Journal of the Operational Research Society (2010) 61, 179–190

www.palgrave-journals.com/jors/

Enterprise risk management: coping with model risk in a large bank

D Wu[1]* and DL Olson[2]

[1]Reykjavik University, Iceland and Risklab, University of Toronto, Toronto, Canada; and [2]University of Nebraska, Lincoln, NE, USA

Enterprise risk management (ERM) has become an important topic in today's more complex, interrelated global business environment, replete with threats from natural, political, economic, and technical sources. Banks especially face financial risks, as the news makes ever more apparent in 2008. This paper demonstrates support to risk management through validation of predictive scorecards for a large bank. The bank developed a model to assess account creditworthiness. The model is validated and compared to credit bureau scores. Alternative methods of risk measurement are compared.

Journal of the Operational Research Society (2010) **61,** 179–190. doi:10.1057/jors.2008.144
Published online 7 January 2009

Keywords: enterprise risk management; model risk management; credit risk; statistical analysis

1. Introduction

The concept of enterprise risk management (ERM) developed in the mid-1990s in industry, expressing a managerial focus. ERM is a systematic, integrated approach to managing all risks facing an organization (Dickinson, 2001). It has been encouraged by traumatic recent events such as 9/11/2001 and business scandals to include Enron and WorldCom (Baranoff, 2004). A Tillinghast-Towers Perrin survey (Miccolis, 2003) reported that nearly half of the insurance industry used an ERM process (with another 40% planning to do so), and 40% had a chief risk officer. But consideration of risk has always been with business, manifesting itself in medieval coffee houses such as Lloyd's of London, spreading risk related to cargos on the high seas. Businesses exist to cope with specific risks efficiently. Uncertainty creates opportunities for businesses to make profits. Outsourcing offers many benefits, but also has a high level of inherent risk. ERM seeks to provide means to recognize and mitigate risks. The field of insurance was developed to cover a wide variety of risks, related to external and internal risks covering natural catastrophes, accidents, human error, and even fraud. Financial risk has been controlled through hedge funds and other tools over the years, often by investment banks. With time, it was realized that many risks could be prevented, or their impact reduced, through loss-prevention and control systems, leading to a broader view of risk management.

The subprime crisis makes companies increasingly stringent about the effective functions of ERM. The failure of the credit rating mechanismtroubles companies who needs timely signals about the underlying risks of their financial assets. Recent development in major credit ratings agencies such as Standard & Poor's (S&P) and Moody's have integrated ERM as an element of their overall analysis of corporate creditworthiness. This paper demonstrates validation of model risk in ERM. A large bank develops scorecard models to assess account creditworthiness. We validate predictive scorecards based on both internal banking and credit bureau data using various statistic measures.

This section introduced the problem of risk in organizations. Section 2 reviews risk modelling, to include balanced scorecard approaches. Section 3 discusses the use of credit rating performance validation models. Section 4 presents data case study of credit scorecards validation. Conclusions are presented in Section 5.

2. Risk modelling

It is essential to use models to handle risk in enterprises. Risk-tackling models can be (1) an analytical method for valuing instruments, measuring risk and/or attributing regulatory or economic capital; (2) an advanced or complex statistical or econometric method for parameter estimation or calibration used in the above; or (3) a statistical or analytical method for credit risk rating or scoring.

The Committee of Sponsoring Organizations of the Treadway Committee (COSO) is an organization formed to improve financial reporting in the US. COSO decided ERM was important for accurate financial reporting in 1999 (Levinsohn, 2004). Smiechewicz (2001) reviewed COSO focuses on ERM. The tools of risk management can include creative risk financing solutions, blending financial, insurance and capital market strategies (AIG, as reported by Baranoff,

*Correspondence: Desheng Dash Wu, RiskLab, University of Toronto, 105 St. George Street, Toronto, Canada M5S 3E6.
E-mail: dash@ru.is

Table 1 Balanced scorecard perspectives, goals, and measures

Perspectives	*Goals*	*Measures*
Financial	Survive	Cash flow
	Succeed	Quarterly sales, growth, operating income by division
	Prosper	Increase in market share, Increase in Return on Equity
Customer	New products	% sales from new products, % sales from proprietary products
	Responsive supply	On-time delivery (customer definition)
	Preferred suppliers	Share of key accounts' purchases, ranking by key accounts
	Customer partnerships	# of cooperative engineering efforts
Internal business	Technology capability	Benchmark *versus* competition
	Manufacturing excellence	Cycle time, unit cost, yield
	Design productivity	Silicon efficiency, engineering efficiency
	New product innovation	Schedule: actual *versus* planned
Innovation and learning	Technology leadership	Time to develop next generation
	Manufacturing learning	Process time to maturity
	Product focus	% products equaling 80% of sales
	Time to market	New product introduction *versus* competition

2004). Capital market instruments include catastrophe bonds, risk exchange swaps, derivatives/options, catastrophe equity puts (cat-e-puts), contingent surplus notes, collateralized debt obligations, and weather derivatives.

Many risk studies in banking involving analytic (quantitative) models have been presented. Crouhy *et al* (1998, 2000) provided comparative analysis of such models. Value-at-risk models have been popular (Alexander and Baptista, 2004; Chavez-Demoulin *et al*, 2006; Garcia *et al*, 2007; Taylor, 2007), partially in response to Basel II banking guidelines. Other analytic approaches include simulation of internal risk rating systems using past data. Jacobson *et al* (2006) found that Swedish banks used credit rating categories, and that each bank reflected it's own risk policy. One bank was found to have a higher level of defaults, but without adversely affecting profitability due to constraining high risk loans to low amounts. Elsinger *et al* (2006) examined systemic risk from overall economic systems as well as risk from networks of banks with linked loan portfolios. Overall economic system risk was found to be much more likely, while linked loan portfolios involved high impact but very low probability of default.

The use of scorecards has been popularized by Kaplan and Norton (1992, 2006) in their balanced scorecard, as well as other similar efforts to measure performance on multiple attributes (Bigio *et al*, 2004; Scandizzo, 2005). In the Kaplan and Norton framework, four perspectives are used, each with possible goals and measures specific to each organization. Table 1 demonstrates this concept in the context of bank risk management.

