Bank i Kredyt 47(3), 2016, 227-250

BANK I KREDYT
BANK & CREDIT

Rank-order statistics for validating discriminative power of credit risk models

Lukasz Prorokowski*

Submitted: 10 December 2015. Accepted: 19 April 2016.

Abstract

This paper provides practical insights into common statistical measures used to validate a model's discriminatory power for the probability of default (PD), loss liven default (LGD) and exposure at default (EAD). The review of available rank-order statistics is not based on analysing empirical data. Thus, the study has more of an informative value without delivering empirical evidence. When there is an alternative model available for comparison, this paper proposes to use the cumulative accuracy curve and the accuracy ratio to assess the rank-order ability for PD models given their popularity in practice. When there is no model available for comparison, due to the limited techniques in this area, this paper proposes to compare the confidence intervals in order to prove that a rating system has any discriminative power. For the LGD/EAD/slotting models, this paper recommends using a graph to check the rank-order ability. No statistical test is recommended. Focusing on enhancing practical implications for the financial industry, this paper advises banks on the existing CRR self-attestation requirements.

Keywords: credit risk, rank-order statistics, PD/LGD/EAD validation; CRR (CRD IV), AIRB banks

JEL: C18, C52, G10, G21, G28, G32

* H.L. Prorokowski LLC, San Francisco; e-mail: lukas.prorokowski@gmail.com.

1. Introduction

This paper provides practical insights into common statistical measures used to validate a model's discriminatory power for the probability of default (PD), loss given default (LGD), exposure at default (EAD) and slotting models. Assessing the discriminative power of wholesale credit risk models remains a priority for banks and regulators – Kraft, Kroisandt and Müller (2002). Against this backdrop, the Prudential Regulation Authority (PRA) has recently published the requirements for a validation process that should include standards of accuracy of calibration and discriminative power that are used to establish whether the models adhere to the standards of objectivity, accuracy, stability and conservatism (PRA 2013). These turgidly expressed standards are in line with the Capital Requirements Regulations (CRR) that ultimately aim to strengthen the ability of banks to predict and rank-order risk.[1]

Complementing the above information it should be noted that regulators increasingly demand banks to develop evidence that "the logic and quality of a rating system (including the quantification process) adequately discriminates between different levels of, and delivers accurate estimates of PD, Expected Loss, LGD and conversion factors" (Financial Conduct Authority 2015, p. 37). In reality, there is some ambiguity of how to assess the model's discriminatory power for regression models, such as EAD/LGD models. Some banks do not conduct back-testing of the model's discriminatory power for the regression problem (Linker, Van Baal 2008). In terms of regression models, discriminatory power is referred to as the capability to differentiate between high or low losses in selected model drivers.

Although the regulators have been demanding that banks identify factors in credit risk that have discriminatory power for over a decade now (BCBS 2001), no formal regulatory definition of the predictive or discriminatory power was given. As noted by Sobehart and Keenan (2004), credit risk model's performance is evaluated on an ex post basis that differentiates between the default and non-default state of the obligors in the underlying portfolio. At this point, the current paper defines discriminatory power as the ability to differentiate between defaults and non-defaults, or high and low losses.

As mentioned in the first paragraph, this paper reviews different statistical measures used for assessing the discriminatory power in the model validation phase. In order to reflect the reality of the banking sector and provide practical implications, two validation processes are analysed separately. Firstly, a scenario in which an alternative model is readily available for comparison is considered. Secondly, the case of an annual review of the model is applied to simulate a scenario of no alternative models being available for comparison. Figure 1 shows the logic behind utilising two separate scenarios. These model validation scenarios are designed to portray the bank's reality. At the credit risk model development/redevelopment, banks' analytics departments do not have rival models readily available. Building alternative PD or LGD models for comparison purposes remains problematic due to the organisational setup of the risk departments, the growing complexity of the business and the size of the underlying data. At the annual review process, given tight deadlines for the model sign-off and time constraints, banks are unable to develop rival models to compare the discriminatory power. Against this backdrop, the two validation scenarios presented in Figure 1 serve to deliver practical advice for practitioners involved in credit risk modelling and validation.

[1] Regulation (EU) No 575/2013 of the European Parliament and of the Council on prudential requirements for credit institutions and investment funds amending Regulation (EU) No 648/2012, Official Journal of the European Union, L 176/1, 27 June 2013.

Although a review of rank-order statistics under the two validation scenarios may seem of inferior value for an experienced academic, the issues discussed in this paper are of great importance to practitioners who benefit from understanding what validation measures for discriminatory power are available to them during the model re-development process. To this end, the main purpose of this paper boils down to presenting the chosen statistics for validating the discriminatory power of credit risk models in a practical way. The practical insights are delivered from a risk analyst's perspective with the focus on the Statistical Analysis Software-based (SAS)[2] implementation of rank-order measures.

This paper's scope of application is limited to a basic revision of the solutions and standards for assessing the rank-order ability of the credit risk models (wholesale). The review does not include any analysis conducted on empirical data. Thus, solutions for the standards of accuracy of calibration should form the subject of another study. In order to give a broad picture of the rank-order statistics for validating the discriminative power of credit risk models, this paper is structured as follows:

Section 2 provides a summary of rank-order statistics with a detailed description of their properties. Here the relevant literature is reviewed. This section also simulates different validation scenarios from a risk analyst's perspective (SAS-based). Against this backdrop, Section 2 is especially focused on investigating the use of graphical rank-order statistics and their practical implications. This section also presents distinct credit grading mappings utilised at the banks that participated in the study.

Section 3 provides a review of rank-order statistics that are being used by the credit risk wholesale teams at four major banks. This section attempts to bridge the theory presented in previous sections with the common practice of the banking industry.

Building on a survey of four international banks, this paper also investigates the rank-order statistics used at credit institutions. Table 1 shows the list of the banks participating in this study. The banks have been assured of their anonymity and their details are not disclosed in this paper. The survey was conducted in November 2015 and aimed at investigating the following issues:

– characteristics of the internally developed credit rating scales,
– ways of assessing the discriminatory power of credit risk models,
– rank-order statistics used for low default portfolios,
– rank-order statistics used during the model review process.

2. Review of rank-order statistics

2.1. Summary and application of rank-order statistics

There is a wide variety of grading methodologies in credit risk models (Engelmann, Hayden, Tasche 2003). With this in mind, banks are testing which of the rating systems are most appropriate for their credit risk exposures by employing new statistical measures that assess the quality of the rating systems. This practice is especially important when set against the current regulations. As it transpires, the sub-optimal credit rating systems lead to increased capital charges (Kraft, Kroisandt, Müller 2014).

