A PD Validation Framework for Basel II Internal Ratings-Based Systems

Maurice P. Joseph[1]

Credit Scoring and Credit Control IX
September 2005

Abstract

The need to have available robust measures to compare credit scoring (or rating) systems has developed in importance in recent times, particularly so for large-sized banks seeking accreditation under the Basel II Internal Ratings-Based (IRB) approach. Given the significant amount of resources devoted to building credit risk models by financial institutions worldwide, it is somewhat surprising to see the paucity of publications directed towards the construction of a sound framework for validating and maintaining risk models over time. In this paper we outline some practical applications of a proposed validation framework that supports the Bank of International Settlements (BIS) recommendations on model validation techniques.

We start with a simple Chi-squared test of randomness for rating outcomes and then proceed to apply a generalised assessment process for model discriminatory power (or what we call the Wilkie Combination Test) following on from the pioneering work of David Wilkie ("Measures for comparing scoring systems"). For more advanced testing, we illustrate two approaches for calculating standard errors and confidence intervals for the key model discriminatory measures as well as a likelihood ratio test and the Brier Score. We also demonstrate a novel approach for validating Probability of Default (PD) numbers for a loan portfolio (given dependencies amongst loan obligors) and an approach that increases the capital requirements—if any dependency exists between PD and Loss Given Default (LGD) estimates. Practical numerical examples are enumerated for all of the required formulae, using Visual Basic for Applications (VBA) code, which may assist with the implementation of these measures.

Key Words: Credit Rating, Basel II, Accuracy Ratio, ROC, Brier Score, Internal Ratings, Validation, Calibration, Probability of Default (PD), Combination Tests, LGD, VBA

[1] Dr. Maurice Joseph (Quantitative Analyst) Basel II Project, Commonwealth Bank of Australia mailto:maurice.joseph@cba.com.au or mailto:MauriceJoseph@hotmail.com
VBA code and MS Excel illustrative examples can be downloaded from the following site:
http://www.machinethinking.com/edinburgh/download.html

Probability of Default (PD) Validation Framework

In this paper, we outline a Probability of Default Validation Framework that we believe would satisfy the Internal Ratings-Based (IRB) approach of the Basel II Accord[2] (based on a quantitative testing approach only).

Before determining what this framework involves, it is necessary to appreciate the design framework underpinning the Basel II Capital Requirement equations, in order to better understand our proposed PD validation approach. In essence, these formulae (plus the relevant BIS accord paragraphs) require a bank to hold a **diversified** portfolio fundamentally dependent upon ratings-based grades of obligors. The BIS explanation for this crucial aspect of the underlying framework is perhaps best illustrated by extensively quoting from its own explanatory document on the IRB risk weight formulae:

> The Basel risk weight functions used for the derivation of supervisory capital charges for Unexpected Losses (UL) are based on a specific model developed by the Basel Committee on Banking Supervision (cf. Gordy, 2003). The model specification was subject to an important restriction in order to fit supervisory needs:
>
> The model should be **portfolio invariant**, i.e. the capital required for any given loan should only depend on the risk of that loan and must not depend on the portfolio it is added to. This characteristic has been deemed vital in order to make the new IRB framework applicable to a wider range of countries and institutions. Taking into account the actual portfolio composition when determining capital for each loan - as is done in more advanced credit portfolio models - would have been a too complex task for most banks and supervisors alike. The desire for portfolio invariance, however, makes recognition of institution-specific diversification effects within the framework difficult: diversification effects would depend on how well a new loan fits into an existing portfolio. As a result the Revised Framework was calibrated to well diversified banks. Where a bank deviates from this ideal it is expected to address this under Pillar 2 of the framework. If a bank failed at this, supervisors would have to take action under the supervisory review process (pillar 2).
>
> In the context of regulatory capital allocation, portfolio invariant allocation schemes are also called **ratings-based**. This notion stems from the fact that, by portfolio invariance, obligor-specific attributes like probability of default, loss given default and exposure at default suffice to determine the capital charges of credit instruments. If banks apply such a model type they use exactly the same risk parameters for EL (Expected Losses) and UL, namely PD, LGD and EAD.[3]

Thus, the emphasis on the outcome of meaningful distribution of exposures across grades (in terms of PD outcomes) ultimately stems from the underlying derivation of the relevant capital requirement equations. Without a diversified underlying portfolio the Basel II framework is virtually rendered unworkable and hence arguments by Hamerle, Rauhmeier and Rosch (2003) that the discriminatory measures cannot be the sole measures of a rating system are placed in their proper context, as per their example of an extreme portfolio with just two rating grades, whereas the BIS guidelines insists on a minimum of seven pass grades and at least one default grade (paragraph 404 of the Accord). We could also add, that if supervisors failed to take necessary action under pillar two for indiscretions of this fundamental assumption, then pillar three (or Market Discipline) will undoubtedly capture such indiscretions over time (and the potential consequences thereof, could be quite severe under this ultimate "safety-net" design feature of the Basel II Accord).

---

[2] **Basel Committee on Banking Supervision (2004):** International Convergence of Capital Measurement and Capital Standards - A Revised Framework.

[3] **Basel Committee on Banking Supervision (2005)**, "An Explanatory Note on the Basel II IRB Risk Weight Functions", July 2005 p. 4

Regular model validation is necessary for IRB compliance, including monitoring of performance and stability, review of model relationships and testing of its outputs against expected outcomes. Furthermore, the Accord makes special mention on the usage of any statistical models for the rating process, with a requirement for a rigorous statistical process to be undertaken that will cover both out-of-time and out-of-sample performance tests. However, it could be argued that any such rigorous process should also be performed for *any* rating model, regardless of whether it is internal (or external) and/or statistical (or judgemental) in nature.

We propose a robust and reliable PD validation framework that should encompass at least the following three "pillars" or parts:

1. The PD validation framework should encompass a **comprehensive** set of statistical tests (many of which should already be extensively utilised by practitioners) but also having just *one main outcome* that is a synthesised number with the attributes of a reliable and valid statistical discrimination and this number effectively "rates the ratings".
2. *Pre-set trigger levels and pre-defined **standards*** for what constitutes an adequate and meaningful outcome from the tests.
3. An ongoing basis of **regular monitoring** (in effect a PD Validation "system") that will adequately forewarn bank management, regulators and the general public of any major deviations from the system's "steady-state".

Apart from regulatory requirements, such a framework would also make significant economic sense as per the Roger Stein contention: "… a conservative estimate of the additional profit that a bank could expect using a model five points of accuracy ratio better than its competition would be around five basis points per dollar granted, if the bank were using the cut-off approach, and eleven basis points per dollar granted under the pricing approach."[4] Thus, for example, a large bank with approximately €200B in consolidated assets that underwrote about €17B in loans (new and renewal) could therefore expect to generate an additional €18.7M in profits in one year if it improved its rating model accuracy by five points (eg. ROC statistic improved from 70% to 75%).

In summary, the key principles as enunciated by BIS for IRB validation are[5]:

- *The Bank has primary responsibility for validation.*
- *Validation is fundamentally about assessing the predictive ability of a bank's risk estimates and the use of ratings in credit processes.*
- *Validation is an iterative process.*
- *There is no single validation method.*
- *Validation should encompass both quantitative and qualitative elements.*
- *Validation processes and outcomes should be subject to independent review.*

The absence of the qualitative testing component in this paper only reflects the need to limit the scope of this topic and not the importance of this crucial aspect, which embraces data quality issues, risk rating design, risk rating operations, human resources, technology and corporate governance issues, for example.

[4] **Roger M. Stein**, The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing, Journal of Banking & Finance 29 (2005), pp1213–1236 (see page 1232).
[5] **Bank of International Settlements (2005)** Working Paper No. 14, p. 4

1. Simple Statistical Tests

We can demonstrate the use of the differing levels of PD Validation measures using a similar dataset to that used by van Deventer and Imai in their chapter titled "Internal Ratings and Approaches to Testing Credit Models"[6]. This dataset enables one to apply a battery of PD Validation tests in order to gauge differences between distinctly different rating style models.

Table 1

Hypothetical Small Sample of Bank Rating Models

Rating Methodology -->			**1**	**1**		**2**	**2**		**3**	**4**
Number	Company	Default (or Bankruptcy) Index	Internal Rating Grade	Internal Rating Numerical Rating	Internal Rating Default Probablity	External Rating Grade	External Rating Numerical Rating	External Rating Default Probablity	Model **1** Default Probabilty	Model **2** Default Probabilty
1	a	**0**	B	9	0.20%	A-	9	0.04%	0.14%	0.45%
2	b	**0**	B	9	0.20%	A-	9	0.04%	0.85%	0.65%
3	c	**0**	B	9	0.20%	A-	9	0.04%	0.16%	0.26%
4	d	**0**	B	9	0.20%	A-	9	0.04%	0.27%	0.33%
5	e	**1**	B	9	0.20%	A-	9	0.04%	3.56%	1.79%
6	f	**0**	B	9	0.20%	BBB	8	0.36%	0.79%	0.65%
7	g	**0**	B	9	0.20%	BBB	8	0.36%	0.39%	0.43%
8	h	**0**	C	8	0.30%	BBB	8	0.36%	0.46%	0.59%
9	I	**0**	C	8	0.30%	BBB	8	0.36%	0.57%	0.75%
10	j	**1**	C	8	0.30%	BBB	8	0.36%	4.25%	2.54%
11	k	**0**	B	9	0.20%	A-	9	0.04%	0.18%	0.16%
12	l	**0**	C	8	0.30%	A-	9	0.04%	0.23%	0.21%
13	m	**0**	C	8	0.30%	BBB	8	0.36%	0.35%	0.32%
14	n	**0**	C	8	0.30%	BBB	8	0.36%	0.36%	0.32%
15	o	**0**	D	7	1.00%	BBB	8	0.36%	0.34%	0.31%
16	p	**0**	D	7	1.00%	BB	7	1.26%	0.55%	0.50%
17	q	**1**	D	7	1.00%	B+	6	3.64%	0.98%	0.88%
18	r	**0**	D	7	1.00%	BB	7	1.26%	1.06%	0.95%
19	s	**0**	D	7	1.00%	BB	7	1.26%	0.70%	0.63%
20	t	**0**	E	6	3.00%	BB	7	1.26%	0.76%	0.68%
21	v	**1**	E	6	3.00%	BB	7	1.26%	0.80%	0.72%
22	w	**1**	E	6	3.00%	B+	6	3.64%	1.56%	1.92%
23	x	**0**	E	6	3.00%	B+	6	3.64%	2.03%	1.83%
24	y	**1**	E	6	3.00%	B+	6	3.64%	3.45%	4.29%
25	z	**0**	F	5	7.00%	B+	6	3.64%	3.20%	2.88%
26	aa	**0**	F	5	7.00%	B/NR	5	10.31%	4.03%	3.63%
27	ab	**0**	F	5	7.00%	B/NR	5	10.31%	2.79%	2.51%
28	ac	**1**	F	5	7.00%	B/NR	5	10.31%	3.65%	5.78%
29	ad	**1**	F	5	7.00%	B/NR	5	10.31%	5.12%	4.61%
30	ae	**1**	F	5	7.00%	B/NR	5	10.31%	6.78%	6.10%
Total Defaults		**9**								

Table 1 above illustrates a sample of obligors rated using 4 different models: Internal Bank Model, External Rating Agency Model and two Probability of Default models (assume Model 1 is a Reduced-Form derivative model and Model 2 could represent a competing model, say a KMV-style model). We use an internal mapping process (see section below) to ascribe PD outcomes to both the Internal and External Rating Models. We also make the assumption that the Not Rated category under the External Rating Model can still be ascribed a PD (a conservative and generic value of 10.31% in this example). The Internal and External Numerical ratings are for convenience only. In terms of score equivalents, the inverse of the PD's can be used (as by convention, the higher the rating score, the better is the underlying credit) but scored comparisons are not essential for the approaches illustrated (except for ROC curve comparisons across several models).

