Güttler, André

Working Paper

Using a Bootstrap Approach to Rate the Raters

Working Paper Series: Finance & Accounting, No. 132

Provided in Cooperation with:
Faculty of Economics and Business Administration, Goethe University Frankfurt

Suggested Citation: Güttler, André (2004) : Using a Bootstrap Approach to Rate the Raters, Working Paper Series: Finance & Accounting, No. 132, Johann Wolfgang Goethe-Universität Frankfurt am Main, Fachbereich Wirtschaftswissenschaften, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/27774

JOHANN WOLFGANG GOETHE-UNIVERSITÄT
FRANKFURT AM MAIN

FACHBEREICH WIRTSCHAFTSWISSENSCHAFTEN

André Güttler

Using a Bootstrap Approach to Rate the Raters

No. 132
October 2004

WORKING PAPER SERIES: FINANCE & ACCOUNTING

André Güttler[†]

USING A BOOTSTRAP APPROACH TO RATE THE RATERS[*]

No. 132
October 2004

ISSN 1434-3401

† André Güttler, University of Frankfurt, Mertonstrasse 17, 60054 Frankfurt, Germany, guettler@finance.uni-frankfurt.de, +49 69 798 23143.

* I would like to thank Patrick Behr, Axel Eisenkopf, Hergen Frerichs, Gunther Löffler, Lars Norden, Peter Raupach, Mark Wahrenburg, participants in the EFMA conference 2004 in Basel, participants in the Augustin Cournot Doctoral Days 2004 in Strasbourg and a referee for their helpful comments.

Abstract

This paper compares the accuracy of credit ratings of Moody's and Standard&Poor`s. Based on 11,428 issuer ratings and 350 defaults in several datasets from 1999 to 2003 a slight advantage for the rating system of Moody's is detected. Compared to former research the robustness of the results is increased by using nonparametric bootstrap approaches. Furthermore, robustness checks are made to control for the impact of Watchlist entries, staleness of ratings and the effect of unsolicited ratings on the results.

Keywords: Credit rating agencies, Validation, Bootstrap

JEL Classification: G15, G23

1 Introduction

Investors and regulators use credit ratings by private rating agencies to economize on the resources they devote to credit risk evaluation. Starting in a small niche at the beginning of the last century, the rating industry has grown to a billion dollar business which is dominated by three international rating agencies, Moody's, Standard & Poor's (S&P) and Fitch Ratings. Only these three agencies, and the minor DBRS, are so called NRSROs, i.e. statistical ratings organizations that have been officially recognized by the SEC, and they therefore play an important role in the regulation of investment activities by investment funds and banks in the US. Since an agency needs to be able to provide international coverage in order to be accepted as an authority by international investors, yet cannot gain access to the all-important US market without the NRSRO status, the regulation of rating agencies in the US increases the market barriers substantially. Market entry is also made more difficult for rating agencies by the necessity of reputation-building. Investors only believe in ratings if their quality has been proven over several years. These two factors largely account for the oligopolistic market structure in the rating business, which yields huge rating fees and wide margins for the three big players. Nowadays, the market value of the three big rating agencies – Moody's, for example, was worth about $10B in September 2004 – often exceeds the value of the companies they rate. In contrast to the widely regulated banking sector, the more and more powerful rating agencies are not regulated or even analyzed on a regular basis. No benchmarking of the quality of default predictions by independent parties takes place. The market participants rely on the agency's annual default reports as the only source of performance verification.

This paper benchmarks the quality of default predictions by Moody's and S&P, a topic which is motivated by a number of factors. First, the benchmarking of external credit rating agencies has huge economic relevance, given that the quality of credit ratings is an important factor in determining the level of regulatory capital for banks using the standardized approach of Basel II (BASEL COMMITTEE ON BANKING SUPERVISION 2004), and also in investment regulation (as reflected, for example, in the fact that mutual funds in the US are not allowed to hold non-investment grade rated assets). Indirectly, risk adequate credit ratings are also important for the optimal

allocation of capital to bond issuers since the (re)financing costs of bonds depend very much on the credit rating of the bond. Risk inadequate credit ratings would, ceteris paribus, result in exaggeratedly high (low) required yields in the case of an exaggeratedly bad (good) rating and would therefore undermine the optimal allocation of capital. Second, the differences in power between rating systems have vast economic effects. Based on simulations, STEIN and JORDÃO (2003) demonstrate that for a medium sized bank with assets of $50B a 5% more powerful rating system in respect to a certain validation measure results in a profitability increase of $3.4MM to $6.2MM per year.[1] The results of independent benchmarking should therefore be of interest not only to investors but also to banks and regulatory authorities.

The difficulty of comparing different rating systems, regardless of whether they stem from banks or external rating agencies, lies in the scarcity of data. For validation purposes, rating and default data for at least two rating instances are required. Since banks do not provide this kind of data to the public for business policy and legal reasons, the data compiled by rating agencies is the best alternative. But even when rating and default data for two or more rating bodies are available, the mapping of the issuers or debtors is far from trivial. In many countries, e.g. Germany, there are no identification numbers for private firms.[2] Even for the publicly traded companies with an external rating, no single data source with rating and default data is available.

Considering the difficulty of obtaining the required data, it is not surprising that to date there is only one paper, by KRÄMER and GÜTTLER (2003), which assesses the different quality of default predictions by two rating bodies. KRÄMER and GÜTTLER (2003) consider Moody's and S&P, and conclude that Moody's outperforms S&P in most of the analyzed validation criteria for ratings of 1,927 issuers as at the end of 1998 and a four-year realization phase from 1999 to 2002 with 209 defaults. The rest of the literature is based either on rating data for several rating bodies but no default data (e.g. CANTOR and PACKER 1997), or on rating and default data for only one rating body (e.g. ENGELMANN et al. 2003), or on rating and default data for one rating body plus equity-based measures of default risk (LÖFFLER forthcoming).

The contribution of this paper to the existing literature, and especially to the paper of KRÄMER and GÜTTLER (2003), is threefold: the dataset of this paper is broader than

the former study because it includes 11,428 issuer ratings and 350 defaults in four datasets from 1999 to 2003 (four is the maximum number of datasets for a period of five years, given the need to form pairs of rating observations for one year and default observations of the subsequent year). By using one-year realization periods instead of the four-year realization periods used by KRÄMER and GÜTTLER (2003) to determine the accuracy of ratings is in accordance with market practice, i.e. the default reports of Moody's and S&P, and it conforms to the requirements of the internal ratings-based approaches of Basel II (BASEL COMMITTEE ON BANKING SUPERVISION 2004). Besides, the calculation of bootstrapped confidence intervals obviates the need to use parametric or otherwise potentially problematic test statistics. Bootstrapping also makes it possible to compute confidence intervals for all validation measures which are available. Finally, robustness checks are made to control for the impact of Watchlist entries, staleness of ratings and the effect of unsolicited ratings on the results.

