An Overview and Framework for PD Backtesting and Benchmarking

Gerd Castermans [a] David Martens [b] Tony Van Gestel [a,d] Bart Hamers [a]
Bart Baesens [c,b]

[a] *Credit Risk Modeling, Group Risk Management, Dexia Group*
Square Meeus 1, 1000 Brussel, Belgium
{Gerd.Castermans;Tony.Vangestel;Bart.Hamers}@dexia.com

[b] *Department of Decision Sciences & Information Management, K.U.Leuven*
Naamsestraat 69, B-3000 Leuven, Belgium
{David.Martens;Bart.Baesens}@econ.kuleuven.be

[c] *University of Southampton, School of Management, United Kingdom*
Highfield Southampton, SO17 1BJ, United Kingdom
Bart@soton.ac.uk

[d] *Department of Electrical Engineering, ESAT-SCD-SISTA, K.U.Leuven*
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium

Abstract

In order to manage model risk, financial institutions need to set up validation processes so as to monitor the quality of the models on an ongoing basis. Validation can be considered from both a quantitative and qualitative point of view. Backtesting and benchmarking are key quantitative validation tools, and the focus of this paper. In backtesting, the predicted risk measurements (PD, LGD, EAD) will be contrasted with observed measurements using a workbench of available test statistics to evaluate the calibration, discrimination and stability ofthe model. A timely detection of reduced performance is crucial since it directly impacts profitability and risk management strategies. The aim of benchmarking is to compare internal risk measurements with external risk measurements so to better gauge the quality of the internal rating system.This paper will focus on the quantitative PD validation process within a Basel II context.

We will set forth a traffic light indicator approach that employs all relevant statistical tests to quantitatively validate the used PD model, and document this complete approach with a real-life case-study.The set forth methodology and tests are the summary of the authors' statistical expertise and experience of world wide observed business practices.

1 Introduction

A backtesting procedure typically evaluates the following characteristics of a rating system: calibration, discrimination and stability. Calibration refers to the mapping of a rating to a quantitative risk measure (PD in our case). A rating system is considered well calibrated if the (ex-ante) estimated risk measures deviate only marginally from what has been observed ex-post. Discrimination measures how well the rating system provides an ordinal ranking of the risk measure considered. E.g., in a default context, discrimination measures to what extent defaulters were assigned low ratings and non-defaulters high-ratings. Finally, stability measures to what extent the population that was used to construct the rating system is similar to the population on which it is currently being used. Due to population drift or new bank strategies, populations may change over time such that the rating system is no longer appropriate and becomes outdated. Benchmarking is another quantitative validation method which aims at assessing the consistency of the estimated parameters and models with those obtained using other estimation techniques, and potentially using other data sources (such as other institutions or ECAIs).

When backtesting these three characteristics of a rating system, one will typically employ statistical tests to decide upon significant differences between estimated and observed measurements. The results of these tests will then indicate potential problems and actions that need to be undertaken. An example calibration test may look like:

$$\begin{aligned} &H_0\text{: the PD of a rating is underestimated (PD} > PD_{cal}\text{).}\\ &H_1\text{: the PD of a rating is overestimated (PD} \leq PD_{cal}\text{).} \end{aligned} \tag{1}$$

The statistical test will then yield a p-value which will be used to make a decision given a pre-specified significance level. High p-values indicate that H_0 is likely and should be accepted and low p-values vice versa. Choosing the appropriate significance level depends on how conservative one wants to be. A traffic light indicator approach may be used to present the p-values in a user-friendly way [4]. A color is then assigned to each p-value depending upon the statistical severity of the measured difference.

A red flag indicates a statistically very significant (95-100%) difference between the observed and estimated measurements, potentially caused by a severe problem which may necessitate reconsidering the estimation of the measure considered. An orange flag indicates a performance difference which needs to be closely monitored in the future (90-95%). A yellow color indicates a minor difference which could be interpreted as an early warning of potentially decreasing performance (70-90%). A green color indicates no significant difference between the observed and estimated measurements (10-70%). A dark green color indicates an even higher degree of a lack of significant difference (0-10%). The preferred color is light green. It is rather straightforward that this is preferred to the yellow, orange, and red colors. However, it is also preferred to the dark green color. Dark green is a sign that the estimations are overly conservative, which will result in an increased capital requirement. Note that when using multiple statistical tests in a backtesting procedure, the occurrence of a few dark green, yellow, orange or even red flags is statistically normal. A financial institution may use more or less colors depending on its preference.

