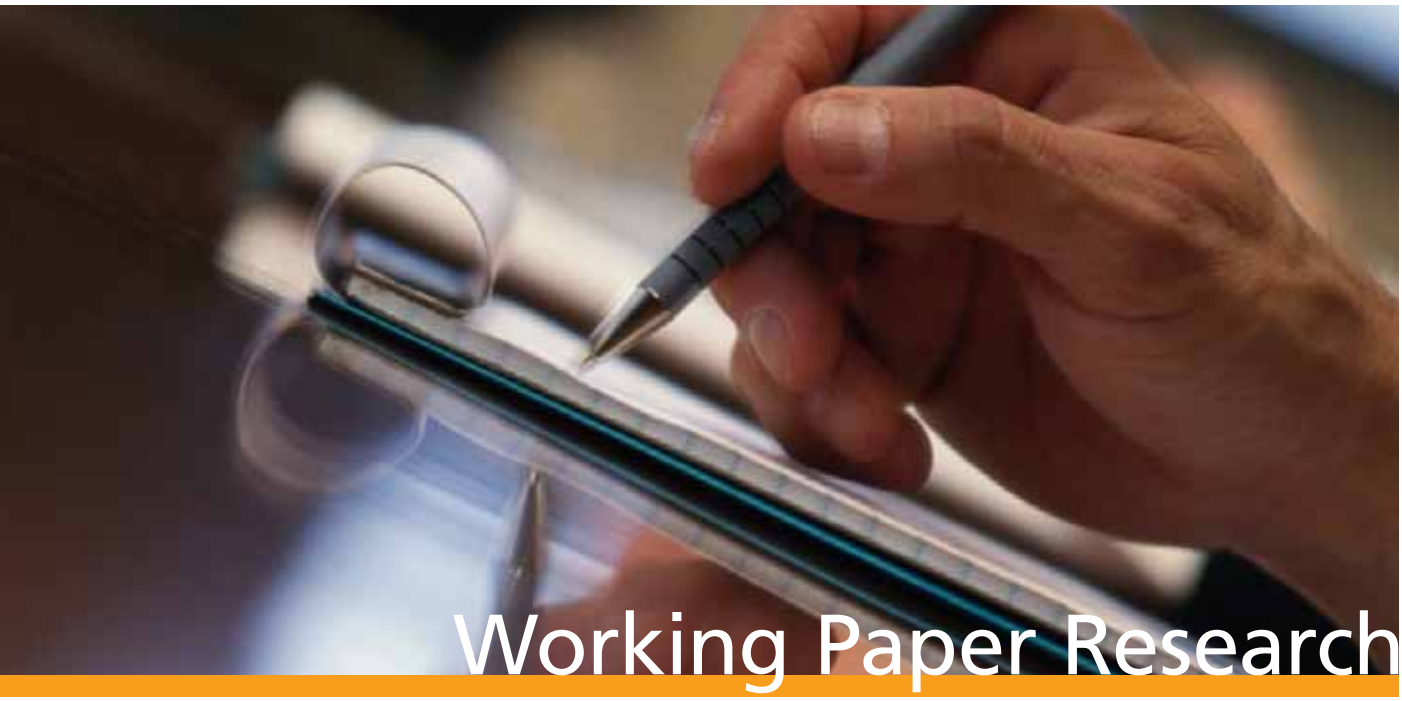
The performance of credit rating systems in the assessment of collateral used in Eurosystem monetary policy operations



Working Paper Research

by François Coppens, Fernando González and Gerhard Winkler

September 2007 No 118

National Bank OF BELGIUM
Eurosystem

Editorial Director

Jan Smets, Member of the Board of Directors of the National Bank of Belgium

Statement of purpose:

The purpose of these working papers is to promote the circulation of research results (Research Series) and analytical studies (Documents Series) made within the National Bank of Belgium or presented by external economists in seminars, conferences and conventions organised by the Bank. The aim is therefore to provide a platform for discussion. The opinions expressed are strictly those of the authors and do not necessarily reflect the views of the National Bank of Belgium.

Orders

For orders and information on subscriptions and reductions: National Bank of Belgium,
Documentation - Publications service, boulevard de Berlaimont 14, 1000 Brussels

Tel +32 2 221 20 33 - Fax +32 2 21 30 42

The Working Papers are available on the website of the Bank: http://www.nbb.be

ISSN: 1375-680X (print)
ISSN: 1784-2476 (online)

Abstract

The aims of this paper are twofold: first, we attempt to express the threshold of a single "A" rating as issued by major international rating agencies in terms of annualised probabilities of default. We use data from Standard & Poor's and Moody's publicly available rating histories to construct confidence intervals for the level of probability of default to be associated with the single "A" rating. The focus on the single A rating level is not accidental, as this is the credit quality level at which the Eurosystem considers financial assets to be eligible collateral for its monetary policy operations. The second aim is to review various existing validation models for the probability of default which enable the analyst to check the ability of credit assessment systems to forecast future default events. Within this context the paper proposes a simple mechanism for the comparison of the performance of major rating agencies and that of other credit assessment systems, such as the internal ratings-based systems of commercial banks under the Basel II regime. This is done to provide a simple validation yardstick to help in the monitoring of the performance of the different credit assessment systems participating in the assessment of eligible collateral underlying Eurosystem monetary policy operations. Contrary to the widely used confidence interval approach, our proposal, based on an interpretation of p-values as frequencies, guarantees a convergence to an ex ante fixed probability of default (PD) value. Given the general characteristics of the problem considered, we consider this simple mechanism to also be applicable in other contexts.

JEL classification: G20, G28, C49.

Keywords: credit risk, rating, probability of default (PD), performance checking, backtesting.

Corresponding author:

NBB, Microeconomic Information Department, e-mail: francois.coppens@nbb.be

TABLE OF CONTENTS

1. INTRODUCTION

To ensure the Eurosystem's requirement of high credit standards for all eligible collateral, the ECB's Governing Council has established the so-called Eurosystem Credit Assessment Framework (ECAF) (see European Central Bank 2007). The ECAF comprises the techniques and rules which establish and ensure the Eurosystem's requirement of high credit standards for all eligible collateral. Within this framework, the Eurosystem has specified its understanding of high credit standards as a minimum credit quality equivalent to a rating of "A",[1] as issued by the major international rating agencies.

In its assessment of the credit quality of collateral, the ECB has always taken into account, inter alia, available ratings by major international rating agencies. However, relying solely on rating agencies would not adequately cover all types of borrowers and collateral assets. Hence the ECAF makes use not only of ratings from (major) external credit assessment institutions, but also other credit quality assessment sources, including the in-house credit assessment systems of national central banks,[2] the internal ratings-based systems of counterparties and third-party rating tools (European Central Bank, 2007).

This paper focuses on two objectives. First, it analyses the assignation of probabilities of default to letter rating grades as employed by major international rating agencies and, second, it reviews various existing validation methods for the probability of default. This is done from the perspective of a central bank or system of central banks (e.g. the Eurosystem) in the special context of its conduct of monetary policy operations in which adequate collateral with "high credit standards" is required. In this context, "high credit standards" for eligible collateral are ensured by requiring a minimum rating or its quantitative equivalent in the form of an assigned annual probability of default. Once an annual probability of default at the required rating level has been assigned, it is necessary to assess whether the estimated probability of default issued by the various credit assessment systems conform to the required level. The methods we review and propose throughout this paper for these purposes are deemed to be valid and applicable not only in our specific case but also in more general cases.

The first aim of the paper relates to the assignation of probabilities of default to certain rating grades of external rating agencies. Ratings issued by major international rating agencies often act as a benchmark for other credit assessment sources whose credit assessments are used for comparison. Commercial banks have a natural interest in the subject because probabilities of default are inputs in the pricing of all sorts of risk assets, such as bonds, loans and credit derivatives (see e.g. Cantor et al. (1997), Elton et al. (2004), and Hull et al. (2004)). Furthermore, it is of crucial importance for regulators as well. In the "standardised approach" of the New Basel Capital Accord, credit assessments from external credit assessment institutions can be used for the calculation of the required regulatory capital (Basel Committee on Banking Supervision (2005a)). Therefore, regulators must have a clear understanding of the default rates to be expected (i.e. probability of default) for specific rating grades (Blochwitz and Hohl (2001)). Finally, it is also essential for central banks to clarify what specific rating grades mean in terms of probabilities of default since most central banks also partly rely on ratings from external credit institutions for establishing eligible collateral in their monetary operations. Although it is well known that agency ratings may to some extent also be dependent on the expected severity of loss in the event of default (e.g. Cantor and Falkenstein (2001)), a consistent and clear assignment of probabilities of default to rating grades should be theoretically possible because we infer from the rating agencies´ own definitions of the meanings of their ratings that their prime purpose is to reflect default

1 Note that we focus on the broad category "A" throughout this paper. The "A"-grade comprises three sub-categories (named A+, A, and A- in the case of Standard & Poor's, and A1, A2, and A3 in the case of Moody's). However, we do not differentiate between them or look at them separately, as the credit threshold of the Eurosystem was also defined using the broad category.

2 At the time of publication of this paper, only the national central banks of Austria, France, Germany and Spain possessed an in-house credit assessment system.

probability (Crouhy et al. (2001)). This especially holds for "issuer-specific credit ratings", which are the main concern of this paper. Hence a clear relation between probabilities of default and rating grades definitely exists, and it has been the subject of several studies (Cantor and Falkenstein (2001), Blochwitz and Hohl (2001), Tiomo (2004), Jafry and Schuermann (2004) and Christensen et al. (2004)). It thus seems justifiable for the purposes of this paper to follow the definition of a rating given by Krahnen et al. (2001) and regard agency ratings as "the mapping of the probability of default into a discrete number of quality classes, or rating categories" (Krahnen et al. (2001)).

We thus attempt to express the threshold of a single "A" rating by means of probabilities of default. We focus on the single A rating level because this is the level at which the ECB Governing Council has explicitly defined its understanding of "high credit standards" for eligible collateral in the ECB monetary policy operations. Hence, in the empirical application of our methods, which we regard as applicable to the general problem of assigning probabilities of default to any rating grades, we will restrict ourselves to a single illustrative case, the "A" rating grade. Drawing on the above-mentioned earlier works of Blochwitz and Hohl (2001), Tiomo (2004) and Jafry and Schuermann (2004), we analyse historical default rates published by the two rating agencies Standard & Poor's and Moody's. However, as default is a rare event, especially for entities rated "A" or better, the data on historically observed default frequencies shows a high degree of volatility, and probability of default estimates could be very imprecise. This may be due to country-specific and industry-specific idiosyncrasies which might affect rating migration dynamics (Nickel et al. (2000)). Furthermore, macroeconomic shocks can generally also influence the volatility of default rates, as documented by Cantor and Falkenstein (2001). As discussed by Cantor (2001), Fons (2002) and Cantor and Mann (2003), however, agency ratings are said to be more stable in this respect because they aim to measure default risk over long investment horizons and apply a "through the cycle" rating philosophy (Crouhy et al. (2001) and Heitfield (2005)). Based on these insights we derive an ex ante benchmark for the single "A" rating level. We use data of Standard & Poor's and Moody's publicly available rating histories (Standard & Poor's (2005), Moody's (2005)) to construct confidence intervals for the level of probability of default to be associated with a single "A" rating grade. This results in one of the main contributions of our work, i.e. the statistical deduction of an ex ante benchmark of a single "A" rating grade in terms of probability of default.