Rank-order statistics evaluate the ordering of categories that is inherent in the nature of ordinal variables, rather than simple statistical independence. For example, credits that have low ratings

[2] SAS is a software suite for advanced analytics and is the most commonly used tool at banks and other credit institutions to interrogate data and perform statistical analysis.

today should on average prove to be more risky in the future, as opposed to the credits that have high ratings. Conversely, the lower value of the final rating on the credit rating scale denotes better credit quality. At the banks participating in this study, the internally developed credit rating scale for wholesale exposures is referred to as the master grading scale (MGS). The creditworthiness of a debtor is evaluated on the MGS that ultimately informs about the likelihood of a debtor's default. With this in mind, the rank of MGS = 27 denotes entities in default, whereas the rank of MGS = 1 is assigned to the debtors of the lowest probability of default. The participating banks have adopted the MGS framework as part of the move to Basel II. For the purpose of clarity, an internally developed credit rating scale is described as the MGS throughout the paper.

Since this paper defines the discriminatory power through the prism of defaults and non-defaults, Table 2 presents internally developed credit rating scales (MGS) at the participating banks mapped to the external ratings provided by Moody's and S&P. The mapping processes differ across the participating banks and entity types, as well as the bank-specific implied alphabet rating mappings. However, the general rule boils down to the lowest MGS value being assigned to the entities of the lowest probability of default (sovereigns, local authorities) and resulting in the best external rating.

Banks should understand that measures of the linear correlation between two credit ratings are not eligible for rank-ordering credit grades. In an example provided in Table 2, an entity graded on a master grading scale with the ranking of MGS = 2 is not twice as risky as the entity with a rating of MGS = 1. Therefore, in order to accommodate the need of accurate ordering comparison, banks should rely on the non-parametric rank-order correlation coefficients.

Firstly, it should be noted that all rank-order statistics described in this section measure the extent to which there is an inherent ordering between two grading points (MGS X and MGS Y), without accounting for the distance (true accuracy). Secondly, it is assumed that all statistical measures are appropriate only when both MGS points lie on an ordinal scale. Table 3 presents a brief review of the rank-order statistics that fall within the scope of this paper.

Complementing Table 3, banks should note that the Spearman correlation coefficient is defined as the Pearson correlation between the ranked variables (Myers, Well 2003). Goodman-Kruskal's gamma is a similar symmetrical statistic that measures the rank correlation coefficient, but ignores tied pairs (Goodman, Kruskal 1954). The tied pairs refer to the pairs of observations that have equal values of X or equal values of Y. Furthermore, with reference to Table 3, banks should note that larger values of Somers' d suggest that a model has better predictive power (Somers 1962). Additionally, Kendall and Gibbons (1990) as well as Newson (2006) have shown that the confidence intervals for Spearman's rho are less trustworthy and less interpretable than confidence intervals for Kendall's tau parameters. Therefore, as evidenced in the study by Newson (2002), Spearman's rho is not a suitable rank-based measure of correlation. All in all, the remaining rank-order measures differ only in their treatment of ties in the utilised pairs. Table 4 shows various types of pairs used in the rank-order statistics. From a practical point of view, this table is especially important for risk modellers, as it illustrates the ways of calculating the number of pairs based on the frequency distributions over three rating grades (AAA, BBB, C).

Unlike the correlation coefficient ranks presented in Table 5, the CAP validation technique provides a graphical illustration of the discriminatory power of a given rating process. To obtain a CAP curve, all debtors are first-ordered by their respective scores from riskiest to safest, that is, from the debtor with the lowest score and a bad credit rating (MGS = 27) to the debtor with the highest score and a good credit rating (MGS = 1). For a fraction x of the total number of debtors, the CAP curve is

constructed by calculating the percentage of the defaulters whose rating scores are equal to or lower than the maximum score of the fraction x. Following the methodology presented by Rezac and Rezac (2011), this is done for x ranging from 0% to 100%.

A perfect rating model will assign the lowest scores to the defaulters. In this case, presented in Figure 2, CAP is increasing linearly and then staying flat at the value of 1. For a random model without any discriminative power, the fraction x of all debtors with the lowest rating scores will contain x% of all defaulters. A real rating system will be placed somewhere between these two extremes (Hong 2009). The quality of a rating system is measured by the accuracy ratio (AR), which is also known as the Gini coefficient (Calabrese 2009). The AR is denoted by the following formula:

$$AR = \frac{a_R}{a_R + a_P} \tag{1}$$

where, a_R denotes the false ratings and a_P groups true ratings.

A visual inspection of Table 5 reveals that the proportion of the rejected defaulters vs. the proportion of all rejects can be easily derived. Rezac and Rezac (2011) state that one can see from the CAP curve that, for example, a decision to reject 70% of the defaulters implies a rejection of 40% of all credit applications.

As far as the ROC curve is concerned, Figure 3 shows possible distributions of rating scores for defaulting and non-defaulting debtors. For a perfect rating model, the left and right distribution should be separated. In the banking industry, however, a perfect rating system is not possible. Thus, banks are advised to use a cut-off value (point C in Figure 3) in order to ensure that debtors graded below the cut--off point are flagged as potential defaulters and the debtors with the rating scores above C are flagged as non-defaulters.

Figure 3 presents the hit rate $HR(c)$, which encompasses the number of defaulters predicted correctly with the cut-off point C. The total number of non-defaulters is denoted by N_D. The hit rate $HR(c)$ is defined as:

$$HR(c) = \frac{H(c)}{N_D} \tag{2}$$

where $H(c)$ denotes the number of defaulters that were classified correctly by the rating model and N_D stands for the total population of defaulters. The false alarm rate $FAR(c)$, derived from Table 5, is defined as:

$$FAR(c) = \frac{F(c)}{N_{ND}} \tag{3}$$

where $F(c)$ denotes the number of non-defaulters that were classified incorrectly as defaulters and N_D stands for the total population of non-defaulters in the credit portfolio.

The CAP curve and the accuracy ratio are closely related to the two concepts commonly used in research on income inequality: the Lorenz curve and the Gini coefficient. The Lorenz curve shows how much of a population's cumulative income accrues to each cumulative proportion of the population, ordered from the poorest to the richest, and thus shows how equally income is distributed across the population (Zenga 2007).

In credit risk validation, banks should apply the CAP curve as a measurement of the discriminatory power of credit rating systems. Furthermore, in order to avoid any ambiguity, this paper recommends using the accuracy ratio or Somers' d instead of the Gini coefficient in this setting. It is proved by Hamerle, Rauhmeier, Rösch (2003) and Engelmann, Hayden, Tasche (2003) that for the binary response, the accuracy ratio (*AR*) and the area under the ROC curve (*AUC*) are equivalent measures:

$$AR = 2(AUC - 0.5) \tag{4}$$

The above equation means that the accuracy ratio can be calculated directly from the area below the ROC curve and vice versa. To this point, both summary statistics contain the equivalent information (Engelmann, Hayden, Tasche 2003; Satchell, Xia 2006).

To sum up, either the accuracy ratio or the area under the curve (*AUC*) is a random variable itself. It is dependent on the portfolio structure, the number of defaulters, the number of rating grades etc. The metric is influenced by what it is measuring and should not be interpreted without knowledge of the underlying portfolio. Therefore, both the accuracy ratio and the area under the curve are not comparable across models developed from different samples (Blochwitz et al. 2004). The accuracy ratio is not invariant across samples and its outcome should only be presented as an indicator and not as an absolute measure of performance. The outcomes of the performance measures strongly depend on the structure of the true default probabilities in the underlying portfolio. Thus, the magnitudes are not interpretable regarding the discriminatory power of the rating system and the rating systems cannot be compared across time and portfolios.