The remainder of the paper is organized as follows: section 2 describes the dataset and the mapping of the two rating systems used by Moody's and S&P, respectively. Section 3 provides the methodological background of the validation measures. Section 4 gives a review of the bootstrap approach. The following section reports the empirical results, and section 6 concludes.

2 Data

To analyze two different rating systems it is essential to have a dataset of multiple rated issuers over an identical time-span. First, this is necessary to avoid misleading results due to sample selection. For example, comparing a rating agency that rates all issuers in a certain region with another agency that only rates a certain sector of issuers is not adequate. Roughly speaking, if this sector is not representative of the whole population of issuers, it is like comparing apples and oranges. Second, due to the calculation of validation measures, a rating agency with a homogenous set of good rated issuers may have a worse validation measure than a rating agency with a high risk portfolio of rated issuers (HAMERLE et al. 2003).

Since no database has actually been established with historical data for two or more rating agencies, manual mapping of data sources for rating and default data is required. The rating data is obtained from Bloomberg while default data is received directly from Moody's and S&P. Hence, the analysis is based on four datasets of borrowers, mostly industrial firms and financial institutions, which have a senior, unsecured, long term credit rating by both Moody's and S&P as of the end of the years 1999, 2000, 2001 and 2002 (see Figure 1). In the following, the year preceding each of these dates is referred to as the estimation period. Between 2000 and 2003 all defaults by these firms are recorded to construct realization periods each with a length of one year.[3] Default data are obtained from the annual default reports of Moody's and S&P. In this study a default is defined as such if Moody's and S&P reported a default in the year following the year of the respective estimation period. The default date is defined as the first default date if there are different default dates in the two default reports. In total, 350 multiple default reports were published in the years 2000–2003 (see Table 1). In all years, S&P reported a higher number of defaults than Moody's. This reflects the fact that rated issuers with a long-term or issuer rating by S&P numbered 6,848, as against the 4,813 rated issuers with an issuer rating by Moody's.[4]

Table 1: Observations of defaulted issuers

	2000	2001	2002	2003	Total
Published by Moody's	111	160	122	68	461
Published by S&P	110	189	163	93	555
Published by both	88	122	92	48	350

The table shows the number of defaults published in the annual default reports of Moody's and S&P with a long-term or issuer rating in Bloomberg. A multiple default is defined as such if both agencies published a default. If different default dates are reported, only the earliest of these default dates is taken into account.

The number of defaults should not be affected by the default definitions of the rating agencies, since they are quite similar. Moody's and S&P define a default according to the following three categories (MOODY'S 2004) and S&P (STANDARD&POORS 2004):

1) first occurrence of missed or delayed payments of interest and/or principal

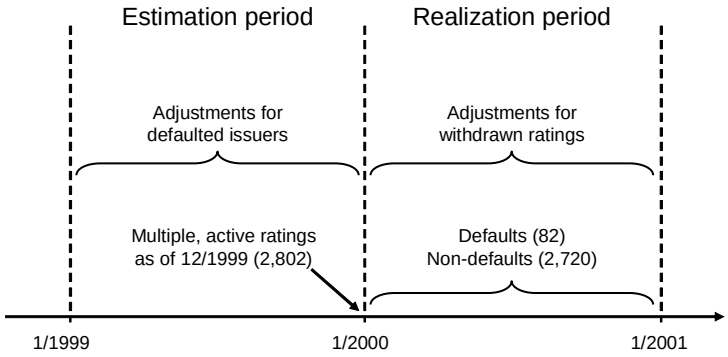
2) bankruptcy as defined by Chapter 11 of the US bankruptcy code

3) distressed exchange, which diminishes the value of financial obligations

But besides these three categories, there are some differences in the definitions of default between Moody's and S&P. For example, S&P, unlike Moody's, does not claim a default when an interest payment missed on the due date is made within the grace period, whereas Moody's does not disclose defaults by issuers with an unsolicited rating.

Sample adjustments were made for issuers that defaulted in the year before and in the estimation period and which were given a withdrawn rating during the estimation period by at least one rating agency (see Figure 1). This was done to include only active rated issuers with multiple ratings. Four pairs of data were built, each of them with rating data at the end of one estimation period and one year with default data, i.e., the realization period. After all adjustments had been made, among the 2,802 issuers covered in the 1999 estimation period, 82 defaults were detected in the corresponding realization period 2000, which translates into a one-year default rate of 2.93% (for the pairs 2000/2001, 2001/2002 and 2002/2003 the corresponding figures were: 2,874, 110, 3.83%; 2,873, 87, 3.03%; 2,879, 45, 1.56%). These four years include periods with record numbers of defaults, e.g., 2002 with a default rate of almost 4%, and also periods with healthy economic conditions, such as 2003 with a default rate of 1.56%. Additionally, all computations were also made on the aggregate number of 11,428 issuer ratings and 350 defaults over the four years, yielding an average default frequency of 3.1%.

Figure 1: Construction of the first pair of estimation period (1999) and subsequent realization period (2000)



The two rating agencies under observation apply significantly different rating approaches. Moody's long-term ratings measure the total expected credit loss over the life of the security, i.e. they are an assessment of both the likelihood that the issuer will default on a security, and the amount of loss after a default occurs (ESTRELLA 2000). This is often called the expected loss approach. In contrast, S&P only assesses the probability of default.

Table 2: Regional distribution of issuers

Country	12/99	12/00	12/01	12/02
US	71.00%	68.12%	66.19%	66.40%
Great Britain	5.00%	5.74%	5.85%	5.00%
Japan	4.61%	5.15%	5.22%	5.28%
Canada	2.89%	2.99%	3.13%	3.02%
Australia	1.61%	1.57%	1.95%	2.08%
France	1.71%	1.57%	1.57%	1.67%
Netherlands	1.29%	1.64%	1.81%	2.05%
Germany	1.61%	1.74%	1.74%	1.60%
Argentina	1.00%	1.08%	1.01%	1.01%
Mexico	0.79%	0.77%	0.49%	0.31%
Other	8.50%	9.64%	11.04%	11.57%
Number of Issuers	2,802	2,874	2,873	2,879

The table shows the regional distribution of issuers of the four estimation periods using rating data from Bloomberg. All quantities refer to the end of the respective year. Only multiple rated issuers, i.e. with a valid long-term unsecured credit rating by both Moody's and S&P are included.

Conventionally in the literature (e.g. CANTOR et al. 1997), the different rating scales of the two rating agencies are mapped to a single numeric scale, with better ratings corresponding to lower numbers: Aaa = AAA = 1, Aa1 = AA+ = 2, … B3 = B- = 16 and C = 17 (see Appendix A). All issuers with a rating worse than B3 or B- are lumped together in the mapped rating class 17. This is done because only a few issuers are rated in these low classes. The default frequencies of the own sample are calculated by dividing the number of defaults by the number of ratings for each rating grade for the respective realization periods. Hence, average default frequencies are calculated over the four periods 2000, 2001, 2002 and 2003. In comparison with the long-term historical averages of Moody's and S&P, which are calculated over more than 20 years, they are clearly higher, especially for the non-investment grade rating classes. This is mainly due to the inclusion of the years 2001 and 2002 with record numbers of defaults.