The second aim of this paper is to explore validation mechanisms for the estimates of probability of default issued by the different rating sources. In doing so, it presents a simple testing procedure that verifies the quality of probability of default estimates. In a quantitative validation framework the comparison of performance could be based mainly on two criteria: the discriminatory power or the quality of calibration of the output of the different credit assessment systems under comparison. Whereas the "discriminatory power" refers to the ability of a rating model to differentiate between good and bad cases, calibration refers to the concrete assignment of default probabilities, more precisely to the degree to which the default probabilities predicted by the rating model match the default rates actually realised. Assessing the calibration of a rating model generally relies on backtesting procedures.[3] In this paper we focus on the quality of the calibration of the rating source and not on its discriminatory power.[4]

Analysing the significance of deviations between the estimated default probability and the realised default rate in a backtesting exercise is not a trivial task. Realised default rates are subject to statistical fluctuations that could impede a straight forward assessment of how well a rating system estimates probabilities of default. This is mainly due to constraints on the number of observations available owing to the scarcity of default events and the fact that default events may not be independent but show some degree of correlation. Non-zero default correlations have the effect of

[3] To conduct a backtesting examination of a rating source the basic data required is the estimate of probability of default for a rating grade over a specified time horizon (generally 12 months), the number of rated entities assigned to the rating grade under consideration and the realised default status of those entities after the specified time horizon has elapsed (i.e. generally 12 months after the rating was assigned).

[4] For an exposition of discriminatory power measures in the context of the assessment of performance of a rating source see, for example, Tasche (2006).

amplifying variations in historically observed default rates which would normally prompt the analyst to widen the tolerance of deviations between the estimated average of the probabilities of default of all obligors in a certain pool and the realised default rate observed for that pool. In this sense, two approaches can be considered in the derivation of tests of deviation significance: tests assuming uncorrelated default events and tests assuming default correlation.

There is a growing literature on probability of default validation via backtesting (e.g. Cantor and Falkenstein (2001), Blochwitz et al. (2003), Tasche (2003), Rauhmeier (2006)). This work has been prompted mainly by the need of banking regulators to have validation frameworks in place to face the certification challenges of the new capital requirement rules under Basel II. Despite this extensive literature, there is also general acceptance of the principle that statistical tests alone would not be sufficient to adequately validate a rating system (Basel Committee on Banking Supervision (2005b)). As mentioned earlier, this is due to scarcity of data and the existence of a default correlation that can distort the results of a test. For example, a calibration test that assumes independence of default events would normally be very conservative in the presence of correlation in defaults. Such a test could send wrong messages for an otherwise well calibrated rating system. However, and given these caveats, validation by means of backtesting is still considered valuable for detecting problems in rating systems.

We briefly review various existing statistical tests that assume either independence or correlation of defaults (cf. Brown et al. (2001), Cantor and Falkenstein (2001), Spiegelhalter (1986), Hosmer and Lemeshow (2000), Tasche (2003)). In doing so, we take a closer look at the binomial model of defaults that underpins a large number of tests proposed in the literature. Like any other model, the binomial model has its limitations. We pay attention to the discreteness of the binomial distribution and discuss the consequences of approximation, thereby accounting for recent developments in statistics literature regarding the construction of confidence intervals for binomially distributed random variables (for an overview see Vollset (1993), Agresti and Coull (1998), Agresti and Caffo (2000), Reiczigel (2004) and Cai (2005)).

We conclude the paper by presenting a simple hypothesis testing procedure to verify the quality of probability of default estimates that builds on the idea of a “traffic light approach” as discussed in, for example, Blochwitz and Hohl (2001) and Tiomo (2004). A binomial distribution of independent defaults is assumed in accordance with the literature on validation. Our model appears to be conservative and thus risk averse. Our hypothesis testing procedure focuses on the interpretation of *p*-values as frequencies, which, contrary to an approach based on confidence intervals, guarantees a long-run convergence to the probability of default of a specified or given level of probability of default that we call the benchmark level. The approach we propose is flexible and takes into account the number of objects rated by the specific rating system. We regard this approach as an early warning system that could identify problems of calibration in a rating system, although we acknowledge that, given the fact that default correlation is not taken into account in the testing procedure, false alarms could be given for otherwise well-calibrated systems. Eventually, we are able to demonstrate that our proposed “traffic light approach” is compliant with the mapping procedure of external credit assessment institutions foreseen in the New Basel Accord (Basel Committee on Banking Supervision (2005a)).

The paper is organised as follows. In Section 2 the statistical framework forming the basis of a default generating process using binomial distribution is briefly reviewed. In Section 3 we derive the probability of default to be associated with a single “A” rating of a major rating agency. Section 4 discusses several approaches to checking whether the performance of a certain rating source is equivalent to a single “A” rating or its equivalent in terms of probability of default as determined in Section 3. This is done by means of their realised default frequencies. The section also contains our proposal for a simplified performance checking mechanism that is in line with the treatment of external credit assessment institutions in the New Basel Accord. Section 5 concludes the paper.

2. A STATISTICAL FRAMEWORK – MODELLING DEFAULTS USING A BINOMIAL DISTRIBUTION

The probability of default itself is unobservable because the default event is stochastic. The only quantity observable, and hence measurable, is the empirical default frequency. In search of the meaning of a single "A" rating in terms of a one year probability of default we will thus have to make use of a theoretical model that rests on certain assumptions about the rules governing default processes. As is common practice in credit risk modelling, we follow the "cohort method" (in contrast to the "duration approach", see Lando and Skoedeberg (2002)) throughout this paper and, furthermore, assume that defaults can be modelled using a binomial distribution (Nickel et al. (2000), Blochwitz and Hohl (2001), Tiomo (2003), Jafry and Schuermann (2004)). The quality of each model's results in terms of their empirical significance depends on the adequacy of the model's underlying assumptions. As such, this section briefly discusses the binomial distribution and analyses the impact of a violation of the assumptions underlying the binomial model.[5] It is argued that postulating a binomial model reflects a risk-averse point of view.[6]

We decided to follow the cohort method as the major rating agencies document the evolution of their rated entities over time on the basis of "static pools" (Standard & Poor's 2005, Moody's 2005). A static pool consists of N_Y rated entities with the same rating grade at the beginning of a year Y. In our case N_Y denotes the number of entities rated "A" at the beginning of year Y. The cohort method simply records the number of entities D_Y that have defaulted by the year end out of the initial N_Y rated entities (Nickel et al. (2000), Jafry and Schuermann (2004)).

It is assumed that D_Y, the number of defaults in the static pool of a particular year Y, is binomially distributed with a "success probability" p and a number of events N_Y (in notational form: $D_Y \approx B(N_Y; p)$). From this assumption it follows that each individual ("A"-rated) entity has the same (one year) probability of default "p" under the assumed binomial distribution. Moreover the default of one company has no influence on the (one year) defaulting of the other companies, i.e. the (one year) default events are independent. The number of defaults D_Y can take on any value from the set $\{0,1,2,...N_Y\}$. Each value of this set has a probability of occurrence determined by the probability density function of the binomial distribution which, under the assumptions of constant p and independent trials, can be shown to be equal to:

$$b(n_Y; N_Y; p) = P(D_Y = n_Y) = \binom{N_Y}{n_Y} p^{n_Y} (1-p)^{N_Y - n_Y} \tag{1}$$

The mean and the variance of the binomial distribution are given by

$$\begin{aligned} \mu_{D_Y} &= N_Y p \\ \sigma_{D_Y}{}^2 &= N_Y p(1-p) \end{aligned} \tag{2}$$

---

5 For a more detailed treatment of binomial distribution see e.g. Rohatgi (1984), and Moore and McCabe (1999).

6 An alternative distribution for default processes is the "Poisson distribution". This distribution has some benefits, such as the fact that it can be defined by only one parameter and that it belongs to the exponential family of distributions which easily allow uniformly most powerful (UMP) one and two-sided tests to be conducted in accordance with the Neyman-Pearson theorem (see the Fisher-Behrens problem). However, in this paper we have opted to follow the mainstream literature on validation of credit systems which rely on binomial distribution to define the default generating process.

As indicated above, a clear distinction has to be made between the "probability of default" (PD) (i.e. the parameter *p* in formula (1)) and the "default frequency". While the probability of default is the fixed (and unobservable) parameter *"p"* of the binomial distribution, the default frequency is the observed number of defaults in a binomial experiment, divided by the number of trials $\left(df_Y = \frac{n_Y}{N_Y} \right)$. This default frequency varies from one experiment to another, even when the parameters *p* and N_Y stay the same. It can take on values from the set $df_Y \in \left\{ \frac{0}{N_Y}, \frac{1}{N_Y}, \frac{2}{N_Y}, ..., 1 \right\}$. The value observed for one particular experiment is the observed default frequency for that experiment.
The mean and variance of the default frequency can be derived from formula (1):

$$\mu_{df_Y} = p$$
$$\sigma_{df_Y}{}^2 = \frac{p(1-p)}{N_Y} \qquad (2')$$

The probability density function can be derived from (1) by setting $f_Y = \frac{n_Y}{N_Y}$:

$$P\left(df_Y = f_Y\right) = \binom{N_Y}{f_Y N_Y} p^{f_Y N_Y} \left(1-p\right)^{(1-f_Y)N_Y} \qquad (3)$$

As $f_Y \in \left\{ \frac{0}{N_Y}, \frac{1}{N_Y}, \frac{2}{N_Y}, ..., 1 \right\}$ this distribution is discrete.

THE BINOMIAL DISTRIBUTION ASSUMPTIONS

It is of crucial importance to note that formula (1) is derived under two assumptions. First, the (one year) default probability should be the same for every "A"-rated company. Secondly, the "A"-rated companies should be independent with respect to the (one year) default event. This means that the default of one company in one year should not influence the default of another "A"-rated company within the same year.

The constant "p"

It may be questioned whether the assumption of a homogeneous default probability for all "A"-rated companies is fulfilled in practice (e.g. Blochwitz and Hohl (2001), Tiomo (2004), Hui et al. (2005), Basel Committee on Banking Supervision (2005b)). The distribution of defaults would then not be strictly binomial. Based on assumptions about the distribution of probability of defaults within rating grades, Blochwitz and Hohl (2001) and Tiomo (2004) use Monte Carlo simulations to study the impact of heterogeneous probabilities of default on confidence intervals.

The impact of a violation of the assumption of a uniform probability of default across all entities with the same rating may, however, also be modelled using "mixed binomial distribution", of which "Lexian distribution" is a special case. Lexian distribution considers a mixture of "binomial subsets",

each subset having its own PD. The PDs can be different between subsets. The mean and variance of the Lexian variable *x*, which is the number of defaults among *n* companies, are given by[7]

$$\begin{aligned} \mu_x &= n\bar{p}, \\ \sigma_x^2 &= n\bar{p}(1-\bar{p}) + n(n-1)\,\mathrm{var}(p) \end{aligned} \tag{4}$$

Where $\bar{p}$ is the average value of all the (distinct) PDs and var(*p*) is the variance of these PDs.

Consequently, if a mixed binomial variable is treated as a pure binomial variable, its mean, the average probability of default would still be correct, whereas the variance would be underestimated when the "binomial estimator" *np*(1-*p*) is used (see the additional term in (4)). The mean and the variance will be used to construct confidence intervals. An underestimated variance will lead to narrower confidence intervals for the (average) probability of default and thus to lower thresholds. Within the context of this paper, lower thresholds imply a risk-averse approach.