2.2. Validation scenarios and rank-order statistics

Having discussed the properties of different rank-order statistics, this section advises on appropriate measures that can be applied under the two validation scenarios described in the Introduction:

– an alternative credit risk model is available for comparison purposes;

– an alternative credit risk model is not available for comparison purposes.

When a rival model is available, Table 6 provides a summary of rank-order statistics for the probability of default (PD), loss liven default (LGD), exposure at default (EAD) and slotting models.

Complementing Table 6, it should be noted that for a PD model, the rank-ordering test confirms the first function by analysing whether the model rank-orders discriminate well between debtors. However, this test does not ensure the right magnitude of default rates. In the case when the realised event is binary, the CAP curve or the ROC curve provide an equivalent graphical illustration of the discriminatory power of a PD model (Tasche 2009). Given the CAP curve is more frequently used by many financial institutions, CAP and its corresponding summary statistics (accuracy ratio) are suggested in this paper to compare the discriminatory power between more than one PD model on the same dataset.

When comparing the rank-order ability between two PD models, a test method proposed by DeLong, DeLong, Clarke-Pearson (1988) can be applied to check whether there is a statistical difference between two rating systems by comparing areas below the ROC curve (AUC 1 and AUC 2).

For LGD models, both the academics and practitioners are focused on the accuracy (distance) between the realised value and the model prediction for model validation. In spite of this fact, the rank--order measures are still proposed in this paper based on the different LGD framework:

– LGD model with constant prediction;

– LGD model with continuous prediction.

In the first case, an LGD model with two benchmarks for each segment is analysed. Such defined LGD models are commonly used in banks for specialised lending (e.g. shipping LGD or project finance LGD). Often an LGD model would have a regulatory benchmark. For example, a shipping LGD model can define two benchmarks: 35% LGD for exposure fully collateralised; and 45% LGD for ship collateral with some negative characteristics. These two segments are the drivers that should be tested for their discriminatory power to distinguish high and low losses. With this in mind, the current paper recommends using a simple graph as a minimum requirement to illustrate the discriminatory power between different segments. When the number of segments is above 5, Somers' d is also suggested to report for increased information purposes. Given such a small sample size, no statistical test is recommended, as advised by Siegel (1957) and Fritz, Morris, Richler (2012).

In the second case, a model returns continuous LGD values instead of categorical estimates (e.g. large corporate LGD or sovereign LGD). Against this backdrop, this paper proposes to compare the discriminatory power by employing the graphical approach. Based on the segments within the model, a graph constitutes a sufficient solution to compare the discriminatory power. When the number of segments is above 5, Somers' d is also suggested to report for increased information purposes. Given such a small sample size, banks should interpret Somers' d with caution and no statistical test is recommended.

For an EAD model, the rank-ordering relies on assessing the discriminatory power of the credit conversion factors (CCF) based on different product types. Introduced by Basel II regulations, the CCFs convert the amount of a free credit line and off-balance sheet exposures to an EAD amount. Depending on the limited number of product types (normally smaller than 10), a graph is suggested as a minimum to compare the discriminatory power. This suggestion is justified by the conclusions on using graphs for discriminating between credit scores reached by Rezac and Rezac (2011).

For a slotting model, there is a limited number of slots, for example: strong, good, satisfactory, weak. Therefore, a graphical approach is suggested as a necessary minimum to assess the discriminatory power. Somers' d is also recommended to report for information purposes, but the results should be interpreted with caution.

When a rival model is not available, Table 7 provides a summary of rank-order statistics for the probability of default (PD), loss liven default (LGD), exposure at default (EAD) and slotting models.

Since the CAP and ROC curves are not comparable across time and portfolio, for a PD model, banks can only report the confidence interval of the accuracy ratio (AR) to test if a rating system has any discriminative power. In this case, the null hypothesis would be AR (or Gini coefficient) = 0.

For an LGD model with a constant prediction for each segment, a graph is recommended for sense--checking the discriminatory power between model drivers. Given the small sample size of segments (normally below 10), no statistical test is recommended. This assumption is in line with the findings delivered by Van der Burgt (2008).

When the model predicted LGDs are continuous instead of categorical, depending on the model assumptions, two methods are proposed in this paper to compare the model's discriminatory power:

1. Graphical approach. Based on the segments/buckets within the model, a graph is suggested as a sense-check exercise. Given the small sample size of segments (normally below 10), no statistical test is recommended.

2. Simulation approach. For the continuous LGD that is predicted from a mixed model (such as Tobit model), Monte Carlo methods should be used to generate predicted LGDs. In order to account for the random variation from the model estimation, Monte Carlo methods should also be used to calculate the Somers' d (between the realised LGD and the predicted LGD) for each simulation. Then, as suggested by Luo and Shevchenko (2013), the 2.5^{th} and 97.5^{th} percentile confidence intervals should be used in order to test if the model framework has any discriminative power.

For an EAD model, in order to assess the discriminatory power of the credit conversion factors based on each product type, a graph is suggested as a minimum to assess the rank-order abilities of an EAD model. Given the small number of product types, no statistical test is recommended.

For a slotting model, based on the limited number of slots in the model, a graph is suggested to sense-check the model's discriminatory power. Given the small sample size of slots (normally below 10), no statistical test is recommended.

3. Use of rank-order statistics in practice

This section investigates the types of rank-order statistics that are being used by the credit risk wholesale teams to access the model's discriminatory powers. The report is especially focused on solutions for PD/LGD/EAD/slotting models. The findings reported in this section are based on the survey of four internationally present banks (Table 1). As shown in the previous sections, there are many possibilities to assess the discriminatory powers of the aforementioned models. The rank-order statistics examples include:

– Gini coefficient,
– accuracy ratio,
– ROC curve,
– Somers' d,
– Spearman's tau,
– Kolmogorov-Smirnov test,
– Kendall's tau,
– other validation statistics.

3.1. Accessing discriminatory powers of models

For PD models, Bank 1 has retained the accuracy ratio to assess the performance of a PD model. For LGD models, Bank 1 uses the following formula:

$$\text{LGD} = \text{TR} \times (\text{LGRD} + \text{costs RC}) + \text{costs IC} \tag{6}$$

At Bank 1, the LGD models are made of three parts:

1. Loss given recovery default (LGRD) – calculates the loss rate for a defaulted credit reaching the final and most severe possible status with the recovery of a claim.

2. Transition rate (TR) – assesses the probability that a Basel-defined default (90 days past due, contagion related to another default of the same client) will evolve to a recovery status.

3. Cost components – combines the managing and legal fees, funding cost and recuperated punitive interests. Hereto, Bank 1 makes a difference between the costs for recovery defaults (costs RC) and those for every Basel default (costs IC). The cost components are updated annually based on the most recent figures.