The sample is dominated by US issuers (see Table 2), which is common for studies with credit ratings. This is due to the fact that Moody's and S&P are both US-domiciled, and did not start to expand into other markets until the 1990s. Markets outside the US are growing, as one can easily see even in these four years. In the selected samples of multiple rated issuers the ratio of US-rated issuers decreases from 71% to 66.4% at the end of 2002.

3 Validation measures

3.1 Receiver Operating Characteristic curve

The Receiver Operating Characteristic (ROC) curve is a commonly used validation technique with its origin in signal detection theory, psychology and medicine. For validation purposes the size of the area below an ROC curve is of special interest (SOBEHART and KEENAN 2001 and ENGELMANN et al. 2003).

The construction of an ROC curve can best be explained through two possible distributions of continuous scores for non-defaulting and defaulting issuers. For well-designed rating systems the distribution of the non-defaulting issuers should have better scores on average than the distribution of defaulting issuers. To decide which issuers will survive during the next period and which issuers will default, the decision-maker might introduce a threshold, C, and classify each issuer with a score lower than C as a defaulter and each issuer with a higher score as a non-defaulter. If the score is below the cut-off value and the issuer subsequently defaults, the decision was correct. Otherwise the decision-maker incorrectly classified a non-defaulter as a defaulter. If the score is above the cut-off value and the issuer does not default, the classification was accurate. Otherwise, a defaulter was incorrectly assigned to the non-defaulters group. Using the notation of SOBEHART and KEENAN (2001), the hit rate HR(C) is defined as:

$$HR(C) = \frac{H(C)}{N_D} \tag{1}$$

where H(C) is the number of defaulters predicted correctly with the cut-off value C, and N_D is the total number of defaulters in the sample. The false alarm rate FAR(C) is defined as:

$$\mathrm{FAR}(C) = \frac{F(C)}{N_{ND}} \tag{2}$$

where F(C) is the number of false alarms, i.e. the quantity of non-defaulters that were classified mistakenly as defaulters by using the cut-off value. The total number of non-defaulters in the sample is denoted by N_{ND}. Hence, the ROC curve is constructed as follows. For all cut-off values C that are contained in the range of the scores the quantities HR(C) and FAR(C) are calculated. The ROC curve is a plot of HR(C) versus FAR(C). The larger the area under the ROC curve, which is defined as AUC:

$$\mathrm{AUC} = \int_0^1 \mathrm{HR}(\mathrm{FAR})\,\mathrm{d}(\mathrm{FAR}) \tag{3}$$

the better the rating model's performance is. The area under the ROC curve is 0.5 for a random model without discriminative power, 1 for an ideal model and between 0.5 and 1 for any rating model in practice. By construction, the ROC curve and the AUC only measure the refinement of default predictions. The concept of refinement pertains to how spread out the default predictions of a rating system are (DEGROOT and FIENBERG 1983). In other words, the more predictions of 0% or 100% are assigned to issuers, the more refined the rating system is.

3.2 Scoring rules

Scoring rules are a second class of validation measures (WINKLER 1996). Their role is to provide summary measures to evaluate probabilities in the light of what actually happens. Besides sharpness, the degree of calibration is also taken into account by these measures since they are based on probabilities of default. Calibration is based on the level of agreement between a rating system's default predictions and the actual observed relative frequency of default (DEGROOT and FIENBERG 1983). A rating system is said to be well calibrated if among those issuers for whom its default

prediction is a certain probability x, the long-run relative frequency of default is also x.[5]

The well-known Brier score (BRIER 1950) is often denoted in the form:

$$B = 1/n \sum_{i=1}^{n} (\theta_i - q_i)^2 \tag{4}$$

where i indicates an issuer (i = 1,…,n), θ is a binary variable (1 if a default occurs, 0 otherwise) and q denotes the probability that a default will occur. This score takes its optimum value of B = 0 when the only predicted probabilities of default are 0 and 1 and the predictions are always correct. It takes its worst value of B = 1 when the only predicted probabilities of default are 0 and 1 and always the opposite of what has been predicted occurs.

The logarithmic score is given by:

$$L = 1/n \sum_{i=1}^{n} LN(|q_i + \theta_i - 1|) \tag{5}$$

The logarithmic score is always negative, with closeness to zero signaling good performance. The spherical score is defined by:

$$S = 1/n \sum_{i=1}^{n} \frac{|q_i + \theta_i - 1|}{\sqrt{q_i^2 + (1-q_i)^2}} . \tag{6}$$

This rule is always positive, with large values indicating an improvement in performance.

4 Bootstrapping

A fundamental task of quantitative research is to make inferences about a population characteristic based on an estimator using a sample drawn from that population. Bootstrapping differs from the traditional parametric approach in that it employs a large number of repetitive computations to estimate the shape of a statistic's sample distribution. This allows drawing inferences in cases where such assumptions are untenable or no parametric statistics exist. Bootstrapping relies on the analogy between the sample and the population from which the sample was drawn. The central idea is

that it may sometimes be better to derive conclusions about the characteristics of a population strictly from the sample at hand, rather than making unrealistic assumptions about that population. Bootstrapping involves resampling the data with replacement many times in order to generate an empirical estimate of the entire sampling distribution of a statistic. Although each resample will have the same number of elements as the original sample (n = size of the sample = size of the resample), through replacement each resample could have some of the original data points represented in it more than once, and some not represented at all. Therefore, each of these resamples will likely be slightly and randomly different from the original sample.

In this study, nonparametric confidential intervals are calculated for the four validation measures described in section 3. Whereas statistics for the comparison of AUC (DELONG et al. 1988) and Brier scores (REDELMEIER et al. 1991) of different rating systems can be found in the literature, no such statistics are available for the comparison of logarithmic or spherical scores. The bootstrap approach therefore makes it possible to use all these available validation measures with comparable results for all of them. Besides, no assumptions about the distribution are necessary.

Essentially, bootstrapping follows five basic steps:

1) Construct an empirical probability distribution from the sample by placing a probability of 1/n at each issuer (x_1, x_2, …, x_n).

2) Draw a random sample of size n with replacement.

3) Calculate the statistic of interest from this resample.

4) Repeat steps 2 and 3 B times. B should be at least 1,000 to estimate confidence intervals around the statistic of interest (EFRON 1987).

5) Construct a probability distribution from the B statistics of interest by placing a probability of 1/B at each single statistic. This distribution is the bootstrapped estimate of the sampling distribution of the statistic of interest. This distribution can be used to make inferences about the population characteristic.

