Benchmarking discriminatory power of credit risk rating models

At one time the bank manager looked you in the eye and determined your credit-worthiness. Now, this process has become more scientific and global. **D H LIYANA ARACHCHIGE F Fin** looks at internal credit rating models and methods of benchmarking them.

D H Liyana Arachchige F Fin National Australia Bank, Melbourne

Benchmarking internal credit risk models is one of the key issues faced by the banking industry with the implementation of New Basel II framework (The Basel Committee on Banking Supervision, 2006). Currently, credit risk rating models are being enhanced, purchased or built and validated by various banks across the globe.

The main consideration in validation is how well the internal rating systems can rate the risk of an obligor defaulting within a given span of time from the time of rating. Industry focus is currently on two major dimensions, namely, the discriminatory power of internal rating models and strength of their calibration to the cycle average or long-term default rates (The Basel Committee on Banking Supervision, 2005).

One challenge faced by many model validation personnel is to find a proper benchmark for the discriminatory power of their models. According to Jordao and Stein (2003) discriminative power of a rating model can have substantial economic impact on the portfolio, and thus finding an appropriate benchmark for its models is rather important to a lending institution.

Arguably, there is an obvious advantage of using information inherent to rating models to benchmark them. Every rating model has an implicit or explicit expected discriminatory power, usually based on the cycle average default rates.

The current benchmarks known to the author do not use expected performance of rating models as the standard to compare against. A lending institution can utilise its own long-term default experience to benchmark the performance of internal rating models.

This approach also has the added advantage of looking at calibration strength of a model together with its discriminatory power, as the expected power of a rating model is intertwined with default rates calibrated to the long-run default history.

A benchmark of this kind is also easy to calculate and to explain, as the performance of models against the expectation, to senior management of an organisation on whose shoulders ultimate approval authority for validation measures often rests.

In this article, we demonstrate how we can easily develop such a benchmark through a simple modification to a popular discriminatory power measure. We also compare the outcome of a test based on this modified measure to two other benchmarking approaches through a publicly available set of data.

This simple comparison shows that

www.finsia.edu.au www.finsia.ac.nz

the proposed benchmark leads to a more stable test, the failure of which could at least be partially attributable to a change in default experience. Through this comparison, it could also be established that the outcome of the proposed method is statistically related to a measure of prediction error of the relevant rating model.

CURRENT BENCHMARKS

The benchmarks that are currently being used in the industry can be broadly classified as:

1. Discriminatory Power-Test-Based Benchmarks
A measure of statistical significance built into a test, i.e. Kolmogorov-Smirnov test;

2. External Models-Based Benchmarks
Benchmarking the portfolio's discriminatory power to the discriminatory power of an external model;

3. Fixed Value-Based Benchmarks
Using a Gini Coefficient, for example, of 50% as a 'reasonable' benchmark for the portfolio.

Each approach has some merit under certain circumstances. The first category does not require a special benchmark other than a preset significance level, usually associated with a two-tailed test. If the test fails to exceed the critical value at the set significance level, the rating model does not meet the desired discriminatory power. Such tests are generally believed to be stringent, sensitive to the number of rating grades, population sizes and the naturally occurring population shifts. This may be one reason for the acknowledged popularity (Rauhmeier and Scheule, 2005) of a class of measures based on 'Area Under the Curve' (AUC) covering discriminatory power statistics known as Gini Coefficient, PowerStat or Accuracy Ratio.

One can also use a simple model to benchmark a regular rating model. But this involves rating the portfolio with two models. The third approach is the simplest although the benchmark could inadvertently be set very low or very high depending on available information on the rating model or similar models.

As an example, a discriminatory power of the sample used for model development can be set as a benchmark for a poorly built model with deteriorating power. Thus, there is a need for new ways to look at the issue of benchmarking discriminatory power.

In the next section of this article, we first develop a benchmarking measure based on Accuracy Ratio and then, using basic statistical logic, construct a benchmark test. In the third section of the article, the methodology developed in the second section will be applied to a set of data based on a hypothetical model.

With same set of data, we then compare two of the above three approaches with the proposed benchmarking method and demonstrate the usefulness of the proposed method. In the concluding section, we draw attention to the intuitively appealing nature of the proposed method and the need for further investigation with real data.