At Bank 1, the LGD model performance is first assessed for each part separately. Since LGRD has a continuous target variable, the main indicator of the LGRD's performance is the Pearson's R^2 (and associated F-test). Transition models target a binary variable: the debtor being either in the recovery or cured stage. Therefore, to assess the performance in this case, Bank 1 retains the accuracy ratio. Even if assessed separately, once all different parts of the LGD model are put together, the final outcome is also assessed. In doing so, Bank 1 uses the same indicator as for the LGRD, because the global LGD model has a continuous target variable. Finally, for the EAD models, Bank 1's main performance indicator is the Pearson's R^2 and associated F-test.

Bank 3 and Bank 4 are at the stage of implementing the Somers' d statistics for the LGD model and use the accuracy ratio for PD models.

3.2. Low default portfolio

The surveyed banks utilise various rank-order statistics for the low default portfolios. Bank 1, for portfolios containing less than 100 defaults, uses the accuracy ratio value. However, Bank 1 exercises caution while considering this statistic and uses it only for the general guidance. Interestingly, Bank 2 benchmarks its low default portfolios to internal and external ratings historically assigned to the debtors. Bank 3 and Bank 4 use the same rank-order statistics as for the PD/LGD/EAD models, but the minimum number of defaults is 10. If the portfolio has less than 10 defaults, no statistical tests are recommended at Bank 4.

3.3. Model review

When conducting a model review, banks need to access the model's discriminatory powers. Problems emerge when there is only one model being reviewed. Here, Bank 1 is using the same performance indicators for a model review as for the PD/LGD/EAD models. Bank 2 considers alternatives at every review. Bank 4 is still looking for the best solution in this area. Currently, Bank 3 and Bank 4 calculate the confidence intervals of the rank-order statistics to see whether they have any discriminatory power.

It has emerged from the qualitative query that banks are still investigating the use of various rank-order statistics for model validation purposes. Although these statistical measures have been available for decades, the surveyed banks are currently at the development stage of implementing the appropriate rank-order statistics to assess the discriminatory power of credit risk models. Therefore, the accuracy ratio remains the first choice among other solutions in both the case where the surveyed

banks have an alternative model for comparison (model development/redevelopment phase) and the case where a rival model is not available (annual review process). At this point, the qualitative findings are in line with the recommendations made in the previous section. All in all, applying rank-order statistics at the model review phase continues to be problematic for the participating banks.

4. Conclusions

This paper has reviewed available rank-order statistics (including cumulative accuracy curve and receiver operating characteristic curve) and recommends appropriate measures for practitioners to assess the discriminatory power of PD/LGD/EAD and slotting models used in credit risk analytics. These statistics are discussed separately in terms of two conditions, when there is an alternative model or there is no rival model for comparison. The recommendations made in this paper are also set against the qualitative findings that report on the use of rank-order statistics at four global banks.

Where there is an alternative model readily available to the risk analyst for comparison purposes, this paper proposes to use the cumulative accuracy curve and the accuracy ratio to assess the rank--order ability for PD models given their popularity in practice. It has emerged from the qualitative query that the accuracy ratio remains the most commonly used rank-order measure for the discriminatory power of credit risk models. Recognising the bank's limitations in applying rank-order correlation coefficients, the graphical approach is also recommended for LGD/EAD and slotting models.

Where there is no rival model available to the risk analyst for comparison purposes, due to limited techniques in this area, this paper proposes to use the graphical approach in order to prove that a rating system has any discriminatory power. Under this scenario, the use of graphs should be extended to cover LGD/EAD and slotting models. Following the established theories by recent studies, no statistical test is recommended.

This paper has emphasised several risk-modelling issues that revolve around the application of rank-order statistics. Firstly, the two popular graphical summary statistics, namely the accuracy ratio and the area under the ROC curve contain the same information. When one of the assessing variables is binary, the accuracy ratio is equivalent to Somers' d. In addition to this, the paper shows that for the two variables being continuous, the Goodman-Kruskal's gamma, Kendall's tau (a) (b) and Somers' d provide equivalent information.

Secondly, reviewing the theoretical background for the use of rank-order statistics, this paper reports that numerous versions and uses of the Gini coefficient exist in the academic literature and banking practice. Against this backdrop, this paper recommends using the accuracy ratio (or Somers' d) instead of the Gini coefficient in order to avoid any ambiguity.

Thirdly, both the accuracy ratio and the AUC are influenced by what they are measuring, and hence should not be interpreted without detailed knowledge of the underlying portfolio. With this in mind, this paper shows that these measures are not comparable across models developed from different samples and the magnitudes are not interpretable regarding the discriminatory power of the rating system.

Complementing the review of the available rank-order statistics, this paper concludes that a risk analyst willing to compare the discriminatory power of different models applied to the same data should not limit the comparison to the coefficient of correlation only. With this in mind, bank

practitioners are advised to state the confidence interval of the difference between two credit rating models. In doing so, greater insights are gained into the quality of a credit rating system. A simple comparison of the two confidence intervals can be misleading because potential correlation of both rating systems is neglected in such circumstances. This paper reports that DeLong, DeLong and Clarke-Pearson (1988) provide a method of comparing the discriminatory power of two models built around the same data. The scholars have developed a way of comparing two rating systems by constructing a statistical test for the difference between the areas under the curve (AUC 1 and AUC 2). To carry out the test on the difference between the two rating systems, banks should state that the null hypothesis refers to the equality of both areas below the ROC curve. Then, in the next step, banks should evaluate test statistics which are asymptotically distributed. Regarding the minimum sample size for this test, Engelmann, Hayden and Tasche (2003) simulated the real data example and proved that the results can be reliable also for a small portfolio.

This study is limited to the assessment of the discriminatory power (rank-order ability) of the wholesale credit risk models. In terms of accuracy of calibration, new research into the model development and validation standards is recommended in order to inform about current credit risk modelling challenges. Furthermore, reference to the concept of a low default portfolio is evident in the current paper. However, issues centred on developing appropriate risk rating systems for the low default portfolios are not investigated in detail. Therefore, a new study is needed to complement the findings reported in this paper by investigating the rank-order statistics for the low default portfolios. This is especially important as regulators expect banks to provide any confidence in statistical measures of the discriminatory power for portfolios characterised by insufficient default experience. At this point, it is worth assessing the appropriateness of a common practice of benchmarking a bank's credit rating system (comparing MGS scores using an implied alphabet) with the external rating systems (S&P, Moody's). This practice has been affected by the fact that the rating agencies started to withdraw their ratings for the common components of the low default portfolios (e.g. housing associations, universities, local authorities, and municipalities). As a result, the low default portfolios are rarely externally rated since 2013.

To sum up, by focusing on bridging theory and practice, this research has more informative value than scientific merit. However, by delivering practical insights into the use of rank-order statistics, this paper advises banks on the existing CRR self-attestation requirements in the area of model validation. In doing so, the current paper suggests the stylised responses to the relevant rules in the CRR self-attestation exercise (Table 11 and Table 12). Thus, banks are provided with a self-assessment tool to test their regulatory compliance in the area of validating their own credit risk models.

