# *Quantitative Validation of Rating Models for Low Default Portfolios through Benchmarking*

Markus Ricke,
Georg von Pföstl

*The new capital adequacy framework (Basel II) is one of the most fiercely debated topics the financial sector has seen in the recent past. Following a consultation process that lasted several years, the regulations formally took effect on January 1, 2007. The advanced approaches (the advanced internal ratings-based, or A-IRB, approach and the advanced measurement approach, or AMA) are scheduled to become operational on January 1, 2008. The new framework allows banks to use the IRB approach for the calculation of the assessment base for credit risk. Use of the IRB approach is subject to regulatory approval, which can only be obtained if the internal rating systems meet certain requirements. One of these requirements is that the models employed must have good predictive power. Banks must review this predictive power once a year by performing a qualitative and quantitative validation of the models. The statistical methods used to perform quantitative validation require a significant amount of default data to derive valid statements about the model, but such data are typically scarce in the case of rating models for so-called low default portfolios (LDPs), i.e. portfolios for which banks have little default history. In this paper, we first deal with the general problems of LDPs under the IRB approach and cover the problems of validating rating models for LDPs. We then present an alternative method for the quantitative validation of such models, based on the idea of benchmarking. Finally, we provide an example of the application of the proposed validation method.*

*JEL classification: G20, C19*
*Keywords: rating models, validation, benchmarking, low default portfolios*

## 1 Introduction

Following a consultation process that lasted several years, the Basel Committee on Banking Supervision (BCBS) published the revised framework "International Convergence of Capital Measurement and Capital Standards" (Basel II) in June 2004. The Capital Requirements Directive, comprising the recast EU directives 2006/48/EC and 2006/49/EC, transposed the Basel II provisions into EU law. These directives, in turn, were transposed into Austrian law by amending the Austrian Banking Act (*Bankwesengesetz* – BWG) in August 2006 and by publishing the new Solvency Regulation (*Solvabilitätsverordnung* – SolvaV) and Disclosure Regulation (*Offenlegungsverordnung* – OffV) in October 2006. The Basel II revised international capital framework finally entered into force in Austria on January 1, 2007.[1]

The new framework allows banks to use the IRB approach for the calculation of the assessment base for credit risk (IRB approach under Article 22b Austrian Banking Act), subject to regulatory approval, which can only be obtained if the internal rating systems meet a number of requirements that are defined under Article 37 ff. of the Solvency Regulation.

One of these requirements stipulates that banks must demonstrate that their rating models have good predictive power, and that the model must be quantitatively and qualitatively validated on an annual basis

[1] *By exercising areas of national discretion, Austrian banks can postpone the application of the new regulations to January 1, 2008.*

(Articles 41 and 59 of the Solvency Regulation). The statistical methods typically used to perform quantitative validation require a significant amount of default data to derive valid statements about the model, which may be problematic in the case of rating models for low default portfolios (LDPs), i.e. portfolios for which banks have little default history, e.g. sovereigns.

Therefore, this paper presents an alternative method for the quantitative validation of rating models that can be used to assess the predictive power of rating models for typical LDPs such as exposures to sovereigns or banks. The method presented is based on a method used in Hornik et al. (2006), i.e. a benchmarking concept in which the results of an internal rating model are compared with the results obtained from other methods or with external data. This paper covers the comparison with external data.

The paper first deals with the problems of LDPs under the IRB approach (section 2). Section 3 discusses the problems involved in the quantitative validation of rating models for LDPs, and section 4 presents an alternative method for the quantitative validation of rating models for LDPs based on a benchmarking concept. Section 5 shows an example of the application of the suggested validation method. Section 6 concludes.

## 2 Low Default Portfolios under the IRB Approach

Low default portfolios (LDPs) are portfolios with only few or no defaults. A portfolio may be LDP for different reasons, e.g.:[2]

– it may be a portfolio with few customers – either globally (e.g. sovereigns) or at an individual bank level;
– it may reflect a globally low default rate for certain customer groups (e.g. banks);
– it may reflect a low default rate for certain customer groups in certain time periods;
– it may have a short default history because the bank is a recent market entrant for a given portfolio.

Based on these different reasons, LDPs are often subdivided into the following types:[3]

– *Long-term versus short-term:* Long-term LDPs may be attributed to generally low default rates of certain borrower groups or a small number of borrowers. LDPs are short term, however, if the lack of sufficient default data is due to a bank's recent entry into a new market segment.
– *Systemic versus institution-specific:* In the case of systemic LDPs, all banks face the problem of having few or no default data, while in the case of institution-specific LDPs, data are unavailable only for the bank in question.