CONCEPTUAL FRAMEWORK

Let us assume that a bank has a simplified statistical model given below as (1) for its credit rating model. This model predicts the realisations of Bernoulli variable y_i, which can take either of two values, 1 or 0, with associated probabilities p_i and (1- p_i). The values of probabilities p_i and (1- p_i) depend on whether the obligor i in the cohort, with a set of explanatory variables, x_{ij} and associated model parameters, b_j did or did not default during the observation period following the rating. E represents mathematical expectation.

$$p_i = E(y_i) = \sum_j^m b_j x_{ij} + \varepsilon_i \qquad j = 1,.\cdots,m \qquad (1)$$

Here, we change the notation, p_i to z_i to absorb the effect of calibration. Now, for the expectation of z_i, using b_j with a hat for the estimated value of b_j, we can write;

$$E(z_i) = \hat{z}_i = \sum_j^m \hat{b}_j x_{ij} \qquad j = 1,.\cdots,m \qquad (2)$$

This provides the following error sum of squares (SSE) for the data sample from the cohort used for the development of the model at time T_1.

$$SSE = f(z_i - \hat{z}_i)^m \qquad (3)$$

This simplified functional form denoted by f represents a goodness-of-fit measure for the model where m is usually equal to 2. Now we assume that the above model is applied to the portfolio at time T_2 and the realisations, z_i at time T_2 can still be treated as essentially similar to the realisations that we used to build the model. In other words, the developed model is time invariant in the short term.

Following normal practice, we now can group the estimated z_i values into R rating grades based on the increasing order of quality. This makes the obligor defaults binomially distributed at rating grades level. In each rating grade r, we have n_r obligors with d'_r predicted defaults and d_r observed defaults such that:

$$d_r = \sum^{n_r} z_i \quad \Rightarrow \quad d'_r = \sum^{n_r} \hat{z}_i \qquad (4)$$

Following Engelmann, Hayden and Tasche (2003), we can define accuracy ratio, G, for this scenario in the popular manner. We use trapezoidal rule-based computation to estimate area under the curve.

$$G = \frac{((\sum_{r=1}^{R} \tilde{n}_r (\sum_{q=1}^{r-1} d_q + 0.5 d_r)) / \sum_{r=1}^{R} d_r \; - \; 0.5) \sum_{r=1}^{R} n_r}{0.5 (\sum_{r=1}^{R} n_r \; - \; \sum_{r=1}^{R} d_r)} \qquad (5)$$

where n_r with a tilde represents the proportion of obligors in rating grade, r.

www.finsia.ac.nz www.finsia.edu.au

Now, we propose to change the above definition to include a comparison between observed and predicted defaults as stated in equation (6). The area between accuracy profiles for observed and predicted defaults is illustrated in Figure 1 as the area $(A - A')$.

$$G^* = \frac{((\sum_{r=1}^{R} \tilde{n}_r (\sum_{q=1}^{r-1} d_q + 0.5 d_r))/\sum_{r=1}^{R} d_r \; - \; (\sum_{r=1}^{R} \tilde{n}_r (\sum_{q=1}^{r-1} d'_q + 0.5 d'_r))/\sum_{r=1}^{R} d'_r)}{(0.5(\sum_{r=1}^{R} n_r \; - \; \sum_{r=1}^{R} d_r)/\sum_{r=1}^{R} n_r)} \quad (6)$$

Assuming that the observed and predicted defaults are only very small proportions of total number of obligors, as usual for many good portfolios, in order to overcome the impact of normalising the areas in equation 6, this can be approximately expressed in the functional form given below.

$$G^* \approx f(d_r - d'_r)^m + f(d_q - d'_q)^m$$

Now, rearranging all the terms leads to the approximate relationship given in equation 7.

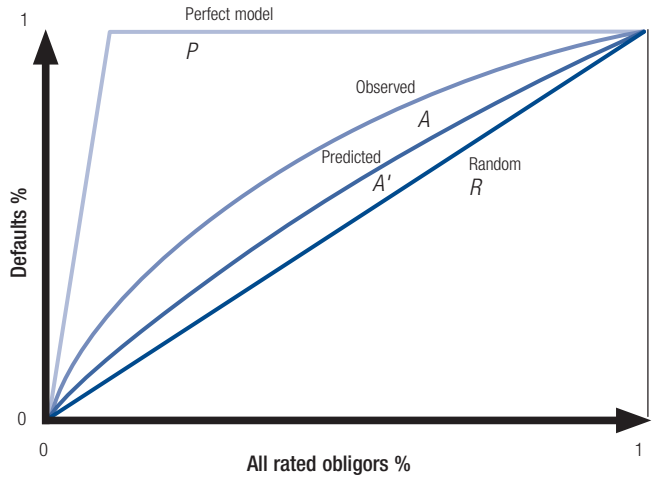
$$G^* \approx f(d_r - d'_r)^m \Rightarrow f(z_i - \hat{z}_i)^m \quad (7)$$

where m equals 1. This relationship would also be tested using the application data in the next section.