Independent trials

Several methods for modelling default correlation have been proposed in literature (e.g. Gordy (1998), Nagpal and Bahar (2001), Servigny and Renault (2002), Blochwitz, When and Hohl (2003, 2005) and Hamerle, Liebig and Rösch (2003)). They all point to the difficulties of measuring correlation.

Although default correlations are low for sufficiently high levels of credit quality such as a single "A" rating, they could be an important factor in performance testing for lower rating grades. Over the period 1981-2005 Standard and Poor's historical default experience (see Table 1) shows that, with the exception of 2001, not more than one company defaulted per year, a fact which indicates that correlation cannot be very high. Secondly, even if we assumed that two firms were highly correlated and one defaulted, the other one will most likely not default in the same year, but only after a certain lag! Given that the primary interest is in an annual testing framework, the possibility of intertemporal default patterns beyond the one year period is of no interest. Finally, from a risk management point of view, providing that the credit quality of the pool of obligors is high (e.g. single "A" rating or above), it could be seen as adequate to assume that there is no default correlation, because not accounting for correlation leads to confidence intervals that are more conservative.[8] Empirical evidence for these arguments is provided by Nickel et al. (2000). Later on we will relax this assumption when presenting for demonstration purposes a calibration test accounting for default correlation.

3. THE PROBABILITY OF DEFAULT ASSOCIATED WITH A SINGLE "A" RATING

In this section we derive a probability of default that could be assigned to a single "A" rating. We are interested in this rating level because this is the minimum level at which the Eurosystem has decided to accept financial assets as eligible collateral for its monetary policy operations. The derivation could easily be followed to compute the probability of default of other rating levels.

Table 1 shows data on defaults for issuers rated "A" by Standard & Poor's (the corresponding table for Moody's is given in Annex 1). The first column lists the year, the second shows the number of

7 See e.g. Johnson (1969)

8 As in the case of heterogeneous PDs, this is due to the increased variance when correlation is positive. Consider, for example, the case where the static pool can be divided into two subsets. Within each subset issuers are independent, but between subsets they are positively correlated. The number of defaults in the whole pool is then a sum of two (correlated) binomials. The total variance is given by $\frac{N}{2}p(1-p) + \frac{N}{2}p(1-p) + 2\sigma_{12}$, which is again higher than the "binomial variance".

"A" rated issuers for that year. The column "Default frequency" is the observed one-year default frequency among these issuers. The last column gives the average default frequency over the "available years" (e.g. the average over the period 1981-1984 was 0.04%).

Table 1: One-year default frequency within Standard and Poor's A-rated class

Year	Number of issuers	Default frequency (%)	Average(1981-YYYY) (%)
1981	494	0,00	0,00
1982	487	0,21	0,11
1983	466	0,00	0,07
1984	471	0,00	0,05
1985	510	0,00	0,04
1986	559	0,18	0,07
1987	514	0,00	0,06
1988	507	0,00	0,05
1989	561	0,00	0,04
1990	571	0,00	0,04
1991	583	0,00	0,04
1992	651	0,00	0,03
1993	719	0,00	0,03
1994	775	0,13	0,04
1995	933	0,00	0,03
1996	1027	0,00	0,03
1997	1106	0,00	0,03
1998	1116	0,00	0,03
1999	1131	0,09	0,03
2000	1118	0,09	0,04
2001	1145	0,17	0,04
2002	1176	0,09	0,04
2003	1180	0,00	0,04
2004	1209	0,00	0,04
Average 1981-2004		0,04	
Standard deviation 1981-2004		0,07	

Source: Standard & Poor's, "Annual Global Corporate Default Study: Corporate defaults poised to rise in 2005"

The average one-year default frequency over the whole observation period spanning from 1981 to 2004 was 0.04%, the standard deviation of the annual default rates was 0.07%.

The maximum likelihood estimator for the parameter *p* of a binomial distribution is the observed frequency of success. Table 1 thus gives for each year between 1981 and 2004 a maximum likelihood estimate for the probability of default of companies rated "A" by S&P, i.e. 24 (different) estimates.

One way to combine the information contained in these 24 estimates is to apply the central limit theorem to the arithmetic average of the default frequency over the period 1981-2004 which is 0.04% according to Table 1. As such, it is possible to construct confidence intervals for the true mean $\mu_{\bar{x}}$ of the population around this arithmetic average. The central limit theorem states that the arithmetic average $\bar{x}$ of *n* independent random variables x_i, each having mean μ_i and variance σ_i^2, is approximately normally distributed with parameters $\mu_{\bar{x}} = \frac{\sum_{i=1}^{n} \mu_i}{n}$ and $\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{n} \sigma_i^2}{n^2}$ (see e.g. DeGroot (1989), and Billingsley (1995)). Applying this theorem to S&P's default frequencies, random variables with $\mu_i = p$ and $\sigma_i^2 = p(1-p)/N_i$, yields the result that the arithmetic average of S&P's default frequencies is approximately normal with mean $\mu_{\bar{x}} = \frac{\sum_{i=1}^{n} p}{n} = p$ and variance

$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{n} \frac{p(1-p))}{N_i}}{n^2}$. If the probability of default "*p*" is not constant over the years then a confidence interval for the average probability of default is obtained. In that case the estimated benchmark would be based on the average probability of default. After estimating *p* and $\sigma_{\bar{x}}^2$ from S&P data ($\hat{p} = 0.04\%$, $\hat{\sigma}_{\bar{x}} = 0.0155\%$ for "A" and $\hat{p} = 0.27\%$, $\hat{\sigma}_{\bar{x}} = 0.0496\%$ for "BBB"), confidence intervals for the mean, i.e. the default probability *p*, can be constructed. These confidence intervals are given in Table 2 for S&P's rating grades "A" and "BBB". Similar estimates can be derived for Moody's data using the same approach. The confidence intervals for a single "A" rating from Moody's have lower limits than those shown for S&P in Table 2. This is due to the lower mean realised default frequency recorded in Moody's ratings. However, in the next paragraph it will be shown that Moody's performance does not differ significantly from that of S&P for the single "A" rating grade.

Table 2: Confidence intervals for the $\mu_{\bar{x}}$ of S&P's "A" compared to "BBB"

(percentages)

Confidence level	Lower	Upper
S&P A		
95,0	0,01	0,07
99,0	0,00	0,08
99,5	0,00	0,09
99,9	0,00	0,10
S&P BBB		
95.0	0,17	0,38
99,0	0,13	0,41
99,5	0,12	0,43
99,9	0,09	0,46

A similar result is obtained when the observations for the 24 years are "pooled". Pooling is based on the fact that the sum of independent binomial variables with the same *p* is again binomial with parameters $\sum D_Y \approx B\left(\sum N_Y ; p\right)$ (see e.g. DeGroot (1989)). Applying this theorem to the 24 years of S&P data (and assuming independence) it can be seen that eight defaults are observed among 19,009 issuers (i.e. the sum of all issuers rated single "A" over the 1981-2004 period). This yields an estimate for *p* of 0.04% and a binomial variance of 0.015%, similar to the estimates based on the central limit theorem.

The necessary condition for the application of the central limit theorem or for pooling is the independence of the annual binomial variables. This is hard to verify. Nevertheless, several arguments in favour the above method can be brought forward. First, a quick analysis of the data in Table 1 shows that there are no visible signs of dependence among the default frequencies. Second, and probably the most convincing argument, the data in Table 1 confirms the findings for the confidence intervals that are found in Table 2. Indeed, the last column in Table 1 shows the average over 2, 3, ..., 24 years. As can be seen, with a few exceptions, these averages lie within the confidence intervals (see Table 2). For the exceptions it can be argued (1) that not all values have to be within the limits of the confidence intervals (in fact, for a 99% confidence interval one exception is allowed every 100 years, and for a 95% interval it is even possible to exceed the limits every 20 years) and (2) that we did not always compute 24-year averages although the central limit theorem was applied to a 24-year average. When random samples of size 23 are drawn from these 24 years of data, the arithmetic average seems to be within the limits given in Table 2. The third argument in support of our findings is a theoretical one. In fact, a violation of the independence assumption would change nothing in the findings about the mean $\mu_{\bar{x}}$. However, the variance would

no longer be correct as the covariances should be taken into account. Furthermore, dependence among the variables would no longer guarantee a normal distribution. The sum of dependent and (right) skewed distributions would no longer be symmetric (like the normal distribution) but also skewed (to the right). Assuming positive covariances would yield wider confidence intervals. Furthermore, as the resulting distribution will be skewed to the right, and as values lower than zero would not be possible, using the normal distribution as an approximation would lead to smaller confidence intervals. As such, a violation of the independence assumption implies a risk-averse result.

An additional argument can be brought forward which supports our findings: First, in the definition of the "A" grade we are actually also interested in the minimum credit quality that "A-grade" stands for. We want to know the highest value the probability of default can take to be still accepted as equivalent to "A". Therefore we could also apply the central limit theorem to the data for Standard & Poor's BBB. Table 2 shows that in that case the PD of a BBB rating is probably higher than 0.1%.

We can thus conclude that there is strong evidence to suggest that the **probability of default** for the binomial process that models the observed default frequencies of Standard & Poor's "A" rating grade is between 0.00% and 0.1% (see Table 2). The average point estimate is 0.04%. For reasons mentioned above, these limits are conservative, justifying the use of values above 0.04% (but not higher than 0.1%). An additional argument for the use of a somewhat higher value for the average point estimate than 0.04% is the fact that the average observed default frequency for the last five years of Table 1 equals 0.07%.

TESTING FOR EQUALITY IN DEFAULT FREQUENCIES OF TWO RATING SOURCES AT THE SAME RATING LEVEL

The PD of a rating source is unobservable. As a consequence, a performance checking mechanism cannot be based on the PD alone. In this section it is shown that the central limit theorem could also be used to design a mechanism that is based on an average observed default frequency.[9]

Earlier on, using the central limit theorem, we found that the 24-year average of S&P's default frequencies is normally distributed:

$$\bar{x}^{S\&P} \approx N(\mu_{\bar{x}^{S\&P}}; \sigma_{\bar{x}^{S\&P}}) \tag{5}$$

with $\mu_{\bar{x}^{S\&P}}$ and $\sigma_{\bar{x}^{S\&P}}$ estimated at 0.04% and 0.0155% respectively.
In a similar way, the average default frequency of any rating source is normally distributed:

$$\bar{x}^{rs} \approx N(\mu_{\bar{x}^{rs}}; \sigma_{\bar{x}^{rs}}) \tag{6}$$

The formulae (5) and (6) can be used to test whether the average default frequency of the rating source is at least as good as the average of the benchmark by testing the statistical hypothesis

$$H_0 : \mu_{\bar{x}^{rs}} < \mu_{\bar{x}^{SaP}} \text{ against } H_1 : \mu_{\bar{x}^{rs}} \geq \mu_{\bar{x}^{SaP}} \tag{7}$$

Although seemingly simple, such a performance checking mechanism has several disadvantages. First, assuming, for example, 24 years of data for the rating source, the null hypothesis cannot be rejected if the annual default frequency is 0.00% on 23 occasions and 0.96% once ($\bar{x}^{rs} = \frac{23 \times 0.00\% + 1 \times 0.96\%}{24} = 0.04\%$, *p*-value is 50%). In other words, extreme values for the

9 This is only possible when historical data are available, i.e. when a *n*-year average can be computed.

observed default frequencies are allowed (0.96%). Second, the performance rule is independent of the static pool size. A default frequency of 0.96% on a sample size of 10,000 represents 96 defaults, while it is only 2 defaults for a sample of 200. Third, requiring 24 years of data to compute a 24-year average is impractical. Other periods could be used (e.g. a 10-year average), but that is still impractical as 10 years of data must be available before the rating source can be backtested. Taking into account these drawbacks, two alternative performance checking mechanisms will be presented in Section 0.