[6] Also available at **http://www.kamakuraco.com/**; *(NB: You will need to register your details first before accessing the research material)* [WP16] Donald van Deventer and Xiaoming Wang, "Measuring Predictive Capability of Credit Models under the Basel Capital Accords: Conseco and Results from the United States, 1963-1998." 1/8/03 – See Appendix A

According to Altman[7] the starting point for any classification of risk is a universal rating equivalent scale. Using S&P PD equivalents in figure 1 below he highlights one such rating scale with suitable risk level descriptions that could be used as a mapping scale. We use a similar mapping approach as per Tables 2 and 3 below (but with a reversal of the rating numbers to match our illustrative data example).

Figure 1

The Starting Point is Establishing a Universal Rating Equivalent Scale for the Classification of Risk

	CREDIT GRADES	RISK LEVEL	PD (bp)	S&P
Performing	1	Minimal	0-1	AAA
	2	Modest	2-4	AA
	3	Average	5-10	A
	4	Acceptable	11-50	BBB
	5	Acceptable with care	51-200	BB
	6	Management Attention	201-1000	B
Substandard	7	Special Mention	1000+	CCC
	8	Substandard	Interest Suspense	CCC / CC
	9	Doubtful	Provision	CC / C
	10	Loss	Default / Loss	D

41

Table 2

Internal Rating Grade	**Mapping Process**	Rating Number	Median Rating **PD**	Rating Explanation in terms of Credit Quality
A		**10**	**0.03%**	Highest Credit Quality (eg. Sov/AAA)
B		**9**	**0.20%**	Excellent
C		**8**	**0.30%**	Good
D		**7**	**1.00%**	Average
E		**6**	**3.00%**	Marginal
F		**5**	**7.00%**	Poor
G		**4**	**15.00%**	Near Default
H		**1**	**>21.00%**	Default (with small chance of recovery)

Table 3

External Rating Grade	**Mapping Process**	Rating Number	Cohort Method Rating **PD**	Rating Explanation in terms of Repayment Capacity or Credit Quality
A+		**10**	**0.030%**	High Credit Quality
A-		**9**	**0.044%**	Strong Repayment Capacity
BBB		**8**	**0.360%**	Adequate Repayment Capacity
BB		**7**	**1.262%**	Average
B+		**6**	**3.635%**	Marginal
B/NR		**5**	**10.309%**	Poor (also covers Not Rated (NR) category)
CCC		**4**	**30.922%**	Near Default
D		**1**	**100%**	Default

Tables 2 and 3 above illustrate an internal bank PD mapping process that is required to convert ordinal ratings into cardinal PD equivalents (PD Master Scale). Such an exercise is always fraught with difficulty as the PD cardinal scale is inevitably a superior one. Nevertheless, the mapping process can assist the PD validation process when it comes to comparing divergent rating models. Models 1 and 2 are already in a cardinal scale format and hence one can merely decide on how many relevant grades to allocate the 30 individual obligors into.

[7] Edward I. Altman "Managing Credit Risk: The Challenge for the New Millennium", Presentation Slides, Stern School of Business, NYU, 2005, No. 41.

For this exercise we place the obligors into quintile groups as per van Deventer and Imai, but it is still possible to allocate them into decile groups despite the relatively small sample. Under the simple statistical test approach the null hypothesis is that the ratings (from any of the models) are no better than random chance. If this were true, then the number of defaults in each rating grade would be the same percentage as for the entire sample (in this case nine defaults out of thirty obligors). The chi-squared test algorithm requires calculating the expected defaults per grade and comparing them with the actual defaults per grade (simply tallying up the individual defaults by grade and also the total obligors by grade). The expected defaults per grade are derived by multiplying the Total Defaults divided by the Total Number of obligors by the actual number of obligors in this grade. For example, in the Internal-rating model if we have nine defaults out of thirty obligors in total, and only eight obligors in the B Grade, then the expected number of defaults for Grade B is 2.4 (i.e., 9/30 x 8). The expected default number of 2.4 can then be meaningfully compared to the actual defaults in this grade (of one default). The number of degrees of freedom for this test is the total number of row inputs (less one) times the total number of column inputs (less one). For example, with five rating categories (rows) and two column outcomes (Actual Defaults and Expected Defaults) we will have four degrees of freedom (i.e. (5 – 1) x (2 – 1)).

Tables 4 to 7 below illustrate a simple random test outcome using a Chi-Squared test (available with Microsoft® Office Excel). *Note that for the Internal bank table, the tally for grades B and C differ from van Deventer and Imai (who inadvertently count seven for each) although this does not affect the test outcomes. Also, the use of the Excel Chi-Squared Test by the authors results in a probability output (not the required Chi-Squared value as shown below) and therefore their estimated Chi-Squared Probabilities are invalid (although the final ranking outcomes remain the same).*

Tables 4

Chi-squared Test

Internal Bank Ratings		**Internal Bank Ratings**	Grade Numbers	Actual Defaults	Expected Defaults	$(A-E)^2/E$
B	Best	9	8	1	2.4	0.8166667
C		8	6	1	1.8	0.3555556
D		7	5	1	1.5	0.1666667
E		6	5	3	1.5	1.5
F	Worst	5	6	3	1.8	0.8
		Totals	**30**	**9**	**9**	**3.6388889**

degrees of freedom =	4	= (No. Rows - 1) x (No. Columns - 1) eg. (5 - 1) x (2 -1)= 4
Chi-squared Value =	3.6389	= Sum of $(A-E)^2/E$
Chi-squared Probability =	**45.7076%**	=CHIDIST(Chi-squared Value, degrees of freedom)

Table 5

Chi-squared Test

	External Bank Ratings	Grade Numbers	Actual Defaults	Expected Defaults	$(A-E)^2/E$
Best	9	7	1	2.1	0.5761905
	8	8	1	2.4	0.8166667
	7	5	1	1.5	0.1666667
	6	5	3	1.5	1.5
Worst	5	5	3	1.5	1.5
	Totals	**30**	**9**	**9**	**4.5595238**

degrees of freedom =	4
Chi-squared Value =	4.5595
Chi-squared Probability =	**33.5548%**

Table 6

Chi-squared Test

	Quintiles	**Model1 Ratings**	**Range**	Grade Numbers	Actual Defaults	Expected Defaults	$(A-E)^2/E$
Best	20%	0.00%	0.35%	6	0	1.8	1.8
	40%	0.35%	0.65%	6	0	1.8	1.8
	60%	0.65%	1.01%	6	2	1.8	0.0222222
	80%	1.01%	3.47%	6	2	1.8	0.0222222
Worst	100%	3.47%	6.78%	6	5	1.8	5.6888889
			Totals	**30**	**9**	**9**	**9.3333333**

degrees of freedom =	4
Chi-squared Value =	9.3333
Chi-squared Probability =	**5.3287%**

Table 7

Chi-squared Test

	Quintiles	**Model2 Ratings**	**Range**	Grade Numbers	Actual Defaults	Expected Defaults	$(A-E)^2/E$
Best	20%	0.00%	0.33%	6	0	1.8	1.8
	40%	0.33%	0.64%	6	0	1.8	1.8
	60%	0.64%	0.91%	6	2	1.8	0.0222222
	80%	0.91%	2.61%	6	3	1.8	0.8
Worst	100%	2.61%	6.10%	6	4	1.8	2.6888889
			Totals	**30**	**9**	**9**	**7.1111111**

degrees of freedom =	4
Chi-squared Value =	7.1111
Chi-squared Probability =	**13.0132%**

Table 8

Summary Table

Rating model	**Chi-Square Value**	**Statistical Significance**	**P-Test**	**Rankings**
Model1 Ratings	9.3333	5.33%	94.67%	**1**
Model2 Ratings	7.1111	13.01%	86.99%	**2**
External Bank Ratings	4.5595	33.55%	66.45%	**3**
Internal Bank Ratings	3.6389	45.71%	54.29%	**4**

The Chi-squared Probability Test measures the statistical significance of the chi-squared statistic being better than random chance. In terms of the null hypothesis, that the Actual defaults and the Expected Defaults are statistically equivalent, we can therefore unequivocally reject the null hypothesis for Model 1 and perhaps Model 2 as well, but the Internal and External models have much less confidence that they are significantly different from random chance. In relative terms, we note that Model 1 has the most significant outcome and is therefore ranked first, whilst the Internal Model is the worst from the Chi-Squared test.

2. Combination Tests

We adopt an extension of the Wilkie[8] proposal for the equivalence of measures for comparing rating or credit scoring systems (see Hand and Henley 1997) under the simplified assumption that scores are normally distributed. Wilkie demonstrated that when using five different measures, namely: the mean differences (D), percentage of cumulative "Goods" for the cumulative 50% of "Bads" (PH), maximum deviation (KS), Gini[9] coefficient (G), and Information Statistic (I), under the assumption of normality of score distributions, that all of these measures can be collapsed to just one statistic. We expand on this approach to reflect both the area under the ROC curve statistic, and the Kullback-Leibler statistic, whilst also widening the applicable range of outcomes but still utilising an averaging process (for functions of one statistic) via an interpolation of the normal distribution outcomes. By adding in the ROC statistic, we are also effectively replicating the Accuracy Ratio (as either measure can be derived from the other). We are not too concerned with this aspect given that the ROC has well documented usage in other sciences, and that having a bias towards a preferred industry measure is perhaps a useful in-built measurement stabiliser.