Considering the fast pace of regulatory change, this paper aims to highlight the practical and CRR-compliant ways of assessing banks' internal credit risk models. With this in mind, the use of rank-order statistics and their application in practice is discussed from a practitioner's perspective. The paper attempts to further augment its practicality by focusing on the regulatory framework for model validation. Adding practical implications to the model validation processes, Table 11 provides additional guidance on addressing the aforementioned regulations by sketching a framework for the CRR self-attestation exercise in the area of validating the discriminative powers of credit risk models (LGD framework). Table 12 provides similar guidance for a PD framework.

References

BCBS (2001), *The New Basel Capital Accord, consultative document BCBS,* 31 May, Banking Committee on Banking Supervision, Bank for International Settlements, http://www.bis.org/publ/bcbsca03.pdf, accessed on 11 September 2015.

Blochwitz S., Hamerle A., Hohl S., Rauhmeier R., Rosch D. (2004), *Myth and reality of discriminatory power for rating systems*, 27 July, http://ssrn.com/abstract=2350369, accessed on 11 September 2015.

Calabrese R. (2009), *The validation of credit rating and scoring models*, Swiss Statistics Meeting Paper, 29 October.

DeLong E.R., DeLong D.M., Clarke-Pearson D.L. (1988), Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*, 44(3), 837–845.

Engelmann B., Hayden E., Tasche D. (2003), *Measuring the discriminative power of rating systems*, Deutsche Bundesbank Discussion Paper, Series 2: Banking and Financial Supervision, 01/2003.

Financial Conduct Authority (2015), *Prudential sourcebook for investment firms*, IFPRU Chapter 4. Credit risk, https://www.handbook.fca.org.uk/handbook/IFPRU/4.pdf, accessed on 11 September 2015.

Fritz C.O., Morris P.E., Richler J.J. (2012), Effect size estimates: current use, calculations, and interpretation, *Journal of Experimental Psychology: General*, 141, 2–18.

Goodman L.A., Kruskal W.H. (1954), Measures of association for cross classifications, *Journal of the American Statistical Association*, 49(268), 732–764.

Hamerle A., Rauhmeier R., Rösch D. (2003), *Uses and misuses of measures for credit rating accuracy*, 28 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2354877, accessed on 11 September 2015.

Hayden E., Tasche D. (2003), Testing rating accuracy, *Risk Magazine*, 03 January.

Hong C.S. (2009), Optimal threshold from ROC and CAP curves, *Communications in Statistics*, 38(10), 2060–2072.

Kendall M.G., Gibbons J.D. (1990), Rank correlation methods, Oxford University Press.

Kraft H., Kroisandt G., Müller M. (2002), *Assessing the discriminatory power of credit scores*, Humboldt Universitat zu Berlin Working Paper, 15 August, http://edoc.hu-berlin.de/series/sfb-373-papers/2002-67/PDF/67.pdf, accessed on 11 September 2015.

Kraft H., Kroisandt G., Müller M. (2014), Redesigning ratings: assessing the discriminatory power of credit scores under censoring, *Journal of Credit Risk*, 10(4), 71–94.

Linker D., Van Baal G. (2008), *Backtest report LGD model 2007 VW Bank*, *Capgemini Annual LGD Model Review Report,* Volkswagen Pon Financial Services.

Luo X., Shevchenko P.V. (2013), *Markov chain Monte Carlo estimation of default and recovery: dependent via the latent systematic factor*, CSIRO Mathematics Working Paper, 10 April.

Myers J.L., Well A.D. (2003), *Research design and statistical analysis*, Lawrence Erlbaum.

Newson R. (2002), Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences, *The Stata Journal*, 2(1), 45–64.

Newson R. (2006), Efficient calculation of Jackknife confidence intervals for rank statistics, *Journal of Statistical Software*, 15(1), 1–10.

PRA (2013), *Internal ratings based approaches*, Prudential Regulation Authority, Bank of England Supervisory Statement, SS11/13.

Rezac M., Rezac F. (2011), How to measure the quality of credit scoring models, *Czech Journal of Economics and Finance*, 61(5), 486–507.

Satchell S., Xia W. (2006), *Analytic models of the ROC curve: applications to credit rating model validation*, Quantitative Finance Research Centre Paper, 181, University of Technology Sydney.

Siegel S. (1957), Nonparametric statistics, *The American Statistician*, 11, 13–19.

Sobehart J., Keenan S. (2004), The score for credit, *Risk Magazine*, 2 March.

Somers R.H. (1962), A new asymmetric measure of association for ordinal variables, *American Sociological Review*, 27(6), 799–811.

Tasche D. (2009), *Estimating discriminatory power and PD curves when the number of defaults is small,* http://arxiv.org/pdf/0905.3928v2.pdf, accessed on 11 September 2015.

Van der Burgt M. (2008), Calibrating low-default portfolios, using the cumulative accuracy profile, *Journal of Risk Model Validation*, 1(4), 17–33.

Zenga M. (2007), *Inequality curve and inequality index based on the ratio between lower and upper arithmetic means*, Universita delgi Studi di Milano-Bicocca Working Paper, http://dipeco.economia.unimib.it/web/pdf/iniziative/Zenga.pdf, accessed on 11 September 2015.