This rule can, however, be used to test whether the average default frequencies of S&P and Moody's are significantly different. Under the null hypothesis

$$H_0 : \mu_{\bar{x}^{S\&P}} = \mu_{\bar{x}^{Moody's}} \tag{8}$$

the difference of the observed averages is normally distributed, i.e. (assuming independence)

$$\bar{x}^{S\&P} - \bar{x}^{Moody's} \approx N(0; \sqrt{\sigma_{\bar{x}^{S\&P}}^{\ 2} + \sigma_{\bar{x}^{Moody's}}^{\ 2}}) \tag{9}$$

Using an estimate of the variance, the variable $\frac{\bar{x}^{S\&P} - \bar{x}^{Moody's}}{\sqrt{s_{\bar{x}^{S\&P}}^{\ 2} + s_{\bar{x}^{Moody's}}^{\ 2}}}$ has a *t*-distribution with 46 degrees of freedom and can be used to check the hypothesis (8) against the alternative hypothesis $H_1 : \mu_{\bar{x}^{S\&P}} \neq \mu_{\bar{x}^{Moody's}}$.

Using the figures from S&P and Moody's ($\hat{p} = 0.04\%, \hat{\sigma}_{\bar{x}} = 0.0155\%$ for S&P's "A" and $\hat{p} = 0.02\%$, $\hat{\sigma}_{\bar{x}} = 0.0120\%$ for Moody's "A"), a value of 0.81 is observed for this *t*-variable. This *t*-statistic has an implied *p*-value (2-sided) of 42% so the hypothesis of equal PDs for Moody's & S&P's "A" grade cannot be rejected. In formula (9) S&P and Moody's "A" class were considered independent. Positive correlation would thus imply an even lower *t*-value.

PERFORMANCE CHECKING: THE DERIVATION OF A BENCHMARK FOR BACKTESTING

To allow performance checking, the assignment of PDs to rating grades alone is not enough. In fact, as can be seen from S&P data in Table 1, the observed annual default frequencies often exceed 0.1%. This is because the PD and the (observed) default frequencies are different concepts. A performance checking mechanism should, however, be based on "observable" quantities, i.e. on the observed default frequencies of the rating source.

In order to construct such a mechanism it is assumed that the annually observed default rates of the benchmark may be modelled using a binomial distribution. The mean of this distribution, the probability of default of the benchmark, is estimated at $\hat{p} \in [0.0\%, 0.1\%]$ (with an average of 0.04%). The other binomial parameter is the number of trials *N*. To define the benchmark *N* is taken to be the average size of S&P's static pool or *N* = 792 (see Table 1). This choice may appear somewhat arbitrary because the average size over the period 2000-2004 is higher (i.e. 1,166), but so is the average observed default frequency over that period (0.07%). If the binomial parameters were based on this period, then the mean and the variance of this binomial benchmark would be higher, and so confidence limits would also be higher.

In Section 4.1 below two alternatives for the benchmark will be used:

1. A fixed upper limit of 0.1% for the benchmark probability of default.
2. A stochastic benchmark, i.e. a Binomial distribution with parameters *p* equal to 0.1% and *N* equal to 792.

4. CHECKING THE SIGNIFICANCE OF DEVIATIONS OF THE REALISED DEFAULT RATE FROM THE FORECAST PROBABILITY OF DEFAULT

As realised default rates are subject to statistical fluctuations it is necessary to develop mechanisms to show how well the rating source estimates the probability of default. This is generally done using statistical tests to check the significance of the deviation of the realised default rate from the forecast probability of default. The statistical tests would normally check the null hypothesis that "the forecast probability of default in a rating grade is correct" against the alternative hypothesis that "the forecast default probability is incorrect".

As shown in Table 1, the stochastic nature of the default process allows for observed default frequencies that are far above the probability of default. The goal of this section is to find upper limits for the observed default frequency that are still consistent with a PD of 0.1%.

We will first briefly describe some statistical tests that can be used for this purpose. The first one is to test a realised default frequency for a rating source against a fixed upper limit for the PD, this is the "Wald test" for single proportions. The second test will assess the significance of the difference between two proportions or, in other words, two default rates that come from two different rating sources. We will then proceed to a test that considers the significance of deviations between forecast probabilities of default and realised default rates of several rating grades, the "Hosmer-Lemeshow test". In some instances, the probability of default associated with a rating grade is considered not to be constant for all obligors in that rating grade. The "Spiegelhalter test" will assess the significance of deviations when the probability of default is assumed to vary for different obligors within the rating grade. Both the Hosmer-Lemeshow and the derived Spiegelhalter test can be seen as extensions of the Wald test. Finally, we introduce a test that accounts for correlation and show how the critical values for assessing significance in deviations can be dramatically altered in the presence of default correlation.

THE WALD TEST FOR SINGLE PROPORTIONS

For hypothesis testing purposes, the binomial density function is often approximated by a normal density function with parameters given by (2) or (2') in Section 2 (see e.g. Cantor and Falkenstein (2001), Nickel et al. (2000)).

$$df_Y \approx N\left(p; \sqrt{\frac{p(1-p)}{N_Y}} \right) \tag{10}$$

When testing the null hypothesis H_0: "the realised default is consistent with a specified probability of default value lower than p_0 or benchmark" against H_1: "the realised default is higher than p_0", a Z-statistic

$$Z = \frac{df - p_0}{\sqrt{\frac{df\,(1-df)}{N_Y}}} \tag{11}$$

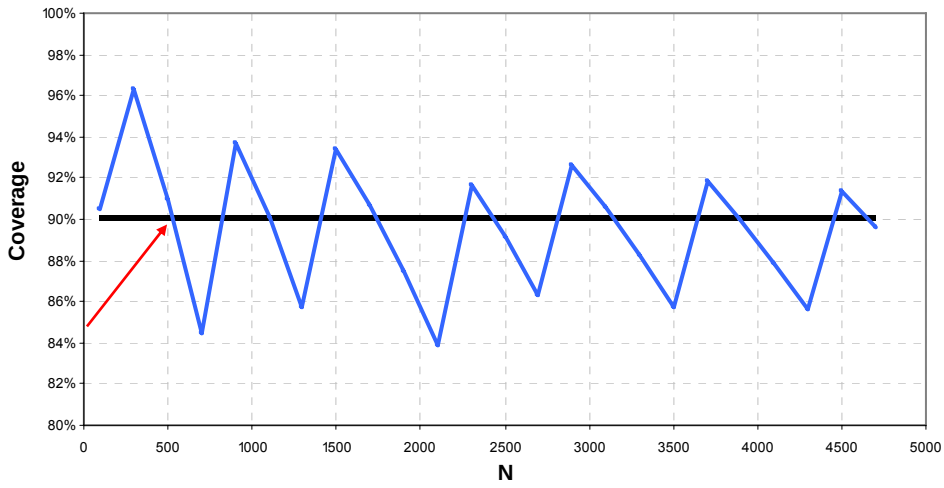
can be used, which is compared to the quantiles of the standard normal distribution.

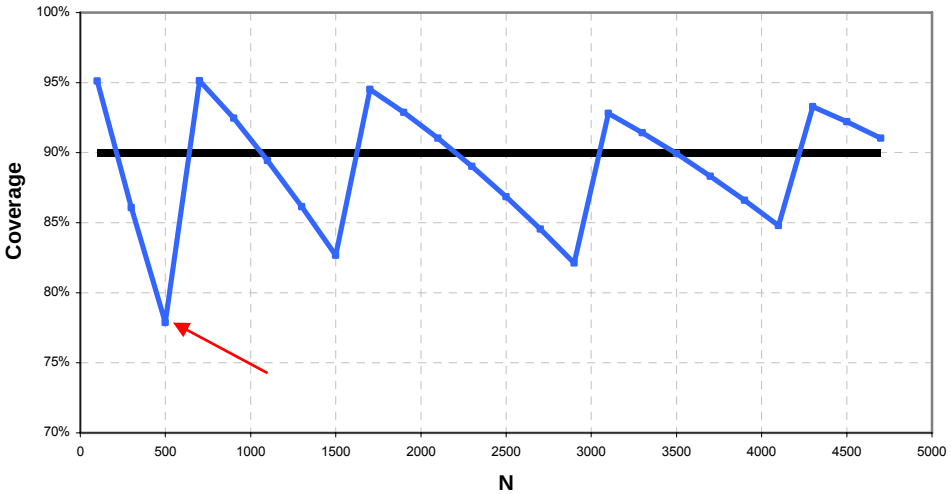
The quality of the approximation depends on the values of the parameters $N_{Y,}$, the number of rated entities with the same rating grade at the beginning of a year Y, and p, the forecast probability of default (see e.g. Brown et al. 2001). A higher N_Y results in better approximations. For the purpose of this paper, N_Y is considered to be sufficiently high. The low PD values for "A" rated companies

(lower than 0.1%) might be problematic since the quality of the approximation degrades when *p* is far away from 50%. In fact, the two parameters interact, the higher N_Y is, the further away from 50% *p* can be. Low values of *p* imply a highly skewed (to the right) binomial distribution, and since the normal distribution is symmetric the approximation becomes poor. The literature on the subject is extensive (for an overview see Vollset (1993), Agresti and Coull (1998), Newcombe (1998), Agresti and Caffo (2000), Brown et al. (2001), Reiczigel (2004), and Cai (2005)). Without going into more details, the problem is briefly explained in a graphical way.

In Figure 1 the performance of the Wald interval is shown for several values of *N,* once for *p* = 0.05% and once for *p* = 0.10%. Formula (10) can then be used to compute the upper limit (df_U) of the 90% one-sided confidence interval. As the normal distribution is only an approximation for the binomial distribution, the cumulative binomial distribution for this upper limit will seldom be exactly equal to 90%, i.e. $B(df_U \times N_Y ; N_Y ; p) = P(D_Y \leq df_U \times N_Y) \neq 90\%$

Figure 1. The performance of the Wald interval for different values of N, and for p =0.1% (left) and p = 0.05% (right)





The zigzag line shows, for different values of N, the values for the cumulative binomial distribution in the upper limit of the Wald interval. For p = 0.1% and N = 500 this value seems to be close to 90%. However for p = 0.05% and N = 500 the coverage is far below 90%. This shows that for p=0.05% the 90% Wald confidence interval is in fact not a 90% but only a 78% confidence interval, meaning that the Wald confidence interval is too small and that a test based on this approximation (for p = 0.05% and N = 500) is (too) conservative. The error is due to the approximation of the binomial distribution (discrete and asymmetric) by a normal (continuous and symmetric) one. Thus it is to be noted that, the higher the value of N, the better the approximation becomes, and that in most cases the test is conservative.[10]

Our final traffic light approach will be based on a statistical test for differences of proportions. This test is also based on an approximation of the binomial distribution by a normal one. In this case, however, the approximation performs better as is argued in the next section.