This framework of measures was proposed as a means to link intangible assets to value creation for shareholders. Scorecards provide a focus on strategic objectives (goals) and measures, and have been appliedin many businesses and governmental organizations with reported success. Papalexandris *et al* (2005) and Calandro and Lane (2006) both have proposed use of balanced scorecards in the context of risk management. Specific applications to finance (Anders and Sandstedt, 2003; Wagner, 2004), homeland security (Caudle, 2005), and auditing (Herath and Bremser, 2005) have been proposed.

Model risk pertains to the risk that models are either incorrectly implemented (with errors) or that make use of questionable assumptions, or assumptions that no longer hold in a particular context. It is the responsibility of the executive management in charge of areas that develop and/or use models to determine to what models this policy applies.

Lhabitant (2000) summarized a series of cases where model risk led to large banking losses. These models vary from trading model in pricing-stripped mortgage-backed securities to risk and capital models in deciding on the structured securities to decision models in issuing a gold card. Table 2 summarizes some model risk events in banking.

Sources of model risk arise from the incorrect implementation and/or use of a performing model (one with good predictive power) or thecorrect implementation/use of a non-performing model (one with poor predictive power). To address these risks, vetting of a statistical model is comprised of two main components: vetting and validation (Sobehart and Keenan, 2001). Vetting focuses on analytic model components, includes a methodology review, and verifies any implementation, while validation follows vetting and is an ongoing systematic process to evaluate model performance and to demonstrate that the final outputs of the model are suitable for the intended business purpose.

3. Performance validation in credit rating

Performance validation/backtesting focuses in credit rating on two key aspects: discriminatory power (risk discrimination) and predictive accuracy (model calibration) (Wu and

Table 2 Model risk events in banking

Model	*Trading and position management models*	*Decision models in retail banking*	*Risk and capital models*
Model risk	Booking with a model that does not incorporate all features of the deal, booking with an unvetted or incorrect model, incorrect estimation of model inputs (parameters), incorrect calibration of the model, etc	Incorrect statistical projections of loss, making an incorrect decision (eg lending decision) or incorrectly calculating and reporting the Bank's risk (eg default and loss estimation) as a result of an inadequate model, etc	Use of an unvetted or incorrect model, poor or incorrect estimation of model parameters, testing limitations due to a lack of historic data, weak or missing change control processes, etc

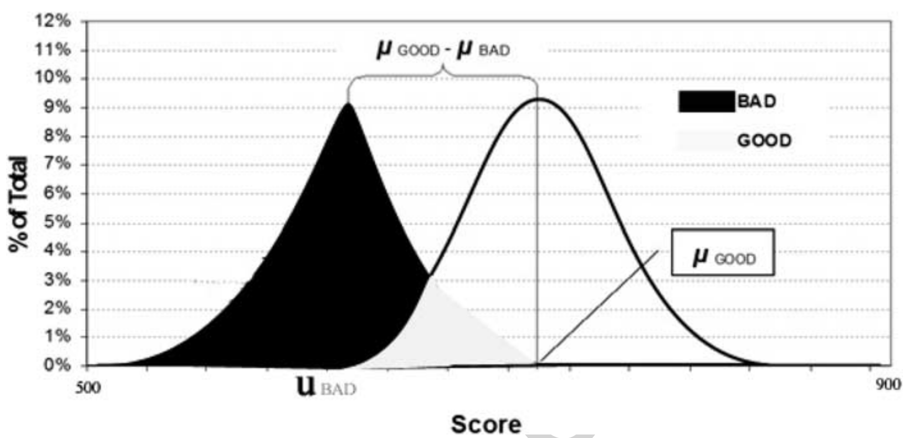

Figure 1 Illustration of divergence.

Olson, 2008). Discriminatory power generally focuses on the model's ability to rank-order risk, while predictive accuracy focuses on the model's ability to predict outcomes accurately (eg probability of defaults, loss given defaults, etc). Various statistic measures can be used to test the discriminatory power and predictive accuracy of a model (Sobehart and Keenan, 2001). Commonly used measures in credit rating include the divergence, Lorenz curve/CAP curve and the Kolmogorov–Smirnov (KS) statistic (Sobehart and Keenan, 2001).

The *divergence* measures the ability of a scoring system to separate good accounts from bad accounts (we informally define good and bad accounts as well as other concepts from credit scoring in the Appendix, but essentially good accounts are those that do not default, while bad accounts are those that do). This statistic is the squared difference between the mean score of the good and bad accounts divided by their average variance:

$$(\text{MeanGood} - \text{MeanBad})^2/((\text{VarGood} + \text{VarBad})/2)$$

The higher the divergence, the larger the separation of scores between good and bad accounts (see Figure 1). Ideally, 'good' accounts should be highly concentrated in the high score ranges and conversely, 'bad' accounts should be highly concentrated in the low score ranges.

Lorenz curve: The Lorenz curve can be produced after a sample of accounts has been scored by the model and then rank ordered based upon the score. If the model is predictive, the score of accounts or customers likely to exhibit the behaviour that is being predicted will trend towards one end of the distribution. The Lorenz curve is a variation of CAP curve in Figure 1. The predictive power of a model can be visually reviewed by tracing through the entire cumulative rank ordered customer distribution (on the x-axis) and comparing it to the distribution of customers that exhibited the behaviour to be predicted (on the y-axis). If a large proportion of the customers displaying the behaviour to be predicted is captured within a relatively small proportion of the entire population, the model is considered predictive. Normally, in addition to the curve of the respective models (namely, the Custom and Beacon Scores), two curves are included on the graph to act as baselines; first, the random line and second, the curve with perfect information.

