and  $\Delta$  define an indifference region.

Munk and Pfluger (1999) advocate this formulation in order to allow for the notion of equivalence to be interpreted as a combined measure of several parameters (e.g. a combination of the  $\text{PD}_i$ s,  $i=1 \dots l$ ). As a consequence, this implies in the calibration problem that very good *marginal* equivalence (e.g. the postulated  $\text{PD}_1$  is very close to the postulated  $\text{PD}_1$ ) should allow larger indifference regions for other  $\text{PD}_i$ s. Conceptually, I believe this point is hard to justify in the validation of CRMs. If miscalibration

---

<sup>48</sup> It can be shown that the degree of conservatism depends on  $\rho_B$ .

<sup>49</sup> The reason for the discrepancy with the TOST relates to the impossibility of making the variance go to 0 as in Berger and Hsu (1996).

<sup>50</sup> The form of the set is not an ellipse, commonly found in multivariate analysis, but rather a figure known as the Limaçon of Pascal.

<sup>51</sup> Note also that the DRAM should be seen just an approximation to reality, so that, even if all borrowers in a rating can be seen as having the same PD, small deviations from the asymptotic model assumptions may in practice force the true **PD** to depart from the theoretical one.

<sup>52</sup> Other confidence set approaches to calibration testing are also possible. Some of them are, however, dominated by the multivariate TOST. (Munk and Pfluger (1999))

were necessarily derived from a systematic erroneous estimation of all the  $PD_i$ s, that indeed could be the case. Nevertheless, the view of this paper is that miscalibration could be rather very much 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 the purposes of calibration testing, ellipsoidal alternatives are seen here more as a practical complication.<sup>53</sup>

## 5. The mapping test

As noted in the introduction, BCBS(2005b) stresses the need for the development of tests on the adequacy of a mapping established between two different rating scales. Usually one rating scale (say A) is already assumed to be correctly calibrated and the appropriateness of the calibration of the other scale (say B) remains to be checked. A mapping test should then test whether mapped ratings between the two scales possess similar PDs. As a consequence of this study's approach to place the desired conclusion in the alternative hypothesis, the bank or supervisor should thus be satisfied that the mapped PDs are sufficiently close but not necessarily equal.

The results of the previous section can be easily adapted to focus on the mapping test. The easiest way to accomplish this is to define appropriate values for  $u_i$ s and  $l_i$ s based on the  $PD_i$ s of the assumed correctly calibrated scale A. More concretely, under this view, a mapping test for the rating scale B is merely a calibration test with choices  $u_i$ s and  $l_i$ s that respect  $PD_i^A \leq u_i \leq PD_{i+1}^A$  and  $PD_{i-1}^A \leq l_i \leq PD_i^A$ , for every rating  $i=1\dots I$ . Similarly to the calibration case, both a one-sided and two-sided versions of the mapping test are possible. The same power concerns and limitations of the calibration tests apply here as well.

The previous approach assumes information on  $PD^A$  is known. This may not always be the case as the latest updated  $PD^A$  may not have been validated yet at the time the mapping test is to be conducted. Also, in more general contexts than of in Basel II, ratings may not have explicit PDs associated to them. A mapping test can still be conducted based on data on rating default rates of both scales A and B. The approach is similar in spirit to the tests of the next section. The cost is that larger variances have to be considered due to the additional uncertainty on  $PD^A$ .

## 6. Tests of rating discriminatory power

One of the most traditional measures of discriminatory power is the area under the ROC curve (AUROC). 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 structure<sup>54</sup>. To the extent that the CRM should not be held accountable for the effect of default dependency between borrowers, the traditional

---

<sup>53</sup> 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.

<sup>54</sup> It is a function of the distribution of borrowers along the ratings too.

measure of discrimination becomes distorted.<sup>55</sup> The next proposition shows explicitly the dependency of AUROC on the asset correlation parameters.

**Proposition:** Given a CreditMetrics (Gupton et. al. (1997)) style model (of which Basel II model and DRAM are particular versions) endowed with a matrix of asset correlations ( $\rho_{ij}$ ) between borrowers of ratings  $i$  and  $j$ ,  $i, j = 1 \dots I$ . Let  $P(i, j)$  and  $P(i)$  be the probability of two random borrowers having ratings  $i$  and  $j$  and one random borrower having rating  $i$ , respectively. Then:

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

Proof: Appendix B.