The Wald test for differences of proportions

To check the significance of deviations between the realised default rates of two different rating systems, as opposed to just testing the significance of deviations of one single default rate against a specified value p_0, a Z-statistic also can also be used.
If we define the realised default rate and the number of rated entities of one rating system (1) as df^1 and N_1 respectively and of another rating system (2) as df^2 and N_2 respectively, we can test the null hypothesis H_0: $df^1 = df^2$ (or $df^1 - df^2 = 0$) against H_1: $df^1 \neq df^2$. To derive such a test of difference in default rates we need to pool the default rates of the two rating systems and compute a pooled standard deviation of the difference in default rates in the following way,

$$df^{pooled} = \frac{N_1 df^1 + N_2 df^2}{N_1 + N_2} \tag{12}$$

$$\sigma_{df^1 - df^2} = \sqrt{df^{pooled}(1 - df^{pooled})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \tag{13}$$

Assuming that the two default rates are independent, the corresponding Z-statistic is given by

$$Z = \frac{df^1 - df^2}{\sqrt{df^{pooled}(1 - df^{pooled})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \tag{14}$$

The value for the Z-statistic may be compared with the percentiles of a standard normal distribution.

Since the binomial distributions considered have success probabilities that are low (< 0.1%) they are all highly skewed to the right. Taking the difference of two right skewed binomial distributions, however, compensates for the asymmetry problem to a large extent.

[10] The authors are well aware of the fact that the Poisson distribution (discrete and skewed, just like the binomial) is a better approximation than the normal distribution. However the normal approximation is more convenient for differences of proportions (because the difference of independent normal variables is again a normal variable, a property that is not valid for Poisson distributed variables).

Figure 2: Performance of the Wald interval for differences of proportions

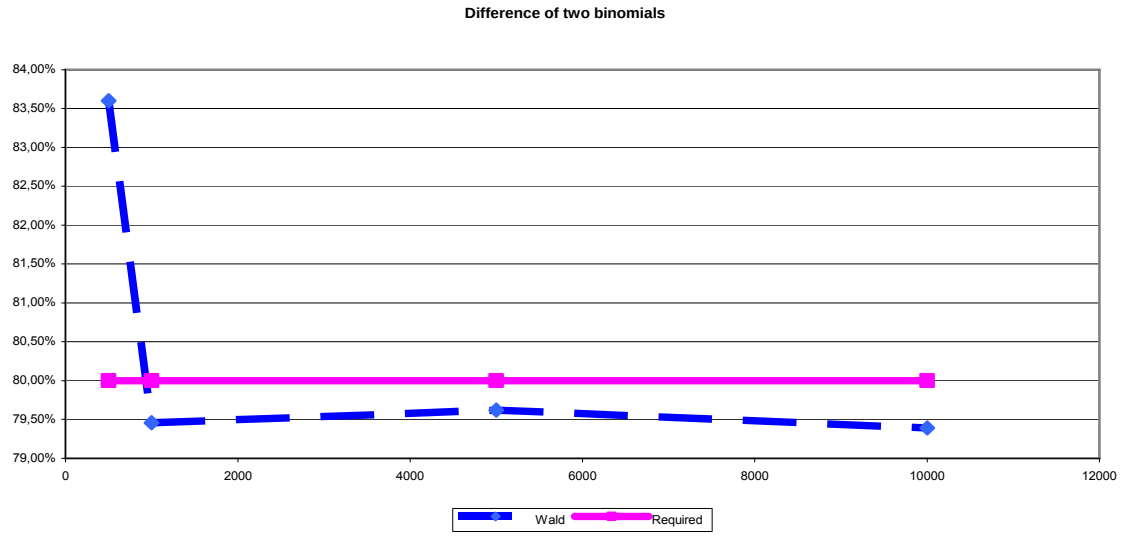


Figure 2 illustrates the performance of the Wald approximation applied to differences of proportions. For several binomial distributions (i.e. (*N*, *p*) = (500, 0.20%), (1000, 0.20%), (5000, 0.18%) and (10,000, 0.16%)) the 80% confidence threshold for their difference with respect to the binomial distribution with parameters (0.07%, 792) is computed using the Wald interval. Then the exact confidence level of this "Wald threshold" is computed.[11]

The figure shows that for the difference between the binomials with parameters (792, 0.07%) and (500, 0.20%) the 80% confidence threshold resulting from the Wald approximation is in fact an 83.60% confidence interval. For the difference between the binomials with parameters (792, 0.07%) and (1000, 0.20%) the 80% confidence threshold resulting from the Wald approximation is in fact a 79.50% confidence interval, and so on.

It can be seen that the Wald approximations for differences in proportions perform better than the approximations in Figure 1 for single proportions (i.e. the coverage is close to the required 80%). From this it may be concluded that hypothesis tests for differences of proportions, using the normal approximation, work well, as is demonstrated by Figure 2. Thus they seem to be more suitable for our purposes in this context.

THE HOSMER-LEMESHOW TEST *(1980, 2000)*

The binomial test (or its above mentioned normal/Wald test extensions) is mainly suited to testing a single rating grade, but not several or all rating grades simultaneously. The Hosmer-Lemeshow test is in essence a joint test for several rating grades.

Assume that there are *k* rating grades with probabilities of default p_1, …., p_k. Let n_i be the number of obligors with a rating grade *i* and d_i be the number of defaulted obligors in grade *i*. The statistic proposed by Hosmer-Lemeshow (HSLS) is the sum of the squared differences of forecast and observed numbers of default, weighted by the inverses of the theoretical variances of the number of defaults.

$$HSLS = \sum_{i=1}^{k} \frac{(n_i p_i - d_i)^2}{n_i p_i (1 - p_i)} \qquad (15)$$

[11] The values that were chosen for the parameters will become clear in Section 0

The Hosmer-Lemeshow statistic is χ^2 distributed with *k* degrees of freedom under the hypothesis that all the probability of default forecasts match the true PDs and that the usual assumptions regarding the adequacy of the normal distribution (large sample size and independence) are justifiable.[12] It can be shown that, in the extreme case, when there is just one rating grade, the HSLS statistic and the (squared) binomial test statistic are identical.

THE SPIEGELHALTER TEST *(1986)*

Whereas the Hosmer-Lemeshow test, like the binomial test, requires all obligors assigned to a rating grade to have the same probability of default, the Spiegelhalter test allows for variation in PDs within the same rating grade. The test also assumes independence of default events. The starting point is the mean square error (MSE) also known as the Brier score (see Brier 1950)

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - p_i)^2 \tag{16}$$

where there are 1, …, *N* obligors with individual probability of default estimates p_i. y_i denotes the default indicator, *y*=1 (default) or *y*=0 (no default).

The MSE statistic is small if the forecast PD assigned to defaults is high and the forecast PD assigned to non-defaults is low. In general, a low MSE indicates a good rating system.
The null hypothesis for the test is that "all probability of default forecasts, p_i, match exactly the true (but unknown) probability of default" for all *i*. Then under the null hypothesis, the MSE has an expected value of

$$E\left[MSE\right] = \frac{1}{N}\sum_{i=1}^{N} p_i(1-p_i) \tag{17}$$

and

$$\mathrm{var}\left[MSE\right] = \frac{1}{N^2}\sum_{i=1}^{N} p_i(1-p_i)(1-2p_i)^2 \tag{18}$$

Under the assumption of independence and using the central limit theorem, it can be shown that under the null hypothesis the test statistic

$$Z = \frac{MSE - E\left[MSE\right]}{\sqrt{\mathrm{var}\left[MSE\right]}} \tag{19}$$

follows approximately a standard normal distribution which allows a standard test decision (see Rauhmeier and Scheule (2005) for practical examples).

CHECKING DEVIATION SIGNIFICANCE IN THE PRESENCE OF DEFAULT CORRELATION

Whereas all the tests presented above assume independence of defaults, it is also important to discuss tests that take into account default correlation. The existence of default correlation within a pool of obligors has the effect of reinforcing the fluctuations in default rate of that pool. The

[12] If we use the HSLS statistic as a measure of goodness of fit when building the rating model using "in-sample" data then the degrees of freedom of the χ^2 distribution are *k*-2. In the context of this paper, we use the HSLS statistic as backtesting tool on "out of sample" data which has not been used in the estimation of the model.

tolerance thresholds for the deviation of realised default rates from estimated values of default may be substantially larger when default correlation is taken into account than when defaults are considered independent. From a conservative risk management point of view, assuming independence of defaults is acceptable, as this approach will overestimate the significance of deviations in the realised default rate from the forecast rate. However, even in that case, it is necessary to determine at least the approximate extent to which default correlation influences probability of default estimates and their associated default realisations.

Most of the relevant literature models correlations on the basis of the dependence of default events on a common systematic random factor (cf. Tasche (2003) and Rauhmeier (2006)). This follows from the Basel II approach underlying risk weight functions which utilise a one factor model.[13] If D_N is the realised number of defaults in the specified period of time for a 1 to N obligor sample:

$$D_N = \sum_{i=1}^{N} 1\left[\sqrt{\rho}X + \sqrt{1-\rho}\varepsilon_i \le \theta\right] \tag{20}$$

The default of an obligor i is modelled using a latent variable $AV_i = \sqrt{\rho}X + \sqrt{1-\rho}\varepsilon_i$ representing the asset value of the obligor. The (random) factor X is the same for all the obligors and represents systemic risk. The (random) factor ε_i depends on the obligor and is called the idiosyncratic risk. The common factor X implies the existence of (asset) correlation among the N obligors.
If the asset value AV_i falls below a particular value θ (i.e. the default threshold) then the obligor defaults. The default threshold should be chosen in such a way that $E\left[D_N\right] = Np$. This is the case if $\theta = \Phi^{-1}(p)$ where Φ^{-1} denotes the inverse of the cumulative standard normal distribution function and p the probability of default (see e.g. Tasche (2003)). The indicator function 1[] has the value 1 if its argument is true (i.e. the asset value is below θ and the obligor defaults) and the value 0 otherwise (i.e. no default). The variables X and ε_i are normally distributed random variables with a mean of zero and a standard deviation of one (and as a consequence AV_i is also standard normal). It is further assumed that idiosyncratic risk is independent for two different borrowers and that idiosyncratic and systematic risk are independent. In this way, the variable X introduces the dependency between two borrowers through the factor ρ, which is the asset correlation (i.e. the correlation between the asset values of two borrowers). Asset correlation can be transformed into default correlations as shown, for example, in Basel Committee on Banking Supervision (2005b).

Tasche 2003 shows that on a confidence level α we can reject the assumption that the actual default rate is less than or equal to the estimated probability of default whenever the number of defaults D is greater than or equal to the critical value given by

$$D_\alpha = q + \frac{2q-1}{2N} - \frac{q(1-q)}{\varphi\left(\dfrac{\sqrt{\rho}\phi^{-1}(1-\alpha) - \phi^{-1}(PD)}{\sqrt{(1-\rho)}}\right)} \times \frac{(1-2\rho)\phi^{-1}(1-\alpha) - \sqrt{\rho}\phi^{-1}(PD)}{2N\sqrt{\rho(1-\rho)}} \tag{21}$$

where

$$q = \phi^{-1}\left(\frac{\sqrt{\rho}\phi^{-1}(\alpha) + \phi^{-1}(PD)}{\sqrt{1-\rho}}\right), \varphi(x) = \frac{d\Phi(x)}{dx}, \tag{22}$$

[13] See Finger (2001) for an exposition.

and Φ^{-1} denotes the inverse of the cumulative standard normal distribution function and ρ the asset correlation. However, the above test, which includes dependencies and a granularity adjustment, as in the Basel II framework, shows a strong sensitivity to the level of correlation.[14]

It is interesting to see how the binomial test and the correlation test as specified above behave under different assumptions. As can be seen in Tables 3 and 4, the critical number of defaults that can be allowed before we could reject the null hypothesis that the estimated probability of default is in line with the realised number of defaults, goes up as we increase the level of asset correlation among obligors for every level of sample size from 0.05 to 0.15.[15] The binomial test produces consistently lower critical values of default than the correlation test for all sample sizes. However, the test taking into account correlation suffers from dramatic changes in the critical values, especially for larger sample sizes (i.e. over 1,000).