Although the lack of default data for LDPs makes it difficult to develop and validate rating models as well as estimate and validate risk parameters for these portfolios, statutory provisions do not contain requirements specifically applicable to LDPs. Consequently, many banks have raised concerns that LDPs may be generally excluded from IRB treatment. In a response to industry questions, the

[2] *See BBA and ISDA (2005).*

[3] *See CEBS (2006, p. 101).*

BCBS published a newsletter in September 2005.[4] The BCBS's core statement is that the relative lack of historical data should not automatically preclude LDPs from the use of IRB approaches. Rather, greater reliance should be placed on alternative external and internal data sources for LDPs. If data richness is still not given, alternative techniques for estimation and validation should be used.[5] Moreover, given an insufficient data base and therefore a larger uncertainty in parameter estimation, banks would have to increase the margin of conservatism added to the risk parameters.[6]

Rating is about bringing borrowers into an order with respect to their default probability. To this end, a discrete scale with various rating grades is typically used. Statistical procedures such as logistic regression are often used to develop a rating model; however, they require a minimum amount of default data. Given the lack of such data, such procedures cannot be applied to LDPs. Instead, expert models, i.e. models where the rating criteria are chosen and weighted by experts, are typically employed.

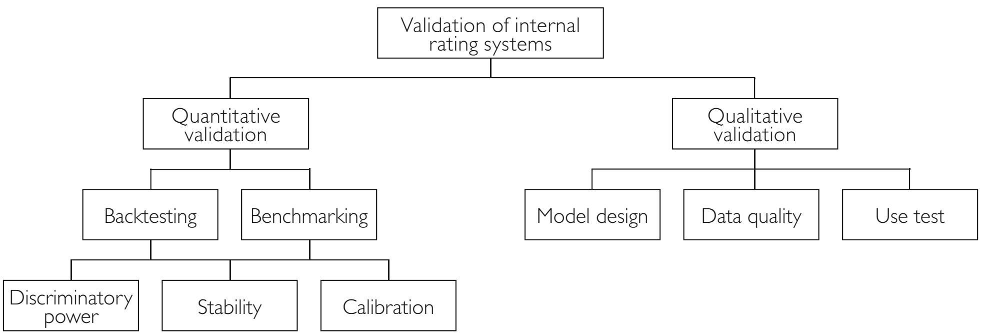
The use of expert models for LDPs is permitted in principle.[7] However, the use of an expert model does not exempt banks from the obligation to validate the model regularly by means of quantitative techniques. This poses problems for many banks, as the methods traditionally employed for the quantitative validation of a rating model require a certain number of defaults, which do not exist in the case of LDPs. The next section therefore presents an alternative technique for the quantitative validation of rating models for LDPs.

## 3 Validation of Low Default Portfolios

According to Deutsche Bundesbank (2003) and OeNB and FMA (2004), the validation of a rating model has to comprise the measures depicted in chart 1.

Chart 1

**Validation of a Rating Model**




*Source: Deutsche Bundesbank (2003, p. 60).*

[4] See BCBS (2005a).

[5] These statements can be also found in the CP 10 consultation paper of the Committee of European Banking Supervisors (CEBS) published in April 2006, see CEBS (2006).

[6] In national legislation, this issue is addressed under Article 47 para 6 of the Solvency Regulation.

[7] For parameter estimates, however, such a method is not permissible, as Article 47 para 1 of the Solvency Regulation explicitly demands that parameter estimates not be based purely on judgmental considerations but also on empirical results.

Quantitative validation refers to the use of statistical procedures to examine the discriminatory power[8] and the accuracy of calibration[9] of the rating model as well as the stability of the rating results, while qualitative validation refers to the data quality, the model design, and the internal use of the rating results in the bank's risk management. Quantitative validation can be performed on the basis of internal data (backtesting) or external data (benchmarking).

Quantitative validation through backtesting is possible only to a very limited extent for LDPs, since the number of defaults in the bank's portfolio is typically so low that performing statistical tests does not lead to any reasonable results. Should the LDP be institute specific, this problem may be solved by using data from other banks. However, if the LDP is systemic, quantitative validation through benchmarking (comparing own data with other banks' data) is not possible.