Appendix

Figure 1
Two validation processes

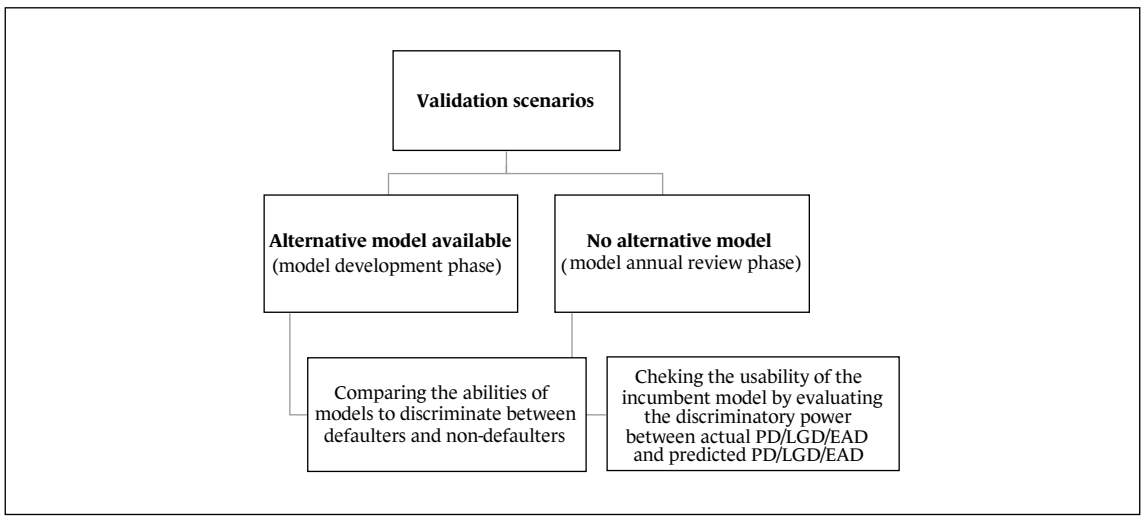
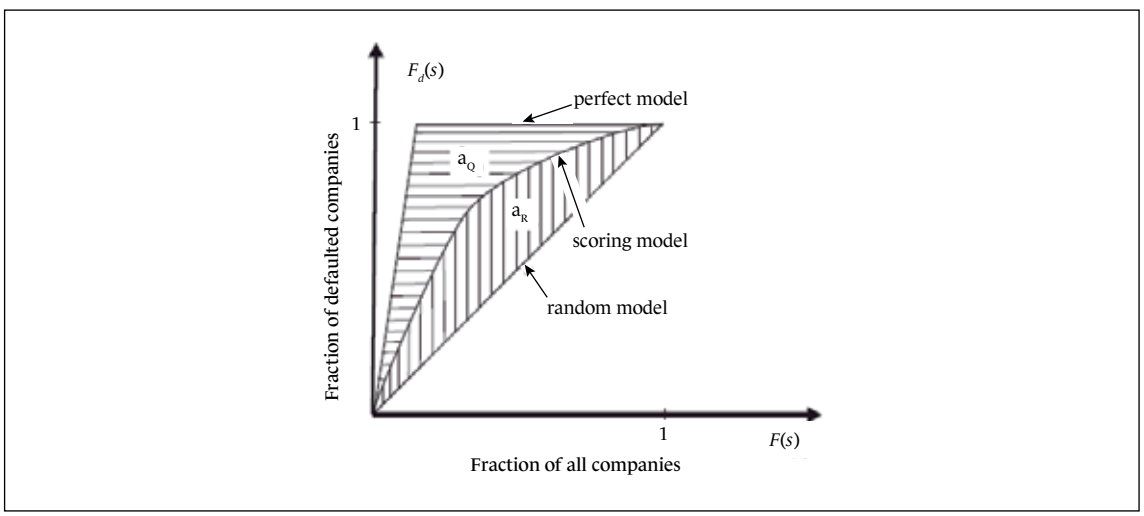
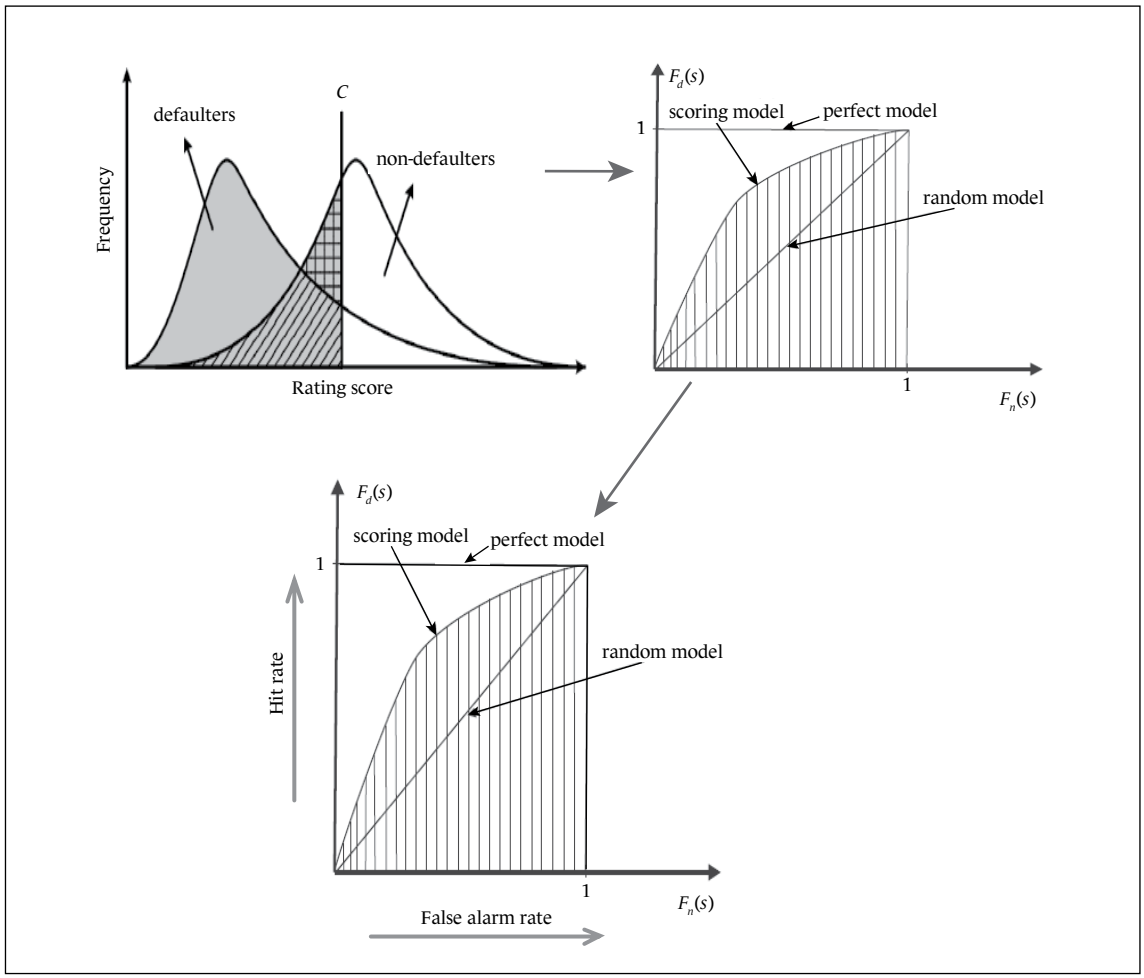


Figure 2
Cumulative accuracy profiles



Source: Calabrese (2009).

Figure 3
Distribution of rating scores for defaulting and non-defaulting debtors



Source: Engelmann, Hayden, Tasche (2003).

Table 1
Banks participating in this research

Bank	Description
Bank 1	Tier-1 universal bank, Europe
Bank 2	Tier-1 universal bank, North America
Bank 3	Tier-1 universal bank, Europe
Bank 4	Tier-2 universal bank, Europe