The percentile method, a simple, but very intuitive method, is used to construct confidence intervals for the area under the ROC curve and the three scoring rules.[6] If someone draws 1,000 resamples and calculates the statistic of interest 1,000 times, a 5% two-sided confidence interval would be based on the 25th-lowest and the 25th-highest value of these statistics. A test for comparing two quantities of a statistic of interest requires the calculation of B numbers of the statistic for Moody's and S&P, and their differences. The H_0-hypothesis proposes that the mean difference of the respective statistic does not equal zero. The H_0-hypothesis has to be rejected if the calculated confidence interval on a significance level of α does not cover zero.

5 Empirical results

Aside from the paper by KRÄMER and GÜTTLER (2003), which also employs rating and default data, earlier studies made use of rating data only. Hence, the comparison of ratings was given special importance. First, in previous studies it was important to analyze whether so called "rating shopping" is observable. Rating shopping characterizes issuers who seek an inflated rating (JEWELL and LIVINGSTON 1999). Only if the requested rating is favorable does the company publish it. If it is not favorable, the issuer pays for the rating but does not release it. Second, a matter of special interest was whether the default risk is more difficult to assess in some business sectors, e.g. the banking sector, than in others (MORGAN 2002).

Table 3 provides an overview of the distribution (the detailed distribution of ratings is given in Appendix B) and the direction of split ratings of the own dataset and three other studies. One criterion for the choice of these three studies was that all of their observation periods occur after the introduction of additional rating classes by Moody's in April 1982, following the lead taken by S&P seven years earlier. Cantor and Packer (1997) analyze rating differences of 4,399 straight bond, US dollar-denominated, public offerings by US corporations from 1983-1993. Perry (1985) observes ratings of 218 recently issued unsubordinated bonds for non-financial corporations for May, 1982. Ederington and Yawitz (1987) present results for 388 industrial bonds at the end of 1982. The results of these three studies are quite similar: Moody's and S&P agree in

41.8 to 45.3% of all cases. In 5 to 18.3% they disagree by two notches (a notch signifies the step from one rating grade to the next, e.g. from BB to BB+) or more. The results for the own, more recent sample with rating data from the period 1999 to 2002 fall nearly in line with the earlier research. 40% are identically rated by Moody's and S&P whereas they disagree by two or more notches for 19.1% of all issuers.

Table 3: Distribution and direction of split ratings by Moody's and S&P

Panel I: Distribution of split ratings

	Own sample	CANTOR et al. (1997)	PERRY (1985)	EDERINGTON et al. (1987)
Identically rated	39.98%	45.33%	41.74%	41.75%
One notch	40.93%	42.12%	44.50%	39.95%
Two notches	14.01%	10.00%	4.13%	13.66%
Three notches	3.34%	2.21%	0.46%	4.12%
> three notches	1.74%	0.34%	0.46%	0.52%

Panel II: Analysis of split ratings (Moody's rating is lower than S&P's)

	Own sample	CANTOR et al. (1997)	PERRY (1985)	EDERINGTON et al. (1987)
All split ratings	62.25%	54.44%	62.20%	56.64%
Investment grade	52.42%	59.58%	65.22%	61.94%
Non-investment grade	75.19%	30.68%	33.33%	45.07%

The split ratings for the own sample are given as averages of the four one-year periods 1999 - 2002. The analysis of split ratings in Panel II, where the rating by Moody's is lower than the respective rating by S&P, gives ratios according to S&P's rating in the case of investment grade and non-investment grade ratings.

Greater differences between this dataset and earlier studies exist in a further analysis of the split ratings. Over all split ratings, Moody's assigns lower ratings. This is a stable result in all empirical studies. In the other three studies this is attributable to the lower ratings for investment grade rated issuers. In this study Moody's also assigns slightly lower ratings for investment grade rated issuers. But the main effect stems from the non-investment grade issuers. Whereas the other three studies find that in this range of ratings Moody's assigns higher ratings, this study discovers an obvious majority of lower ratings. A possible explanation might be a change to more stringent rating standards. BLUME et al. (1998) show for S&P that there was a trend towards more

stringent rating standards for a panel of issuers over the time period 1978 to 1993. Unfortunately, their study does not also contain rating data from Moody's. The results presented in Table 3, Panel II, would seem to be plausible if the shift towards more stringent rating standards over the last decade was even more pronounced at Moody's than at S&P.

Table 4 presents the results of the four validation measures in Panels I and II, while Panel III shows the differences between Moody's and S&P as well as one-sided significance levels based on bootstrapped confidence intervals. The percentile bootstrap method was used for computing these intervals. Whereas the AUC of the ROC curve is based on the mapped numerical ratings, the three validation scores are calculated by using the default frequencies of the own sample as shown in Appendix A. Columns two to five contain the results for the four one-year periods. Column six presents the results for the aggregation of all four periods. Since positive differences in Panel III indicate an advantage for Moody's, and all four validation measures favor Moody's, this rating agency leads for the aggregated dataset with all issuer ratings and defaults. For the spherical score this difference is significant on the 10% level. This advantage for Moody's results mainly from its superior performance in the years 2001 and 2002. In 2001 the differences in the Brier and the spherical score are not equal to zero on the 10% level. For 2002 the differences are significant on the 5% level for the logarithmic score and on the 1% level for the Brier and the spherical score. For the years 1999 and 2000 there are no significant results. S&P tends to perform better in these two years.

Table 4: Differences of risk assessments

Estimation	12/1999	12/2000	12/2001	12/2002	all
Realization period	1/2000-12/2000	1/2001-12/2001	1/2002-12/2002	1/2003-12/2003	
Panel I: Validation measures for issuers rated by Moody's					
AUC	0.9039	0.8916	0.8828	0.9298	0.8963
Brier score	0.0254	0.0316	0.0258	0.0164	0.0248
Logarithmic score	-0.0950	-0.1161	-0.0993	-0.0645	-0.0937
Spherical score	0.9729	0.9661	0.9724	0.9828	0.9736
Panel II: Validation measures for issuers rated by S&P					
AUC	0.9016	0.8953	0.8764	0.9256	0.8948
Brier score	0.0248	0.0310	0.0269	0.0180	0.0252
Logarithmic score	-0.0946	-0.1134	-0.1030	-0.0685	-0.0949
Spherical score	0.9737	0.9665	0.9712	0.9807	0.9730
Panel III: Differences between Moody's and S&P					
AUC	0.0023	-0.0037	0.0064	0.0042	0.0016
Brier score	-0.0007	-0.0006	0.0011*	0.0016***	0.0004
Logarithmic score	-0.0004	-0.0027	0.0037	0.0040**	0.0012
Spherical score	-0.0007	-0.0004	0.0012*	0.0021***	0.0005*

The validation measures are calculated by using the default frequencies of the own sample (see Appendix A). Panel III presents the differences of these measures. For convenience, these differences are multiplied by -1 for the Brier score. Positive differences show an advantage for Moody's. One-sided significance levels are given as ***, **, and * representing 1%, 5%, and 10% respectively using the percentile bootstrap method of calculating nonparametric confidence intervals.