BCBS (2005a) as well as BBA and ISDA (2005) therefore define the term benchmarking more broadly to comprise methods such as the comparison of internal ratings with ratings by rating agencies and with proxies for default risk derived from market prices. In the next section, we will present a possible technique for this kind of comparison.

## 4 Benchmarking of Rating Models for Systemic Low Default Portfolios

A rating is an ordinal variable, i.e. the borrowers are ranked by their default probability, typically using a discrete scale with different rating grades to which the borrowers are allocated.[10]

In the following, we present a benchmarking approach for the quantitative validation of rating models for systemic LDPs,[11] where the ordinal structure of the results of a rating model, i.e. the ranking of the borrowers by their default probability, is compared with the ranking of rating agencies or with proxies for default risk observable in the capital market.[12]

The literature often suggests Spearman's rank correlation coefficient, Somer's D or Kendall's $\tau$ as measures of the strength and direction of association between two ordinally scaled variables.[13] However, Emond and Mason (2002) have shown that these measures have certain weaknesses and have therefore suggested an enhanced coefficient, $\tau_x$, which e.g. Hornik et al. (2006) use.

[8] *The discriminatory power of a rating model refers to its ex ante ability to distinguish borrowers who will default from those who will not default.*

[9] *The calibration of a rating model denotes the assignment of probabilities of default to the different rating grades.*

[10] *In a further step – calibration – a default probability is assigned to the individual rating grades.*

[11] *It should be mentioned at this point that, on the one hand, benchmarking analysis requires sufficiently large multi-rater panels. On the other hand, these panels have to be complete, which is usually not the case in practice as not all agencies assess all borrowers concerned. This contribution does not deal with the problem of incomplete panels, as all borrowers in the example are assessed by all agencies. See Hornik et al. (2006) for a treatment of this problem.*

[12] *The approach presented in this paper thus can be seen as an examination of the discriminatory power under the assumption that the rating agencies and/or capital market players are in a position to distinguish borrowers who will default from those who will not default in the future. It is also possible to examine the calibration of rating models for LDPs through benchmarking, but this is not the object of this paper.*

[13] *See BCBS (2005b).*

To calculate $\tau_x$ for a sample with $n$ borrowers, an $n \times n$ matrix is first created for each variable,[14] the elements of which are determined as follows for variable $a$:

$a_{xy} = 1$ if borrower $x$ is ranked ahead of or even with borrower $y$;

$a_{xy} = -1$ if borrower $x$ is ranked behind borrower $y$; and

$a_{xy} = 0$ for all diagonal elements of the matrix.

Based on this matrix, $\tau_x$ can be calculated for variables $a$ and $b$ with the following formula:

$$\tau_x = \frac{\sum_{x=1}^{n}\sum_{y=1}^{n} a_{xy} b_{xy}}{n(n-1)} \qquad (1)$$

$\tau_x$ can range between –1 and +1, with higher values of representing a higher degree of association.

External ratings and proxies for default risk derived from market prices, e.g. bond spreads[15] or credit default swap (CDS) spreads,[16] are typically recommended as benchmarks.[17] The benchmarks implicitly assume that the ranking of the borrowers by the external rating agencies and/or capital market investors is perfect.

The ratings of the large rating agencies and the level of bond spreads and/or CDS spreads are closely linked.[18] Nevertheless, there are some important differences between these measures of the default risk of a borrower. One of these differences lies in the stability of the measure.

Rating agencies emphasize that their ratings are through the cycle (TTC).[19] This means that the rating should reflect the borrower's long-term creditworthiness irrespective of the business cycle.[20] Short-term, possibly only temporary, changes in default risk are not considered, as the agencies tend to focus on the stability of the rating.[21]

The market-based proxies for a borrower's default risk, by contrast, are typically point-in-time (PIT) measures. This means that they react to changes in the economic environment and therefore fluctuate more strongly than TTC ratings.