It can be easily shown that the above modified accuracy ratio is equivalent to a comparison between the accuracy ratios for observed defaults and model predicted defaults with the same perfect model.

The next step is to calculate a standard error for a modified accuracy ratio. For this, we should first consider the relationship between ROC curve and Accuracy Profile.

FIGURE 1 **PERFECT, OBSERVED, PREDICTED AND RANDOM OUTCOMES — CUMULATIVE PERCENTAGES**



Engelmann, Hayden and Tasche (2003) show that for the observed model, area under the ROC curve and the respective accuracy profile can be linked by the following equation.

$$G = 2((\sum_{r=1}^{R}(n_r - d_r)(\sum_{q=1}^{r-1} d_q + 0.5 d_r))/\sum_{r=1}^{R}(n_r - d_r)\sum_{r=1}^{R} d_r) - 0.5) \quad (8)$$

The above relationship can also be written as below.

$$G = 2\,(L - 0.5) \quad (9)$$

Let us state the predicted outcome in the following simplified form.

$$G' = \frac{2\,g'(L' - 0.5)}{g} \quad (10)$$

where

$$g = \sum_{r=1}^{R}(n_r - d_r) \;;\; g' = \sum_{r=1}^{R}(n_r - d'_r)$$

Using equation (9) and obtaining its variance lead to the equation (11).

$$V(G) = 4\,V(L) \quad (11)$$

where $V(.)$ refers to the variance of a variable.

Now, assuming error-free measurement of g, we can state the following:

$$V(G') = \frac{4(V(g'L') + (1/4)V(g') - 2Cov(g'L'))}{g^2} \quad (12)$$

where Cov(.) refers to the covariance between two variables. Note that here we made a simplifying assumption about the covariance term.

www.finsia.edu.au www.finsia.ac.nz

If we make simplifying assumptions about the form of g', L' and their correlation, by following Popoulis (1984) to derive the variance of a product of two random variables, we can write resulting variance formula as given in equation 13.

$$V(g'L') = (V(g')L'^2)+V(L')g'^2+2g'L'\sqrt{V(L')}\sqrt{V(g')}) \quad (13)$$

This now leads to the following result for the variance of G', V(G').

$$V(G') = \frac{4(V(g')L'^2+V(L')g'^2+2g'L'\sqrt{V(L')}\sqrt{V(g')}+(1/4)V(g')-2\sqrt{V(L')}\sqrt{V(g')})}{g^2} \quad (14)$$

where $V(g')$ is assumed equal to the variance of a comparable binomial variable. This formula results in a slightly larger value than what the assumption of constant g and g' would produce. However, if computational burden is not desired, one can use the simplifying assumption of constant g and g'.

Using standard statistical procedures, we can derive the following variance equation for the variance between two areas under the curves. Note that we use a variance of accuracy ratio as this allows us to incorporate the uncertainty of the predictive model into the variance equation.

$$SE(G\text{-}G') = \sqrt{V(G) + V(G') - 2Cov(G,G')} \quad (15)$$

If we can assume asymptotic normality, this standard error can be used to calculate a test statistic as described by Engelmann, Hayden and Tasche (2003).

APPLICATION TO A HYPOTHETICAL RATING SYSTEM

Now we can illustrate the methodological framework presented above with a published data set. Our data set is reported by Vazza, Aurora & Schneck (2005) in Standard & Poor's Annual 2005 Global Corporate Default study. In this application, we assume that Standard & Poor's rating agency uses a latent rating model, which has been calibrated to their long-term default rate for the period from 1981 to 2005. We also assume that this rating model outputs their annual ratings. Let us now suppose that we want to back-test their rating model after the latest calibration in 2005.

If our theoretical framework holds true, we should be able to establish a data-based relationship for the difference between observed and predicted default outcomes and the corresponding difference between their respective accuracy profiles. Hence, the test statistic Z defined below can be considered to provide the information about predicted and observed defaults in the following manner.

$$\sum_r (d_r - d'_r) \Rightarrow (A - A') \Rightarrow Z$$

To compute $V(L)$ and $V(L')$, we follow Hanley and McNeil (1982) as they state a formula which only uses area under the curve, number of non-defaults and the number of defaults. We adjusted their formula to reflect rating quality increase along the horizontal axis of accuracy profile. We also adapt their test statistic Z (Hanley and McNeil, 1983), defined as below.

$$Z = \frac{(A - A')}{\sqrt{V(G) + V(G') - 2Cov(G,G')}}$$

In the calculation of variance of (G – G'), we cannot obtain a proper covariance term as we do not have sufficient information to derive this term.