An overview of the PD backtesting and benchmarking statistics, as well as their main advantages and disadvantages are listed in Table 2. These statistical tests and their properties will be discussed throughout this paper. Finally, note that in portfolios having only a limited number of exposures, or where only a few defaulters are observed each year, using statistical tests to do backtesting may never yield significant results. In such cases, it is advised to do either data aggregation, e.g. by merging ratings together, or consider longer time horizons in order to improve the power and quality of backtesting.

2 Data description

The dataset used in this paper originates from a real life dataset covering 24 years (1983 to 2006). The last 10 years (1997 to 2006) are used to run the backtest on, the first 14 years (1983 to 1996) are used as a reference test. There are about 5000 observations per year (1 observation per counterpart) with an average default rate of 198 bps in the backtest dataset. Counterparts are rated with 6 PD rating classes, ranging from A (lowest PD, 0 defaults 0bps) to F (highest PD, 500 defaults 1480 bps). Sector information is available in the dataset, where 5 sectors are defined: banks, insurers, corporates, public and real estate.

	nb obs	A	B	C	D	E	F
A	7766	95%	5%	0%	0%	0%	0%
B	11104	3%	91%	6%	0%	0%	0%
C	9279	0%	5%	88%	5%	1%	1%
D	5055	0%	1%	8%	79%	11%	2%
E	6851	0%	0%	0%	6%	82%	12%
F	2184	0%	0%	0%	1%	10%	89%

Table 1
Migration matrix reflecting the average rating migrations over the years 1997-2006.

3 PD Calibration

Correct calibration of a PD rating system means that the calibrated PD estimates are accurate and conform to the observed default rates [3]. Hence, when backtesting PD calibration, one will typically start with a matrix contrasting the estimated PD with the observed default rates for each rating and time period considered. Similar matrices may be constructed depicting the number of observations and defaults in order to better gauge the importance of potentially significant differences. The bottom rows of the tables may aggregate the information of longer time periods (e.g. 5 years). Depending on data availability and portfolio size, the matrix may also be constructed for aggregated ratings and for the total portfolio. Once the matrix has been constructed, one can start testing for significant differences between estimated and observed numbers. Different test statistics have been suggested to do this. In what follows, we will discuss the most frequently used.

Characteristic	What?	Statistics	Advantages	Disadvantages
Level 2 : Risk Estimate ***Calibration***	Estimated risk measure (PD) vs observed risk measure	Vasicek Conf. $[0, q(\alpha)]$ with $q(\alpha) = \Phi(\frac{\Phi^{-1}(\hat{PD})+\sqrt{\rho}\Phi^{-1}(\alpha)}{\sqrt{1-\rho}})$	► Default correlation / economic condition ► In spirit of Basel II	► Infinitely granular portfolio assumption → Monte Carlo extensions for small samples ► Choosing correlation
		Binomial	► Sample variability ► Well known	► No default correlation / economic condition
Level 1 : Scorecard ***Discrimination***	Ordinal ranking among risk measures	ROC	► Sample distribution not required to be same as population distribution.	► Hard to define a minimum value that determines acceptable discriminatory power
			► Confidence intervals and tests available for AUC measures.	► Importance of only limited relevant cut-off range
		DeLong	► Sample variability	► Complex to calculate
		Cumulative notch difference	► Easy calculation and intuitive	► Differences equally weighted
Level 0 : Data ***Stability***	Population used in development vs population used upon deployment	$SI = \sum_{i=1}^{m} (R_i - O_i) \ln \frac{R_i}{O_i}$	► Easy calculation and intuitive	► Hard to define cut-offs
		$\chi^2 = \sum_{i=1}^{m} \frac{(OD_i - TD_i)^2}{TD_i}$	► Statistically well founded	► Bias towards less categories
		$\gamma = \frac{n_C - n_D}{n_C + n_D}$	► Ties no problem ► For ordinal data ► Confidence intervals and tests available	► Hard to define cut-offs ► Insensitive to rating shifts
		$\kappa = 1 - \frac{1-P_o}{1-P_e}$	► Sensitive to rating shifts	► Care needed with binary ratings

Table 2
An overview of relevant statistical tests for quantitative PD validation.