On the one hand, the result for the aggregated dataset that Moody's performs better than S&P confirms the outcome of KRÄMER and GÜTTLER (2003) who analyze a single estimation at the end of 1998 with a four-year realization period from 1999-2002. Based on four-year default frequencies the AUC is 0.833 for Moody's (0.819 for S&P), the Brier score is 0.066 (0.0686), the logarithmic score is -0.2005 (-0.2056) and the spherical score is 0.9051 (0.9019). All these validation measures are in favor of Moody's. Using the described test statistics (see section 4) they find evidence that Moody's outperforms S&P significantly on the 10% significance level for the AUC and on the 1% significance level for the Brier score. But on the other hand, this study

demonstrates that the superiority of Moody's is not stable over time but that in 1999 and 2000 no differences in the quality of default predictions are observable.

Three robustness checks are made to further validate the results of the base case of Table 4 (see Table 5 for the differences and the significance levels as well as Appendix D for the validation measures). For these checks different default rates are calculated because a different mapping procedure is employed and further sample adjustments are necessary (see Appendix C).

First of all, Watchlist entries are incorporated. As part of the rating monitoring process an issuer is often placed on a formal rating review, which is called the Watchlist or credit watch. These additions to the Watchlist signal to the market participants that a rating change will come soon with a high probability but that the rating analysts need some more time to assess the intensity and sometimes also the direction of the forthcoming rating change. Watchlist announcements are often made after M&A activities or corporate restructuring plans have been published. As one example among others, HAND et al. (1992) provide evidence for significant abnormal stock returns after announcements of additions to the S&P Watchlist. HAMILTON and CANTOR (2004) find that the accuracy of default predictions is significantly better if they include Watchlist information. Hence, Watchlist additions are an important source of information. To incorporate them, the credit rating (or the assigned default rate) is downgraded by one notch for a negative Watchlist entry and upgraded by one notch for a positive one. To give an example of this adjustment: an issuer with a mapped rating of 8 is upgraded by one notch to 7. Panel I of Table 5 presents the results for the Watchlist additions. Moody's still performs better than S&P for the years 2001 and 2002 as well as for the overall dataset. The difference of the spherical score for the overall dataset is even significantly different from zero on the 5% level. Unlike the base case of Table 4, S&P is significantly better than Moody's in the year 2000 according to all four validation measures on the 10% level.

Table 5: Robustness checks

Estimation	12/1999	12/2000	12/2001	12/2002	all
Realization period	1/2000-12/2000	1/2001-12/2001	1/2002-12/2002	1/2003-12/2003	
Panel I: Differences between Moody's and S&P (Watchlist anticipation)					
AUC	-0.0047	-0.0100**	0.0042	0.0002	-0.0032
Brier score	-0.0012	-0.0014*	0.0016*	0.0032***	0.0006
Logarithmic score	-0.0029	-0.0044*	0.0046*	0.0065***	0.0010
Spherical score	-0.0010	-0.0013*	0.0021**	0.0045***	0.0011**
Panel II: Differences between Moody's and S&P (ratings younger than 2 years)					
AUC	-0.0099	-0.0066	-0.0013	0.0114*	-0.0023
Brier score	-0.0019	-0.0009	0.0012	0.0035***	0.0006
Logarithmic score	-0.0071	-0.0042	0.0023	0.0090***	0.0004
Spherical score	-0.0019	-0.0006	0.0017	0.0045***	0.0011*
Observations	1,505	1,501	1,404	1,801	6,211
Panel III: Differences between Moody's and S&P (without unsolicited ratings)					
AUC	-0.0012	-0.0068	0.0037	0.0031	-0.0010
Brier score	-0.0010	-0.0008	0.0015*	0.0026***	0.0006*
Logarithmic score	-0.0017	-0.0040*	0.0040	0.0056***	0.0010
Spherical score	-0.0010	-0.0005	0.0019**	0.0034***	0.0010**
Observations	2,704	2,759	2,757	2,797	11,017

The validation measures (see Appendix D) are calculated by using the default frequencies (see Appendix C) of the reduced sub-samples. For convenience, these differences are multiplied by -1 for the Brier score. Positive differences show an advantage for Moody's. One-sided significance levels are given as ***, **, and * representing 1%, 5%, and 10% respectively using the percentile bootstrap method of calculating nonparametric confidence intervals.

The second robustness check is designed to control for staleness in ratings. Staleness in ratings means that the link between the rating and the factors that influence its determination might not truly reflect how the decisions are made by the rating agency (AMATO and FURFINE 2003). This could be due to monitoring costs, the unavailability of qualified staff and the oligopolistic structure of the rating market, all of which raise doubts as to whether suitable resources are in fact allocated to examining all rated firms on a permanent basis. Panel II of Table 5 shows the results for a reduced sample where only ratings younger than two years are included. Both ratings by

Moody's and S&P for a specific issuer must fulfill this criterion. The end of the estimation period is taken as a reference point for defining a rating as "young", e.g. for the first estimation period 1999 a young rating is one that was published no earlier than January 1, 1997. The high ratio of stale ratings can be recognized by the sharp decrease in the size of the sub-samples to almost half the original samples. For the overall sample Moody's still dominates S&P with regard to prognostic power. But as regards the individual years, only in 2002 does this superiority hold for Moody's. For the years 1999 to 2001 there is no clear winner.

Naturally, issuers initiate and pay for their credit ratings. But there are also issuers on the international bond market who do not actively seek credit ratings. Some of them still get rated and these unwanted ratings are usually called unsolicited ratings. Unsolicited ratings were excluded from this study because of their downward bias (POON 2003). There is no data available for both rating agencies because Moody's, unlike S&P, does not publish this kind of information together with its credit ratings. Therefore, unsolicited ratings were estimated by inferring that S&P ratings with a pi indication (meaning that the rating was based on public information) signify an unsolicited rating. Due to the elimination of issuers with a pi rating the sub-samples are smaller than the original samples but the average reduction is smaller than 5%. The results are somewhat stronger than in the base case. For the overall sample the difference between the spherical scores is significantly different on the 5% level and the difference between the Brier scores is different from zero on the 10% level. There is also strong support for the superiority of Moody's for the year 2002 and superiority in two scores for 2001. In 2000, S&P shows a better performance according to a significant difference on the 10% level for the logarithmic score.