Table 2
Credit grading mappings

Static mapping used for sovereign entities or low default portfolios								
Master gradings	Bank 1		Bank 2		Bank 3		Bank 4	
MGS	S&P	MD	S&P	MD	S&P	MD	S&P	MD
1	AAA	Aaa	AAA	Aaa	AAA	Aaa	AAA	Aaa
2	AAA	Aaa	AA+	Aa1	AA+	Aa1	AA+	Aa1
3	AAA	Aaa	AA+	Aa1	AA+	Aa1	AA+	Aa1
4	AA+	Aa1	AA	Aa2	AA	Aa2	AA	Aa2
5	AA	Aa2	AA	Aa2	AA	Aa2	AA	Aa2
6	AA-	Aa3	AA-	Aa3	AA-	Aa3	AA-	Aa3
7	A+	A1	A+	A1	A+	A1	A+	A1
8	A-	A3	A	A2	A	A2	A-	A3
9	BBB+	Baa1	A-	A3	A-	A3	BBB+	Baa1
10	BBB	Baa2	BBB+	Baa1	BBB+	Baa1	BBB	Baa2
11	BBB	Baa2	BBB	Baa2	BBB	Baa2	BBB	Baa2
12	BBB-	Baa3	BBB-	Baa3	BBB-	Baa3	BBB-	Baa3
13	BBB-	Baa3	BB+	Ba1	BB+	Ba1	BB+	Ba1
14	BB+	Ba1	BB+	Ba1	BB+	Ba1	BB	Ba2
15	BB	Ba2	BB	Ba2	BB	Ba2	BB	Ba2
16	BB-	Ba3	BB-	Ba3	BB-	Ba3	BB-	Ba3
17	BB-	Ba3	BB-	Ba3	B+	B1	B+	B1
18	B+	B1	B+	B1	B+	B1	B+	B1
19	B+	B1	B	B2	B	B2	B	B2
20	B	B2	B	B2	B	B2	B	B2
21	B-	B3	B-	B3	B-	B3	B-	B3
22	B-	B3	B-	B3	B-	B3	CCC+	Caa1
23	CCC+	Caa1	CCC+	Caa1	CCC+	Caa1	CCC+	Caa1
24	CCC+	Caa1	CCC+	Caa1	CCC+	Caa1	CCC	Caa2
25	CCC	Caa2	CCC	Caa2	CCC	Caa2	CCC-	Caa3
26	CCC	Caa2	CCC-	Caa3	CCC-	Caa3	CC	Caa3
27	D	C	D	C	D	C	D	C

Table 3
Review of rank-order statistics

Statistic	Brief description
Spearman's rho	The calculation is based on deviations between two ranked variables. Thus, this measure is more sensitive to errors and discrepancies in data
Goodman – Kruskal's gamma	The estimator of gamma is based only on the number of concordant and discordant pairs of observations
Kendall's tau (a)	Kendall's tau is also based on the concordant and discordant pairs without adjustments for ties. Given its symmetric nature, no distinction is made between the independent and dependent variable
Kendall's tau (b)	Kendall's tau (b) is the most widely used measure of the different versions of tau. Compared with tau (a) and gamma, it makes adjustments for tied ranks
Somers' d	Somers' d is also a similar ordinal association indicator, which differs from tau (b) in that it makes a correction for tied pairs on the independent variables. Therefore, it becomes an asymmetric statistics

Table 4
Calculation of different types of pairs

Type of pair	Number of pairs	Symbolic designation
Concordant	$a(e+f)+b(f)$	P
Discordant	$c(d+e)+b(d)$	Q
Tied on X only	$ad+be+ef$	Xo
Tied on Y only	$a(b+c)+bd+d(e+f)+ef$	Yo
Tied on both X and Y	$1/2\ \{a(a-1)+b(b-1)+c(c-1)+d(d-1)+e(e-1)+f(f-1)\}$	Z
Total	$1/2\ \{n(n-1)\}$	Tn

a, b, c – non-default frequency distributions; *d, e, f* – default frequency distributions.

Table 5
Summary of the rank-order correlation coefficients

Statistic	Ties	Symmetric / asymmetric	Range	Available in SAS	Interpretation	Formula
Spearman's rho	No	Symmetric	[-1, 1]	Yes	The sign and absolute value of the Spearman correlation indicates the direction and magnitude of association between X (the independent variable) and Y (the dependent variable)	$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ $D_i = x_i - y_i$ – the difference between ranks n – sample size
Kendall's tau (a)	No	Symmetric	[-1, 1]	Yes	It tells the proportionate excess of concordant over discordant pairs among all possible pairs	$tau\_a = \frac{P-Q}{Tn}$ P – the number of concordant pairs Q – the number of discordant pairs Tn – the function of not tied pairs
Kendall's tau (b)	Yes	Asymmetric	[-1, 1]	Yes	No operational interpretation is available, which means that one cannot express in words (in terms of probabilities, errors in prediction, etc.) the meaning of a value equals to 0.6	$tau\_b = \frac{P-Q}{\sqrt{(P+Q+Xo)(P+Q+Yo)}}$ P – the number of concordant pairs Q – the number of discordant pairs X/Y – target features
Goodman and Kruskal's gamma	No	Symmetric	[-1, 1]	Yes	It tells the proportionate excess of concordant over discordant pairs among all pairs which are fully discriminated or fully ranked	$gamma = \frac{P-Q}{P+Q}$ P – the number of concordant pairs Q – the number of discordant pairs
Somers' d	Yes	Asymmetric	[-1, 1]	Yes	It expresses the proportionate excess of concordant pairs over discordant pairs among pairs not tied on the independent variable	$d_{yx} = \frac{P-Q}{P+Q+Yo}$ $d_{xy} = \frac{P-Q}{P+Q+Xo}$ P – the number of concordant pairs Q – the number of discordant pairs X/Y – target features

Table 6
Recommended rank-order statistics when alternative model is available

Model	Characteristic	Example	Current statistics for rank-order ability	Recommend statistics (minimum)	Additional test (for information only)	Rational	Limitations
PD	Compare the continuous PD with the binary realised default/ non-default	Large corporate PD	(1) CAP (2) Spearman's rho / Pearson correlation coefficient	CAP and accuracy ratio (Somers' d)	Statistical test between two rating systems by using ROC curves	(1) For binary response, Somers' d is equivalent to AR ratio (also called Gini by some firms) (2) CAP provides a graphical illustration of the discriminatory power and is more intuitive to understand (3) CAP and AR/Gini is also the most popular rank-order used by many financial institutions, such as Moody's and EY	(1) These are the proposed measures for assessing the model's rank-order ability, not the accuracy between realised values and predicted values (2) These are not applicable for LDP (3) CAP & ROC curves and their corresponding summary statistics AR and AUC are not comparable across time and across portfolio
LGD (benchmark)	For each segment, compare the model benchmark (a fixed value) with the continuous realised LGD	Shipping model (35% /45% LGD)	Spearman's rho	Graph comparison for each segment	Somers' d*	(1) When there is a constant (flat) prediction for each segment, each segment (or the way of classification) is the driver of offering discriminatory power of the model (2) Given the small number of segments (usually smaller than 10), Somers' d is just calculated for information only	
LGD (continuous)	Compare the continuous predicted LGD with the continuous realised LGD	Large corporate LGD		Graph comparison for each segment	Somers' d*	Given the small number of segments (usually smaller than 10), Somers' d is just calculated for information only	
EAD model	Compar the predicted CCF factors (by product type) vs. realised continuous CCF	Bank EAD	Spearman's rho/ Pearson correlation coefficient	Graph comparison for each product type	Somers' d*	A graph is suggested as a minimum to compare the average realised CCF vs. predicted CCF for different product types (a fixed prediction for each product type)	
Slotting model	Compare the ordinal bucket mapping with the continuous realised expected loss	Property slotting model	Somers' d	Graph comparison for each segment	Somers' d*	Based on the slots within the model, a graph is suggested as a minimum to compare the discriminatory power (as in the LGD benchmark model)	

* The minimum required number of defaults is 5.