3.1 Vasicek test

Given the estimated probability of default, $\hat{PD}$ and asset correlation ρ, the Vasicek one factor model yields asymptotically the following quantile for the observed default rate [14]:

$$q(\alpha) = \Phi(\frac{\Phi^{-1}(\hat{PD}) + \sqrt{\rho}\Phi^{-1}(\alpha)}{\sqrt{1-\rho}}) \tag{2}$$

with $\Phi(x)$ the cumulative standard normal distribution and Φ^{-1} its inverse. An $\alpha\%$ confidence interval can then be constructed as follows: $[0, q(\alpha)]$. When the observed default rate falls outside this confidence interval, then a statistically significant difference is concluded according to the Vasicek test. The asset correlation ρ can be derived from the Basel II Accord. In case multiple asset correlations ρ are relevant, one can argue via the contamination principle that the maximum ρ is applicable. Because the Basel II correlations are known to be conservative, one may opt to make the statistical test more strict by using half the correlation.

A disadvantage of the Vasicek test is that an infinitely granular portfolio is assumed. For finite samples, one may use Monte Carlo simulation to calculate a more precise confidence interval [2].

3.2 Binomial test

The binomial test contrasts the forecast default rate of a rating, $\hat{PD}$ versus the observed default rate, DR using following hypothesis test [3,7]: H_0: $PD = \hat{PD}$ vs H_1: $PD > \hat{PD}$.

When assuming that defaults occur independently and H_0 is true, the number of defaulters follows a normal distribution $PD \sim N(\hat{PD}, \frac{\hat{PD}(1-\hat{PD})}{n})$, for growing n. One then arrives at the following test statistic to test H_0 versus H_1:

$$z = \frac{DR - \hat{PD}}{\sqrt{\frac{\hat{PD}(1-\hat{PD})}{n}}} \sim N(0,1), \tag{3}$$

which is standard normal distributed. One then compares the computed z-value against a cut-off, based on the desired confidence level, and makes a decision on whether to accept or reject H_0. This approximation by $N(0,1)$ is applied when $n > 1000$ and $n \times PD \times (1 - PD) \geq 9$, and if not by Poisson (which is often experienced in the case of low default portfolios). For a low number of observations the binomial distribution is effectively calculated.

Fig. 1 shows the critical value of the binomial test for a reference PD of 1.5% assuming various significance levels and values of n. It can be seen that the critical value decreases with a growing number of observations, which makes it interesting to perform backtesting on an aggregated level.

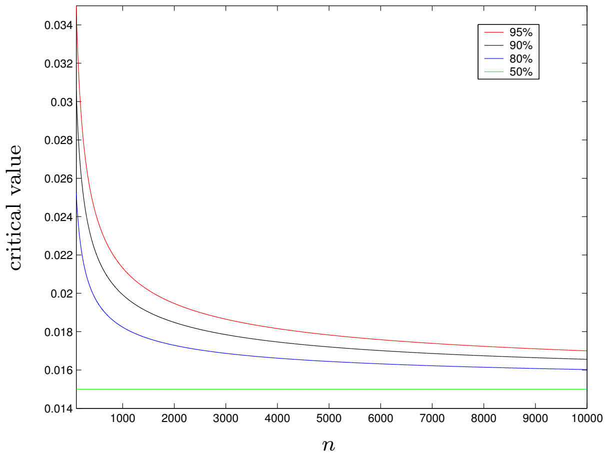

Fig. 1. Critical value of the binomial test for a reference PD of 1.5%.