Thomas, Edelman and Crook (2004) find Wilkie's concept interesting, but suggest that practitioners do *not* assume that credit scores are normally distributed. However, when Wilkie was proposing this concept he was making the point that even though the scores of "Goods" (or non-defaults) and "Bads" (or defaults) in any population may be *approximately* normally distributed, the irregularities of sampling errors and the discreteness of integer scores meant that different measures could rank scoring systems in a different sequence and that this is unlikely to be a problem, *in practice*. In fact, it could be argued that this is a desirable feature—instead of having several uniformly consistent outcomes, we can now obtain slightly different views based on the approximation towards normality of data outcomes and also from the idiosyncratic measurement features of each comparison statistic. Hence, we are making use of a "group or crowd vote" for a decision (instead of relying on just one measure which could lead to conflicting results for comparing systems).[10] The wisdom of reliance on a group approach—with differing measurement attributes—can help introduce the qualities of stability, integrity, objectivity, accuracy, and an appropriate level of conservatism into the measurement process (as per the BIS[11] requirements).

Although the ascribed meaning of each score outcome is relatively arbitrary it does serve the purpose of ascribing meaningful differences in measurement for any PD ratings outcome. We can also illustrate the different measurement approaches across the entire assessment range as per the standardised chart approach in Figure 2 below. This chart illustrates how the different measures combine to produce a consensus viewpoint of the underlying rating model. Appendix (C) illustrates the underlying formulae and indicates the VBA functions required to derive the Combination Test statistics and validation scores.

---

[8] See Chapter 4 in "Readings in Credit Scoring" by Thomas, Lyn. C., David B. Edelman and Jonathon N. Crook (2004)

[9] The Gini statistic is very similar to the more commonly used term, Accuracy Ratio (AR), but involves multiplying the AR by the Number of Goods divided by the Total Number of Obligors.

[10] See for example the popular publication by **James Surowiecki, (2004)**, "The Wisdom of Crowds – Why the Many are Smarter than the Few and How Collective Wisdom shapes Business, Economies, Societies and Nations", Little, Brown. *In his introduction, Surowiecki makes the interesting point about British scientist Francis Galton, who in 1906, attended a farm exhibition and noted a guessing competition for the weight of a slaughtered and dressed ox, from 800 (diverse) entrants. He assumed that the crowd would be way off mark, yet when he acquired the written data, after the competition, he discovered to his surprise that the average guess from the crowd was 1,197 lbs - compared to the actual weight of 1,198 lbs. Surowiecki claims that this phenomenon is widespread and that it only requires a diversity of independent votes and an integration forum to be available, in order to derive the wisdom from any crowd votes.*

[11] See remarks by Mr. Nicholas Le Pan, Chairman of the Basel Accord Implementation Group, 6th Annual Global Association of Risk Professionals 2005 – Inaugural Address – Feb 1, NYC p. 6.

An example of the expanded Wilkie combination approach (with linear interpolation of statistical outcomes for standardised normal numbers) is shown in table 9 below:

Table 9

COMBINED APPROACH TABLE

PROBABILITY OF DEFAULT RATING VALIDATION

Validation Range		*Statistics :*	1.20	86.38%	43.77%	58.34%	79.17%	1.42	0.71
Lower to Upper Limits		Meaning	Mean Difference	CND% > 50% CD: (1-PH)	K-S Statistic	Accuracy Ratio	ROC Statistic	Information Statistic	Kullback-Leibler
0	1	*Random*	0.00	50.00%	0.00%	0.00%	50.00%	0.0000	0.0000
1	2	*Doubtful*	0.25	59.87%	9.95%	14.00%	57.00%	0.0625	0.0313
2	3	*Poor*	0.50	69.15%	19.74%	27.60%	63.80%	0.2500	0.1250
3	4	*Marginal*	0.75	77.34%	29.23%	40.40%	70.20%	0.5625	0.2813
4	5	*Satisfactory*	1.00	84.13%	38.29%	52.00%	76.00%	1.0000	0.5000
5	6	*Good*	1.25	89.44%	46.80%	62.30%	81.15%	1.5625	0.7813
6	7	*Very Good*	1.50	93.32%	54.67%	71.10%	85.55%	2.2500	1.1250
7	8	*Strong*	1.75	95.99%	61.84%	78.40%	89.20%	3.0625	1.5313
8	9	*Very Strong*	2.00	97.72%	68.27%	84.30%	92.15%	4.0000	2.0000
9	10	*Excellent*	2.25	98.78%	73.94%	90.08%	95.04%	5.0625	2.5313
10	11	*Excellent*	2.50	99.38%	78.87%	94.20%	97.10%	6.2500	3.1250
11	12	*Excellent*	2.75	99.70%	83.09%	97.14%	98.57%	7.5625	3.7813
12	13	*Superior*	3.00	99.87%	86.64%	98.91%	99.46%	9.0000	4.5000
		Validation Scores :	5.80	5.42	5.64	5.62	5.62	5.74	5.75
						Average Validation Score:	5.66	*or Good*	

Table 9 illustrates the use of a standardised table based on a Normal Distribution assumption for combining different ratings measures into an average PD Validation score outcome.

Figure 2

PD Validation Rating Methods Comparison

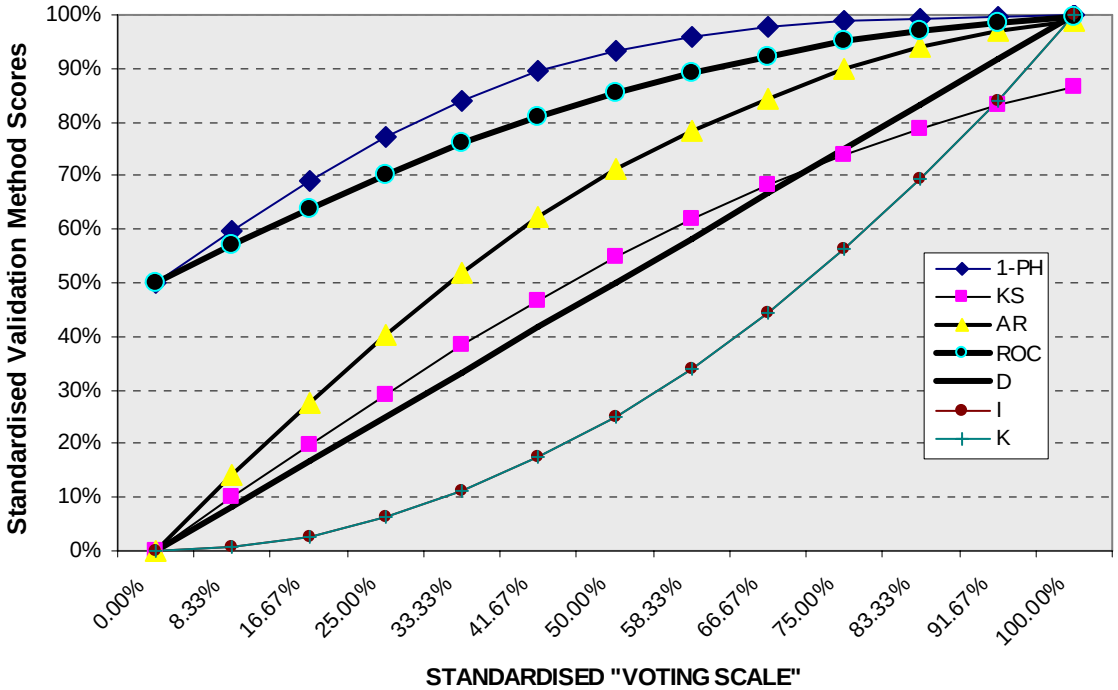

Figure 2 illustrates the individual contributions by differing methods under the standardised table based on a Normal Distribution assumption in PD Validation outcomes. These different measures combine to effectively "vote" on a consensus validation score outcome.

Applying the combined approach to the test data reveals the following outcomes:

Table 10

Internal Bank Ratings	Defaults	Non-Defaults
9	1	7
8	1	5
7	1	4
6	3	2
5	3	3
Totals	9	21

Validation Measures	Statistic	Validation Score
Mean Differences	0.86186	4.44746
CND% > 50%CDef (1-PH)	0.80952	4.53185
K-S Statistic	0.42857	5.53636
Accuracy Ratio	0.44444	4.34866
ROC Statistic	0.72222	4.34866
Information Statistic	0.84336	4.64197
Kullback-Leibler Stat.	0.43338	4.69546
Average Validation Score:		4.65006
Internal Bank Ratings:		*Satisfactory*

Table 11

External Bank Ratings	Defaults	Non-Defaults
9	1	6
8	1	7
7	1	4
6	3	2
5	3	2
Totals	9	21

Validation Measures	Statistic	Validation Score
Mean Differences	1.00651	5.02604
CND% > 50%CDef (1-PH)	0.85714	5.29805
K-S Statistic	0.47619	6.10368
Accuracy Ratio	0.49735	4.80478
ROC Statistic	0.74868	4.80478
Information Statistic	1.04837	5.08600
Kullback-Leibler Stat.	0.54828	5.17166
Average Validation Score:		5.18500
External Bank Ratings:		*Good*

Table 12

Pr(X%)	Model 1 Quintiles	Defaults	Non-Defaults
0.3%	0.2	0	6
0.6%	0.4	0	6
1.0%	0.6	2	4
3.5%	0.8	2	4
6.8%	1	5	1
	Totals	9	21

Validation Measures	Statistic	Validation Score
Mean Differences	1.71184	7.84737
CND% > 50%CDef (1-PH)	0.95714	7.89539
K-S Statistic	0.57143	7.34435
Accuracy Ratio	0.76190	7.69733
ROC Statistic	0.88095	7.69733
Information Statistic	1.25765	5.45805
Kullback-Leibler Stat.	1.43336	7.75905
Average Validation Score:		7.38555
Model 1 Quintiles:		*Strong*

Table 13

Pr(X%)	Model 2 Quintiles	Defaults	Non-Defaults
0.3%	0.2	0	6
0.6%	0.4	0	6
0.9%	0.6	2	4
2.6%	0.8	3	3
6.1%	1	4	2
	Totals	9	21

Validation Measures	Statistic	Validation Score
Mean Differences	1.49733	6.98931
CND% > 50%CDef (1-PH)	0.88095	5.74724
K-S Statistic	0.57143	7.34435
Accuracy Ratio	0.69841	6.85696
ROC Statistic	0.84921	6.85696
Information Statistic	0.70422	4.32392
Kullback-Leibler Stat.	1.00133	6.64023
Average Validation Score:		6.39414
Model 2 Quintiles:		*Very Good*

Table 14

Summary Table

Rating Model	Validation Score	Description	Rankings
Model 1 Quintiles	**7.3856**	Strong	**1**
Model 2 Quintiles	**6.3941**	Very Good	**2**
External Bank Ratings	**5.1850**	Good	**3**
Internal Bank Ratings	**4.6501**	Satisfactory	**4**

We note that the outcome ranks for the combination approach match the simpler chi-squared tests. These outcomes are also in agreement with van Deventer and Imai for their ROC component measure. However, we have now separated the rankings with distinct descriptors (these were pre-set as per our framework requirement) and a regulator or bank management could now more readily identify with the strength of the rating differences. There is a discernible difference, for example, between a "Strong" PD rating validation and one that is merely "Satisfactory". Furthermore, if we were to use decile bandings for Models 1 and 2 instead of quintile grades then the PD Validation scores would increase for each (7.539 and 6.997 respectively) but the rating descriptors and ranking outcomes would remain the same.[12]

Thus, from the application of this Combination Test to the entire PD rating outcomes we would have adequately satisfied parts one and two of our proposed PD Validation framework. It would now require an independent credit risk control unit to be set up (as per paragraphs 441 and 442 of the Accord) within, for example, an Enterprise Risk Metrics structure of a large bank to ensure that the crucial part three requirement (for an ongoing system) is in place to effectively monitor the rating outcomes at regular intervals (at least annually and perhaps quarterly, for example). However, we may also wish to explore the rating outcomes in greater depth and thus, we propose more advanced or sophisticated measures, which would assist any institution seeking an advanced-IRB status to perhaps move beyond just the minimalist requirements. These tests, which are outlined in the next section of this paper, necessarily involve comparisons of alternative test and range outcomes, and are thus especially useful in any "champion versus challenger' assessment, or for tests involving model "out-of-sample" and "out-of-time" comparisons.