The remainder of this section describes alternative proposals of tests of rating discriminatory power built upon the DRAM distribution. The qualifying term *rating* is added purposely to the traditional expression *discriminatory power* to emphasize that the property desired to be concluded or 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 be, therefore, a strict subset of the  $H_1$  of the two-sided calibration test. 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 monotony along the rating dimension is also useful to corroborate the assumptions of some methods of PDs 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 stated below.

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

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

Viewing  $PD_{i+1} - PD_i$  as the unknown parameter to be estimated by  $DR_{i+1} - DR_i$  for every rating  $i$ , the previous test can be seen as one involving testing homogeneous inequalities about normal means<sup>56</sup>. So, similarly to the one-sided calibration test, a size- $\alpha$  likelihood-ratio critical region can be derived.

---

<sup>55</sup> 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.

Reject  $H_0$  (i.e. validate the model) if

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

It is worth noting above that, opposed to the calibration tests, there is no need for the configuration of an indifference region, as the desired  $H_1$  conclusion is already defined by strict inequalities. On the other hand, here the critical region and therefore the decision itself to validate the model does depend on the unknown parameter  $\rho_B$ . The Basel II case ( $\rho_B = \rho_W$ ) represents the extreme liberal situation where just an observed increasing behaviour of the yearly average default rates along the rating dimension is sufficient to validate the model (regardless of the confidence level  $\alpha$ ) whereas the case  $\rho_B = 0$  places the strongest requirement in terms of the incremental increase of the default rate averages along the rating dimension<sup>57</sup>. In practical situations the bank or supervisor may want to find what is highest value of  $\rho_B$  such that the general test still validates the model and then check whether this value conforms to his 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 two factors<sup>58</sup>. First, the fact that now the underlying normal variables may have smaller variances ( $\text{Var}(DR_{i+1} - DR_i) = 2(\rho_W - \rho_B) / (1 - \rho_W) < \text{Var}(DR_i) = \rho_W / (1 - \rho_W)$ , if  $\rho_B / \rho_W > 1/2$ ) contributes to an increase in power in that case. 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 power<sup>59</sup>. The resulting dominating force is to be determined by the particular choices of  $\rho_B$  and  $\rho_W$ . In general, the same comments on possible strategies for power improvement and their limitations apply here as well.<sup>60</sup>

It is also worthwhile to discuss the situation where the bank or supervisor is satisfied by the *general level* of rating discrimination except for a particular pair of consecutive ratings. Suppose the supervisor or the bank 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<sup>61</sup>. From a supervisory standpoint, a suspicion of regulatory arbitrage may for instance motivate the concern. I denote a test to examine this issue as a focal test of rating discriminatory power, whose hypotheses are stated as follows.<sup>62,63</sup>

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

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

<sup>56</sup> 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.

<sup>57</sup> This is again intuitive as low values of  $\rho_B$  mean that *ex-post* information about one rating does not contain much information about others ratings.

<sup>58</sup> Similarly to the calibration case, the power expression can be easily derived.

<sup>59</sup> Therefore, not necessarily validating rating discriminatory power is easier than validating (one-sided) calibration, as one could think a priori because of the less stringent nature of  $H_1$  in the former. In fact it may indeed be harder if  $\rho_B / \rho_W < 1/2$ .

<sup>60</sup> In particular, similarly to the calibration case, restricting the space of parameters based on ordinal information about the PDs can be similarly tried, although now the way to do that is not so straightforward.

<sup>61</sup> Suspicion of regulatory arbitrage may derive from a situation where large credit risk exposures are apparently rated with slightly better rating so that the resulting capital charge of Basel II is diminished.

<sup>62</sup> This discussion of this section is easily generalized to the situation where a set larger than two ratings is to have its rating discriminatory power verified.

<sup>63</sup> It can be shown that replacing  $PD_1 = PD_2$  by  $PD_1 \geq PD_2$  leads to the same test.