As stated above, the binomial test assumes that defaults are independent. This assumption may not always be fulfilled in practice since defaults will typically be correlated. When default correlation is present, the binomial test may be too conservative hereby giving an elevated type I error which is the probability of incorrectly detecting a performance difference when no difference is present [3]. Several extensions have been proposed to the binomial test, to take into account the asset correlation ρ [12] and the volatility of the estimated default rate due to fluctuations and shocks [5]. One may also opt to combine both in order to model the joint effect of both asset correlation and volatility:

$$z = \frac{DR - \hat{PD}}{\sqrt{\frac{\hat{PD}(1-\hat{PD})}{n(1-\rho^2)} + \frac{\sigma^2}{n^2}\sum_{t=1}^{T} n_t^2}} \sim N(0,1), \qquad (4)$$

Given all these extensions, the question still remains which significance levels to use to decide whether the observed default rates differ significantly from the predicted ones. The financial regulators should clearly give guidance on this. E.g., the Hong Kong Monetary Authority (HKMA) advised financial institutions to use significance levels of 95% and 99.9% [7].

		1997			**2006**			**Total**		
	PDcal	DR	p		DR	p		DR	p	
A	0,00%	0,00%	50,00%	🟩	0,00%	50,00%	🟩	0,00%	50,00%	🟩
B	0,01%	0,00%	45,11%	🟩	0,00%	43,92%	🟩	0,00%	96,65%	🟥
C	0,28%	0,27%	65,81%	🟩	0,33%	34,11%	🟩	0,08%	15,15%	🟩
D	1,30%	0,52%	27,94%	🟩	1,36%	63,17%	🟩	1,30%	0,65%	■
E	4,60%	5,47%	70,31%	🟨	4,87%	70,75%	🟨	5,54%	97,45%	🟥
F	14,38%	16,18%	66,44%	🟩	14,62%	43,41%	🟩	12,70%	74,78%	🟨

Table 3
Calibration: comparing the estimated and observed default rates.

4 PD Discrimination

The major aim of backtesting PD discrimination is to verify whether the model still correctly distinguishes between defaulters and non-defaulters, or provides a correct ordinal ranking of default risk such that defaults get assigned low ratings and non-defaulters high ratings. In order to do this, the performance measures used for model evaluation, such as the Area Under the receiver operating characteristic Curve (AUC), Accuracy Ratio and the DeLong test, are used to compare the reference model with the new models.

The receiver operating characteristic curve (ROC) is a 2-dimensional graphical illustration of the hit rate $F_D(s)$ on the Y-axis versus the false alarm rate $F_{ND}(s)$ on the X-axis for all possible values of s. One often calculates the area under the receiver operating characteristic curve (AUC). Higher AUC values indicate that the classifier is on average more to the north-west region of the graph. Solid statistical tests exist for the comparison of AUC values coming from different rating or scoring models, in the form of the DeLong test. Other measures for PD discrimination includes the Accuracy Ratio (AR), which is linearly related to the AUC as follows [3]: $AR = 2 \times AUC - 1$. In this backtesting setup the AUC is compared with the model's AUC with DeLong test [6], which once more will define the traffic light color, as shown in Table 4.

	Number of Obs.	Number of defaulters	AR	p	
AR 2006	5677	97	0,81	47,6%	🟩
AR 2005	5462	108	0,80	61,8%	🟩
AR 2004	5234	111	0,83	17,6%	🟩
AR 2003	5260	123	0,79	81,5%	🟨
AR 2002	5365	113	0,79	78,3%	🟨
AR 2001	5354	120	0,75	98,9%	🟥
AR 2000	5306	119	0,82	25,0%	🟩

Table 4
Monitoring PD discrimination using the accuracy ratio.

5 PD Stability

The aim of backtesting stability is to check whether internal or external changes in the environment will impact the rating model in an unfavourable way, hereby making it no longer appropriate. To backtest the stability of the population, the following three-step approach may be adopted: (1) Check whether the population on which the model is currently being used is to similar to the population that was used to develop the model. (2) If differences occur in step (1), verify the stability of the individual variables in the model. (3) Check the stability of the ratings.

5.1 Stability Index (SI)

In order to check the stability of the population, a comparison will be made between the population observed at year t, and the population used to construct the model. In order to better understand the evolution of the population, one may also compare the population in year t with that of year $t-1$.