[12] At the extreme, if we took every available data point for Models 1 and 2 (as opposed to grouping them in quintiles) we would for example improve the ROC statistic from 88.09% for Model 1 up to 90.48%, and from 84.92% for Model 2 up to 89.41%.

3. More Sophisticated Tests

The measurement of average outcomes is only one dimension of a rating validation process. Measurement of the range of outcomes gives a much better overall understanding of the confidence that one can place on the rating results. To do this, we need measures of the variance of the area under the receiver-operating curve (and the accuracy ratio). Several approximations are available for deriving these bounds (as per Engelmann, Hayden and Tasche, 2003). For example, for the Internal Bank Ratings example, the bounds can be reported as follows using our illustrative data set:

3.1 Variance of PD Validation Statistics (AR & ROC)

Table 15

	Internal Bank Ratings			
	Lower Bound @ 95% CI	Upper Bound @ 95% CI	Lower Bound @ 95% CI	Upper Bound @ 95% CI
Method	deLong, deLong & Clarke-Pearson		Hanley & McNeal	
Accuracy Ratio	Calc. AR = 44.44%			
Range:	**1.79%**	**87.10%**	1.84%	87.04%
ROC Statistic	Calc. ROC = 72.22%			
Range:	**50.90%**	**93.55%**	50.92%	93.52%

Table 15 above shows the range boundaries for the calculated ROC and AR statistics are very wide but this is more to do with the small sample used.

3.2 ROC Confidence Interval Equations (Example as per Engelmann, Hayden and Tasche, 2003):

$$\left[\hat{U} - \hat{\sigma}_{\hat{U}}\Phi^{-1}\left(\frac{1+\alpha}{2}\right), \hat{U} + \hat{\sigma}_{\hat{U}}\Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right]$$

$where,$

$$\hat{U} = \frac{1}{(N_D N_{ND})}\sum_{(D,ND)} \upsilon_{D,ND}$$

$$\upsilon_{D,ND} = \begin{cases} 1, if\ S_D < S_{ND} \\ 0.5, if\ S_D = S_{ND} \\ 0, if\ S_D > S_{ND} \end{cases}$$

S_D and S_{ND} represent the distributions of Default Scores and Non-Default Scores respectively. For example, $P(S_D < S_{ND})$ is the probability of a defaulter with score s_D drawn from distribution S_D being lower than a non-defaulter with score s_{ND} drawn from distribution S_{ND}.

$$\hat{\sigma}_{\hat{U}} = \sqrt{\frac{1}{4(N_D-1)(N_{ND}-1)}\left[\hat{P}_{D\neq ND} + (N_D-1)\hat{P}_{D,D,ND} + NL\right]}$$

$$NL = (N_{ND}-1)\hat{P}_{ND,ND,D} - 4(N_D + N_{ND} - 1)\left(\hat{U} - 0.5\right)^2$$

$$\hat{P}_{D\neq ND} = P(S_D \neq S_{ND})$$

$$\hat{P}_{D,D,ND} = P(S_{D,1}, S_{D,2} < S_{ND}) + P(S_{ND} < S_{D,1}, S_{D,2}) - P(S_{D,1} < S_{ND} < S_{D,2}) - P(S_{D,2} < S_{ND} < S_{D,1})$$

$$\hat{P}_{ND,ND,D} = P(S_{ND,1}, S_{ND,2} < S_D) + P(S_D < S_{ND,1}, S_{ND,2}) - P(S_{ND,1} < S_D < S_{ND,2}) - P(S_{ND,2} < S_D < S_{ND,1})$$

$$\Phi^{-1} = Cumulative\ Normal\ Distribution\ function, \alpha = Confidence\ Level$$

3.3 ROC Curve Comparisons

Figure 3

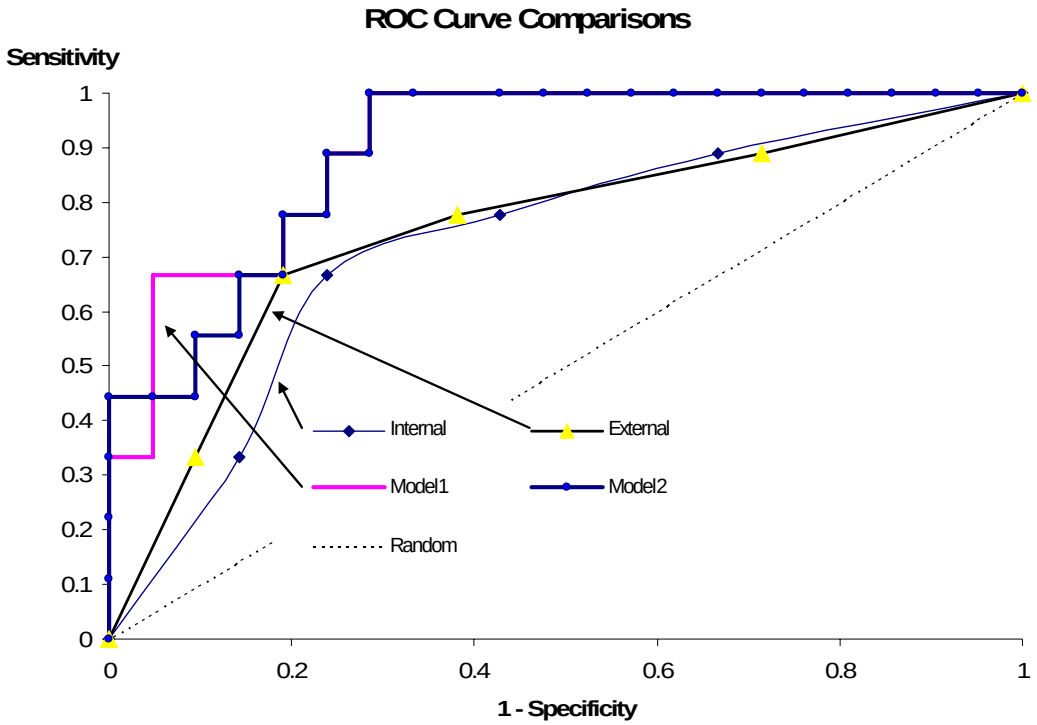

Figure 3 above shows that Model 1 is best in terms of area covered under the ROC.

Aside from a graphical comparison of the ROC curves for each individual model, we can also apply a test comparing each pair of models using a T-statistic for ROC outcome comparisons as per Engelmann, Hayden and Tasche (2003) - see Table 16 below.

3.4 Comparison of Areas under the ROC Curves for Different Rating Models[13]

Table 16

Rating Model (Area under the ROC) Comparisons

		Prob. both same
1	Model 1 Deciles versus Internal Bank Ratings	36.32%
2	Model 1 Deciles versus External Ratings	46.01%
3	Model 2 Deciles versus Internal Bank Ratings	47.04%
4	Model 2 Deciles versus External Ratings	57.67%
5	Model 1 Deciles versus Model 2 Deciles	85.72%
6	External Ratings vs. Internal Bank Ratings	89.17%

From the table above we can see that the area under the ROC comparison for pairs of models reveals that the Internal and External Rating models are most comparable (very high probability of same area under the ROC curve), whilst the two models that are most different to each other are the Internal Bank Rating Model and Model 1. These results tend to confirm the overall ranking patterns from previous tests.

[13] See detailed formulae in Appendix (B).

3.5 Likelihood Ratio Test Statistic[14]

The logistic regression approach makes it possible to test the null hypothesis: that the addition of new information to an existing model will not add any more explanatory power. For example, we can combine two models (an ordinal PD model with a cardinal PD model) and assess if the combined model yields any additional predictive power. In this respect, we are using logistic regression as a testing tool (not as a modelling tool in the usual approach).

Combining the data of two or more models and then gauging the significance of the outcome requires the use of the likelihood ratio test statistic. The likelihood ratio (LR) test statistic can be expressed as:

LR = -2[ln(likelihood of base regression) – ln(likelihood of new regression with added variables)]

The added variables are arranged as extra dummy variables (less one column to allow for k-1 degrees of freedom) for ordinal PD scale models, and simply as a one variable explanatory model for cardinal PD scale models. For example, to derive the base model output for the Internal Rating Model we would create the following data matrix:

Table 17

Bankruptcy Index	Dummy5	Dummy6	Dummy7	Dummy8	Internal Rating	Number
0	0	0	0	0	9	1
0	0	0	0	0	9	2
0	0	0	0	0	9	3
0	0	0	0	0	9	4
1	0	0	0	0	9	5
0	0	0	0	0	9	6
0	0	0	0	0	9	7
0	0	0	0	1	8	8
0	0	0	0	1	8	9
1	0	0	0	1	8	10
0	0	0	0	0	9	11
0	0	0	0	1	8	12
0	0	0	0	1	8	13
0	0	0	0	1	8	14
0	0	0	1	0	7	15
0	0	0	1	0	7	16
1	0	0	1	0	7	17
0	0	0	1	0	7	18
0	0	0	1	0	7	19
0	0	1	0	0	6	20
1	0	1	0	0	6	21
1	0	1	0	0	6	22
0	0	1	0	0	6	23
1	0	1	0	0	6	24
0	1	0	0	0	5	25
0	1	0	0	0	5	26
0	1	0	0	0	5	27
1	1	0	0	0	5	28
1	1	0	0	0	5	29
1	1	0	0	0	5	30

[14] See van Deventer and Imai, op. cit., p. 112.