However, we can compute a covariance term using an approximate correlation between G and G' by using the relationship between distributions of observed and predicted defaults. The correlations thus calculated, together with some basic statistics about the data, are given below in Table 1.

As can be seen from the same table, the number of defaults required for the assumption of asymptotic normality is not always available in annual cohorts. Given the nature of the available data set, we do not dwell on this fact for this exercise.

For the proposed methodology, the following decision rules can be used with a desired significance level. If $(A - A') > 0$ & Z is significant, the model performs better than the expectation. If $(A - A') < 0$ & Z is significant, the rating model performs worse than the expectation. If Z is non-significant, the model meets the expectation.

Figure 2 (overleaf) shows the results of the above application. The Spearman's rank correlation between test statistic Z and the prediction error, defined as the average difference

www.finsia.ac.nz www.finsia.edu.au

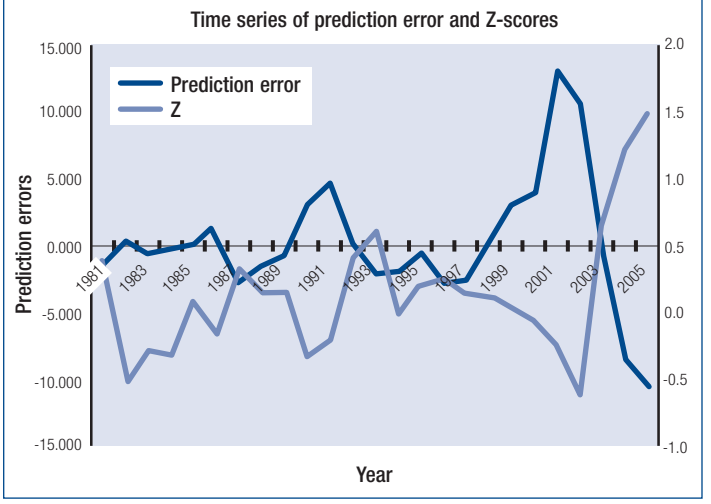
TABLE 1 **SOME RESULTING STATISTICS FOR S&P'S DATA FROM 1981 TO 2005**

Year	Goods (*g*)	Bads (*b*)	1-*A*	1-*A'*	Area (*A*–*A'*)	Correlation (*G*,*G'*)	Std. error (*G*–*G'*)
1981	1390	2	0.040	0.085	0.045	0.730	0.115
1982	1423	17	0.150	0.087	-0.064	0.617	0.116
1983	1453	11	0.103	0.085	-0.018	0.929	0.069
1984	1538	14	0.104	0.085	-0.019	0.955	0.065
1985	1618	18	0.083	0.092	0.009	0.984	0.079
1986	1838	32	0.115	0.103	-0.012	0.991	0.076
1987	2002	19	0.088	0.098	0.011	0.990	0.035
1988	2079	30	0.100	0.108	0.008	0.998	0.046
1989	2117	36	0.094	0.105	0.011	0.927	0.056
1990	2117	36	0.094	0.105	0.011	0.970	0.063
1991	2018	68	0.092	0.081	-0.012	0.915	0.065
1992	2156	30	0.046	0.071	0.025	0.977	0.060
1993	2372	13	0.049	0.072	0.024	0.996	0.039
1994	2617	16	0.095	0.095	0.000	0.982	0.040
1995	2946	30	0.087	0.099	0.012	0.998	0.053
1996	3217	16	0.093	0.102	0.008	0.974	0.035
1997	3545	22	0.099	0.106	0.007	0.980	0.039
1998	3999	53	0.116	0.122	0.006	0.982	0.050
1999	4368	96	0.122	0.123	0.002	0.997	0.046
2000	4508	109	0.125	0.123	-0.001	0.997	0.045
2001	4556	179	0.129	0.117	-0.012	0.988	0.048
2002	4603	172	0.126	0.101	-0.025	0.992	0.040
2003	4742	93	0.080	0.106	0.026	0.960	0.039
2004	5035	37	0.077	0.114	0.036	0.926	0.030
2005	5385	30	0.106	0.123	0.017	0.995	0.011

between predicted and observed defaults, is about -0.75 which is significant at 1% level for 25 observations. Similar relationship with a rank correlation of 0.57 exists between the difference of areas $(A - A')$ and prediction error. Even though the simple average difference as defined above is not a traditional measure of prediction error, it is indicative of the direction of overall deviations. This shows a reasonable relationship between the benchmark based on a modified accuracy ratio and a measure of departures between the observed and expected defaults.