The similarity between difference populations will be measured using a Stability Index (SI), which is closely related to the Kullback-Leibler distance [8]. First, a table is constructed showing the population distribution across the different segments of a rating system. These segments may e.g. be the ratings or score deciles of a scorecard.

The SI can then be computed as follows [13]:

$$SI = \sum_{i=1}^{m} (R_i - O_i) \ln \frac{R_i}{O_i}, \tag{5}$$

whereby m is the number of classes considered, R_i represents the percentage of the reference population in class i and O_i the percentage of the observed population in class i. Higher values of SI indicate more substantial shifts in the population. The following rules of thumb are commonly applied [13], although they may be too rigid, and require expert intuition:

- SI $\leq$ 0.1: no significant shift (e.g. green flag)
- 0.1 $<$ SI $\leq$ 0.25: minor shift (e.g. yellow flag)
- SI $>$ 0.25: major shift (e.g. red flag)

When the analysis of population stability detected significant shifts, it must be further investigated which variables are responsible for these shifts. This can be done by contrasting the variables of the different populations individually using histograms, averages, t-tests, . . . One may also calculate the SI for each variable individually as illustrated in Table 5. The outcome of this analysis may then indicate which variables have changed and should be monitored.

Variable		**%Reference**	**% Sample $t-1$**	**% Sample t**
Name	**Categories**			
days past due	< 8	85	80	78
	8 to 90	12	15	14
	≥ 90	3	5	8
SI reference			0.020 🟩	0.058 🟩
SI year before				0.028 🟩

Table 5
Calculating the SI for individual variables

5.2 γ statistic

The γ statistic is a bivariate measure of association employed with ordinal data [11]. It evaluates the degree of correlation between two sets of rankings coming from two years of observation. The γ statistic can also be used in a benchmarking context, in order to evaluate the correlation between rankings from an internal rating system and an external shadow rating or benchmark, as will be illustrated in Section 6. The degree of correlation is evaluated based on the relative ordering of all possible pairs of obligors. Consider two obligors ranked by both the internal rating system and the benchmark. The two obligors are said to be concordant if the obligor who is rated higher by the internal system, is also rated (scored) higher by the benchmark, and are discordant if the debtor rated (scored) higher by the internal system is rated lower by the benchmark.

The bivariate γ statistic is especially suitable when many ties are present [11]. Let n_C represent the number of concordant pairs, n_D the number of discordant pairs, and n the number of observations. Goodman and Kruskal's gamma, γ is computed as follows:

$$\gamma = \frac{n_C - n_D}{n_C + n_D} \tag{6}$$

γ always ranges between -1 ($n_C = 0$) and +1 ($n_D = 0$). Higher values of γ indicate better correlation between the internal rating system and the external benchmark. Since γ only considers n_C and n_D, it ignores all pairs of ties. Furthermore, as γ is a correlation measure a rating shift of the inputs by one or more notches will not be measured.

In order to test $H_0 : \gamma = 0$, a test statistic which follows a standard normal distribution can be used [11]. Note that for this statistic, no dark green color is used as for stability one can never be overly conservative.

5.3 κ statistic

The κ statistic is used for the evaluation of categorical data [10], and takes takes into account agreement that should be expected by mere chance. The definition is derived from these observed and expected agreements P_o and P_e as given by Eq. 7.

$$\kappa = \frac{P_o - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e} \tag{7}$$

The observed agreement is simply the accuracy, while the expected agreement P_e can be calculated from the marginal distributions. As the original κ statistic is valid for nominal data only, we use the weighted κ, applicable to ordinal data [1]. The statistic has some desirable properties: when there is no agreement, then $\kappa = 0$, when agreement is less than expected by chance, $\kappa < 0$, and minimal agreement corresponds to $\kappa = -1$. When agreement is higher than expected by chance, $\kappa > 0$, with maximal agreement corresponding to $\kappa = 1$. Apart from these general properties which κ shares with correlation coefficients, there are also some practical guidelines reported in the literature [9]: excellent agreement beyond chance corresponds to a $\kappa > 0.75$, while a κ smaller than 0.40 indicates poor agreement beyond chance. In between these values there is a fair to good agreement beyond chance. In order to test H_0: $\kappa = 0$, a test statistic which follows asymptotically a standard normal distribution is used. Finally, note that using the kappa on binary ratings has to be done by care [10].