Table 3. 95% critical values for a benchmark probability of default of 0.10% under different calibration tests

N	Binomial	Correlation = 0.05	Correlation = 0.15
100	1	2	2
500	2	3	3
1000	3	4	5
5000	9	15	21

Table 4. 99.9% critical values for a benchmark probability of default of 0.10% under different calibration tests

N	Binomial	Correlation = 0.05	Correlation = 0.15
100	2	4	4
500	2	6	12
1000	5	10	22
5000	13	37	102

As can be inferred from the above tables, the derivation of critical values of default, taking into account default correlation, is not a straight forward exercise. First, we need to have a good estimate of asset correlation. In practice, this number could vary depending on the portfolio considered. A well-diversified portfolio of retail loans across an extensive region will present very different correlation characteristics than that of a sector concentrated portfolio of corporate names. In practice, default correlations could be seen in the range of 0-5%.[16] Second, the validation analyst should take into account that there should be a consistency between the modelling of correlation for risk measurement in the credit assessment system that is going to be validated and the validation test to derive consistent confidence intervals for such credit system. This consistency

[14] Tasche (2003) also discusses an alternative test to determine default-critical values assuming a Beta distribution, with the parameters of such a distribution being estimated by a method of matching the mean and variance of the distribution. This approach will generally lead to results that are less reliable than the test based on the granularity adjustment.

[15] The $\rho = 0.05$ may be justified by applying the non-parametric approach proposed by Gordy (2002) to data on the historical default experiences of all the rating grades of Standard & Poor's, which yields an asset correlation of ~5%. Furthermore, Tasche (2003) also points out that "$\rho = 0.05$ *appears to be appropriate for Germany*". 24% is the highest asset correlation according to Basle II (see Basel Committee on Banking Supervision (2005a)).

[16] Huschens and Stahl (2005) show evidence that, for a well diversified German retail portfolio, asset correlations are in the range between 0% and 5%, which implies even smaller default correlations.

is in practice difficult to achieve because the correlation dynamics in the validation test may not be in line with those assumed in the rating system.
The binomial test, although conservative, is seen as a good realistic proxy for deriving critical values. It is considered a good early warning tool, free of all the estimation problems seen in tests that incorporate correlation estimates. Therefore, in what remains of this paper we will focus on the binomial distribution paradigm and its extension in the normal distribution as the general statistical framework to derive a simple mechanism for performance checking based on backtesting.

4.1. TWO POSSIBLE BACKTESTING STRATEGIES

In what follows we will concentrate on elaborating two backtesting strategies that focus on a rating level of a single "A" as defined by the main international rating agencies. This is the credit quality level set by the Eurosystem for determining eligible collateral for its monetary policy operations. The single "A" rating is thus considered the "benchmark". The previous section defined this benchmark in terms of a probability of default. As the true probability of default is unobservable, this section presents two alternative backtesting strategies based on the (observable) default frequency:

1. A backtesting strategy that uses a fixed, absolute upper limit for the probability of default as a benchmark.
2. A backtesting strategy that uses a stochastic benchmark. This assumes that the benchmark is not constant as in the first strategy

These alternatives will be summarised in a simplified rule which will result in a traffic light approach for backtesting, much in the same vein as in Tasche (2003), Blochwitz and Hohl (2001) or Tiomo (2004).

A BACKTESTING STRATEGY RELYING ON A FIXED BENCHMARK

Using the central limit theorem we found in Section 3 that the probability of default of the benchmark (p^{bm}) is at most 0.1%. A rating source is thus in line with the benchmark if its default probability for the single "A" rating is at most 0.1%.
Assuming that the rating source's default events are distributed in accordance with a binomial distribution with parameters PD^{rs} and N_Y^{rs}, the backtesting should check whether

$$PD^{rs} \leq 0.1\% \tag{23}$$

Since PD^{rs} is an unobservable variable, (23) can not be used for validation purposes. A quantity that can be observed is the number of defaults in a rating source's static pool within one particular year, i.e. df_Y^{rs}.

The performance checking mechanism should thus check whether observing a value df_Y^{rs} for a random variable which is (approximately) normally distributed $df_Y^{rs} \approx N\left(PD^{rs}; \sqrt{\frac{PD^{rs}(1-PD^{rs})}{N_Y^{rs}}}\right)$ is consistent with (23).

This can be done using a statistical hypothesis test. The null hypothesis that $H_0 : p^{rs} \leq 0.1\%$ must be tested against the alternative hypothesis $H_1 : p^{rs} > 0.1\%$.

Assuming that the null hypothesis of this statistical test, H_0 , is true, the probability of observing the value df_Y^{rs} can be computed. This is the *p*-value of the hypothesis test or the probability of obtaining a value of df_Y^{rs} or higher, assuming that H_0 is true. This *p*-value is given by

$$1-\Phi\left(\frac{df'^{rs}-0.1\%}{\sqrt{\frac{0.1\%(1-0.1\%)}{N_y^{rs}}}}\right) \tag{24}$$

where Φ is the cumulative probability function for the standard normal distribution. Table 5 gives an example for an eligible set of $N_y^{rs}=10{,}000$ companies.

Table 5: Test of credit quality assessment source against the limit of 0.1% for a sample size of 10,000. N denotes the number of defaults.

(percentages)

N	*df'(rs)*	*p-value*	*Probability of "N" if H0 is true*
0	0,00		
1	0,01	99,78	0,35
2	0,02	99,43	0,77
3	0,03	98,66	1,54
4	0,04	97,12	2,80
5	0,05	94,32	4,60
6	0,06	89,72	6,84
7	0,07	82,87	9,22
8	0,08	73,66	11,24
9	0,09	62,41	12,41
10	0,10	50,00	12,41
11	0,11	37,59	11,24
12	0,12	26,34	9,22
13	0,13	17,13	6,84
14	0,14	10,28	4,60
15	0,15	5,68	2,80
16	0,16	2,88	1,54
17	0,17	1,34	0,77
18	0,18	0,57	0,35
19	0,19	0,22	0,14
20	0,20	0,08	0,05
21	0,21	0,03	0,02
22	0,22	0,01	0,01
23	0,23	0,00	0,00
24	0,24	0,00	0,00
25	0,25	0,00	0,00

The first column of the table gives different possibilities for the number of defaults observed in year "*Y*". The observed default frequency is derived by dividing the number of defaults by the sample size. This is shown in the second column of the table. The third column shows the *p*-values computed using formula (24). So the *p*-value for observing at least 15 defaults out of 10,000, assuming that H_0 is true, equals 5.68%. In the same way it follows from the table that if H_0 is true, then the probability of observing at least 18 defaults in 10,000 is 0.57%, or "almost impossible". Or, to put it another way, if we observe 18 defaults or more then it is almost impossible for H_0 to be true.

The last column "probability" computes the theoretical probability for observing a particular number of defaults if H_0 is true. It is the difference between two successive *p*-values. For example, if H_0 is true, then the probability of observing at least one default out of 10,000 equals 99.78%, and the probability of observing at least two defaults is 99.43%. As a consequence, if H_0 is true, the probability of having exactly one default is 0.35%. The column "probability" can thus be used as an exact behavioural rule, i.e. if H_0 is true then one can have

- exactly one default in 10,000 every 0.35 years out of 100 years
- exactly two defaults in 10,000 every 0.77 years out of 100 years
- exactly three defaults in 10,000 every 1.54 years out of 100 years
- ....

Averaging this rule over a 100 year period shows that in the long run the average default frequency will converge to 0.1%. However, such a rule is, of course, too complex to be practical. It is simplified below.
Table 5 can be used as backtesting for a sample of 10,000 obligor names with an ex ante probability of default of 0.10% after fixing a confidence level (i.e. a minimum *p*-value, e.g. 1%): if the size of the static pool is 10,000 then the rating source is in line with the benchmark only if at most 17 defaults are observed (confidence level of 1%) i.e. $df_Y^{rs} \leq 0.17\%$.

This technique has the disadvantage of first having to decide on a confidence level. Moreover, fixing only one limit (0.17% in the case above) does not guarantee a convergence over time to an average of 0.1% or below.

A *p*-value, being a probability, can be interpreted in terms of "number of occurrences". From Table 5 we infer that, if the null hypothesis is true, the observed default frequency must be lower than 0.12% in 80% of cases. In other words, a value above 0.12% should be observable only once every 5 years (i.e. if the realised default frequency should be lower than 0.12% in 80 out of 100 years, then the realised default frequency could be higher than 0.12% in 20 out of 100 years, or once every 5 years), otherwise the rating source is not in line with the benchmark.

This gives a second performance checking rule: a rating source with a static pool of size 10,000 is in line with the benchmark if at most once every five years a default frequency above 0.12% is observed. A default frequency above 0.17% should "never" be observed.

The intervals for other sizes of the static pool are shown in Table 6. The lower value of the "Once in 5y" interval is derived from the 80% confidence limit, the absolute upper limit "Never" is derived from a 99% confidence interval.

Table 6: Backtesting strategy based on a fixed benchmark for different static pool sizes *(percentages)*

	All time	Once in 5y	Never	Average DF
500	*0.00-0.00*	*0.20-0.40*	*>0.40*	*0,06*
1000	*0.00-0.10*	*0.20-0.40*	*>0.40*	*0,10*
2000	*0.00-0.10*	*0.15-0.25*	*>0.25*	*0,08*
3000	*0.00-0.10*	*0.13-0.23*	*>0.23*	*0,08*
4000	*0.00-0.13*	*0.15-0.25*	*>0.25*	*0,09*
5000	*0.00-0.12*	*0.14-0.20*	*>0.20*	*0,08*
6000	*0.00-0.12*	*0.13-0.20*	*>0.20*	*0,08*
7000	*0.00-0.11*	*0.13-0.19*	*>0.19*	*0,08*
8000	*0.00-0.11*	*0.13-0.19*	*>0.19*	*0,08*
9000	*0.00-0.11*	*0.12-0.18*	*>0.18*	*0,07*
10000	*0.00-0.11*	*0.12-0.17*	*>0.17*	*0,07*
50000	*0.00-0.112*	*0.114-0.134*	*>0.134*	*0,07*

The column “average DF” is an estimated average using 4 in 5 occurrences at the midpoint of the first interval and 1 in 5 occurrences at the midpoint of the second. These averages are clearly below the benchmark limit of 0.1%.

Notice, however, that the validation strategies proposed above make use of hypothesis tests for one proportion. As illustrated earlier in Section 4, the Wald approximation performs worse for one proportion than for differences of proportions. Hence an alternative test based on differences of proportions will be developed in the following section.

A BACKTESTING STRATEGY BASED ON A STOCHASTIC BENCHMARK

In the preceding section a performance checking mechanism using a fixed upper limit for the benchmark was derived. That fixed upper limit followed from the central limit theorem and was found to be 0.1%.