As can be expected, S&P's ratings being mainly through the cycle ratings do not seem to produce very large annual Z values that would result in significant departures (ie. Z > 1.64) from the long-term calibration at 5% significance level. However, it may not be wise to disregard high fluctuations towards the end of series which are significant at 10% level. For the larger part of the time series, Z values show that performance of the hypothetical model is at an acceptable strength with slight fluctuations about the expectation.

FIGURE 2 **Z AND SIGNED DEVIATIONS BASED PREDICTION ERRORS**



www.finsia.edu.au www.finsia.ac.nz

TABLE 2 **COMPARISON OF METHODS**

Year	Z	Z test	Max. difference	K-S test	Accuracy ratio	Fixed benchmark
1981	0.39	Pass	0.93	Pass	0.92	Pass
1982	-0.55	Pass	0.65	Pass	0.71	Fail
1983	-0.27	Pass	0.68	Pass	0.80	Fail
1984	-0.30	Pass	0.62	Pass	0.80	Fail
1985	0.11	Pass	0.75	Pass	0.84	Pass
1986	-0.16	Pass	0.69	Pass	0.78	Fail
1987	0.30	Pass	0.75	Pass	0.83	Fail
1988	0.17	Pass	0.69	Pass	0.81	Fail
1989	0.19	Pass	0.68	Pass	0.83	Fail
1990	-0.33	Pass	0.68	Pass	0.83	Fail
1991	-0.18	Pass	0.75	Pass	0.84	Pass
1992	0.41	Pass	0.88	Pass	0.92	Pass
1993	0.61	Pass	0.81	Pass	0.91	Pass
1994	0.00	Pass	0.75	Pass	0.81	Fail
1995	0.23	Pass	0.70	Pass	0.83	Fail
1996	0.24	Pass	0.73	Pass	0.82	Fail
1997	0.17	Pass	0.73	Pass	0.81	Fail
1998	0.13	Pass	0.65	Pass	0.78	Fail
1999	0.04	Pass	0.69	Pass	0.77	Fail
2000	-0.03	Pass	0.67	Pass	0.77	Fail
2001	-0.25	Pass	0.64	Pass	0.77	Fail
2002	-0.62	Pass	0.62	Pass	0.78	Fail
2003	0.67	Pass	0.74	Pass	0.86	Pass
2004	1.20	Pass	0.70	Pass	0.85	Pass
2005	1.50	Fail	0.69	Pass	0.79	Fail

COMPARISON OF BENCHMARKING METHODS

In the introduction, we discussed three approaches used to look at benchmarking issues. Now, we can apply two of these methods to the above data set together with the method proposed in this paper. The lack of more detailed data did not allow the use of a different model as a benchmark model. The Kolmogorov-Smirnov test statistic (denoted in Table 2 as Max. Difference), used to derive the chi-squared statistic and representing the first method, was calculated for each annual cohort of defaults and non-defaults. (Note that the years with a small number of defaults shown in this methodological example should be carefully interpreted.) The critical chi-squared value for a one-tailed large sample K-S test at 5% significance level is 5.99 with two degrees of freedom. The fixed benchmark was taken from the same study done by S&P's rating agency from which the data set was derived. The benchmark was set to their one year global Gini Coefficient of 84%. The modified accuracy ratio method was tested at 10% significance level. Table 2 shows the status of testing against each benchmark as pass or fail.

The analysis presented here shows that the fixed benchmark procedure produces a very conservative benchmark, leading to frequent failures. Irrespective of the 5% or 10% significance level used (ie. the critical value for a one-tailed large sample K-S test at 10% significance level is 4.60), the K-S test based benchmark was always met. However, the K-S test, based on the observed populations, ignores information about the expected behaviour. The proposed test using a similar significance approach compares the observed to the expected.

Even though the benchmarking is not about significance testing, we can find some guidance for decision making in statistical significance. The test statistic Z triggers a failure when the variance of the difference between G and G′ is relatively small and the area between A and A' is relatively large. As S&P's rating outcomes are generally calibrated to the cycle averages and closely monitored, we do not expect them to frequently fail. Given the high default environment around 2001–2002 it may not be unreasonable to expect a

www.finsia.ac.nz www.finsia.edu.au

subsequent calibration issue as indicated by the proposed method.