Table 7
Recommended rank-order statistics when alternative model is not available

Model	Characteristic	Example	Current statistics for rank-order ability	Recommend statistics (minimum)	Rational	Limitations
PD	Compare the continuous PD with the binary realised default / non-default	Large corporate PD model	There is no suggested rank-order statistics in PD/LGD/EAD standard when no alternative model is available	The CI for AR to test if a rating system has any discriminative power at all	(1) There is not much method available in this area (2) The CI for AR (also called Gini by some firms) coefficient reply on the asymptotical normal distribution. The minimum required number of defaults is 10	(1) These are the proposed measures for assessing the model's rank-order ability, not the accuracy between realised values and predicted values (2) These are not applicable for LDP (3) CAP & ROC curves and their corresponding summary statistics AR and AUC are not comparable across time and across portfolios
LGD (benchmark)	For each segment, compare the model benchmark (a fixed value) with the continuous realised LGD	Shipping model (35% /45% LGD)		Graph comparison for each segment	(1) When there is a constant (flat) prediction for each segment, then each segment (or the way of classification) is the driver of offering the discriminatory power of the model (2) Given the small number of segments. No statistical test is recommended here	
LGD (continuous)	Compare the continuous predicted LGD with the continuous realised LGD	Large corporate LGD		Graph comparison for each segment	Based on the segments/buckets within the model, a graph is suggested as a sense-check. No statistical test is recommended here	
				Graph comparison for each segment	Given the small number of segments (usually smaller than 10), Somers' d is just calculated for information only	
EAD model	Compared the predicted CCF factors (by product type) vs. realised continuous CCF	Bank EAD		Graph comparison for each product type	Based on the segments/buckets within the model, a graph is suggested as a sense-check. No statistical test is recommended here	
Slotting model	Compare the ordinal bucket mapping with the continuous realised expected loss	Property slotting model		Graph comparison for each segment	Based on the slots within the model, a graph is suggested as a sense-check. No statistical test is recommended here	

Table 8
What kind of rank-order statistics are used to access the model's discriminatory power?

Bank	Rank-order statistics
Bank 1	PD models: accuracy ratio LGD models: Pearson's tau and associated F-test EAD models: Pearson's tau and associated F-test
Bank 2	PD models: CAP/ Gini coefficient LGD models: collateral-level Kendall tau (b) EAD models: cumulative EAD accuracy ratio / cumulative accuracy profile
Bank 3	PD models: cumulative accuracy profiles / accuracy ratio LGD models: Somers' d / graphical approach EAD models: visual graph check
Bank 4	PD models: accuracy ratio LGD models: Somers' d EAD models: visual graph check

Table 9
What kind of rank-order statistics are used for low default portfolios?

Bank	Rank-order statistics
Bank 1	Accuracy ratio
Bank 2	Benchmarking to internal and external ratings
Bank 3	Accuracy ratio / visual graph check
Bank 4	Accuracy ratio / cumulative accuracy profiles / visual graph check

Table 10
What kind of rank-order statistics are used for model review?

Bank	Rank-order statistics
Bank 1	Accuracy ratio Pearson's tau and associated F-test
Bank 2	Various alternatives considered
Bank 3	Various alternatives considered
Bank 4	In development

Table 11
Model responses for CRR self-attestation (LGD framework)

Regulation	Reference	Suggested response to the regulation (areas for consideration)
PRA SS11/13	13.14i	Geographical discrimination should be considered relevant/not relevant to the underlying portfolio The extent of the coverage of the model should be considered to decide whether further granularity in collateral type is required
PRA SS11/13	10.9	All available internal loss data should be used, which does not suggest additional discrimination in the model, beyond the use of the LGD benchmarks External data should be used only for benchmarking purposes due to limitations of representativeness of the sample and lack of granularity by facility type
CRR	174a	The model under consideration should be a statistical model, so the requirements specified by CRR 174 apply Models should be selected based on their predictive power ('rank-ordering ability') and accuracy ('accuracy of calibration'). The model under consideration should be based on the average realised LGDs with a margin of prudence built into it Accuracy at model development/re-development should be demonstrated and in the latest annual review Analysis should be conducted to check if it is possible to test the model's discriminatory power when defaults are experienced against each of the predicted LGD values The developments tandards should be designed to require input variables to be selected based on a credible assumption that they logically determine the LGDs For the model under consideration the conceptual soundness of the selected input variables should be discussed and empirical evidence for their effectiveness in the model development document should be provided A flat LGD model should be proved reasonable given the paucity of default observations and lack of theoretical support for a differentiation Appropriateness of the model's calibration should be confirmed. The calibration to remain prudent vis-à-vis realisations. Any open cases should be placed in the caveats The model under consideration should exhibit no material biases
PRA SS11/13	17.1b 17.1c	The development standards should define the relevant measures of accuracy of calibration (e.g. McFadden pseudo R-squared; deviation of predicted vs. realised portfolio default rates) and discriminative power (e.g. Spearman rank correlation) A process should be in place to establish whether an LGD model meets the above standards
PRA SS11/13	17.2	Article provides definitions to terms used in PRA SS11/13 17.1d. Compliance assessment not required for this article

PRA SS11/13	17.3	If the model produces only two benchmark LGDs (e.g. 35% and 45%), the granularity should be improved to capture the key drivers of risk influencing recovery. This process should be supported by the available loss experience and by conceptual arguments supporting the proposed drivers and discarding other potential drivers of loss (e.g. collateral coverage prior to default) If it is not possible to benchmark the model's discriminatory power against external benchmarks due to the limitations of these benchmarks, an investigation should be launched to check if alternative models envisaging higher granularity do offer demonstrable improvements over the proposed model
PRA SS11/13	17.5	Explanation should be provided as to why the model under consideration has limited ability to discriminate risks by construct Justification should be provided as to why it is believed the model's discriminatory power is adequate based on the benchmarking carried out on alternative models. Conceptual arguments and available evidence should be presented

Table 12
Model responses for CRR self-attestation (PD framework)

Regulation	**Reference**	**Suggested response to the regulation (areas for consideration)**
CRR	170.1d	The MGS should have a sufficient number of grades (26 non--default grades) to provide sufficient discrimination across the portfolio Quarterly PD analysis packs should be referred to the grade distribution with focus on overrides and exposures across grades
PRA SS11/13	17.2 17.3	Performance of the model under consideration should be explained against the chosen measures of discriminative power in the annual review documentation A sufficient number of default events should be considered across the entire data history in order to assume that sampling errors are small Co-rated samples should be benchmarked against external ratings (e.g. S&P)
PRA SS11/13	7.5	The model should be developed under an agency grade replication approach The number of observations sufficient to provide confidence in statistical measures of discriminative power should be considered The performance in aligning common obligors to be analysed in the model development/redevelopment Model overrides should be regularly monitored from the moment of model implementation and reviewed as part of the annual review process