Table 17 above shows that each new Dummy variable represents a rating. For example, rating 9 contains all zeroes as its representation is derived from all other variables. The Dummy8 variable receives a value whenever Internal-rating 8 is present (and all other dummy variables thus receive a zero value). The logistic regression is then run with the bankruptcy (or default indicator) variable as the dependent variable and Dummy variables 5 to 8 only as the explanatory or independent variables. Using any statistical tool that can do a logistic regression (we use a user-defined function within Excel), you can then derive the required statistics, as per the following table.

Table 18

Logistic Regression Model Output

No. of Observations =	30	**Internal Ratings Model**
No. of 0's =	21	**P-Value**
G-Statistic =	**6.41765**	**17.005%**
LogLikelihood =	**-15.11710**	
-2 x LogLikelihood =	30.23421	

	LREGR Coefficients	Standard Errors	T-Stats	P-Value
Constant	-1.7917595	1.0801234	-1.658847	10.9642%
Dummy5	2.19722458	1.4142135	1.55367243	13.2832%
Dummy6	2.19722458	1.4142135	1.55367243	13.2832%
Dummy7	0.40546511	1.5545632	0.26082254	79.6366%
Dummy8	-0.1541507	1.5197116	-0.1014342	92.0015%

Table 18 above highlights the key output statistics from the likelihood ratio test.

Thus, to compare the Internal Rating Model with Model 1 for example, the data would need to be arranged as follows:

Table 19

Bankruptcy Index	Dummy5	Dummy6	Dummy7	Dummy8	Model **1** Default Probabilty	Internal Rating
0	0	0	0	0	0.140	9
0	0	0	0	0	0.850	9
0	0	0	0	0	0.160	9
0	0	0	0	0	0.270	9
1	0	0	0	0	3.560	9
0	0	0	0	0	0.790	9
0	0	0	0	0	0.390	9
0	0	0	0	1	0.460	8
0	0	0	0	1	0.570	8
1	0	0	0	1	4.250	8

Table 19 above shows a ***truncated*** version of how we can combine the Internal Ratings Model with Model 1, which because of it has cardinal PD numbers; we can simply input just one extra column to the data matrix and then run it through a logistic regression model.

Table 20

Logistic Regression Model Output

No. of Observations	30		**(Internal Ratings and Model 1)**	
No. of 0's =	21	**P-Value**	**Difference**	**P-Test**
G-Statistic =	**19.88025**	**0.131607%**	**13.46260**	**0.9223%**
LogLikelihood =	**-8.38581**			
-2 x LogLikelihood =	16.77161			

	Coefficients	**Standard Errors**	**T-Stats**	**P-Value**
Constant	-5.166943534	2.825525176	-1.82866661	7.9905%
Dummy5	-3.024020102	2.950290834	-1.02499051	31.5588%
Dummy6	2.5672162	2.593131383	0.99000622	33.2048%
Dummy7	2.202451168	2.71937809	0.80990987	42.5950%
Dummy8	-0.432752558	4.154425702	-0.10416664	91.7903%
Model 1 Default Probabilty	2.051680693	0.8567241	2.39479745	2.4791%

Table 20 above shows the output from a logistic regression model for a combined Internal Ratings Model and Model 1. However, the likelihood test requires a subtraction of the statistics derived above from the base model (shown previously) to determine the effect of additional variables. This statistic can be readily derived using the Excel function CHIDIST (Difference, k-1 degrees of freedom) or CHIDIST (13.4626, 4) = 0.9223%. The value for the Difference = -2[ln(likelihood of base regression 0 – ln(likelihood of new regression with variable j added)] = -2[-15.11710 x -1 x -8.38581] = 13.46258.

This process can be repeated for every possible paired combination and we ultimately derive the following summary outputs and rankings:

Table 21

Rating Model Log Likelihood Test Comparisons

		P-Values
1	Model 2 Deciles versus Internal Bank Ratings	0.05%
2	Model 1 Deciles versus Internal Bank Ratings	0.13%
3	Model 1 Deciles versus Model 2 Deciles	0.18%
4	Model 2 Deciles versus External Ratings	0.73%
5	Model 1 Deciles versus External Ratings	0.94%
6	External Ratings vs. Internal Bank Ratings	31.26%

Table 21 above shows the ranking outcomes (after combining all possible combinations of model pairs) using the log likelihood test statistics. This test is for the null hypothesis that the combined model has no explanatory power. We can see that combining Model 2 with the Internal Bank Ratings model is essentially better than random chance with 99.95% confidence. Also, combining Model 1 with the Internal Model is almost as good too, at the 0.13% significance level. However, combining the external and the Internal Models will yield only about 69% confidence that the explanatory power is better than random chance.

3.6 Brier Score Concept

The BIS[15] suggests the concept of a Brier Score as a suitable PD validation technique. The Brier score, which is also known as the mean square error (MSE), is derived as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2$$

Where, in a portfolio of *i =1…N* obligors we have the probability of defaults ($\hat{\pi}_i$) that are forecasted at the beginning of a period and the actual default ($y_i = 1$) or non-default ($y_i = 0$) outcomes for each obligor that are observed at the end of the period. The MSE is a measure of accuracy and thus quantifies the deviation of the forecasts and actual observations. The higher the forecasts' accuracy, then the smaller will be the MSE outcome.

In addition, Rauhmeier and Scheule (2005) suggest a method for decomposing the statistic into an overall **calibration** and **variance** components of the forecast error as follow:

Table 22

MSE Components	Formulae
*Overall **Calibration**:*	$\left(\bar{y} - \bar{\hat{\pi}}\right)^2$ ⇨
***Variance** of forecast error:*	
***Uncertainty** (or the variance for both the default and non-default classes)*	$+ s_y^2$
***Refinement** (or variance of the probability of default forecasts)*	$+ s_{\hat{\pi}}^2$
The cross-product interaction:	$-2 s_y s_{\hat{\pi}}$
***Association**: The Bravais-Pearson correlation coefficient between y (0 or 1 outcomes) and π (default probabilities) across all obligors.*	$r_{y\hat{\pi}} = \frac{n\left(\sum_{i=1}^{n}\hat{\pi}_i Y_i\right) - \left(\sum_{i=1}^{n}\hat{\pi}_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n}\hat{\pi}_i^2 - \left(\sum_{i=1}^{n}\hat{\pi}_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}$

Measures the fit between the mean default rates: $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

And the mean of the probability forecast: $\bar{\hat{\pi}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_i$

The MSE can also be further decomposed as follows to reveal the useful properties of **refinement** and **discrimination**:

Table 23

MSE Components	Formulae
***Refinement** (same as above)*	$s_{\hat{\pi}}^2$

[15] BIS Working Paper No. 14, p. 46

Discrimination I *(The squared differences between average forecasted probabilities and actual outcomes for defaults and non-default classes)*	$+\sum_y \frac{N_y}{N}\left(\bar{\hat{\pi}}_y - y\right)^2$
Discrimination II *(The squared differences between average forecasted probabilities for defaults and non-default classes to the overall average)*	$-\sum_y \frac{N_y}{N}\left(\bar{\hat{\pi}}_y - \bar{\hat{\pi}}\right)^2$

The MSE can be used to test the rating system's accuracy in forecasting default probability, using the following null hypothesis and test statistic that is (approximately) distributed as a standard normal distribution:

$$H_o : \hat{\pi}_i = \pi_i = E(MSE_{\pi_i=\hat{\pi}_i}) = \frac{1}{N}\sum_{i=1}^{N} \pi_i\left(1-\pi_i\right)$$

$$Z_s = \frac{MSE - E\left(MSE_{\pi_i=\hat{\pi}_i}\right)}{Var\left(MSE_{\pi_i=\hat{\pi}_i}\right)^{0.5}}$$

3.6.1 Application to example dataset

Calculations of the Brier Score components for the three external rating systems are shown in the tables below:

Table 24

	Internal Rating	External Rating	Rating Model 1	Rating Model 2
Brier Scores =	28.0150%	27.3022%	27.6813%	27.9632%

Table 25

Summary Table

Rating Model	Brier Score	Rankings
External Rating	**27.30%**	**1**
Rating Model 1	**27.68%**	**2**
Rating Model 2	**27.96%**	**3**
Internal Rating	**28.01%**	**4**

The two tables above illustrate the Brier Score applied to the sample dataset. Oddly enough, the Brier Score results show that the External Rating has the lowest score, indicating that it is the best predictor. However, from previous analysis we would suspect that this conclusion is possibly inappropriate. It could be that adding in the Not Rated probability of default from the previous PD Mapping exercise could be the culprit for this odd rank. However, the Brier Score can be decomposed to reveal some useful properties that may not be readily evident in its first application.

Table 26

Decomposition I of MSE	Internal Rating	External Rating	Rating Model 1	Rating Model 2
Overall Calibration =	0.0773952	0.0748528	0.0784504	0.0794000
Variance of Y (Uncertainty) =	0.2100000	0.2100000	0.2100000	0.2100000
Variance of pi (**Refinement**) =	0.0006743	0.0013243	0.0004229	0.0003531
$-2S_yS_{pi}$ =	-0.0237993	-0.0333529	-0.0188482	-0.0172217
Association (Bravais-Pearson) =	**33.2783%**	**39.4432%**	**63.9884%**	**58.7707%**
Association (KendallsTau B) =	27.2040%	30.1763%	49.7566%	48.8580%
***MSE* =**	28.0150%	27.3022%	27.6813%	27.9632%
Difference Check =	0.000000	0.000000	0.000000	0.000000
Std Devn of pi (**Refinement**) =	0.02596716	0.03639095	0.020565089	0.01879047

Table 26 above highlights the much higher association exists for models 1 and 2 (cross-checked using Kendall's Tau B statistic).

Table 27

Decomposition II of MSE	Internal Rating	External Rating	Rating Model 1	Rating Model 2
Refinement =	0.00067429	0.00132430	0.00042292	0.00035308
Discrimination**1** =	0.27954991	0.27190341	0.27656291	0.27940059
Discrimination**2** =	0.00007467	0.00020603	0.00017317	0.00012195
***MSE* =**	28.0150%	27.3022%	27.6813%	27.9632%
Difference Check =	0.000000	0.000000	0.000000	0.000000

Table 27 above highlights the Discrimination 1 property is also higher for both Models 1 and 2 compared to the External rating model, which is indicative of a superior rating model. Also, if Discrimination 2 property is lower (applicable for Models 1 & 2 by comparison to the External Rating Model) then this too is a sign of a superior rating model.