When choosing the benchmark, this circumstance has to be taken into account. If the model to be validated is a TTC model, external ratings would appear to be appropriate as a benchmark. In the case of a PIT

[14] *In this case, the variables are the internal rating and the proxies for the default probability used for comparison, e.g. an external rating.*

[15] *A bond spread is the difference in yield between a risky bond and a (nearly) risk-free bond with the same maturity; it is typically higher the higher the default risk of the bond issuer is.*

[16] *A CDS is a contract to hedge against credit risks, i.e. the protection seller agrees to pay compensation to the protection buyer in the amount of a potential loss in the event of a prespecified credit event. In exchange, the protection buyer pays the protection seller a fee, the so-called CDS spread (in percent of the nominal amount of the exposure) for the hedging period. The higher the probability of the credit event is, the higher the fee is.*

[17] *Zhu (2004) showed that bond spreads and CDS spreads move together in the long run, but that this relationship does not always hold in the short run. The level of both measures is influenced not only by default risk but also by other factors such as liquidity, taxes or risk premiums requested by investors; see e.g. Elton et al. (2001) or Amato and Remolona (2003).*

[18] *See Amato and Remolona (2003).*

[19] *See Cantor (2001) and Standard & Poor's (2006).*

[20] *Several empirical studies investigated whether the ratings of the big rating agencies are really independent of the state of the economy; see Nickell et al. (2000), Bangia et al. (2002), Amato and Furfine (2004), and Löffler (2006).*

[21] *See Fons et al. (2002).*

Table 1

**Rating Grades and CDS Spreads of Sovereigns**

| Borrower | Internal rating model | S&P | Moody's | Fitch | CDS spreads |
|---|---|---|---|---|---|
| Brazil | 8 | BB+ | Ba2 | BB+ | 71 |
| Hungary | 3 | BBB+ | A2 | BBB+ | 19 |
| Mexico | 4 | BBB | Baa1 | BBB | 34 |
| Poland | 3 | A− | A2 | BBB+ | 8 |
| Russia | 5 | BBB+ | Baa2 | BBB+ | 42 |
| South Korea | 2 | A− | A3 | A+ | 16 |
| Turkey | 9 | BB− | Ba3 | BB− | 148 |
| Ukraine | 9 | BB− | B1 | BB− | 131 |
| Venezuela | 10 | BB− | B2 | BB− | 251 |

*Source: Standard & Poor's, Moody's, Fitch, Deutsche Bank (2007).*

model, however, a market-based proxy should be used as a benchmark.[22] However, due to the high fluctuation of the market-based proxies compared to internal ratings, which are normally updated only once a year, the benchmarking result may depend strongly on the valuation date.

Notwithstanding the different rating philosophies discussed, the various ratings should mirror the same risk parameter. Thus, it has to be considered whether the ratings are to be regarded exclusively as PD estimates or whether they focus on expected loss. In addition, the different ratings should refer to the same time horizon. We are aware of the fact that the benchmarks proposed do not always fulfill these requirements. Nevertheless, they are proposed since "better" benchmarks for LDPs are often not available in practice.

The next section demonstrates the application of the presented method to a simple example. The results of a fictitious internal rating model for sovereigns are compared with the ratings of the three big rating agencies Standard and Poor's (S&P), Moody's, and Fitch and with CDS spreads observable in the capital market.

## 5 Example for the Application of the Proposed Benchmarking-Based Method for the Validation of Rating Models

This section uses an example to illustrate the application of the method presented in section 4 in more detail. To this end, the (fictitious) results of an internal rating model for sovereigns are compared with the ratings of the rating agencies S&P, Moody's and Fitch on the one hand and with (CDS) spreads observable in the capital market on the other hand.

Table 1 presents the ratings and the CDS spreads of the individual sovereigns.[23] The results of the internal rating model are fictitious values on a rating scale of 1 to 12, with 1 being the best rating. The CDS

[22] *Based on interviews, Treacy and Carey (1998) discovered that the internal rating models of (big U.S.) banks are typically PIT rating models. In addition, Weber et al. (1999) found out that the ratings of the models of larger German banks fluctuate more strongly than the external ratings of the respective borrowers, which might serve as evidence that the internal models of large German banks are PIT rather than TTC models.*

[23] *In general, validation should be performed with a sample that is as large as possible so that the results are not distorted by individual outliers. However, for the sake of clarity, only ratings of nine sovereigns are considered in the example.*