Kolmogorov–Smirnov (K–S) *test*: Ideally, the bad curve should increase more quickly at the low score ranges, where these accounts should be found if the model is accurately rank ordering. Conversely,a low percentage of good accounts should be found in the low score range and then show a higher concentration in the high score range (see Figure 2). The K–S statistic identifies the maximum separation (percentage)

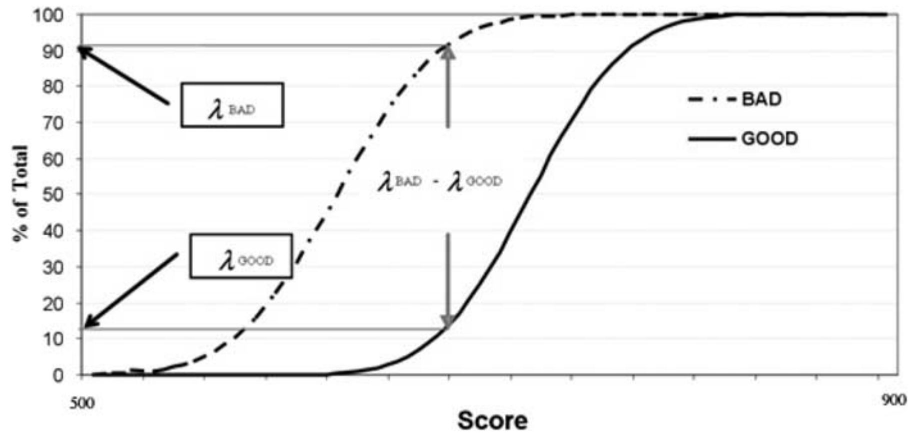

Figure 2 Illustration of K–S statistics.

Table 3 Scorecard performance validation January 1999–June 1999

		Scorecard	*Beacon*	*Beacon/Empirical*	*Scorecard (No Bureau score)*	*Bureau* 1	*Bureau* 2
Good	*N*	26 783	25 945	26 110	673	26 783	26 783
	Mean	250	734	734	222	42	208
	Std. dev	24	55	55	22	9	21
Bad	*N*	317	292	296	21	317	317
	Mean	228	685	685	204	40	188
	Std. dev	23	55	55	13	9	22
Total	*N*	27 100	26 237	26 406	694	27 100	27 100
	Mean	249	733	733	221	42	207
	Std. dev	24	55	55	22	9	21

between the cumulative percentage of goods *versus* bads at any given score. It may also be used to provide a cut-off score to assess applicants. The K–S statistic ranges from 0 to 100%.

Population stability index (PSI): This index gauges the discrepancy between the original development population used to generate the model and the population consisting of all the current applicants. It is used to measure comparatively the distribution of the scores between the two populations in order to detect any shifts in the samples. Assume p_i, $q_i, i = 1, \ldots, m$ are the ranges of scores for a more recent sample and for chosen benchmark, respectively. The PSI is calculated as follows:

$$PSI = \sum_{i=1}^{m} (p_i - q_i) \ln(p_i/q_i)/100$$

The following indices may be used as guidelines: an index of 0.10 or less is indicative of no real change between the samples; a score between 0.10 and 0.25 indicates some shift; and an index greater than 0.25 signifies a definite change that should be further analysed.

4. Case study: credit scorecard validation

The section aims to validate the predictive scorecard that is currently being used in a large Ontario bank. The names of this bank cannot be revealed due to confidentiality clauses. From the perspective of checking model risk, the whole process starts with a review of the bank background and raw data demonstration. This process will continue with a detailed validation through analysis of various statistic measures and population distributions and stability. This bank has a network of branches with a total of more than 8000 branches and 14 000 ATM machines operating across Canada. This bank successfully conducted a merger of two brilliant financial institutions in 2000 and became Canada's leading retail banking organization. It has also become one of the top three online financial service providers by providing online services to more than 2.5 million online customers. The used scorecard system in retail banking strategy will then need to be validated immediately due to this merger event. This scorecard system under evaluation predicts the likelihood that a 60–120-day delinquent account (mainly on personal secured and unsecured loans and lines of credit) will cure within the subsequent 3 months.

Table 4 Scorecard performance validation July 1999–December 1999

		Scorecard	*Beacon*	*Beacon/Empirical*	*Scorecard (No Bureau score)*	*Bureau* 1	*Bureau* 2
Good	*N*	20 849	20 214	20 302	547	20 849	20 849
	Mean	248	728	728	222	42	206
	Std. dev	24	54	54	23	9	21
Bad	*N*	307	296	297	10	307	307
	Mean	231	691	692	208	40	191
	Std. dev	23	55	55	12	9	22
Total	*N*	21 256	20 510	20 599	557	21 156	21 156
	Mean	248	727	727	222	42	206
	Std. dev	24	54	54	22	9	21

Table 5 Scorecard performance validation January 2000–June 2000

		Scorecard	*Beacon*	*Beacon/Empirical*	*Scorecard (No Bureau score)*	*Bureau* 1	*Bureau* 2
Good	*N*	23 941	23 254	23 361	580	23 941	23 941
	Mean	246	723	723	223	41	205
	Std. dev	24	54	54	21	9	21
Bad	*N*	533	490	495	38	533	533
	Mean	225	683	683	216	38	187
	Std. dev	21	51	51	16	9	20
Total	*N*	24 474	23 744	23 856	618	24 474	24 474
	Mean	245	723	723	222	41	204
	Std. dev	24	54	54	20	9	21

Table 6 Summary for performance samples

Time	*Statistic*	*Scorecard*	*Beacon*	*Beacon/Empirical*	*Scorecard*	*Bureau* 1	*Bureau* 2
January 1999–June 1999	KS value	39	37	37	44	14	36
	Divergence	0.869	0.792	0.79	0.877	0.07	0.814
	Bad%	1.17	1.11	1.12	3.03	1.17	1.17
July 1999–December 1999	KS value	33	26	26	42	10	29
	Divergence	0.528	0.45	0.435	0.624	0.04	0.498
	Bad%	1.45	1.44	1.44	1.8	1.45	1.45
January 2000–June 2000	KS value	38	33	33	26	14	34
	Divergence	0.843	0.606	0.598	0.147	0.078	0.789
	Bad%	2.18	2.05	2.07	6.15	2.18	2.18

Three time slots, that is, January 1999 to June 1999, July 1999 to December 1999, and January 2000 to June 2000, across six samples have been created and compared (see Tables 3–5). These are yielded by breaking-up funded accounts into three time slots based on their limit issue date for six samples: 'Scorecard data', 'Beacon', 'Beacon/Empirical', 'Scorecard without Bureau data', 'Bureau 1' and 'Bureau 2'. Tables 3–5 give the sample size, mean and standard deviation of these six samples for three time slots. Bad accounts in these tables include cases 90 days delinquent or worse, accounts closed with a 'NA (non-accrual)' status or that were written-off. Good cases are those that do not meet the bad definition. The bad definition is evaluated at 18 months. Specified time periods refer to month-end dates. For the performance analyses, the limit issue dates will be considered, while the population analyses will use the application dates. 'Scorecard' sample is our modelling sample and a combination of both 'Beacon/Empirical' and 'Scorecard without Bureau data'. 'Beacon' sample is designated as benchmarking sample for validation. But for deeper validation, we employ another two samples, that is, 'Bureau 1' and 'Bureau 2', from available Bureau score data. 'Bureau 1' and 'Bureau 2' are homogeneous to existing 'Scorecard data' in terms of bad and good account numbers. The homogeneity