Table 28

Capital Requirements	Internal Rating	External Rating	Rating Model 1	Rating Model 2
PD =	2.18%	2.64%	1.99%	1.82%
LGD =	45%	45%	45%	45%
EAD =	1000	1000	1000	1000
M =	2.5	2.5	2.5	2.5
Avg Basel Capital Requirement =	**$0.67**	**$0.88**	**$0.57**	**$0.52**

Table 28 above shows the impact of averaging the potential outcomes using the Basel II risk weight equations, subject to a minimum PD of 0.03% (BIS Corporate minimum value). These capital requirement numbers depend upon the refinement properties of the model and the average Probability-of-Default (PD) values, given the assumptions of Loss-Given-Default (LGD) @ 45%, Exposure-At-Default (EAD) @ 1000 and Maturity (M) of 2.5 years.

Table 29

Summary Table

Rating Model	K Required	Rankings
Rating Model 2	**$0.52**	**1**
Rating Model 1	**$0.57**	**2**
Internal Rating	**$0.67**	**3**
External Rating	**$0.88**	**4**

Table 29 above shows the ultimate impact from the properties of refinement, discrimination I and discrimination II on the average Basel capital requirements such that we now have a ranking **reversal** of the previous PD Validation rankings based on Brier Score alone.

4. Calibration Tests of PD Outcomes

4.1 Traffic Lights Approach to PD Calibration

Tasche (2003) suggests a Traffic Lights approach for the PD Calibration component of the required PD Validation Framework. This methodology accounts for any potential **default** *correlations* (and **asset** *correlations*) and allows one to test the hypothesis that the observed default rate for any grade is within some tolerable limit as per the long-term default rate required for that rating grade. The approach assumes that the numbers of defaults are distributed as per the Basel II-Vasicek model, that the number of obligors in the portfolio exceeds at least 10, and that the portfolio is homogeneous in nature. The observed number of defaults is D, the predicted annual default probability is P (i.e., the overall or master-scale probability of default for the given rating grade) and N is the number of obligors in the rating grade. The Traffic Lights approach has pre-set critical values (D_G for "Green" and D_Y for "Yellow" and D_R for "Red") and these have been approximated by the means of the granularity methodology (as per Gordy, 2003). The convenience of this approach is that it neatly calculates the number of obligors required for the threshold limits that could otherwise be derived using the Binomial Test with correlations.[16] However, under the Traffic Lights approach one no longer needs to use the bivariate standard normal distribution function (with a derived default correlation parameter) as per the Binomial test. We show a hypothetical use of the approach in Table 30 below with a "Green" outcome for decile grade 0.4; a "Yellow" outcome for decile grade 0.5; and a "Red" outcome for decile grade 0.6. The colour of the outcome is derived by reference to the tables on the right-hand side for each grade's critical limits.

Table 30

Year:	2002		
Grade	0.4		
Observed number of defaults:	8	D	
Number of obligors:	83	N	
Actual probability:	9.6%	p' = D/N	
Critical default probability:	10.0%	p	
Rho:	12.1%	(asset correlation)	

Critical limits (using the granularity adjustment approach):

Lower Bound	Defaults	Upper Bound	Confidence Levels
0	<= D <=	19	<=95.0%
20	< D <=	36	<=99.9%
37	< D <=	83	>99.9%

Year:	2002		
Grade	0.5		
Observed number of defaults:	10	D	
Number of obligors:	77	N	
Actual probability:	13.0%	p' = D/N	
Critical default probability:	2.0%	p	
Rho:	16.4%	(asset correlation)	

Critical limits (using the granularity adjustment approach):

Lower Bound	Defaults	Upper Bound	Confidence Levels
0	<= D <=	5	<=95.0%
6	< D <=	16	<=99.9%
17	< D <=	77	>99.9%

Year:	2002		
Grade	0.6		
Observed number of defaults:	15	D	
Number of obligors:	93	N	
Actual probability:	16.1%	p' = D/N	
Critical default probability:	1.0%	p	
Rho:	19.3%	(asset correlation)	

Critical limits (using the granularity adjustment approach):

Lower Bound	Defaults	Upper Bound	Confidence Levels
0	<= D <=	4	<=95.0%
5	< D <=	14	<=99.9%
15	< D <=	93	>99.9%

Table 30 above shows an approach suggested by Dirk Tasche whereby after setting the threshold limit parameters, one tests the default rates derived over a period against the expected PD for that grade. Correlation effects are taken into consideration using the Basel II asset correlation factors.

[16] BIS WP #14 op. cit. pp. 47-52

4.2 Adjusting for PD and LGD Correlation Effects

The **Single Risk Factor** Capital Charge by Tasche (2003) can be utilised for coping with any known LGD and PD interaction effects. The single risk factor (V) is based on the historical interaction of the PD outcomes and the known LGD rates and is calculated as the volatility of the LGD time series for the bank(s). Compared with the current Corporate Basel II Capital Requirements, we can see from table 31 below, that given the example Credit Risk Factors (PD, LGD and M), a bank would have to increase its capital charge proportionately more than current requirements if the single risk factor (V) is set at 25% by its regulator.

Table 31

TABLE 1 as per Tasche (2003)

PD =	LGD =	M =	Basel II Capital Charge	Single Risk Factor Capital Charge	SRF/BCC
0.03%	45.0%	2.5	1.155%	**1.266%**	109.61%
0.10%	45.0%	2.5	2.372%	**2.651%**	111.73%
0.25%	45.0%	2.5	3.958%	**4.498%**	113.65%
0.50%	45.0%	2.5	5.569%	**6.417%**	115.24%
0.75%	45.0%	2.5	6.622%	**7.694%**	116.18%
1.00%	45.0%	2.5	7.385%	**8.629%**	116.84%
2.00%	45.0%	2.5	9.188%	**10.881%**	118.42%
3.00%	45.0%	2.5	10.275%	**12.277%**	119.48%
5.00%	45.0%	2.5	11.988%	**14.552%**	121.39%
7.50%	45.0%	2.5	13.863%	**17.158%**	123.77%
10.00%	45.0%	2.5	15.447%	**19.486%**	126.15%
15.00%	45.0%	2.5	17.723%	**23.185%**	130.82%
20.00%	45.0%	2.5	19.059%	**25.829%**	135.53%
100.00%	45.0%	2.5	0.000%	**21.006%**	
SRF:	Volatility of LGD:		V = 25%		

Table 31 above shows the Single Risk Factor capital Charge using a set of formulae as set out below. The SRF/BCC column above indicates the increased amount of capital required given the degree of correlation 'v' between LGD and PD (or the SRF) and is set at 25% in the above example. Column 1 has a range of PD values, column 2 a set Foundation approach LGD value of 45%, Column 3 has a Maturity value of 2.5 years and column 4 is the Corporate curve Basel Capital Charge as a percentage of the EAD. The Single Risk Factor Capital charge is derived using the equations set forth by Tasche (see VBA code below) and the final column is the ratio of the two capital requirement approaches.

From a Basel II perspective the SRF approach has the distinct advantage of continuing to treat the estimates of PD and LGD as independent factors. It remains a point of contention as to whether or not banks would still need to estimate *extreme* stressed values of these credit risk factor estimates—given that the regulator could now impose its own estimate of V to regulate maximum capital levels to adequately cover any abnormal recovery rates known to occur during recessionary periods (see Fyre 2000 and also Altman, Brady, Resti and Sironi 2002).

Conclusion

Undoubtedly, model validation will be an iterative process in which banks and regulators will periodically refine and improve their discrimination tools and methodologies over time. Nevertheless, we recommend the use of this proposed PD Validation framework for Basel II Internal Ratings-Based models as a starting point towards a requisite global standard.

Appendix (A) - VBA Code[17]

```
Function SingleRiskFactorCapitalCharge(PD, LGD, m, V)
'Main routine for Single Risk Factor Capital Charge alternative to the Basel 2
'Capital Charge
    Alpha = LGD * (1 - V) / V
    beta = (1 - LGD) * (1 - V) / V
'Correlation as per the Basel 2 methodology (for Corporates)
    rho = ((0.12 * (1 - Exp(-50 * PD)) / (1 - Exp(-50))) + _
        (0.24 * ((1 - (1 - Exp(-50 * PD)) / (1 - Exp(-50))))))

Dim W(5), TArray(5)
'Weights
    W(1) = 0.2369268851: W(2) = 0.4786286705: W(3) = 128 / 225
    W(4) = W(2): W(5) = W(1)
'Abscissa values
    TArray(1) = -0.9061798459: TArray(2) = -0.5384693101: TArray(3) = 0
    TArray(4) = -TArray(2): TArray(5) = -TArray(1)

'Set value for variable a1
    A1 = PD / (2 * Sqr(1 - rho))
'Sum the integration approximation
    sum = 0
    A = 0.999 'set by Basel 2
    X = InverseCDF(A)

For i = 1 To 5
        sum = sum + W(i) * FofX(TArray(i), PD, X, rho, Alpha, beta)
Next i

    ValueAtRisk = A1 * sum
    TL = ValueAtRisk
    EL = PD * LGD
'Incorporate Basel 2 Maturity Adjustment and EAD adjustment factors
    B = (0.11852 - 0.05478 * Log(PD)) ^ 2
    EADAdjustment = ((1 - 1.5 * B) ^ -1) * (1 + (m - 2.5) * B)
    SingleRiskFactorCapitalCharge = (TL - EL) * EADAdjustment
End Function
```

```
Function FofX(Ti, PD, X, rho, Alpha, beta)
'Called by the h sub-function routine of the capital charge routine
    Numerator = SND((InverseCDF((PD * (Ti + 1) / 2) + 1 - PD) -
Sqr(rho) * X) / Sqr(1 - rho))
    Denominator = SND(InverseCDF((PD * (Ti + 1)) / 2 + 1 - PD))
    LossFunction = MyBetaInverseCumDist((Ti + 1) / 2, Alpha, beta)
    FofX = Val(Numerator) / Val(Denominator) * LossFunction
End Function
```

[17] **MyBetaInverseCumDist**, **InverseCDF** and **SND** are specialized functions that calculate the cumulative Inverse Beta Distribution, the Inverse of the Cumulative Normal Distribution and the Standard Normal Density function of z respectively and are available on the website address on the front page of this paper. *NB: there is a typographical error in the paper, which prevents one from deriving the correct end result, but the VBA code above has the corrected formulae embedded inside the FofX function.*

```
Function MyGranularAdjustment(Alpha#, p#, n&, rho#)
'Created March 2005 for PD Validation Traffic Lights Test
'Mimics Tasche / Gordy Granular adjustment for limits of confidence
'interval for PDs
'Based on the article by Dirk Tasche (Deutsche Bundesbank)
'dirk.tasche@bundesbank.de
'See Tasche, Dirk (2003), A Traffic lights approach to PD Validation,
'Working Paper
'Output is the threshold level of defaults for a given confidence level
'Inputs:
'alpha is the confidence level required eg. 95% (double or #)
'p is the probability of default (double or #)
'n is the number of obligors for the granular grade (Long Integer or &)
'rho is the correlation as per Basel 2 (double or #)
    Dim C#, X#, r#, Phi#
'Transformations
    C = InverseCDF(p)
    X = InverseCDF(1 - Alpha)
    r = InverseCDF(Alpha)
    Bphi = myN((Sqr(rho) * r + C) / Sqr(1 - rho)) 'q(alpha,R)
    frac = (C - Sqr(rho) * X) / Sqr(1 - rho)
    Phi = SND(frac)
'Main equation
    MyGranularAdjustment = (n * Bphi) + 0.5 * (2 * Bphi - 1 - (Bphi / Phi) * (1 -
    Bphi) * (Sqr((1 - rho) / rho) * X + frac))
End Function
```

Appendix (B) – Formulae for Derivation of the Test Statistic for Comparison of Two Areas under the ROC Curve

The test statistic (T) is asymptotically χ^2-distributed with one degree of freedom, for confidence level α to test the null hypothesis of equality of both areas below the ROC curve.