6 Concluding remarks

It has been examined whether the default predictions of Moody's or S&P are more accurate. Using a dataset with more than 11,428 issuer ratings and 350 defaults in four datasets from 1999 to 2003 the performance of Moody's and S&P is assessed by bootstrapping the differences revealed by several validation measures. The evidence suggests that Moody's performs slightly better than S&P, which supports the results of the first study in this field by KRÄMER and GÜTTLER (2003). Although in this study Moody's does not lead in all observed periods for the aggregated data of all estimation and realization periods, Moody's nonetheless appears to perform significantly better. The overall result is driven primarily by the noticeable differences in the predictive power of defaults in favor of Moody's in the estimation periods 2001 and particularly 2002. Despite the fact that not all periods support the superiority of Moody's, the results of the base case are robust for variations of the mapping procedure as well as for different adjustments to account for staleness in ratings and for unsolicited ratings.

One possible criticism of the study is that the results, based on a relatively short period with rating and default data, are driven by chance. However, there is no more data available at the moment, because default data prior to 1999 is not published by S&P. Therefore, the maximum amount of rating and default data has been used.

Another potential problem might be the mapping of the different rating scales of Moody's and S&P. As a consequence Moody's ratings are, on average, lower than the ratings of S&P. Of course this could drive results, especially in a period with a high number of defaults. Irrespective of the fact that this mapping is done in every study which assesses split ratings, it should be noted that the superiority of Moody's performance over that of S&P is particularly noticeable in the last period, with rating data as of the end of 2002 and default data as of 2003. Since default rates are lower in 2003 than in any other year during the observation period, the "conservatism" of Moody's should not bias the results.

A further limitation of the study is that it compares two rating agencies with different rating approaches. Whereas S&P's ratings incorporate only probabilities of default, Moody's additionally accounts for expected recovery rates. Empirical evidence

suggests that ratings by Moody's and recovery rates are negatively correlated (MOODY'S 2004). Since recovery rates in 2000 to 2002 were exceptionally low this might have led to the more pessimistic ratings by Moody's. As a consequence, these lower ratings should give Moody's an advantage during a recession with a high number of defaults. In answer to this point, it should be pointed out that, as mentioned above, Moody's performs best in the last period with average recovery rates. Furthermore, by using validation scores which compare probabilities of default with the observed outcome, the inclusion of recovery rates into ratings has no influence since observed default frequencies are used for the estimation of probabilities of defaults.

What are the implications of this paper, or, to put it another way, what can be done with benchmarking results of external rating agencies? First of all, even after this study, there remains a constant need to rate the raters, and not only the two big ones. Thus, a neutral authority like the Bank for International Settlements should introduce a central rating and default database incorporating all rated companies worldwide to allow frequent benchmarking of the accuracy of the different rating agencies' default predictions. Results should be published to increase the transparency of the market for ratings. Perhaps this would be enough to raise the level of competition among the existing rating agencies and to increase the effort exerted by the rating agencies to analyze the issuers carefully. If public pressure were not enough to force low ranked rating agencies to make more effort, regulatory measures should be considered as the next forceful step. On the other hand, rating agencies with comparatively high performance should be privileged in any process involving recognition by national supervisors, should such a process become necessary.

REFERENCES

Amato, J.D. and C.H. Furfine, C.H. (2003): "Are credit ratings procyclical?", BIS Working Papers, No. 129.

BASEL COMMITTEE ON BANKING SUPERVISION (2004): International Convergence of Capital Measurement and Capital Standards. Basel.

BLUME, M.E., F. LIM and A.C. MACKINLAY (1998): "The declining credit quality of U.S. corporate debt: Myth or reality?", Journal of Finance, August, pp. 1389-1414.

BRIER, G. (1950): "Verification of forecasts expressed in terms of probability", Monthly Weather Review, pp. 1-3.

CANTOR, R. and F. PACKER (1997): "Differences of opinion and selection bias in the credit rating industry", Journal of Banking and Finance, October, pp. 1395-1417.

DEGROOT, M.H. and S.E. FIENBERG (1983): "The comparison and evaluation of forecasters", Statistician, March/June, pp. 12-22.

DELONG, E., D. DELONG and D. CLARKE-PEARSON (1988): "Comparing the areas under two or more correlated Receiver Operating Characteristic curves: A nonparametric approach", Biometrics, September, pp. 837-845.

EDERINGTON, L.H. and J.B. YAWITZ (1987): "The bond rating process", in: E.I. Altman and M.J. McKinney (eds.): Handbook of Financial Markets and Institutions 6, New York: John Wiley & Sons, Chapter 23, pp. 49-51.

EFRON, B. (1987): "Better bootstrap confidence intervals", Journal of the American Statistical Association, March, pp. 171-185.

ENGELMANN, B, E. HAYDEN and D. TASCHE (2003): "Testing rating accuracy", Risk, January, pp. 82-86.

ESTRELLA, A. (2000): "Credit ratings and complementary sources of credit quality information", BIS Working Papers, No. 3.

HAMERLE, A., R. RAUHMEIER and D [illegible]CH (2003): "Uses and misuses of measures for credit rating accuracy", Working Paper, University of Regensburg.

HAMILTON, D.T and R. CANTOR (2004): "Rating transitions and defaults conditional on Watchlist, Outlook and rating history, Moody's Special Comment, January.

HAND, J.R.M., R.W. HOLTHAUSEN and R.W. LEFTWICH (1992): “The effect of bond rating agency announcements on bond and stock prices”, Journal of Finance, June, pp. 733-752.

JEWELL, J. and M. LIVINGSTON, M. (1999): “A comparison of bond ratings from Moody’s, S&P and Fitch IBCA”, Financial Markets, Institutions, and Instruments, August, pp. 1-45.

KRÄMER, W. and A. GÜTTLER (2003): “Comparing the accuracy of default predictions in the rating industry: The case of Moody’s vs. S&P”, Technical Report-Series 23, SFB 475.

LÖFFLER, G. (forthcoming): “Ratings versus equity-based measures of default risk in portfolio governance”, Journal of Banking and Finance.

MOODY’S (2004): “Default and recovery rates of corporate bond issuers”, Moody’s Special Comment, January.

MORGAN, D.P. (2002): “Rating banks: Risk and uncertainty in an opaque industry”, American Economic Review, September, pp. 874-889.

PERRY, L.G. (1985): “The effect of bond rating agencies on bond rating models”, Journal of Financial Research, Winter, pp. 307-315.

POON, W.P.H. (2003): “Are unsolicited credit ratings biased downward?”, Journal of Banking and Finance, April, pp. 593-614.

REDELMEIER, D., D. BLOCK and D. HICKAM (1991): “Assessing predictive accuracy: How to compare Brier scores”, Journal of Clinical Epidemiology, November, pp. 1141-1146.

SOBEHART, J. and S. KEENAN (2001): “Measuring default accurately”, Risk, March, pp. 31-33.

STANDARD&POORS (2004): “Corporate defaults in 2003 recede from recent highs”, January.

STEIN, ROGER M. and F. JORDÃO (2003): “What is a more powerful model worth?”, Moody's KMV, Technical Report, August.

WINKLER, R. (1996): “Scoring rules and the evaluation of probabilities”, Test, June, pp. 1-60.