An examination of Table 1 could also prompt the idea that the benchmark is not fixed but stochastic. Thus we will develop an alternative backtesting strategy in this section, based on a stochastic benchmark. In fact, in Section 3 we concluded that the benchmark can be defined as

$$df^{bm} \approx N\left(PD^{bm}; \sqrt{\frac{PD^{bm}(1-PD^{bm})}{N^{bm}}} \right) \tag{25}$$

where PD^{bm} was estimated at 0.04% and N^{bm} was estimated at 792.

On the other hand, the rating source’s default frequency is also normally distributed,

$$df_Y^{rs} \approx N\left(PD^{rs}; \sqrt{\frac{PD^{rs}(1-PD^{rs})}{N_Y^{rs}}} \right) \tag{26}$$

If one assumes a stochastic benchmark, there is no longer an upper limit for the PD of the rating source. The condition on which to base the performance-checking mechanism should be that “the rating source should do at least as well as the benchmark”. In terms of a probability of default, this means that the rating source’s PD should be lower than or equal to that of the benchmark. The hypothesis to be tested is thus $H_0 : PD^{rs} \leq PD^{bm}$ against $H_1 : PD^{rs} > PD^{bm}$
where PD^{bm} was estimated at 0.04% and N^{bm} was estimated at 792.

The test is completely different from the one in the preceding section. Indeed we cannot replace PD^{bm} by 0.04% because this is only an estimate of the benchmark’s PD. The true PD of the benchmark is unknown. We should therefore combine the variance of the ex-ante estimated probability of default of the rating source and that of the benchmark in one measure in order to conduct the backtesting.

The difference of two normally distributed variables also has a normal distribution thus, assuming that both are independent[17]:

$$df_Y^{rs} - df^{bm} \approx N\left(PD^{rs} - PD^{bm}; \sqrt{\frac{PD^{rs}(1-PD^{rs})}{N_Y^{rs}} + \frac{PD^{bm}(1-PD^{bm})}{N^{bm}}} \right) \tag{27}$$

[17] If the rating source’s eligible class and the benchmark are dependant then the variance of the combined normal distribution should include the covariance term.

PD^{rs} and PD^{bm} are unknown, but if the null hypothesis is true then their difference should be $PD^{rs} - PD^{bm} \leq 0$. An estimate of the combined variance $\frac{PD^{rs}(1-PD^{rs})}{N_Y^{rs}} + \frac{PD^{bm}(1-PD^{bm})}{N^{bm}}$ is needed. A standard hypothesis test, testing the equality of two proportions, would use a "pooled variance" as estimator. This pooled variance itself being derived from a "pooled proportion" estimator (see e.g. Moore and McCabe (1999), and Cantor and Falkenstein (2001)). The reasoning is that as we test the hypothesis of equal proportions, all observations can be pooled so that there are a total of $N_Y^{rs} + N^{bm}$ observations, among which there are $N^{bm}.df'^{bm} + N_Y^{rs}.df'^{rs}_Y$. The pooled proportion is thus

$$df'^{pooled} = \frac{792 \times 0.04\% + N_Y^{rs}.df'^{rs}_Y}{792 + N_Y^{rs}} \tag{28}$$

and the two variances are then

$$\sigma_{bm}^2 = \frac{df'^{pooled}(1-df'^{pooled})}{792}, \sigma_{rs}^2 = \frac{df'^{pooled}(1-df'^{pooled})}{N_Y^{rs}} \tag{29}$$

and (9) becomes

$$df_Y^{rs} - df^{bm} \approx N\left(0; \sqrt{df'^{pooled}(1-df'^{pooled})\left(\frac{1}{N_Y^{rs}} + \frac{1}{792}\right)}\right) \tag{30}$$

However, as we have an estimate of the benchmark that is based on only 24 past observations, we decided not to touch the variance estimate of the benchmark. So, taking a risk-averse position, we assigned the empirical variance of the benchmark recorded in these 24 observations (1981-2004) untouched, and the hypothesis test uses the distribution given in (31).

$$df_Y^{rs} - df^{bm} \approx N\left(0; \sqrt{\frac{df'^{pooled}(1-df'^{pooled})}{N_Y^{rs}} + 0.07\%^2}\right) \tag{31}$$

Using the observed default frequency (df'^{rs}) as an estimate for the rating source, and using the estimated benchmark values, the *p*-values of the test are given by:

$$1 - \Phi\left(\frac{df'^{rs} - 0{,}04\%}{\sqrt{0.07\%^2 + \frac{df'^{pooled}(1-df'^{pooled})}{N_Y^{rs}}}}\right) \tag{32}$$

The results for an estimated benchmark PD of 0.04% and a static pool of 10,000 companies are shown in Table 7.

Table 7: Test of credit quality assessment source against the limit of stochastic benchmark for a sample size of 10,000, using df'(bm) = 0.04%
(percentages)

	Mean	Stand dev	df'(bm)
BM	*0,04%*	*0,07%*	*0,04%*

Size	10.000	mean

N	*df'(rs)*	*p-value*	*Probability of "'N" if H0 is true*
0	0,0000	71,27	5,10
1	0,0100	66,17	5,31
2	0,0200	60,86	5,43
3	0,0300	55,43	5,43
4	0,0400	50,00	5,34
5	0,0500	44,66	5,17
6	0,0600	39,49	4,90
7	0,0700	34,59	4,59
8	0,0800	30,00	4,23
9	0,0900	25,77	3,83
10	0,1000	21,94	3,44
11	0,1100	18,50	3,04
12	0,1200	15,46	2,65
13	0,1300	12,81	2,29
14	0,1400	10,52	1,95
15	0,1500	8,57	1,65
16	0,1600	6,92	1,37
17	0,1700	5,55	1,14
18	0,1800	4,41	0,93
19	0,1900	3,48	0,75
20	0,2000	2,73	0,61
21	0,2100	2,12	0,48
22	0,2200	1,64	0,38
23	0,2300	1,26	0,30
24	0,2400	0,96	0,23
25	0,2500	0,73	0,18

Table 7 can be used as backtesting for a sample of 10,000 obligor names with an stochastic benchmark, after fixing a confidence level (i.e. a minimum *p*-value, e.g. 1%). A rating source with a static pool of size 10,000 is in line with the benchmark if at most once every five years a default frequency above 0.1% is observed. A default frequency above 0.23% should "never" be observed.

The intervals for other sizes of the static pool are shown in Table 8. The average default frequency seems to be lower than 0.1% for all sizes. As argued in Section 3, a higher average than 0.04% could be justified. Table 8 also shows the results when an estimate of 0.07% is used for the benchmark PD. As in Table 6, the lower value of the "Once in 5y" interval is derived from the 80% confidence limit, the absolute upper limit "Never" is derived from a 99% confidence interval.

Table 8: Backtesting strategy based on a stochastic benchmark for different static pool sizes *(percentages)*

	p(bm)=0.04%			p(bm)=0.07%		
	All time	**Once in 5y**	**Average DF**	**All time**	**Once in 5y**	**Average DF**
500	*0-0*	*0,2-0,6*	*0,08*	*0-0*	*0,2-0,8*	*0,10*
1000	*0-0*	*0,2-0,5*	*0,07*	*0-0,1*	*0,2-0,5*	*0,11*
2000	*0-0,1*	*0,15-0,35*	*0,09*	*0-0,1*	*0,2-0,45*	*0,11*
3000	*0-0,1*	*0,13-0,3*	*0,08*	*0-0,1*	*0,13-0,37*	*0,09*
4000	*0-0,1*	*0,125-0,275*	*0,08*	*0-0,125*	*0,15-0,37*	*0,10*
5000	*0-0,1*	*0,12-0,28*	*0,08*	*0-0,16*	*0,18-0,34*	*0,12*
6000	*0-0,1*	*0,12-0,25*	*0,08*	*0-0,15*	*0,16-0,35*	*0,11*
7000	*0-0,1*	*0,11-0,24*	*0,08*	*0-0,15*	*0,17-0,34*	*0,11*
8000	*0-0,1*	*0,11-0,237*	*0,07*	*0-0,15*	*0,16-0,32*	*0,11*
9000	*0-0,1*	*0,11-0,23*	*0,07*	*0-0,15*	*0,16-0,32*	*0,11*
10000	*0-0,1*	*0,11-0,23*	*0,07*	*0-0,15*	*0,16-0,32*	*0,11*
50000	*0-0,1*	*0,11-0,21*	*0,07*	*0-0,15*	*0,152-0,29*	*0,10*

It appears that with these assumptions (i.e. an ex ante estimate for the probability of default of 0.07% and benchmark sample size of 792 obligors) the backtesting strategy based on the stochastic benchmark is less conservative than that based on a fixed benchmark (cf. Table 6 and 8). For every sample size the confidence intervals for the realised default rates are wider for the stochastic benchmark test. It can also be seen that the confidence intervals are relatively wider for large sample sizes when using the stochastic benchmark.

THE BACKTESTING MECHANISMS AND BASEL II

Under the new rules of Basel II supervisors will be responsible for assigning an eligible External Credit Assessment Institution's (ECAI) credit risk assessment to the risk weights available under the standardised approach. Annex 2 of the revised Basel II framework (2005a) contains the Committee's proposal for a consistent mapping of credit risk assessments into the available risk weights. The mapping mechanism uses two quantifiable parameters:

- a ten-year average of a three-year cumulative default frequency, and
- the two most recent three-year cumulative default frequencies

These measures have to be compared to benchmark values:

- For the two most recent three-year cumulative default frequencies the Basel II documents give two benchmark levels:
  - a monitoring level derived from Monte Carlo simulations and fixed at the 99% quantile which takes a value of 1.0% for the "A" grade, and
  - a trigger level representing the 99.9% quantile, which takes a value of 1.3% for the "A" grade.

Using the 10 year average of the 3-year default frequencies observed over the years 1993-2002 by Standard&Poor's (0.25%) and the average number of issuers over the same period (1,024), the 99% and the 99.9% confidence intervals resulting from tests with a stochastic benchmark are shown for different static pool sizes in Table 9.

Table 9: Confidence intervals based on a stochastic benchmark rule applied to the 3 year default rates
(percentages)

	Conf 99%	Conf 99.9%
500	*1,00*	*1,40*
600	*1,00*	*1,30*
700	*1,00*	*1,28*
800	*0,90*	*1,25*
900	*0,90*	*1,22*
1000	*0,90*	*1,20*
1100	*0,90*	*1,20*
1200	*0,90*	*1,15*
average	***0,94***	***1,25***
p.m. Basel II	**1,00**	**1,30**

This table shows that our stochastic benchmark tests for confidence levels of 99% and 99.9% respectively yield results similar to the threshold values given in Annex 2 of the revised Basel II framework. It appears that our proposed test is slightly more conservative. If, instead of the stochastic benchmark, we were to use the fixed benchmark strategy, the confidence intervals would be even more conservative than those provided by the Basel II rules.

4.2. THE TRAFFIC LIGHT APPROACH, A SIMPLIFIED BACKTESTING MECHANISM

The testing procedures outlined in the previous sections allow

- confidence intervals to be defined for the annual default frequency (i.e. the annual interpretation of the rule) and
- the specification of how often a value should fall within a specific interval (i.e. the multi-period interpretation of the rule) in order to converge to a long-run average of 0.10%.

Moreover, the intervals for the annual default frequency depend on the size of a credit quality assessment source's eligible set (i.e. the static pool).

To apply the tests discussed in 4.1 in practice, a traffic light approach is proposed. Instead of defining exactly how often every possible default frequency may be observed for a certain credit quality assessment source, Tables 6 and 8 can be simplified to a restriction on how often a realised default frequency should fall within one of only three intervals (the three zones (green, orange, and red) of the "traffic light approach").