.

$$T = \frac{\left(\hat{U}_1 - \hat{U}_2\right)^2}{\sigma^2_{\hat{U}_1} + \sigma^2_{\hat{U}_2} - 2\sigma_{\hat{U}_1,\hat{U}_2}}$$

$\hat{U}_1, \hat{U}_2$: ROC statistics for Curves 1 & 2 (see Section 3.1),

$\sigma^2_{\hat{U}_1}, \sigma^2_{\hat{U}_2}$: Variances of Curves 1 & 2 (see Section 3.1),

$\sigma_{\hat{U}_1,\hat{U}_2}$: Covariance between Curves 1 & 2, see below:

$$\sigma_{\hat{U}_1,\hat{U}_2} = \frac{1}{4\left(N_D - 1\right)\left(N_{ND} - 1\right)}\left[\tilde{P}^{12}_{D,D,ND,ND} + \left(N_D - 1\right)\tilde{P}^{12}_{D,D,ND} + \left(N_{ND} - 1\right)\tilde{P}^{12}_{ND,ND,D} - 4\left(N_D + N_{ND} - 1\right)\left(\hat{U}_1 - 0.5\right)\left(\hat{U}_2 - 0.5\right)\right]$$

Where, $\tilde{P}$ is an estimator for the Probability or $\mathrm{P}(\ldots)$ functions below:

$$\tilde{P}^{12}_{D,D,ND,ND} = P\left(S^1_D > S^1_{ND}, S^2_D > S^2_{ND}\right) + P\left(S^1_D < S^1_{ND}, S^2_D < S^2_{ND}\right) - P\left(S^1_D > S^1_{ND}, S^2_D < S^2_{ND}\right) - P\left(S^1_D < S^1_{ND}, S^2_D > S^2_{ND}\right)$$

$$\tilde{P}^{12}_{D,D,ND} = P\left(S^1_{D,1} > S^1_{ND}, S^2_{D,2} > S^2_{ND}\right) + P\left(S^1_{D,1} < S^1_{ND}, S^2_{D,2} < S^2_{ND}\right) - P\left(S^1_{D,1} > S^1_{ND}, S^2_{D,2} < S^2_{ND}\right) - P\left(S^1_{D,1} < S^1_{ND}, S^2_{D,2} > S^2_{ND}\right)$$

$$\tilde{P}^{12}_{ND,ND,D} = P\left(S^1_D > S^1_{ND,1}, S^2_D > S^2_{ND,2}\right) + P\left(S^1_D < S^1_{ND,1}, S^2_D < S^2_{ND,2}\right) - P\left(S^1_D > S^1_{ND,1}, S^2_D < S^2_{ND,2}\right) - P\left(S^1_D < S^1_{ND,1}, S^2_D > S^2_{ND,2}\right)$$

VBA Function: **ROCcurveComparisons**(Defaults1, NonDefaults1, Defaults2, NonDefaults2)

Pseudo-Algorithm for Derivation of Covariance Terms

[Forms part of the Test statistic for Comparison of Two Areas under the ROC Curve]

(*As per Bernd Engelmann*) [Email: bernd.engelmann@quanteam.de]

The following two examples illustrate the estimation of the various probabilities necessary to carry out the tests for the covariance term: $\sigma_{\hat{U}_1,\hat{U}_2}$.

$$\widetilde{P}\left(S_D^1 > S_{ND}^1, S_D^2 > S_{ND}^2\right) = \frac{1}{N_D N_{ND}} \sum_{(D,ND)} \mathbf{1}_{\left\{S_D^1 > S_{ND}^1\right\}} \mathbf{1}_{\left\{S_D^2 > S_{ND}^2\right\}}$$

The sum is over all pairs of defaulters and non-defaulters. The symbol **1** denotes the indicator function. The symbol S_D^1 denotes the score of a defaulter under rating method 1, S_D^2 means the score of a defaulter under rating method 2. An analogous interpretation holds for non-defaulters.

$$\widetilde{P}\left(S_D^1 > S_{ND,1}^1, S_D^2 > S_{ND,2}^2\right) = \frac{1}{N_D N_{ND} N_{ND}} \sum_{(D,ND,ND)} \mathbf{1}_{\left\{S_D^1 > S_{ND,1}^1\right\}} \mathbf{1}_{\left\{S_D^2 > S_{ND,2}^2\right\}}$$

Here, the sum is over all possible triples consisting of one defaulter and two non-defaulters. For example, the meaning of $S_{ND,1}^2$ is score of the first non-defaulter in the triple under rating method 2.

If the above sums are evaluated in a naive way, the computational time can be very large if the sample of rated companies is large. For example, the second sum can be evaluated in an efficient way as follows:

1. Sort the non-defaulters independently under both rating methods according to increasing scores.
2. Define a variable sum = 0
3. Loop over all defaulters:
   - Determine the score of the defaulter under both rating methods
   - Determine the number of non-defaulters with a lower score under rating 1 and the number of non-defaulters with a lower score under rating 2 and multiply these numbers
   - Add the result of the multiplication to sum
4. Divide sum by the denominator in the above formula

By carrying out steps 1-4, one essentially saves one loop over the number of non-defaulters, which is a considerable gain in computational time.

Appendix (C): COMBINATION TEST FORMULAE

D (or Mean Difference) Statistic

$$D = \frac{ABS(\bar{S}_G - \bar{S}_B)}{\sqrt{\frac{(\bar{S}_G * \sigma_G^2) + (\bar{S}_B * \sigma_B^2)}{N_B + N_G}}}$$

$$\bar{S}_B = \sum_i \frac{(S_i * b_i)}{N_B} \qquad \sigma_B^2 = \sum_i \frac{(S_i - \bar{S}_B) * b_i}{N_B}$$

$$\bar{S}_G = \sum_i \frac{(S_i * g_i)}{N_G} \qquad \sigma_G^2 = \sum_i \frac{(S_i - \bar{S}_G) * g_i}{N_G}$$

where,

b_i = Number of Bads (or Defaults) at Score class (i)

g_i = Number of Goods (or Non-Defaults) at Score class (i)

N_B = Total Number of Bads (or Defaults)

N_G = Total Number of Goods (or Non-Defaults)

$\bar{S}_G$ = Average Score of the Goods (or Non-Defaults)

$\bar{S}_B$ = Average Score of the Bads (or Defaults)

σ_G^2 = Variance of the Goods (or Non-Defaults)

σ_B^2 = Variance of the Bads (or Defaults)

S_i = Score for class (i)

ABS() = Absolute value function

VBA Functions:

D = DstatisticByGroups(b_i, g_i)

D = 0.86186403

Score = ScorecardGradeDStat(D)

Score = 4.44745611

	Numerical Example			Average Scores		Variances	
	Scores	Bads	Goods	Bads	Goods	Bads	Goods
i	**S_i**	**b_i**	**g_i**	$\bar{S}_B$	$\bar{S}_G$	σ_B^2	σ_G^2
1	0.2	3	3	0.0666667	0.0285714	0.0237037	0.036398
2	0.4	3	2	0.1333333	0.0380952	0.00148148	0.008846
3	0.6	1	4	0.0666667	0.1142857	0.00197531	0.00209
4	0.8	1	5	0.0888889	0.1904762	0.01234568	0.00216
5	1	1	7	0.1111111	0.3333333	0.03160494	0.029055
	Totals	**9**	**21**	**0.4666667**	**0.7047619**	**0.07111111**	**0.078549**

(1-PH) or the Cumulative Non-Default% given Cumulative Defaults at the 50% Level

$$1-PH = 1-[CP_G(S_i) + \frac{Med_B - S_i}{S_{i+1} - S_i} * p_G(S_{i+1})]$$

$$where \quad S_i < Med_B < S_{i+1}$$

$$with \quad Med_B = S_M + \frac{0.5 - CP_B(S_M)}{CP_B(S_{M+1})}$$

$$and \quad CP_B(S_M) < 0.5 < CP_B(S_{M+1})$$

where,

Med_B = Calculated Median Score for the Bads (or Defaults)

S_M = Score of the Median value

S_{M+1} = Score of the class above the Median

S_i = Score of the class i

S_{i+1} = Score of the class above i

$CP_G(S_i)$ = Cumulative probability of the Goods (or Non-Defaults) Class (below the median score)

$p_G(S_{i+1})$ = Probability of the Goods (or Non-Defaults) Class (above the median score)

$CP_B(S_M)$ = Cumulative probability of the Bads (or Defaults) Class (below the median score)

$CP_B(S_{M+1})$ = Cumulative probability of the Bads (or Defaults) Class (above the median score)

(NB: Scores need to be sorted from Lowest to Highest to derive Median statistic)

VBA Functions:

1-PH = PctCumNonDefaultsFor50PctCumDefaultsByGroups(b_i, g_i)

1-PH = 0.80952381

Score = ScorecardGradeOneMinusPH(1-PH)

Score = 4.53185251

	Numerical Example			Probability		Cumulative		Cumulative Probs	
	Scores	Bads	Goods	Bads	Goods	Bads	Goods	Bads	Goods
i	**S_i**	**b_i**	**g_i**	**p_b**	**p_g**	**B_i**	**G_i**	**CP_B**	**CP_G**
1	0.2	3	3	0.3333333	0.1428571	3	3	33.33%	14.29%
2	0.4	3	2	0.3333333	0.0952381	6	5	66.67%	23.81%
3	0.6	1	4	0.1111111	0.1904762	7	9	77.78%	42.86%
4	0.8	1	5	0.1111111	0.2380952	8	14	88.89%	66.67%
5	1	1	7	0.1111111	0.3333333	9	21	100.00%	100.00%
	Totals	**9**	**21**	**1**	**1**				

Med_B = 0.3

PH = 0.1904762

KS (or the Kolmogorov-Smirnov Statistic)

$$KS = \underset{i}{Max}\left(CP_B(S_i) - CP_G(S_i)\right)$$

where,

Max() = Maximum value function

S_i = Score of the class i

$CP_G(S_i)$ = Cumulative probability of the Goods (or Non-Defaults) Class for score class (i)