Table 2

**Assessment Matrix for the Internal Rating System**

| | Brazil | Hungary | Mexico | Poland | Russia | South Korea | Turkey | Ukraine | Vene-zuela |
|---|---|---|---|---|---|---|---|---|---|
| Brazil | 0 | –1 | –1 | –1 | –1 | –1 | 1 | 1 | 1 |
| Hungary | 1 | 0 | 1 | 1 | 1 | –1 | 1 | 1 | 1 |
| Mexico | 1 | –1 | 0 | –1 | 1 | –1 | 1 | 1 | 1 |
| Poland | 1 | 1 | 1 | 0 | 1 | –1 | 1 | 1 | 1 |
| Russia | 1 | –1 | –1 | –1 | 0 | –1 | 1 | 1 | 1 |
| South Korea | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| Turkey | –1 | –1 | –1 | –1 | –1 | –1 | 0 | 1 | 1 |
| Ukraine | –1 | –1 | –1 | –1 | –1 | –1 | 1 | 0 | 1 |
| Venezuela | –1 | –1 | –1 | –1 | –1 | –1 | –1 | –1 | 0 |

Table 3

**Product Matrix of a (Fictitious) Internal Rating and S&P Rating**

| | Brazil | Hungary | Mexico | Poland | Russia | South Korea | Turkey | Ukraine | Vene-zuela |
|---|---|---|---|---|---|---|---|---|---|
| Brazil | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Hungary | 1 | 0 | 1 | –1 | 1 | 1 | 1 | 1 | 1 |
| Mexico | 1 | 1 | 0 | 1 | –1 | 1 | 1 | 1 | 1 |
| Poland | 1 | 1 | 1 | 0 | 1 | –1 | 1 | 1 | 1 |
| Russia | 1 | –1 | –1 | 1 | 0 | 1 | 1 | 1 | 1 |
| South Korea | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| Turkey | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| Ukraine | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| Venezuela | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |

*Source: Standard & Poor's.*

spreads used are values observed in the capital market.[24]

Based on the data presented in table 1 and following the technique for calculating $\tau_x$ described in section 4, we first create a matrix for each variable (i.e. for the internal rating system, the ratings of the three rating agencies, and the CDS spreads). The columns and rows of the matrix represent the respective sovereigns (borrowers). Table 2 presents the assessment matrix for the internal rating system as an example. If a cell contains 1, the internal rating of the sovereign in that row is better or the same as that of the sovereign in the column. The row for South Korea, for instance, contains 1 in every cell, as this sovereign was assigned the best rating of the nine sovereigns by the internal rating system. –1, however, is assigned if the sovereign in that row has a worse internal rating than the sovereign in the column. By definition, the diagonal is 0.

After an assessment matrix has been created for each of the five variables, the matrix for the internal rating system is multiplied with each of the other matrices in turn. Table 3 presents the product matrix for the internal rating system and the matrix for the S&P ratings as an example. A cell contains 1 whenever the respective cells in both matrices concurrently show 1 or –1. This means that the ranking of the two sovereigns to which the respective cell refers is not opposite in the two variables. For ex-

[24] *The data for the CDS spreads (five-year CDS spreads) are from Deutsche Bank (2007); as at July 6, 2007.*

ample, South Korea receives a better rating than Hungary both from the internal rating system and S&P. However, a value of –1 arises in the product matrix if the ranking of the two compared sovereigns is opposite in the observed variables. Hence, e.g. Russia's creditworthiness is lower than that of Mexico in the internal rating system, while S&P awards a better rating to Russia than to Mexico.

After the product matrices have been created, the indicator $\tau_x$ can be calculated for each product matrix based on formula (1). To compare the results of the internal rating system with those of S&P, for instance, $\tau_x$ is computed as follows:

$$\tau_x = \frac{58}{9(9-1)} \approx 0.81$$

Table 4 presents the (rounded) results for the given example:

Table 4

**(Rounded) Results for $\tau_x$**

| Method | $\tau_x$ |
|---|---|
| Internal rating system and S&P | 0.81 |
| Internal rating system and Moody's | 0.86 |
| Internal rating system and Fitch | 0.83 |
| Internal rating system and CDS spreads | 0.89 |

It is evident that in this fictitious example the result of all four comparisons exceeds 0.8, with the highest $\tau_x$ for the risk measure CDS spread.

The above-mentioned issues – the rating philosophy, the considered risk parameter or the time horizon – have to be considered when interpreting the results.

## 6 Conclusion

This paper has suggested a method for the quantitative validation of rating models for LDPs. One necessary requirement for the application of this method is the existence of an appropriate benchmark. The benchmarks *external ratings* and *bond and/or CDS spreads* presented in this paper are available for typical LDPs, such as sovereign, bank and large corporate exposures, making the method particularly well suited for these LDPs.