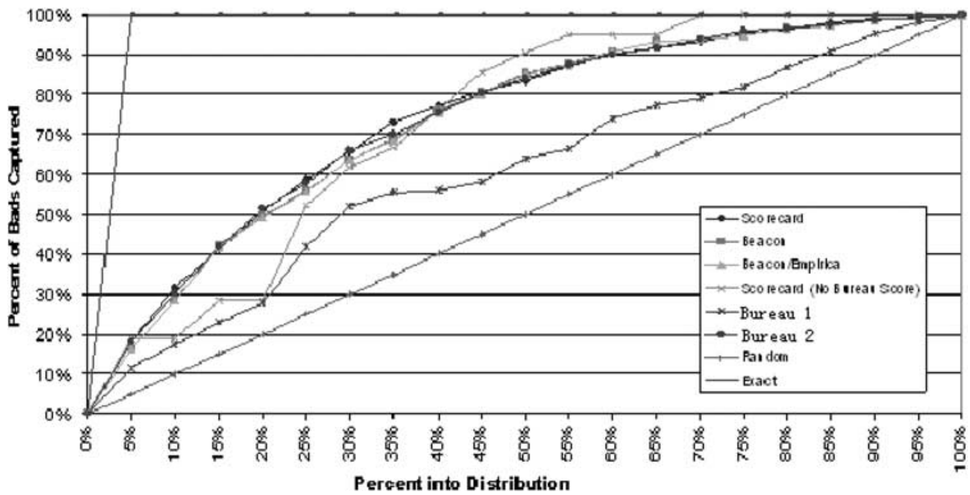

Figure 3 Lorenz curve on January 1999–June 1999 sample.

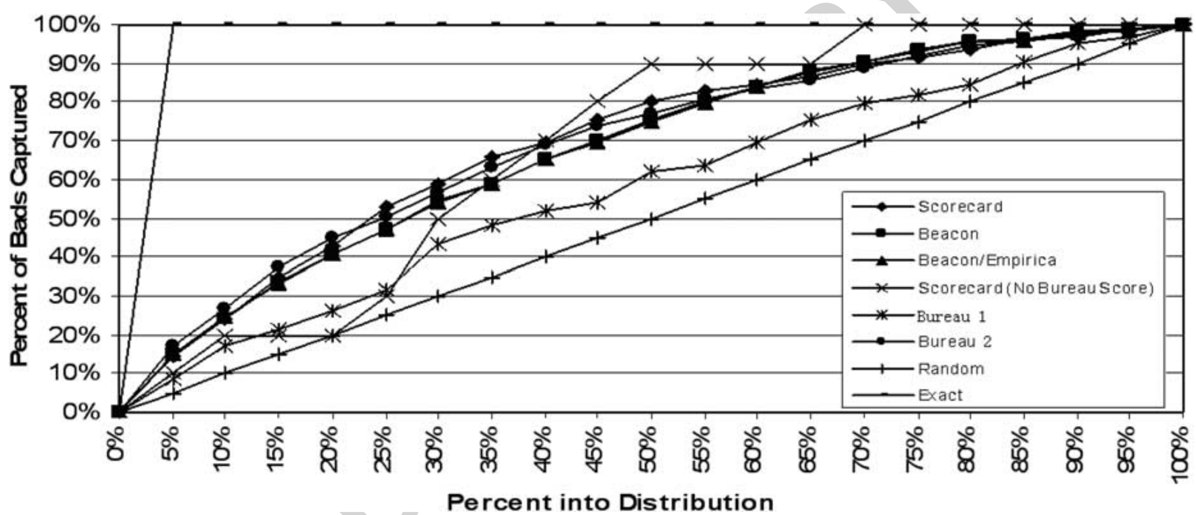

Figure 4 Lorenz curve on July 1999–December 1999 sample.

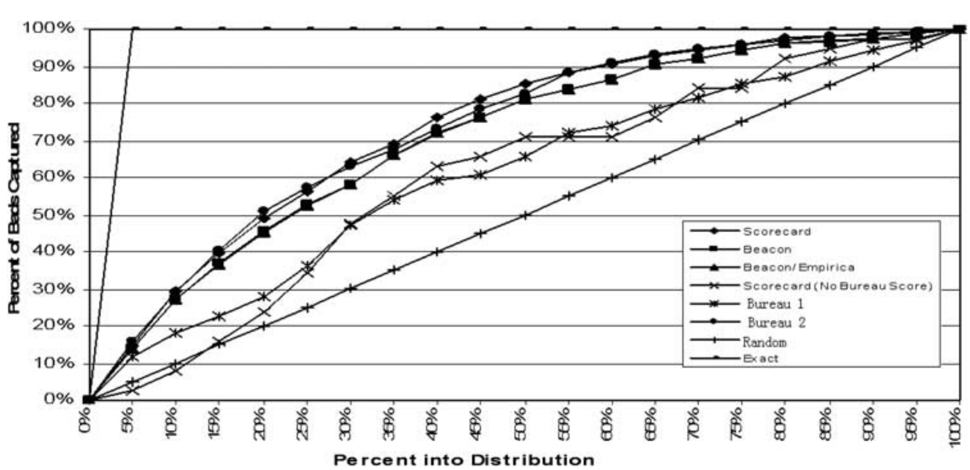

Figure 5 Lorenz curve on January 2000–June 2000 sample.