CONCLUSION

The nature of rating models can vary with the rating philosophy of a financial institution. According to new Basel standards, even if the rating philosophy is mainly point-in-time orientated, the model output should be calibrated to the cycle average default experience.

Thus, it is intuitive to argue that every lending institution would rather like to know whether its models achieve the power represented by a monotonically increasing sequence of cycle average default rates assigned to its rating grades. For such an institution to become price-competitive, the model power should closely match the discriminatory power expectation built into its rating scale through such a sequence of cycle average default rates.

The methodology discussed in previous sections endeavours to benchmark performance of a rating model using its expected and actual default observations. This approach, resulting in a modified accuracy ratio, also incorporates calibration information, which the widely used accuracy ratio is lacking.

While the author acknowledges that the methodology should be further tested with a variety of data sets, the proposed approach results in an intuitively appealing benchmark based on internal expectations of a lending institution. This benchmark represents a measure of error between the expected and observed performance.

References

Basel Committee on Banking Supervision, (2005), *Studies on the Validation of Internal Rating Systems*, May.

Basel Committee on Banking Supervision, (2006), *International Convergence of Capital Measurement and Capital Standards,* June.

Engelmann, B., Hayden E. and Tasche, D., (2003), *Testing Rating Accuracy* www.risk.net January 2003 RISK.

Hanley, J. A. and McNeil, B.J., (1982), "The Meaning and Use of the Area Under a Receiver Operating Characteristics (ROC)", *Curve Diagnostic Radiology* 143, pp. 29–36.

Hanley, J. A. and McNeil, B.J., (1983), "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases", *Diagnostic Radiology* 148, pp. 839–843.

Jordao, F. and Stein, R.M., (2003), *What is a More Powerful Model Worth?,* Technical report #030124, Moody's KMV Company.

Papoulis, A., (1984), *Probability, Random Variables and Stochastic Processes,* McGraw-Hill, New York.

Rauhmeier, R. and Scheule, H., (2005), *Rating Properties and their Implications for Basel II Capital,* www.risk.net March 2005 RISK.

Vazza, D., Aurora D., and Schneck, R., (2005), "Annual 2005 Global Corporate Default Study and Rating Transistions", *Standard & Poor's*, January 2006.

**The author acknowledges Dr Alberto Pambira for his useful comments on an earlier version of this document. However, the author is solely responsible for any errors, and the views expressed in this article. Please contact the author at darshiarachige@optusnet.com.au with any inquiries on the content.* J

Raising international standards

In line with Finsia's vision of 'Raising standards in the finance industry' our international area is responding to significant growth in hosting international students' education here in Australia.

We are meeting growing demands from financial institutes around the world in terms of delivering education on site at the Finsia premises. Finsia International again recently hosted two international students from the State Securities Commission, Vietnam: Pham Nguyen Hoang, Deputy Manager of Training Division, Securities Research and Training Centre (SRTC) and Nguyen Thuy Hoan, Acting Head of Division, Information Division, Securities Research and Training Centre.

Finsia hosts luncheon

Mr Martin Wheatley, CEO Securities & Futures Commission Hong Kong, presented his insights on "Investment Management Opportunities in Hong Kong and China" during a Finsia Boardroom luncheon on 6 March 2007.

We extend our sincere thanks to Giles Gunesekera SF Fin and his colleagues at Principal Global Investors for kindly sponsoring this luncheon which was held in their Sydney office.

KARP study tour

Finsia has entered into an agreement with the Korea Association of Risk Professionals (KARP) to provide a three and a half week financial markets study tour program in Australia for 30 graduate students specialising in finance from Korean universities. Finsia will be responsible for the design and implementation of the study tour program, training facilities and equipment.

The objectives of the intensive study tour will be to:

- Identify the key influences on the development of the Australian finance and banking markets
- Understand the key market regulators of the Australian financial markets and their specific role in maintaining the integrity of the markets
- Understand the financial markets exchanges and their role in market facilitation, supervision and reporting
- Explore new and popular products of the Australian banking and finance markets
- Analyse the changing role of corporate finance and the funding options that are available via debt and equity
- Explore the key issues in mergers and acquisitions and the increased activity in the Australian market place
- Understand the new strategies implemented by financial risk managers in the current environment.