From a mathematical point of view, the development of such a test is much more complex than the tests considered so far in this paper because now the union of the null and the alternative hypotheses do not span the full  $|R|^1$ . This implies that the null distribution of the likelihood ratio (LR) statistic is complicated and depends on the structure of the cone  $C = H_0 \cup H_1$ , whether it is obtuse or acute with respect to norm induced by  $\Sigma^{-1}$ .<sup>64,65</sup> In the first case, the LR statistic follows a  $\chi^2$  *bar* distribution under  $H_0$  (Menendez *et. al.* (1992a)).<sup>66</sup> In the second case, the distribution of the LR statistic is unknown 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  $\Sigma$  adopted in this paper makes the cone  $C$  acute so that the second case is the relevant one.<sup>67</sup> The reduced dominating test takes the form below.

$$H_0: PD_1 = PD_2$$

$$H_1: PD_1 < PD_2$$

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

Reject  $H_0$  (i.e. validate the model)

$$\text{if } \Sigma(DR_2 - DR_1)/Y > z_\alpha (2(\rho_W - \rho_B)/(Y(1 - \rho_W)))^{1/2}$$

The dominance of the focal test by a reduced test is a surprising result and was long considered an anomaly of the LRT (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 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 indeed that 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  $H_1$ .<sup>68</sup> Nevertheless, the practitioner does not have other choice besides using the power dominating test, because, as just mentioned, 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 referred 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.

## 7. Small sample properties

---

<sup>64</sup> See ?? for the definition of these forms of cones.

<sup>65</sup>  $\|x\|_{\Sigma^{-1}} = x^T \Sigma^{-1} x$

<sup>66</sup> Although  $\chi^2$  *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 both the determination of the LRT statistic and the p-values are computer intensive.

<sup>67</sup> This is true because  $\mathbf{a}_i^T \Sigma \mathbf{a}_j \leq 0$  where the  $\mathbf{a}_i$ 's ( $\mathbf{a}_i = (0, \dots, -1, 1, \dots, 0)^T$ ) generate the linear restrictions defining the cone  $C$ . More specifically, it is true that  $\mathbf{a}_i^T \Sigma \mathbf{a}_j = (\rho_B - \rho_W)/(1 - \rho_W)$  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  $\Sigma$  lead to a different conclusion is an interesting question not addressed in this paper.

<sup>68</sup> Perlman and Wu (1999) conclude once again that UMP size- $\alpha$  tests should not be pursued at any cost.

All the tests discussed in this paper are based on an asymptotic distribution of the DRAM, 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)<sup>69</sup>. Due to the strong reliance of the test on the asymptotic marginal normal distributions of the DRAM, it is important to verify how the real marginals compare to the asymptotic ones<sup>70</sup>. The focus on a particular marginal allows then to restrict the attention to the case  $l=1$ <sup>71</sup>. Hence this section conducts Monte-Carlo simulations of the Basel II model at the stage in which idiosyncratic risk is not yet diversified away and for  $l=1$ ,  $N=100$  and  $Y=5$ , unless stated otherwise.<sup>72</sup> Based on a large set of simulated yearly average default rates, the effective significance level is computed as a function of nominal significance level  $\alpha$ , for varying scenarios of parameters<sup>73</sup>.

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

where  $\overline{DR}$  is simulated a number of times and the estimated probability is computed as the empirical frequency of the event.

The effective level measures the real size of the asymptotic size- $\alpha$  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, when the asymptotic power at postulated PD is equal to  $\alpha$ , of an asymptotic size  $\delta$  one-sided test,  $\delta < \alpha$ <sup>74</sup>. From both interpretations, 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 DRAM. Effective levels higher than nominal levels indicate the opposite: a small sample liberal *bias*.

A general important finding of the performed simulations is that the convergence of the lower tails of the ( $\Phi^{-1}$  transformed) average default rate distributions to their normal asymptotic limits is much less smooth than in the case of the upper tails, for realistic values of PDs<sup>75</sup>. 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 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 is approximately only 4%<sup>76</sup>.

---

<sup>69</sup> 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.

<sup>70</sup> Review the form of the critical region in section 4.

<sup>71</sup> 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.

<sup>72</sup> 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 ( $\Phi^{-1}$ ) transformed default rate distributions.

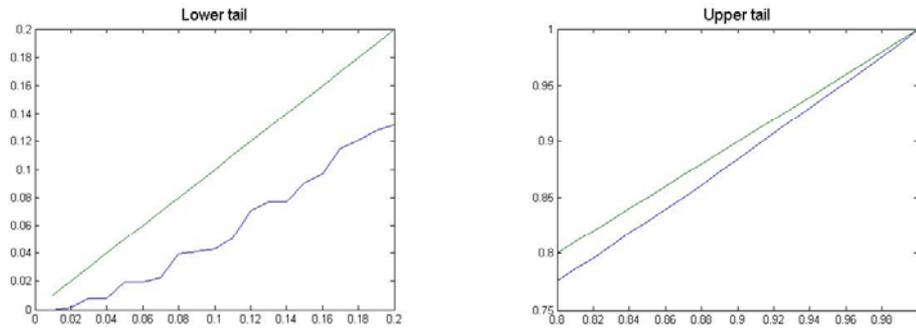
<sup>73</sup> In general 200000 simulations are run for each scenario.