	γ						κ					
	Average			2005-2006				Average			2005-2006			
	γ	σ	p_0	γ	σ	p_0	p	κ	p_0	p ($\neq avg$)	κ	σ	p_0	p
Banks	98,94%	0,12%	0,00% 🟩	99,35%	0,36%	0,00% 🟩	0,00% 🟩	95,32%	0,21%	0,00% 🟩	95,92%	0,60%	0,00% 🟩	0,00% 🟩
Insurers	98,28%	0,28%	0,00% 🟩	99,09%	0,48%	0,00% 🟩	0,00% 🟩	93,30%	0,53%	0,00% 🟩	95,51%	1,07%	0,00% 🟩	0,00% 🟩
Corporates	98,10%	0,10%	0,00% 🟩	98,66%	0,24%	0,00% 🟩	0,00% 🟩	95,61%	0,11%	0,00% 🟩	95,99%	0,29%	0,00% 🟩	0,00% 🟩
Public	98,39%	0,17%	0,00% 🟩	98,78%	0,49%	0,00% 🟩	0,00% 🟩	94,79%	0,31%	0,00% 🟩	96,00%	0,68%	0,00% 🟩	0,00% 🟩
Real estate	98,48%	0,32%	0,00% 🟩	99,68%	0,22%	0,00% 🟩	0,00% 🟩	94,46%	0,48%	0,00% 🟩	97,09%	0,84%	0,00% 🟩	0,00% 🟩
Total	98,55%	0,06%	0,00% 🟩	99,17%	0,12%	0,00% 🟩	0,00% 🟩	96,35%	0,07%	0,00% 🟩	97,18%	0,17%	0,00% 🟩	0,00% 🟩

Table 6
γ and κ statistics for the year 2005-2006, as well as the average over all the year transitions from 1997 until 2006.

5.4 χ^2 test

The χ^2 test compares the observed default frequencies per rating with the theoretical default frequencies per rating under the condition of independence between rating and default frequency [11,13]. The more these values differ, the more the assumption of independence is invalid. The calculation of the theoretical number of defaults TD_i for rating i, under the independence assumption can be derived from the number of observed defaults per rating OD_i, and the total number of observations.

The χ^2 value provides an indication of the extent to which the theoretical and observed default frequencies deviate from another, the calculation of which is shown in Eq. 8.

$$\chi^2 = \sum_{i=1}^{m} \frac{(OD_i - TD_i)^2}{TD_i} \tag{8}$$

	Rating	Ref.	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
						Banks						
nb obs	A	7690	297	329	333	348	381	411	399	420	448	458
	B	9396	346	339	357	343	362	373	404	404	403	435
	C	5428	99	114	117	135	115	108	110	132	139	136
	D	5601	50	95	95	91	69	85	68	59	51	55
	E	4348	51	57	66	67	68	29	24	24	22	25
	F	1370	9	6	22	17	10	17	13	6	4	3
SI	year $t-1$			97,99% 🟥	89,26% 🟨	23,41% 🟩	86,67% 🟨	99,90% 🟥	49,49% 🟩	63,13% 🟩	14,06% 🟩	3,13% 🟩
	Ref.		100,0% 🟥	100,0% 🟥	100,0% 🟥	100,0% 🟥	100,0% 🟥	100,0% 🟥	100,0% 🟥	100,0% 🟥	100,0% 🟥	100,0% 🟥

Table 7
χ^2-based stability over various years for the bank sector, compared with the previous year $t-1$ and reference year.

6 PD Benchmarking

Benchmarking is another quantitative validation method which aims at assessing the consistency of the estimated parameters and models with those obtained using other estimation techniques, and potentially using other data sources (such as other institutions or ECAIs). An example could be a financial institution comparing its internal corporate ratings with those provided by Moody's or S&P's. The purpose of benchmarking is to measure the degree of agreement or disagreement between both ratings or models. From a statistical perspective, this can be done using the ordinal multiclass performance measures discussed in Section 4. Benchmarking works complementary to backtesting and allows to further asses the quality and credibility of the internal rating system and model.