The explanatory power of the results strongly depends on the quality[25] of the benchmark, since the presented method does not directly assess the quality of the results of the internal rating model but rather the association of its results with those of the benchmark. Thus, it can only be concluded from a high $\tau_x$ value that the internal rating model has a high discriminatory power if the benchmark itself has a high discriminatory power. Conversely, a low discriminatory power of the internal rating model cannot be directly inferred from a low result for $\tau_x$. Rather, the reasons for the low $\tau_x$ value – for example a low discriminatory power of the benchmark – should be examined.

[25] *Quality in the sense of discriminatory power.*

## References

**Amato, J. and C. Furfine. 2004.** Are Credit Ratings Procyclical? In: Journal of Banking and Finance 28. 2641–2677.

**Amato, J. and E. Remolona. 2003.** The Credit Spread Puzzle. In: BIS Quarterly Review. December. 51–63.

**Bangia, A., F. Diebold, A. Kronimus, C. Schagen and T. Schuermann. 2002.** Ratings Migration and the Business Cycle, with Application to Credit Portfolio Stress Testing. In: Journal of Banking and Finance 26. 445–474.

**BBA and ISDA – British Bankers' Association, London Investment Banking Association and International Swaps and Derivatives Association. 2005.** Low Default Portfolios. Joint Industry Working Group Discussion Paper.

**BCBS – Basel Committee on Banking Supervision. 2005a.** Validation of Low-Default Portfolios in the Basel II Framework. Basel Committee Newsletter 6. September.

**BCBS – Basel Committee on Banking Supervision. 2005b.** Studies on the Validation of Internal Rating Systems. Working Paper 14.

**Cantor, R. 2001.** Moody's Investors Service Response to the Consultative Paper Issued by the Basel Committee on Banking Supervision "A New Capital Adequacy Framework". In: Journal of Banking and Finance 25. 171–185.

**CEBS – Committee of European Banking Supervisors. 2006.** Guidelines on the Implementation, Validation and Assessment of Advanced Measurement (AMA) and Internal Ratings Based (IRB) Approaches. April.

**Deutsche Bank. 2007.** Finanzmarktmonitoring – CDS Spreads. Deutsche Bank Research. http://www.dbresearch.com/servlet/reweb2.ReWEB?rwkey=u727019
(as of July 6, 2007).

**Deutsche Bundesbank. 2003.** Approaches to the Validation of Internal Rating Systems. In: Monatsbericht September. Frankfurt am Main. 59–71.

**Elton, E., M. Gruber, D. Agrawal and C. Mann. 2001.** Explaining the Rate Spread on Corporate Bonds. In: Journal of Finance 56. 247–277.

**Emond, E. and D. Mason. 2002.** A New Rank Correlation Coefficient with Application to the Consensus Ranking Problem. In: Journal of Multi-Criteria Decision Analysis 11. 17–28.

**Fons, J., F. Cantor and C. Mahoney. 2002.** Understanding Moody's Corporate Bond Ratings and Rating Process. Special Comment. Moody's Investors Services.

**Hornik, K., R. Jankowitsch, M. Lingo, S. Pichler and G. Winkler. 2006.** Validation of Credit Rating Systems Using Multi-Rater Information. Working Paper. November.

**Löffler, G. 2006.** Can Rating Agencies look Through the Cycle? Working Paper. University of Ulm.

**Nickell, P., W. Perraudin and S. Varotto. 2000.** Stability of Rating Transitions. In: Journal of Banking and Finance 24. 203–227.

**OeNB and FMA – Oesterreichische Nationalbank and Finanzmarktaufsicht. 2004.** Leitfadenreihe zum Kreditrisiko: Ratingmodelle und -validierung. Vienna.

**Standard & Poor's. 2006.** Corporate Ratings Criteria. New York.

**Treacy, W. and M. Carey. 1998.** Credit Risk Rating at Large U.S. Banks. In: Federal Reserve Bulletin. November. 897–921.

**Weber, M., J. Krahnen and F. Vossmann. 1999.** Risikomessung im Kreditgeschäft: Eine empirische Analyse bankinterner Ratingverfahren. In: Zeitschrift für betriebswirtschaftliche Forschung. Sonderheft 41. 117–142.

**Zhu, H. 2004.** An Empirical Comparison of Credit Spreads between the Bond Market and the Credit Default Swap Market. BIS Working Paper 160. Basel.