<sup>74</sup> More specifically, it is easy to see that  $\delta = \Phi(-z_\alpha - (u - PD)/(1-\rho_w)^{1/2})$

<sup>75</sup> The intuitive reason for this being that  $\Phi^{-1}(PD) \rightarrow -\infty$  when  $PD \rightarrow 0$ .

<sup>76</sup> There is less mass in the simulated lower tail than in the DRAM distribution.

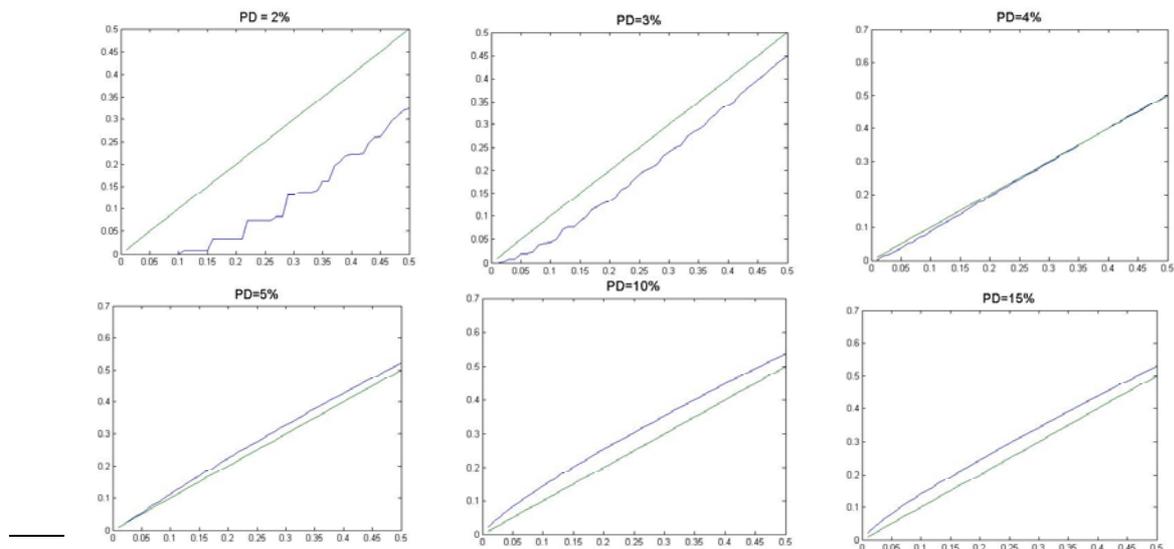
**Graph 1: Lower and upper tails,  
PD=3%,  $\rho_W=0.20$  N=100, Y=5**



The fact that the lower tail is less well behaved is strongly relevant to the discussion of this paper. 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 indeed on the lower tail of the distribution. On the contrary, the upper tail would have been the relevant part of the distribution had the approach of placing the “correctly specified” hypothesis in  $H_0$  been adopted, as in BCBS(2005b). As 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 approach<sup>77</sup>.