When doing benchmarking, one first needs to determine what will be benchmarked. Examples of candidate benchmarking quantities are credit scores, ratings, calibrated risk measures, or even migration matrices. Different types of benchmarking partners can be considered to perform the comparison. Examples are rating agencies, credit bureaus or other data pooling partners. The selection will largely depend on the characteristics of the portfolio. While credit bureaus have traditionally focused on the retail sector, rating agencies are more targeted towards non-retail. In the absence of an external benchmarking party, financial institutions may also consider other alternatives, such as benchmarking internally with e.g. internal rating reviewers that re-rate a sample of obligors on an expert-judgement basis [7], or using independent development teams each developing their own rating system. These could then be compared using a champion-challenger approach. Another alternative would be to compare internal ratings and estimates with market-based proxies for credit quality, such as equity prices or bond spreads, or premiums for credit derivatives.

Although benchmarking is an attractive validation tool, many difficulties arise during its implementation. First, one needs to guarantee the degree of equivalence between the internal and external quantities to be benchmarked. In a PD context, the same definition of default and assessment horizon must be used by both partners. Furthermore, it must be clear whether stressed or unstressed PDs are compared. Also, when contrasting the ratings, the TTC or PIT character of both rating systems must be well understood. Mapping schemes may be defined to map both ratings to a same rating philosophy so as to allow objective comparison. Furthermore, since both parties may use different rating scales depending on the granularity of their risk measurements, a common masterscale must be defined so to be able to objectively compare.

Summarizing, many problems still exist when conducting a benchmarking exercise and in many cases it remains very subjective. Financial supervisors in collaboration with industry partners should work on developing formal benchmarking schemes and mapping methodologies in order to fully leverage the idea behind benchmarking.

7 Conclusion

Quantitative validation consists of backtesting and benchmarking of estimated risk parameters. We have put forward and applied a number of relevant statistics that can be used to validate the appropriateness of the internal measures at the level of data, scorecard and calibration. Although this paper focuses on the quantitative part of validation, the qualitative part is of crucial importance as well and should incorporate the expertise of domain specialists.

Validation is a crucial part of credit risk management, but it is only a diagnostic approach that may reveal some shortcomings of the constructed models. The cure however, is not incorporated within this framework, as this takes place in a subsequent step. Many issues remain for PD validation such as the determination of proper minimum values and insufficient data, which is of particular importance in a low default portfolio context. For the validation of LGD and EAD an even higher need for research into relevant validation procedures exists across the three validation levels, as defined in this paper.

References

[1] J.H. Fleiss abd J. Cohen. The equivalence of weighted kappa and the intraclass correlation as measures of reliability. *Educ. Psychol.Meas*, 33:613–619, 1973.

[2] L. Balthazar. PD estimates for Basel II. *Risk*, pages 584–585, 2004.

[3] Basel Committee on Banking Supervision. Studies on the validation of internal rating systems. Technical Report Working Paper No. 14, Bank for International Settlements, 2005.

[4] S. Blochwitz, S. Hohl, D. Dirk Tasche, and C.S. Wehn. Validating default probabilities on short time series. 2004.

[5] R. Cantor and E. Falkenstein. Testing for rating consistency in annual default rates. *Journal of Fixed Income*, pages 36–51, 2001.

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

[7] Hong Kong Monetary Authority. Validating risk rating systems under the IRB approaches. Technical report, 2006.

[8] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[9] K. Landis. The measurement of observer agreement for categorical data. *Biometrics*, 53:159–177, 1977.

[10] C.A. Lantz and E. Nebenzahl. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*, 49(4):431–434, April 1996.

[11] D.J. Sheskin. *Handbook of parametric and nonparametric statistical procedures.* Chapman and Hall/CRC, 2000.

[12] R. Stein. Are the probabilities right? a first approximation to the lower bound on the number of observations required to test for default rate accuracy. *Moody's KMV, Technical Report 030124*, 2003.

[13] L. Thomas, D. Edelman, and J. Crook, editors. *Credit Scoring and its Applications.* SIAM, 2002.

[14] O. Vasicek. The loan loss distribution. Working, KMV Corporation, 1997.