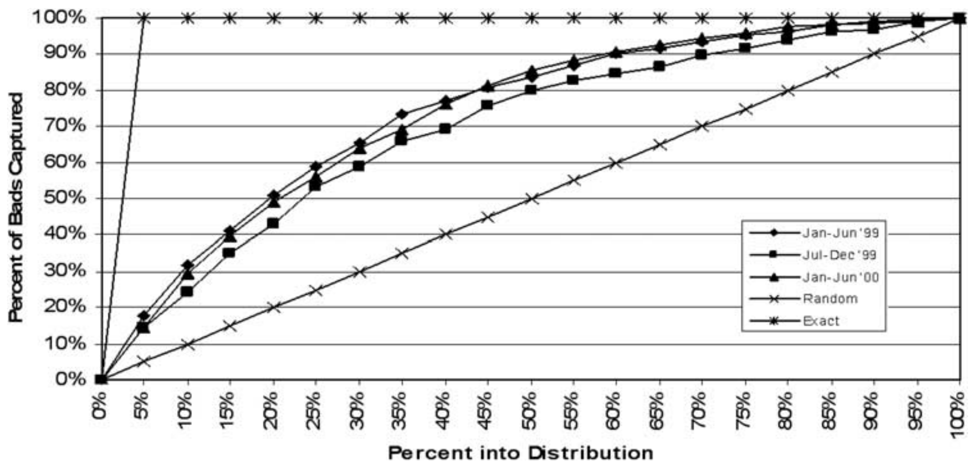

Figure 6 Performance comparison of three time slots for existing scorecard.

in samples will enable our later comparison more relevant. ‘Beacon’ sample has the largest mean and standard deviation values of scores, while ‘Bureau 1’ has the smallest mean and standard deviation values of scores. Our ‘Scorecard’ sample has a moderate mean and standard deviation values close to those of ‘Bureau 2’ while between ‘Beacon/Empirical’ and ‘Scorecard without Bureau data’. For example, Table 3 shows ‘Scorecard’ sample has a mean and standard deviation values of 250 and 24, close to 208 and 21 for ‘Bureau 2’. As time goes, scores of good accounts in our ‘Scorecard’ sample constantly decrease from 250 to 248 to 246, while bad values change from 228 to 231 to 225. The mean score of the total population constantly decreases from 249 to 248 to 245. These population changes will be detected later in the next section in a detailed validation process.

We note that numbers in these tables are rounded off to demonstrate the nature of score values assigned to different customer accounts. This also helps prevent revealing the bank’s business details for security. We will validate for each individual sample the model’s ability to rank order accounts based on creditworthiness. Comparison will be done to the credit bureau scores.

4.1. Statistical results and discussion

In order to validate the relative effectiveness of the Scorecard, we conduct statistic analysis and report results for the following statistical measures: divergence test, Lorenz curve, Kolmogorov–Smirnov (K–S) test, and population stability index. Table 6 presents the computation statistic values for KS value, divergence and bad% (bad ratio) of all performance samples across three time slots in Tables 3–5. The KS value and divergence values are computed using equations from Section 3. Bad% equals the ratio of number of bad accounts divided by the number of total accounts. Again, numbers in Table 6 are rounded off. Using the rounded number values in Tables 3–5, we can easily compute the divergence values close to those in the last row of each table. For example, relating to the Scorecard in Table 3: Mean Good $=250$, Std. dev. Good $=24$, Mean Bad $=228$, Std. dev. Bad $=23$. The difference (Mean Good$-$Mean Bad) is equal to 22 andthe average variances sum is equal to 552.5. The divergence is the fraction $484/552.5=0.876$, which is very close to non-rounded value 0.869.

Two findings are shown from Table 6. First, there is a trend of aging for the ‘Scorecard’ performance. From the third column of Table 6, we see that both the KS value and divergence are downgrading from the original value of 39% and 240, respectively. The KS and divergence statistics determine how well the models distinguished between ‘good’ and ‘bad’ accounts by assessing the properties of their respective distributions. The Scorecard was found to be a more effective assessor of risk for the earlier sample, Jan-99 to Jun-99, then the latest sample, Jan-00 to Jul-00, but was slightly less effective for the Jul-99 to Dec-99 sample. The bad ratio keeps increasing from 1.17 to 1.45 to 2.18%. Again, this demonstrates a hind of model risk and a thorough validation is required. Second, the performance statistics for the selected samples as provided in Table 6 indicate the superiority of the Scorecard as a predictive tool. In all three time slots, the existing ‘Scorecard’ outperforms both the benchmarking ‘Beacon’ model and other two designated Bureau models, that is, ‘Bureau 1’ and ‘Bureau 2’ models. The only model that can ‘beat’ ‘Scorecard’ is the ‘Scorecard without Beacon’ in January 1999–June 1999 and July 1999–December 1999 with divergence being 0.877 and 0.624. However, the divergence and KS values dropped to 0.147 and 26 for the Jan-00 to Jul-00 ‘Scorecard without Beacon’ model. This indicates ‘Scorecard without Beacon’ is not a stable model at all and should never be considered as an alternative tool. Instead, the existing ‘Scorecard’ from a combination of most Beacon data (about 97.44, 96.91 and 97.47% for three periods respectively) and some empirical internal banking data (about 2.6, 3.1 and 2.5% for three periods respectively) provides a

Table 7 Population stability for January 1999 to June 1999

Score range	*FICO development #*	*Jan–Jun 99 #*	*FICO development %*	*Jan–Jun 99 %*	*Proportion change (5)–(4)*	*Ratio (5)/(4)*	*Weight of evidence in (7)*	*Contribution to index (8)×(6)*	*Ascending cumulative of FICO %*	*Ascending cumulative of Jan–Jun 99 %*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
< 170	37 601	1430	7.13	1.96	−0.0517	0.2749	−1.2912	0.0668	7.13	1.96
170–179	25 093	1209	4.76	1.66	−0.0310	0.3483	−1.0546	0.0327	11.89	3.62
180–189	30 742	1888	5.83	2.59	−0.0324	0.4440	−0.8119	0.0263	17.72	6.21
190–199	37 128	3284	7.04	4.50	−0.0254	0.6394	−0.4471	0.0114	24.77	10.71
200–209	42 055	4885	7.98	6.70	−0.0128	0.8398	−0.1746	0.0022	32.74	17.41
210–219	46 355	5735	8.79	7.86	−0.0093	0.8944	−0.1116	0.0010	41.53	25.27
220–229	49 068	6716	9.31	9.21	−0.0010	0.9895	−0.0106	0.0000	50.84	34.48
230–239	48 577	7543	9.21	10.34	0.0113	1.1226	0.1156	0.0013	60.06	44.83
240–249	48 034	8762	9.11	12.02	0.0290	1.3187	0.2767	0.0080	69.17	56.84
250–259	46 023	9121	8.73	12.51	0.0378	1.4328	0.3596	0.0136	77.90	69.35
260–269	40 541	8826	7.69	12.10	0.0441	1.5739	0.4535	0.0200	85.59	81.45
270–279	37 940	8310	7.20	11.40	0.0420	1.5835	0.4596	0.0193	92.78	92.85
> 280	38 050	5216	7.22	7.15	−0.0006	0.9910	−0.0090	0.0000	100.00	100.00
Total	527 207	72 925	100	100				0.2027		

Note: Population stability index (sum of contribution): 0.2027.
The contribution index can be interpreted as follows:
⩽0.10 indicates little to no difference between the FICO development score distribution and the current score distribution.
0.10–0.25 indicates some change has taken place.
⩾0.25 indicates a shift in the score distribution has occurred.