APPENDIX A: Default frequencies across rating grades

The table shows default frequencies of the own sample. They are calculated by dividing the number of ratings by the number of defaults for each rating grade for the respective realization periods. Average default frequencies are calculated over the four periods 2000, 2001, 2002 and 2003. The mapping of the different rating classes of the two rating agencies is indicated in the column "Rating grade". All issuers with a rating worse than B3 or B- are lumped together in the mapped rating class 17. Columns five and six provide the long-term averages of default frequencies for Moody's (MOODY'S 2004) and S&P (STANDARD&POORS 2004).

		Own sample		**Historical averages**	
Rating grade	Mapping	Moody's	S&P	Moody's	S&P
Aaa/AAA	1	0.00%	0.00%	0.00%	0.00%
Aa1/AA+	2	0.00%	0.00%	0.00%	0.00%
Aa2/AA	3	0.00%	0.00%	0.00%	0.00%
Aa3/AA-	4	0.00%	0.00%	0.04%	0.02%
A1/A+	5	0.00%	0.23%	0.00%	0.06%
A2/A	6	0.20%	0.00%	0.03%	0.05%
A3/A-	7	0.10%	0.13%	0.04%	0.04%
Baa1/BBB+	8	0.52%	0.43%	0.19%	0.32%
Baa2/BBB	9	0.11%	0.59%	0.13%	0.34%
Baa3/BBB-	10	0.59%	0.50%	0.45%	0.46%
Ba1/BB+	11	1.19%	0.41%	0.69%	0.64%
Ba2/BB	12	1.23%	1.39%	0.66%	1.15%
Ba3/BB-	13	0.86%	3.15%	2.34%	1.97%
B1/B+	14	3.51%	3.85%	3.22%	3.19%
B2/B	15	6.94%	13.04%	6.54%	8.99%
B3/B-	16	9.24%	23.77%	11.55%	13.01%
C	17	33.51%	41.95%	23.49%	30.85%

APPENDIX B: Distribution of Moody's and S&P Ratings

The distribution of Moody's and S&P's ratings is given as an average over the four one-year periods 1999 - 2002. The numbers are rounded to zero digits if necessary.

		S&P																	
		AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	< B-	Total
Moody's	Aaa	38	7	2	-	0	1	1	0	-	-	-	-	-	-	-	-	-	49
	Aa1	11	13	12	10	1	0	-	-	-	-	0	-	-	-	-	-	-	47
	Aa2	4	11	36	42	25	1	2	0	-	-	-	-	-	-	-	-	-	120
	Aa3	0	4	25	68	54	11	1	-	0	-	-	-	-	1	-	-	-	164
	A1	0	1	4	25	73	46	11	1	1	1	1	1	-	-	-	-	-	163
	A2	1	-	0	12	51	115	56	15	4	1	-	0	-	-	-	-	-	254
	A3	-	-	-	2	11	50	88	54	17	4	1	1	-	-	-	-	-	226
	Baa1	-	-	1	3	2	9	48	117	47	7	3	2	-	-	-	-	-	238
	Baa2	-	-	1	1	1	4	14	62	117	32	4	6	2	-	-	-	-	242
	Baa3	-	-	-	-	-	2	1	11	61	101	21	9	3	1	2	-	-	212
	Ba1	-	-	-	-	-	-	2	4	14	33	32	14	2	3	2	-	0	105
	Ba2	-	-	-	-	-	-	-	0	3	12	29	47	14	2	2	-	-	108
	Ba3	-	-	-	-	-	-	0	0	0	4	21	53	60	31	5	1	1	175
	B1	-	-	-	-	-	0	-	-	0	8	5	32	71	98	18	5	2	238
	B2	-	-	-	-	-	-	0	-	-	-	2	12	47	99	43	9	5	215
	B3	-	-	-	-	-	-	-	-	1	-	1	5	11	44	39	25	14	139
	< B3	-	-	-	-	-	-	-	-	1	1	-	0	4	18	33	34	74	164
	Total	53	36	81	162	216	238	223	265	264	202	119	180	212	295	142	74	96	2,857

APPENDIX C: Default frequencies across rating grades for the robustness checks

The table shows default frequencies of the own sample for different sub-samples of the robustness test. In general, default frequencies are calculated by dividing the number of ratings by the number of defaults for each rating grade for the respective realization periods. Average default frequencies are calculated over the four periods 2000, 2001, 2002 and 2003. Watchlist entries are anticipated by increasing the respective rating by one notch for positive Watchlist entries and by decreasing it for negative entries. The sub-sample with young ratings consists of ratings younger than two years. As a reference point, the end of the estimation period is taken, e.g. for the first estimation period 1999 the rating should not date from earlier than January 1, 1997. Unsolicited ratings are estimated by inferring that S&P ratings with a pi (public information) indication signify an unsolicited rating (POON 2003).

	Watchlist anticipation		**Ratings younger than 2 years**		**Without unsolicited ratings**	
Rating grade	Moody's	S&P	Moody's	S&P	Moody's	S&P
Aaa/AAA	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Aa1/AA+	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Aa2/AA	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Aa3/AA-	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
A1/A+	0.00%	0.00%	0.00%	0.56%	0.00%	0.23%
A2/A	0.00%	0.21%	0.44%	0.00%	0.20%	0.00%
A3/A-	0.33%	0.13%	0.00%	0.29%	0.12%	0.13%
Baa1/BBB+	0.52%	0.46%	0.22%	0.00%	0.59%	0.49%
Baa2/BBB	0.11%	0.38%	0.22%	0.89%	0.12%	0.68%
Baa3/BBB-	0.47%	0.51%	0.23%	0.43%	0.70%	0.57%
Ba1/BB+	0.44%	1.03%	1.58%	0.32%	1.38%	0.49%
Ba2/BB	1.84%	0.85%	2.08%	1.76%	1.39%	1.57%
Ba3/BB-	0.59%	2.83%	1.32%	3.15%	0.93%	3.49%
B1/B+	3.53%	3.66%	3.89%	4.29%	3.79%	4.02%
B2/B	6.83%	10.42%	7.22%	16.54%	7.44%	14.78%
B3/B-	9.49%	21.83%	12.16%	25.79%	9.56%	26.23%
C	33.10%	48.69%	38.84%	48.35%	35.98%	47.34%

APPENDIX D: Validation measures for the robustness checks

The table below shows the differences in the four validation measures of Table 5. For convenience, these differences are multiplied by -1 for the Brier score. Positive differences indicate an advantage for Moody's. One-sided significance levels are given as ***, **, and * representing 1%, 5%, and 10% respectively using the percentile bootstrap method of calculating nonparametric confidence intervals. Watchlist entries are anticipated by increasing the respective rating by one notch for positive Watchlist entries and by decreasing it for negative entries. The sub-sample with young ratings consists of ratings younger than two years. As a reference point, the end of the estimation period is taken, e.g. for the first estimation period 1999 the rating should not date from earlier than January 1, 1997. Unsolicited ratings are estimated by inferring that S&P ratings with a pi (public information) indication signify an unsolicited rating (POON 2003).