$CP_B(S_i)$ = Cumulative probability of the Bads (or Defaults) Class for score class (i)

VBA Functions:

KS = **KolmogorovSmirnovByGroups**(b_i, g_i)

KS = **42.857%**

Score = **ScorecardGradeKS**(KS)

Score = **5.5363605**

	Numerical Example			Cumulative		Cumulative Probs		Differences	
	Scores	Bads	Goods	Bads	Goods	Bads	Goods		
i	**S_i**	**b_i**	**g_i**	**B_i**	**G_i**	**CP_B**	**CP_G**	**$CP_B - CP_G$**	
1	0.2	3	3	3	3	33.333%	14.286%	19.048%	
2	0.4	3	2	6	5	66.667%	23.810%	42.857%	<== Maximum Value
3	0.6	1	4	7	9	77.778%	42.857%	34.921%	
4	0.8	1	5	8	14	88.889%	66.667%	22.222%	
5	1	1	7	9	21	100.000%	100.000%	0.000%	
	Totals	**9**	**21**						

AR (or the Accuracy Ratio Statistic)

$$AR = 1 - 2\sum_{i} p_b(S_i)\left[\frac{CP_G(S_{i-1}) + CP_G(S_i)}{2}\right]$$

with,

$$CP_G(S_0) = 0$$

where,

S_i = Score of the class i

$p_b(S_i)$ = Probability of the Bads (or Defaults) Class for score class (i)

$CP_G(S_i)$ = Cumulative probability of the Goods (or Non-Defaults) Class for score class (i)

$CP_B(S_i)$ = Cumulative probability of the Bads (or Defaults) Class for score class (i)

$CP_G(S_{i-1})$ = Cumulative probability of the Goods (or Non-Defaults) Class for score class (i-1)

VBA Functions:

AR = **GiniByGroups**(b_i, g_i)

AR = **44.444%**

Score = **ScorecardGradeGini**(AR)

Score = **4.348659**

	Numerical Example			Probability		Cumulative		Cumulative Probs		
	Scores	Bads	Goods	Bads	Goods	Bads	Goods	Bads	Goods	
i	**S_i**	**b_i**	**g_i**	**p_b**	**p_g**	**B_i**	**G_i**	**CP_B**	**CP_G**	**p_B ($CP_{GS(i-1)}$ + $CP_{G(Si)}$)/2**
1	0.2	3	3	0.3333333	0.1428571	3	3	33.333%	14.286%	2.381%
2	0.4	3	2	0.3333333	0.0952381	6	5	66.667%	23.810%	6.349%
3	0.6	1	4	0.1111111	0.1904762	7	9	77.778%	42.857%	3.704%
4	0.8	1	5	0.1111111	0.2380952	8	14	88.889%	66.667%	6.085%
5	1	1	7	0.1111111	0.3333333	9	21	100.000%	100.000%	9.259%
	Totals	**9**	**21**	**1**	**1**					**27.778%**

ROC (or the Receiver Operating Characteristic)

$$ROC = 1 - \sum_i p_b(S_i)\left[\frac{CP_G(S_{i-1}) + CP_G(S_i)}{2}\right]$$

with,

$$CP_G(S_0) = 0$$

where,

S_i = Score of the class i

$p_b(S_i)$ = Probability of the Bads (or Defaults) Class for score class (i)

$CP_G(S_i)$ = Cumulative probability of the Goods (or Non-Defaults) Class for score class (i)

$CP_B(S_i)$ = Cumulative probability of the Bads (or Defaults) Class for score class (i)

$CP_G(S_{i-1})$ = Cumulative probability of the Goods (or Non-Defaults) Class for score class (i-1)

VBA Functions:

ROC = **ROCByGroups**(b_i, g_i)

ROC = **72.222%**

Score = **ScorecardGradeROC**(ROC)

Score = **4.348659**

	Numerical Example			Probability		Cumulative		Cumulative Probs		
	Scores	Bads	Goods	Bads	Goods	Bads	Goods	Bads	Goods	
i	**S_i**	**b_i**	**g_i**	**p_b**	**p_g**	**B_i**	**G_i**	**CP_B**	**CP_G**	**p_B ($CP_{GS(i-1)}$ + $CP_{G(Si)}$)/2**
1	0.2	3	3	0.3333333	0.1428571	3	3	33.333%	14.286%	2.381%
2	0.4	3	2	0.3333333	0.0952381	6	5	66.667%	23.810%	6.349%
3	0.6	1	4	0.1111111	0.1904762	7	9	77.778%	42.857%	3.704%
4	0.8	1	5	0.1111111	0.2380952	8	14	88.889%	66.667%	6.085%
5	1	1	7	0.1111111	0.3333333	9	21	100.000%	100.000%	9.259%
	Totals	**9**	**21**	**1**	**1**					**27.778%**

I (or the Information Statistic)

$$I = \sum_i \left(p_b(S_i) - p_g(S_i) \right) * \log \left[\frac{p_b(S_i)}{p_g(S_i)} \right]$$

$$\forall\, i, \quad p_g(S_i) > 0$$

where,

S_i = Score of the class i

$p_b(S_i)$ = Probability of the Bads (or Defaults) Class for score class (i)

$p_g(S_i)$ = Probability of the Goods (or Non-Defaults) Class for score class (i)

log() = natural Logarithm function

VBA Functions:

I = InformationStatisticByGroups(b_i, g_i)

I = 84.336%

Score = ScorecardGradeInfoStat(I)

Score = 4.64196622

	Numerical Example			Probability		
	Scores	Bads	Goods	Bads	Goods	
i	**S_i**	**b_i**	**g_i**	**p_b**	**p_g**	**$(p_b(S_i) - p_g(S_i)) * \log(p_b(S_i) / p_g(S_i))$**
1	0.2	3	3	0.3333333	0.1428571	16.139%
2	0.4	3	2	0.3333333	0.0952381	29.828%
3	0.6	1	4	0.1111111	0.1904762	4.278%
4	0.8	1	5	0.1111111	0.2380952	9.678%
5	1	1	7	0.1111111	0.3333333	24.414%
	Totals	**9**	**21**	**1**	**1**	**84.336%**

KL (or the Kullback-Leibler Statistic)

$$KL = \sum_i p_b(S_i) * \log\left[\frac{p_b(S_i)}{p_g(S_i)}\right]$$

$$\forall\, i, \quad p_g(S_i) > 0$$

where,

S_i = Score of the class i

$p_b(S_i)$ = Probability of the Bads (or Defaults) Class for score class (i)

$p_g(S_i)$ = Probability of the Goods (or Non-Defaults) Class for score class (i)

log() = natural Logarithm function

VBA Functions:

KL = **KLInformationStatByGroups**(b_i, g_i)

KL = **43.338%**

Score = **ScorecardGradeKL**(KL)

Score = **4.69545836**

	Numerical Example			Probability		
	Scores	Bads	Goods	Bads	Goods	
i	**S_i**	**b_i**	**g_i**	**p_b**	**p_g**	**$p_b(S_i) * \log(p_b(S_i) / p_g(S_i))$**
1	0.2	3	3	0.3333333	0.1428571	28.243%
2	0.4	3	2	0.3333333	0.0952381	41.759%
3	0.6	1	4	0.1111111	0.1904762	-5.989%
4	0.8	1	5	0.1111111	0.2380952	-8.468%
5	1	1	7	0.1111111	0.3333333	-12.207%
	Totals	**9**	**21**	**1**	**1**	**43.338%**

References

1. **Altman, Edward I., (2005)** Managing Credit Risk: The Challenge for the New Millennium, Stern School of Business, New York University http://pages.stern.nyu.edu/~ealtman/Presentations.htm
2. **Altman, Edward I., Brooks Brady, Andrea Resti, and Andrea Sironi. (2002).** The Link between Default and Recovery Rates: Implications for Credit Risk Models and Procyclicality. forthcoming, Journal of Business.
3. **Bank of International Settlements (2005)**, Studies on the Validation of Internal Rating Systems, Working Paper No. 14, February 2005
4. **Basel Committee on Banking Supervision (BCBS) (2004)** International Convergence of Capital Measurement and Capital Standards - A Revised Framework. http://www.bis.org/publ/bcbsca.htm
5. **Engelmann, Bernd, Evelyn Hayden and Dirk Tasche (2003),** Testing Rating Accuracy. Risk, January 2003. p. 82-86.
6. **Frye, Jon, (2000)** Depressing recoveries. RISK 13(11), 106–111.
7. **Gordy, Michael B., (2003),** A risk-factor model foundation for ratings-based bank capital rules. - Journal of Financial Intermediation 12, pp. 199 – 232.
8. **Hand, David J. and William E. Henley, (1997),** Statistical Classification Methods in Consumer Credit Scoring: A Review, Journal of the Royal Statistical Society, 160, Part 3. pp 523-541.
9. **Hamerle, Alfred, Robert Rauhmeier and Daniel Rosch, (2003**), Uses and Misuses of Measures for Credit rating Accuracy, University of Regensburg
10. **Rauhmeier, Robert and Harold Scheule (2005**), Rating Properties and their Implications for Basel II Capital, Asia**Risk,** May pp. 51-54 - (www.asiarisk.com.hk)
11. **Stein, Roger M., (2002),** Moody's | KMV - Technical Report #030124 – Benchmarking Default Prediction models – Pitfalls and Remedies in Model Validation, Revised July 13, 2002
12. **Stein, Roger M., (2005)**, The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. Journal of Banking & Finance 29 pp1213–1236
13. **Tasche, Dirk, (2003**), Traffic Lights Approach to the Validation of Default Probabilities, Working Paper, Deutsche Bundesbank, (April 1)
14. **Tasche, Dirk, (2004),** The single risk factor approach to capital charges in case of correlated loss given default rates. (Feb 17)
15. **Tasche, Dirk, (2005)**, Bank of International Settlements - Studies on the Validation of Internal Rating Systems, Working Paper No. 14, February 2005, Chapter Three, "Rating and Probability of Default Validation", pp. 28-59
16. **Thomas, Lyn. C., David B. Edelman and Jonathon N. Crook (2004**), Readings in Credit Scoring – recent developments, advances, and aims, Oxford finance; (first published - In Credit Scoring and Credit Control, (Eds. Thomas, Crook, and Edelman) pp. 123-38. OUP, Oxford 2002)
17. **van Deventer, Donald R., and Kenji Imai (2003),** Credit Risk Models & the Basel Accords, Wiley Finance
18. **Wilkie, A. David (2004),** Chapter 4: **"**Measures for comparing scoring systems", In Readings in Credit Scoring – recent developments, advances, and aims, Eds. Thomas, Lyn. C., David B. Edelman and Jonathon N. Crook (2004), Oxford finance