Table 8 Population stability for July 1999 to December 1999

Score range	*FICO development #*	*July–Dec 99 #*	*FICO development %*	*July–Dec 99 #*	*Proportion change (5)–(4)*	*Ratio (5)/(4)*	*Weight of evidence in (7)*	*Contribution to index (8)×(6)*	*Ascending cumulative of FICO %*	*Ascending cumulative of July–Dec 99 #*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
< 170	37 601	1447	7.13	2.12	−0.0502	0.2968	−1.2146	0.0609	7.13	2.12
170–179	25 093	1352	4.76	1.98	−0.0278	0.4156	−0.8781	0.0244	11.89	4.10
180–189	30 742	2106	5.83	3.08	−0.0275	0.5284	−0.6379	0.0175	17.72	7.18
190–199	37 128	3609	7.04	5.28	−0.0176	0.7498	−0.2880	0.0051	24.77	12.46
200–209	42 055	5452	7.98	7.98	0.0000	0.9999	−0.0001	0.0000	32.74	20.43
210–219	46 355	6169	8.79	9.03	0.0023	1.0265	0.0261	0.0001	41.53	29.46
220–229	49 068	7009	9.31	10.25	0.0095	1.1018	0.0969	0.0009	50.84	39.71
230–239	48 577	7454	9.21	10.91	0.0169	1.1836	0.1685	0.0029	60.06	50.62
240–249	48 034	7908	9.11	11.57	0.0246	1.2699	0.2389	0.0059	69.17	62.19
250–259	46 023	7774	8.73	11.37	0.0264	1.3029	0.2646	0.0070	77.90	73.56
260–269	40 541	7362	7.69	10.77	0.0308	1.4007	0.3370	0.0104	85.59	84.33
270–279	37 940	6716	7.20	9.83	0.0263	1.3654	0.3114	0.0082	92.78	94.16
> 280	38 050	3993	7.22	5.84	−0.0138	0.8094	−0.2114	0.0029	100.00	100.00
Total	527 207	68 351	100	100				0.1461		

Note: Population stability index (sum of contribution): 0.1461.

Table 9 Population stability for January 2000 to June 2000

Score range	FICO development #	Jan–Jun 00 #	FICO development %	Jan–Jun 00 %	Proportion change (5)–(4)	Ratio (5)/(4)	Weight of evidence in (7)	Contribution to index (8)×(6)	Ascending cumulative of FICO %	Ascending cumulative of Jan–Jun 00 %
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
<170	37 601	1928	7.13	2.46	−0.0467	0.3448	−1.0648	0.0498	7.13	2.46
170–179	25 093	1838	4.76	2.34	−0.0242	0.4925	−0.7082	0.0171	11.89	4.80
180–189	30 742	3136	5.83	4.00	−0.0183	0.6859	−0.3770	0.0069	17.72	8.80
190–199	37 128	4784	7.04	6.10	−0.0094	0.8664	−0.1434	0.0013	24.77	14.91
200–209	42 055	6505	7.98	8.30	0.0032	1.0401	0.0393	0.0001	32.74	23.20
210–219	46 355	7212	8.79	9.20	0.0041	1.0462	0.0451	0.0002	41.53	32.40
220–229	49 068	8250	9.31	10.52	0.0122	1.1306	0.1227	0.0015	50.84	42.92
230–239	48 577	8762	9.21	11.18	0.0196	1.2129	0.1930	0.0038	60.06	54.10
240–249	48 034	8769	9.11	11.18	0.0207	1.2276	0.2050	0.0043	69.17	65.28
250–259	46 023	8451	8.73	10.78	0.0205	1.2348	0.2109	0.0043	77.90	76.06
260–269	40 541	7850	7.69	10.01	0.0232	1.3020	0.2639	0.0061	85.59	86.07
270–279	37 940	6736	7.20	8.59	0.0140	1.1939	0.1772	0.0025	92.78	94.67
> 280	38 050	4182	7.22	5.33	−0.0188	0.7391	−0.3024	0.0057	100.00	100.00
Total	527 207	78 403	100	100				0.1036		

Note: Population stability index (sum of contribution): 0.1036.

powerful tool for measuring account creditworthiness. There was a more distinct separation between 'goods' and 'bads' for the above-mentioned first two time slots, that is, Jan-99 to Jun-99 and Jan-00 to Jul-00, than the last: the maximum difference between the 'good' and 'bad' cumulative distributions was 39 and 38%, respectively, *versus* 33% for the remaining sample. Similarly, the divergence values were 0.869 and 0.843, *versus* 0.528 for the less effective sample.

The Lorenz curves corresponding to Table 6 are depicted in Figures 3–6. Figures 3–5 depict Lorenz curves for all six samples across three time periods while Figure 6 draws the performance comparison of three time slots for existing scorecard. Note that all figures are based on best applicant. In a data set that has been sorted by the scores in ascending order with a low score corresponding to a risky account, the perfect model would capture all the 'bads' as quickly as possible. The Lorenz curve assesses a model's ability to effectively rank order these accounts. For example, if 15% of the accounts were bad, the ideal or exact model would capture all these bads within the 15th percentile of the score distribution (the x-axis).

Again, the results indicate that the Scorecard is a good predictor of risk. Figures 3–5 indicate that the Scorecard curve lies above all other curves except 'Scorecard without Beacon',which was deemed as an invalid tool due to its instability. Scorecard performs better than, though not by a significant margin, the Credit Bureau Score. Among the three selected sampling periods, as can be seen from Figure 6, the two periods of Jan-99 to Jun-99 and Jan-00 to Jun-00 highlight a slightly better predictive ability than the period of Jul-99 to Dec-99.