Estimation	12/1999	12/2000	12/2001	12/2002	all
Realization period	1/2000-12/2000	1/2001-12/2001	1/2002-12/2002	1/2003-12/2003	
Panel I: Validation measures for issuers rated by Moody's (Watchlist anticipation)					
AUC	0.9058	0.8949	0.8855	0.9285	0.8984
Brier score	0.0255	0.0311	0.0261	0.0167	0.0248
Logarithmic score	-0.0948	-0.1135	-0.0991	-0.0652	-0.0931
Spherical score	0.9729	0.9665	0.9722	0.9825	0.9735
Panel II: Validation measures for issuers rated by S&P (Watchlist anticipation)					
AUC	0.9105	0.9049	0.8814	0.9283	0.9017
Brier score	0.0243	0.0297	0.0276	0.0199	0.0254
Logarithmic score	-0.0919	-0.1091	-0.1036	-0.0717	-0.0941
Spherical score	0.9739	0.9679	0.9700	0.9780	0.9725
Panel III: Validation measures for issuers rated by Moody's (Ratings younger than 2 years)					
AUC	0.8702	0.8727	0.9119	0.9158	0.8871
Brier score	0.0384	0.0484	0.0343	0.0249	0.0360
Logarithmic score	-0.1401	-0.1696	-0.1177	-0.0908	-0.1279
Spherical score	0.9587	0.9475	0.9625	0.9734	0.9611
Panel IV: Validation measures for issuers rated by S&P (Ratings younger than 2 years)					
AUC	0.8801	0.8793	0.9132	0.9045	0.8895
Brier score	0.0365	0.0476	0.0355	0.0283	0.0366
Logarithmic score	-0.1330	-0.1654	-0.1200	-0.0998	-0.1283
Spherical score	0.9606	0.9480	0.9608	0.9689	0.9600
Panel V: Validation measures for issuers rated by Moody's (without unsolicited ratings)					
AUC	0.9022	0.8894	0.8813	0.9281	0.8944
Brier score	0.0266	0.0329	0.0271	0.0176	0.0260
Logarithmic score	-0.0985	-0.1203	-0.1033	-0.0683	-0.0975
Spherical score	0.9717	0.9647	0.9710	0.9814	0.9722
Panel VI: Validation measures for issuers rated by S&P (without unsolicited ratings)					
AUC	0.9034	0.8962	0.8776	0.9251	0.8954
Brier score	0.0256	0.0321	0.0286	0.0202	0.0266
Logarithmic score	-0.0968	-0.1163	-0.1073	-0.0739	-0.0985
Spherical score	0.9727	0.9652	0.9691	0.9780	0.9713

ENDNOTES

[1] LÖFFLER (forthcoming) provides empirical evidence that in addition to commonly used validation measures like the area under the ROC curve, other factors, e.g. the investment horizon and trading costs, have to be accounted for. Therefore, the economic value of rating information has to be assessed in every specific context.

[2] Huge efforts are being made by banking regulatory authorities to create so called "central credit registers" with unique company numbers, internal ratings and default predictions by banks and default data (ESTRELLA 2000). Once these registers have been introduced, quantitative benchmarking of the quality of default prediction is possible.

[3] Defaults for 1999 are used for the adjustments of the estimation period 1999.

[4] The two rating agencies do not disclose the official numbers of rated issuers. Therefore the numbers are calculated using the issuer-weighted default rate and the number of defaults.

[5] The concept of calibration is also used for single rating classes and for other risk measures, such as the loss given default.

[6] More advanced bootstrap approaches exist, such as the bias corrected and accelerated method (EFRON 1987). The advantages of this method are that it is transformation respecting and that it is second-order accurate.

Working Paper Series: Finance & Accounting

No.133: **Christian Laux/ Volker Laux,** Performance Measurement and Information Production, October 2004

No.132: **André Güttler,** Using a Bootstrap Approach to Rate the Raters, October 2004

No.131: **Holger Daske,** Economic Benefits of Adopting IFRS or US-GAAP – Have The Expected Costs of Equity Capital really decreased?, October 2004

No.130: **Holger Daske/ Günther Gebhardt,** Zukunftsorientierte Bestimmung von Kapitalkosten für die Unternehmensbewertung, September 2004

No.129: **Andreas Gintschel/ Andreas Hackethal,** Multi-Bank Loan Pool Contracts, June 2004

No.128: **Andreas Hackethal/ Alexandre Zdantchouk,** Share Buy-Backs in Germany – Overreaction to Weak Signals?, April 2004

No.127: **Louis John Velthuis**, Value Based Management auf Basis von ERIC, March 2004

No.126: **Reinhard H. Schmidt/ J.D. Von Pischke,** Networks of Micro and Small Enterprise Banks: A Contribution to Financial Sector Development, January 2004

No.125: **Andreas Hackethal/ Reinhard H. Schmidt,** Financing Patterns: Measurement Concepts and Empirical Results, January 2004 (fully revised version of Working Paper No. 33)

No.124: **Holger Daske/ Günther Gebhardt/ Stefan Klein**, Estimating the Expected Cost of Equity Capital Using Consensus Forecasts, January 2004

No.123: **Peter Raupach**, The Cost of Employee Stock Options, November 2004

No.122: **Peter Raupach**, The Valuation of Employee Stock Options - How Good is the Standard?, December 2003

No.121: **Andreas Jobst**, European Securitisation: A GARCH Model of CDO, MBS and Pfandbrief Spreads, November 2003

No.120: **Baris Serifsoy/ Marco Weiss**, Efficient Systems for the Securities Transaction Industry- A Framework for the European Union, November 2003

No.119: **Andreas Jobst**, Verbriefung und ihre Auswirkung auf die Finanzmarktstabilität, October 2003

No.118: **Reinhard H. Schmidt,** Corporate Governance in Germany: An Economic Perspective, August 2003 (erschienen in: "The German Financial System", Krahnen, J.P. und Schmidt, R.H. (Hrsg.), Kapitel 12, Oxford University Press, London (2004))

No.117: **Volker Laux**, The Ignored Performance Measure, October 2003

For a complete list of working papers please visit

www.finance.uni-frankfurt.de

Kontaktadresse für Bestellungen:

Professor Dr. Reinhard H. Schmidt
Wilhelm Merton Professur für
Internationales Bank- und Finanzwesen
Mertonstr. 17
Postfach 11 19 32 / HPF66
D-60054 Frankfurt/Main

Tel.: +49-69-798-28269
Fax: +49-69-798-28272

e-mail: merton@wiwi.uni-frankfurt.de

http://www.finance.uni-frankfurt.de

Mit freundlicher Unterstützung der Unternehmen der Sparkassen-Finanzgruppe Hessen-Thüringen.