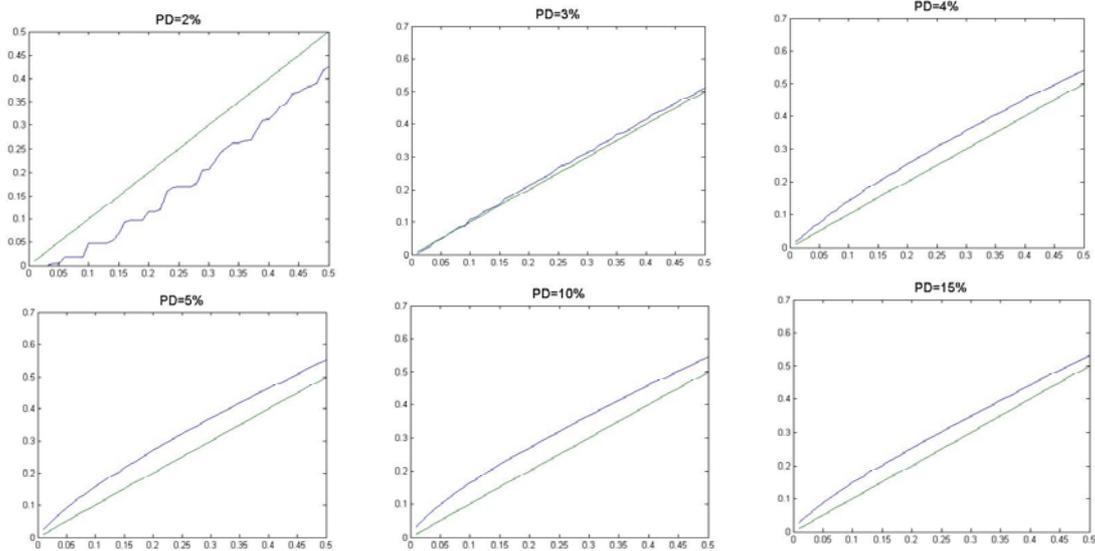
The main numerical findings regarding the small sample power performance of the one-sided calibration test are described in the sequence, based on the analysis of the simulated lower tails. The investigation starts with the effect of the true PD on the effective confidence level, for two different values of  $\rho_W$ , 0.15 and 0.20. 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 bias is approximately null as the test matches its theoretical limiting values. On the other hand, in the region  $10\% < PD < 15\%$ , as PD increases, the blue line comes a bit closer back to the green one, i.e. the test diminishes its liberal bias (but not sufficiently so as to become conservative).

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



<sup>77</sup> Because the worse relative behaviour of the lower tail would not be revealed.

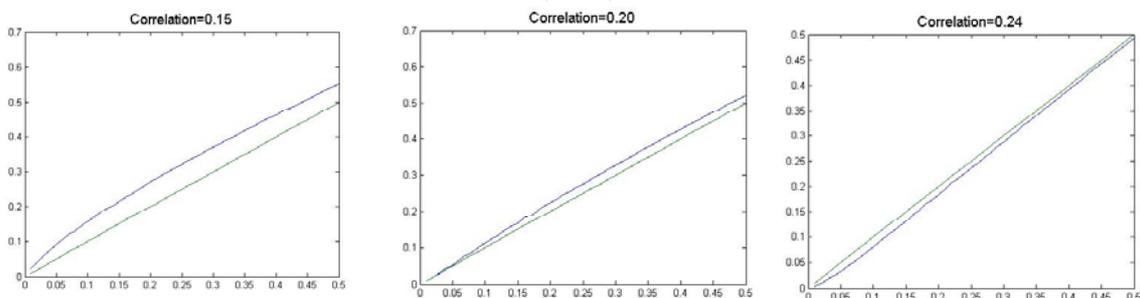
**Graph 3: Effect of PD,  
 $\rho_w=0.15, N=100, Y=5$**



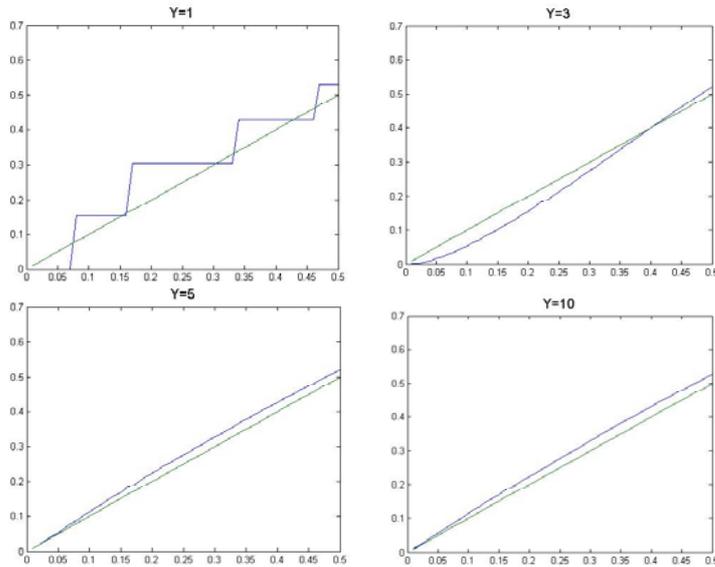
As the asymptotic one-sided test based on DRAM already suffers from problems of lack of power, this section suggests, as possible general recommendations, to consider a real (unmodified) application of the test solely in the cases where the small sample analysis indicate a liberal bias. Indeed, if otherwise 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.  $PD \leq 3\%$ ). As a result, 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  $\alpha$  could be applied to the low PDs.