It is possible that the Scorecard was better able to separate 'good' accounts from 'bad' ones for the earlier sample. On the other hand, the process to clean up delinquent unsecured line of credit accounts starting from mid-2001 may result in more bad observations for the latest sample (those accounts booked between Jan-00 and Jun-00 with a 18-month observation window will catch up with this clean-up process). This can be evidenced by the bad rate of 2.18% for the Jan-00 to Jun-00 sample, compared to 1.45% for the Jul-99 to Dec-99 sample, and 1.17% for the Jan-99 to Jun-99 sample. If most of these bad accounts in the clean-up have a low initial score, the predictive ability of the Scorecard on this cohort will be increased.

4.2. *Population distributions and stability*

We conduct a comparison analysis between the initial sample used to develop the model and subsequent sampling periods, which provides insight into whether or not the scorecard is being used to score a different population. The analyses considered all applicants are included, but outliers have been excluded, that is, invalid scorecard points. We consider four sampling periods for the cumulative and interval population distribution charts: the FICO (see the Appendix for a

Table 10 Total population stability index

Contribution index	*Jan-00*	*Feb-00*	*Mar-00*	*Apr-00*	*May-00*	*Jun-00*	*Jul-00*	*Aug-00*	*Sep-00*	*Oct-00*	*Nov-00*	*Dec-00*	
⩽0.10	0.0959		0.0962			0.0826	0.0999	0.0919	0.0940	0.0926	0.0693	0.0656	
0.10–0.25		0.1097		0.1313	0.1236								
Contribution index	*Jan-01*	*Feb-01*	*Mar-01*	*Apr-01*	*May-01*	*Jun-01*	*Jul-01*	*Aug -01*	*Sep -01*	*Oct-01*	*Nov-01*	*Dec-01*	*Jan-02*
⩽0.10	0.0940	0.0898	0.0787	0.0979	0.0829	0.0615	0.0696	0.0701	0.0907	0.0816	0.0817	0.0915	0.0771

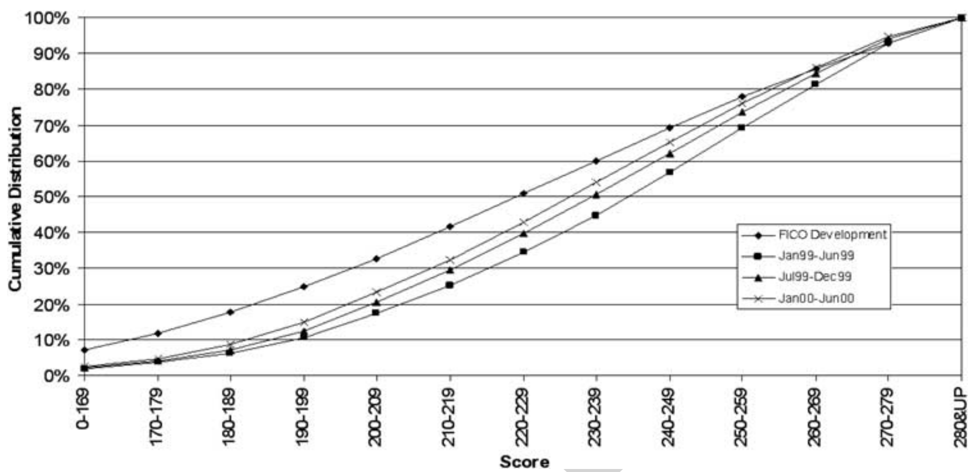

Figure 7 Cumulative population distribution on all applicants.

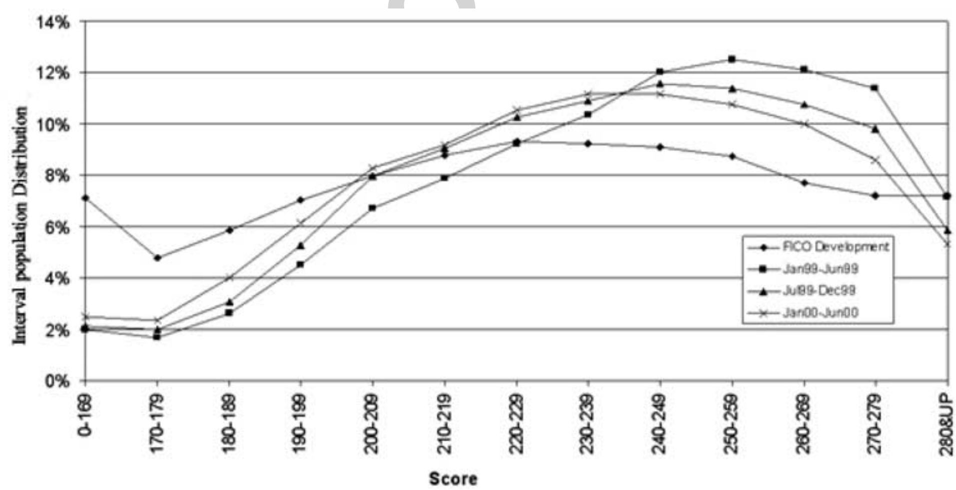

Figure 8 Interval population distribution on all applications.

definition of FICO score) development sample, Jan-99 to Jun-99, Jul-99 to Dec-99, and Jan-00 to Jul-00 (see Figures 7 and 8). From Figure 8, we can see a very notable population shift across the samples where the recent applicants clearly were scoring lower points than before. On the other hand, the development sample was markedly distinct from the three selected samples from three time slots.

We now use the population stability index to estimate the change between the samples. As mentioned in Section 4, a stability index of <0.10 indicates an insignificant shift, 0.10–0.25 requires some investigation and >0.25 means that a major change has taken place between the populations being compared. Tables 7–9 present a detailed score distribution report together with the 6-month population stability index for each of the above three selected sample from three time slots, included funded accounts only. Computation shows that the indexes for the three time slots on funded accounts are greater than 0.1, and the more recent samples scores a lower index than the older samples: 0.2027 for the Jan-99 to Jun-99 sample, 0.1461 for the Jul-99 to Dec-99 sample, and 0.1036 for the Jan-00 to Jun-00 sample. We also compute the monthly population stability which shows the monthly index for total applications (funded or not funded) in the past 2 years starting from Jan-00. This result further confirms on the declining trend with the monthly indexes for the past 20 months all rest within 0.1 (see Table 10).

As indicated in Figures 7 and 8, more of the latest sample accounts were having a lower score compared to the older samples. A tendency of lower score over time has been revealed. All of the three samples from three time slots had a score distribution higher than the Development sample.

The stability indices revealed that the greatest population shift occurred when the Scorecard was originally put in place, then the extent of shift reduced gradually across time. The indexes stayed within 0.1 for the past 20 months.

5. Conclusion and discussion

Maintaining a certain level of risk has become a key strategy to make profits in today's economy. Risk in enterprise can be quantified and managed using various models. Models also provide support to organizations seeking to control enterprise risk. We have discussed risk modelling and reviewed some common risk measures. Using the variation of these measures, we demonstrate support to risk management through validation of predictive scorecards for a large bank. The bank uses a Scorecard based on a combination of most Beacon data and some empirical internal banking data. The scorecard model is validated and compared to credit bureau scores. A comparison of the KS value and the divergence value between Scorecard and Bureau Score in the three different time periods indicated that internal existing scorecard is a better tool than Bureau Score to distinguish the 'bads' from the 'goods'. Vetting and validation of models may encounter many challenges in practice. For example, when retail models under vetting are relatively new to the enterprise, when there are large amounts of variables and data to manipulate and limited access to these data sets due to privacy restrictions, when validation tests are not standardized and there are demands for ability to change the measure if results do not look favourable, these challenges become apparent.

References

Alexander GJ and Baptista AM (2004). A comparison of VaR and CVaR constraints on portfolio selection with the mean-variance model. *Mngt Sci* **50**(9): 1261–1273.

Anders U and Sandstedt M (2003). An operational risk scorecard approach. *Risk* **16**(1): 47–50.

Baranoff EG (2004). Risk management: A focus on a more holistic approach three years after September 11. *J Insur Regul* **22**(4): 71–81.

Bigio D, Edgeman RL and Ferleman T (2004). Six sigma availability management of information technology in the Office of the Chief Technology Officer of Washington, DC. *Total Qual Mngt* **15**(5–6): 679–687.

Calandro Jr J and Lane S (2006). An introduction to the enterprise risk scorecard. *Measur Bus Excel* **10**(3): 31–40.

Caudle S (2005). Homeland security. *Public Perform Mngt Rev* **28**(3): 352–375.

Chavez-Demoulin V, Embrechts P and Nešlehová J (2006). Quantitative models for operational risk: Extremes, dependence and aggregation. *J Bank Fin* **30**: 2635–2658.

Crouhy M, Galai D and Mark R (1998). Model risk. *J Fin Eng* **7**(3/4): 267–288 (reprinted in Model risk: Concepts, calibration and pricing) In: Gibson, R (ed). *Risk Book*, 2000, pp 17–31.

Crouhy M, Galai D and Mark R (2000). A comparative analysis of current credit risk models. *J Bank Fin* **24**: 59–117.

Dickinson G (2001). Enterprise risk management: Its origins and conceptual foundation. *Geneva Pap Risk Insur* **26**(3): 360–366.

Elsinger H, Lehar A and Summer M (2006). Risk assessment for banking systems. *Mngt Sci* **52**(9): 1301–1314.

Garcia R, Renault É and Tsafack G (2007). Proper conditioning for coherent VaR in portfolio management. *Mngt Sci* **53**(3): 483–494.

Herath HSB and Bremser WG (2005). Real-option valuation of research and development investments: Implications for performance measurement. *Mngt Audit J* **20**(1): 55–72.

Jacobson T, Lindé J and Roszbach K (2006). Internal ratings systems, implied credit risk and the consistency of banks' risk classification policies. *J Bank Fin* **30**: 1899–1926.

Kaplan RS and Norton DP (1992). The balanced scorecard—Measures that drive performance. *Harvard Bus Rev* **70**(1): 71–79.

Kaplan RS and Norton DP (2006). *Alignment: Using the Balanced Scorecard to Create Corporate Synergies*. Harvard Business School Press Books: Cambridge, MA.

Levinsohn A (2004). How to manage risk—Enterprise-wide. *Strat Fin* **86**(5): 55–56.

Lhabitant F (2000). Coping with model risk. In: Lore M and Borodovsky L (eds). *The Professional Handbook of Financial Risk Management*. Butterworth-Heinemann: London.

Miccolis J (2003). ERM lessons across industries, http://www.irmi.com/Expert/Articles/2003/Miccolis03.aspx.

Papalexandris A, Ioannou G, Prastacos G and Soderquist KE (2005). An integrated methodology for putting the balanced scorecard into action. *Eur Mngt J* **23**(2): 214–227.

Scandizzo S (2005). Risk mapping and key risk indicators in operational risk management. *Econ Notes Banca Monte dei Paschi di Siena SpA* **34**(2): 231–256.

Smiechewicz W (2001). Case study: Implementing enterprise risk management. *Bank Account Fin* **14**(4): 21–27.

Sobehart J and Keenan S (2001). Measuring default accurately. *Credit Risk Special Rep, Risk* **14**: 31–33.

Taylor N (2007). A note on the importance of overnight information in risk management models. *J Bank Fin* **31**: 161–180.

Wagner H (2004). The use of credit scoring in the mortgage industry. *J Fin Serv Market* **9**(2): 179–183.

Wu D and Olson D (2008). *Enterprise risk management: Credit scoring in finance operations*. Working paper, Joseph L. Rotman School of Management, University of Toronto.

Appendix. Informal definitions

(a) Bad accounts refer to cases 90 days delinquent or worse, accounts closed with a 'NA (non-accrual)' status or that were written-off.

(b) Good accounts were defined as those that did not meet the bad definition.
(c) Credit score is a number that is based on a statistical analysis of a person's credit report, and is used to represent the creditworthiness of that person—the likelihood that the person will pay his or her debts.
(d) A credit bureau is a company that collects information from various sources and provides consumer credit information on individual consumers for a variety of uses.
(e) Custom score refers to the score assigned to existing customers or new applicants.
(f) Beacon score is the credit score produced at the most recognized agency Equifax in Canada.
(g) The FICO score is the credit score from Fair Isaac Corporation, a publicly traded corporation that created the best-known and most widely used credit score model in the United States.

Received November 2007;
accepted September 2008 after one revision