The effects of varying values of correlation and of the number of years under the base case of  $N=100$  are also analyzed in graphs 4 and 5. As the within-rating asset correlation  $\rho_w$  increases, the test evolves from a liberal bias to a small conservative one. Note that this represents the second channel, now through the small sample properties, by which  $\rho_w$  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 of the lower tail although the *direction* of the 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  $\rho_w$   
 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 is based totally in the simulated average default rate distribution of this section, there are 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.

## 8. Conclusion

This study contributes to the CRM validation literature in presenting several new ways of addressing the **PD** validation issue. Firstly, it proposes new statements by which  $H_0$  and  $H_1$  can be formulated in order to control the error of accepting an incorrect model. Secondly, it provides an integrated treatment of all ratings at a time. Finally, it develops a default rate distribution model that leads to a unified framework for testing calibration, mapping and rating discriminatory power. Important practical consequences derive from these proposals as outlined in the following paragraphs.

On calibration testing, some insights on the drivers of power are uncovered for the one-sided version. The feature of increasing differences between consecutive ratings is shown to be generally necessary for the achievement of reasonable levels of power. On the other hand, the effect of the correlation between the ratings, whose calibration is not present in Basel II, is shown to have only a minor effect on power. Also the lower PDs along the rating scale contribute more to the power final figure than higher PDs. A general message of the analysis is, however, that the power can be substantially low in some cases. Regarding this issue, strategies of power improvement are examined suggesting limited efficacy. Additionally, the paper discusses the conceptual problems of applying modern ideas in multivariate equivalence to the two-sided calibration test.

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 dependent on cardinal information of PDs, is less stringent than the requirement of perfect calibration and incorporates the assumption of 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.

All the tests discussed in this paper are based on the DRAM distribution. While DRAM is convenient for testing because it results in a non-degenerate multivariate normal distribution (thanks to the inclusion of additional rating-specific systemic factors), it has one main disadvantage: it is 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 is likely to be the case, for example, for small PDs (e.g.  $PD \leq 3\%$ ) and 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 the proposed test in those cases. On the other hand, when a liberal small sample bias is present, it may counterbalance the nominal conservatism, but caution should always be exercised.

Above all, the bank or the regulator should not demand much from statistical testing of CRMs. Even under the simplifying assumptions of DRAM, the power of these tests 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, it is the view of this paper that the possibility of reliance on qualitative aspects opened by the Basel Committee should not be seen as an excuse for not trying to get as much quantitative feedback as possible from statistical testing, including a quantitative sense of its uncertainty.

## 9. References

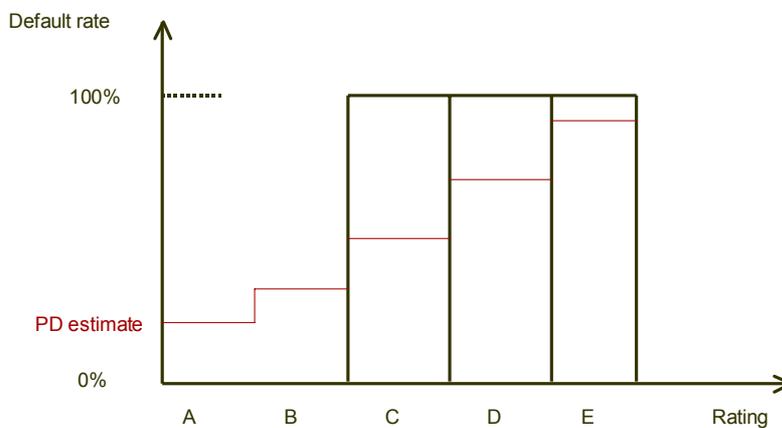
- Balthazar, L. (2004), "PD Estimates for Basel II", *Risk*, April 2004.
- Bamber, D. (1975), "The area above the ordinal dominance graph and the area below the receiver operating graph", *Journal of Mathematical Psychology*, 12, 387-415.
- Basel Committee on Banking Supervision (2004), "International 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 Limacon 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.
- Gupton, G. M., C. C. Finger and M. Bhatia (1997), "CreditMetrics -Technical Document", *New York: J.P. Morgan & Co. Incorporated*.
- 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.
- Mc.Dermott 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. Pflugger (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.
- 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.

## 10. 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 pushed 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 CreditMetrics 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.