1. Depending on the size of a rating source's static pool, two threshold levels are defined that separate these three zones: (1) a monitoring level, and (2) a trigger level.
2. If the annually observed default frequency is (strictly) below the monitoring level, then the rating source is in line with the benchmark and is in the green zone.
3. If the observed default frequency is above (or equal to) the monitoring level and (strictly) below the trigger level, then the rating source is in the orange zone. The rating source is allowed to be in the orange zone only once in five years (on average).
4. If the observed default frequency is above (or equal to) the trigger level, then the rating source is in the red zone.

A practical example of a traffic light approach as defined above, with the monitoring and the trigger levels derived from Table 8, is given in Table 10 below. A similar example of a traffic light approach could also be constructed on the basis of Table 6 (test based on a fixed benchmark rule). The monitoring and trigger levels would be somewhat more conservative than those shown in Table 8.

Table 10: Example of traffic light monitoring and trigger levels based on a stochastic benchmark rule
(percentages)

Size of eligible set	Monitoring Level (orange)	Triggering Level (red)
Up to 500	0,20	1
Up to 1,000	0.20	0.80
Up to 5,000	0.18	0.34
Up to 50,000	0.16	0.28

One way of applying the traffic light approach in practice, assuming that a rating source has a static pool in the vicinity of 500 obligors, could be as follows: if the rating source records an annual realised default rate that is in the red zone (i.e. a realised default rate above 1%) or a default rate that repeatedly falls in the orange zone (i.e. more than once over a period of 5 years), then the analyst may wish to consult with the relevant rating provider to understand why its default experience is considerably worse than the historical default experience of the benchmark rating agencies. The credit quality assessment system provider will be asked to provide additional information to justify its violation of the rules.

If no convincing arguments were brought forward, then the conclusion would be that the rating source's model estimates probabilities of default which are too low and the model must be re-calibrated.[18]

Finally, please note that default frequencies are discrete variables, so the upper limit of the green zone can be far below the lower limit of the orange zone. E.g. for a static pool size below 500 the green zone means no defaults at all, because the lowest non zero is 0.2% (1/500) (see Table 8).

5. SUMMARY AND CONCLUSIONS

In this paper we concentrate on two main goals. First, we are interested in translating the single "A" rating as published by major rating agencies for debt issuers into a quantitative metric, the annual probability of default. In particular we look at the single "A" rating, because that is the minimum level at which the Eurosystem sets it requirement of high credit standards for collateral that can be used in its monetary policy operations. This translation method could be useful for mapping credit assessments of other rating sources to those of the major rating agencies by means of this probability of default metric. Although, the information that is contained in a rating goes beyond the probability of default of an obligor, we present arguments in support of the translation from a rating to a PD. The example presented with the single "A" rating could also be extended to other rating grades. We demonstrate that the probability of default for a single "A" issued by the main rating agencies is at most 0.1%.

Second, we are interested in assessing the quality of the estimated probability of default for a rating grade, in this case the single "A" rating grade, by means of the realised default rate, also called backtesting of the probability of default. We review briefly the main statistical tests that have appeared in the literature focusing on the binomial test and its normal extension, analysing in particular its main underlying assumptions, independence of default events and constant probability of default. We show that the existence of default correlation would imply wider confidence intervals than those derived with the binomial test, in which independence of default events is assumed. However, it is also argued that in practice default correlations would be low, in particular for high

[18] If forecast PDs and ex-post default information are available for every individual borrower in the static pool of the rating source, then the Brier score/Spiegelhalter test, for example, could be used to check the forecasting performance of the rating source's model.

credit quality debtors, and that, from a risk management perspective, it is preferable to rely on a more conservative test, such as the binomial test to derive critical values.

Assuming that the default generating process follows a binomial distribution, the paper proposes two generic backtesting strategies for testing the quality of forecast probabilities of default: first, a backtesting strategy that uses a fixed, absolute upper limit for the probability of default, which in the case of a single "A" rating is derived at 0.10%; and second, a backtesting strategy that relies on an stochastic benchmark, a benchmark probability of default that is not constant, unlike in the first strategy. The second strategy could be justified in cases where there is uncertainty about the level of the benchmark or if the benchmark is expected to move over time. We show that a backtesting strategy based on a stochastic benchmark would produce wider confidence intervals than those obtained using a fixed benchmark. The two strategies are based on one and five-year multi-period tests. The use of a multi-period test is intended to provide a more informative statement about the performance of a rating source as reliance only on annual tests may be misleading due to, for example, problems in the measurement of default with scarce data, situations of unforeseen non-normal stress that increase default rates, or the existence of default correlation.

The backtesting strategies presented are implemented through a traffic light approach in the same vein as in Tasche (2003). Depending on the size of a rating source's static pool, two threshold levels are defined that separate three zones, green, orange and red: (1) a monitoring level, and (2) a trigger level. If the annually observed default frequency is (strictly) below the monitoring level, then the rating source is in the green zone and is in line with the benchmark. If the observed default frequency is above (or equal to) the monitoring level and (strictly) below the trigger level, then the rating source is in the orange zone, which implies that the realised default rate is not compatible with the PD forecast but still in the range of usual statistical deviations. We implement the multi-period rule by allowing the default frequency to be in the orange zone only once in five years (on average). If the observed default frequency is above (or equal to) the trigger level, then the rating source is in the red zone, indicating that the realised default frequency is unequivocally not in line with the PD forecast.

We see the backtesting techniques and strategies described in this paper as early warning tools for identifying performance problems in credit assessment systems. This could be useful in the context of the Eurosystem Credit Assessment Framework, in which various credit assessment sources can be employed to assess the credit quality standards of eligible collateral. In such a setting it is important to guarantee that the different participating credit systems fulfil their rating mandates correctly and comparably. In this sense, this paper puts emphasis on risk management and therefore there is a general preference for backtesting strategies that are more conservative. However, the techniques presented in this paper are by no means the only mechanism that an analyst has at his disposal for validating the functioning of a credit system, but they are an important one. They could be considered as a first step (i.e. early warning) in a more comprehensive process that should take into account also more qualitative elements (see Basel Committee (2005b). The drawbacks shown in this paper as regards problems of measurement, existence of correlation, or the existence of non-normal stress situations should be weighed carefully when assessing credit assessment systems based solely on the results of backtesting tools of the type presented.

ANNEX 1: HISTORICAL DATA ON MOODY'S A-GRADE

YEAR	Issuers	1Y-Def.-Freq.
1981	376	0.00%
1982	387	0.26%
1983	432	0.00%
1984	472	0.00%
1985	524	0.00%
1986	579	0.00%
1987	555	0.00%
1988	553	0.00%
1989	587	0.00%
1990	614	0.00%
1991	609	0.00%
1992	694	0.00%
1993	740	0.00%
1994	880	0.00%
1995	968	0.00%
1996	1071	0.00%
1997	1133	0.00%
1998	1154	0.00%
1999	1173	0.00%
2000	1237	0.00%
2001	1287	0.16%
2002	1301	0.16%
2003	1279	0.00%
2004	1244	0.00%
	Mean	0.02%
Standard Deviation		0.07%

References

Agresti, A. and Coull, B. (1998), Approximate is better than "exact" for interval estimation of binomial proportions, The American Statistician, May.

Agresti, A. and Caffo, B. (2000), Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, The American Statistician, November.

Basel Committee on Banking Supervision (2005a), International Convergence of Capital Measurement and Capital Standards. A Revised Framework.

Basel Committee on Banking Supervision (2005b), Studies on the validation of internal rating systems (revised), Working Paper, Bank for International Settlements.

Billingsley, P. (1995), Probability and Measure, 3rd edition, John Wiley & Sons.

Blochwitz, S. and Hohl, S. (2001), The worst case or what default rates do we have to expect from the rating agencies? Working Paper, Deutsche Bundesbank, Frankfurt.

Blochwitz, S., When, C. and Hohl, S. (2003), Reconsidering ratings, Working Paper. Deutsche Bundesbank, Frankfurt.

Brier, G. (1950), Verification of forecasts expressed in terms of probability, Monthly Weather Review, vol. 78, No. 1, pp. 1-3.

Brown, L. D., Cai, T. and DasGupta, A. (2001), Interval estimation for a binomial proportion, Statistical Science, Vol. 16, No. 2.

Cai, T. (2005), One-sided confidence intervals in discrete distributions, J. Statistical Planning and Inference, Vol. 131.

Cantor, R. (2001), Moody's Investors Service's response to the consultative paper issued by the Basel Committee on Banking Supervision "A new capital adequacy framework", Journal of Banking and Finance, Vol. 25.

Cantor, R. and Falkenstein, E. (2001), Testing for rating consistency in annual default rates, Journal of Fixed Income.

Cantor, R. and Mann, C. (2003), Are corporate bond ratings procyclical? Moody's Special Comment, October.

Cantor, R., Packer, F. and Cole, K. (1997), Split ratings and the pricing of credit risk, Journal of Fixed Income, December.

Christensen, J., Hansen, E. and Lando, D. (2004), Confidence sets for continuous-time rating transition probabilities, Journal of Banking and Finance, Vol. 28.

Crouhy, M., Galai, D. and Mark, R. (2001), Prototype risk rating system, Journal of Banking and Finance, Vol. 25.

DeGroot, M. (1989), Probability and statistics, Second edition, Addison-Wesley Publishing Company.

Elton, E., Martin, J., Gruber, J., Agrawal, D. and Mann, C. (2004), Factors affecting the valuation of corporate bonds, in Cantor, R. (ed.), Recent Research on Credit Ratings, Journal of Banking and Finance, Vol. 28, special issue.

European Central Bank (2007), The implementation of monetary policy in the Euro area. General documentation on Eurosystem monetary policy instruments and procedures.

Finger, C. (2001), The one-factor Creditmetrics model in the new Basel Capital Accord, Riskmetrics journal, Vol. 2(1), pp. 9-18.

Fons, J. (2002), Understanding Moody's corporate bond ratings and rating process, Moody's Special Comment, May.

Gordy, M. (1998), A Comparative Anatomy of Credit Risk Models, Working Paper, Federal Reserve System.

Hamerle, A., Liebig, T., Rösch, D. (2003), Benchmarking Asset Correlations, RISK, Nov., pp. 77-81.

Heitfield, H. (2005), Studies on the Validation of Internal Rating Systems, Working Paper 14, Basel Committee on Banking Supervision.

Hosmer, D., Lemeshow, S. and Klar J. (1988), Goodness of fit testing for multiple logistics regression analysis when the estimated probabilities are small, Biometrical Journal, Vol. 30, pp. 991-924.

Hosmer, D. and Lemeshow, S. (1980), A goodness of test for the multiple logistic regression, Communication in Statistics, A10, 1043-1069.

Hosmer, D. and Lemeshow, S. (2000), Applied logistic regression, Wiley series in Probability and Statistics.

Huschens, S. and Stahl, G. (2005), A general framework for IRBS backtesting, Bankarchiv, Zeitschrift für das gesamte Bank und Börsenwesen, 53, pp. 241-248.

Hui, C.-H., Wong, T.-C., Lo, C.-F. and M.-X. Huang (2005), Benchmarking Model of Default Probabilities of Listed Companies, Journal of Fixed Income, September.

Hull, J., Predescu, M. and White, A. (2004), The relationship between credit default swap spreads, bond yields, and credit rating announcements, in Cantor, R. (ed.), Recent Research on Credit Ratings, Journal of Banking and Finance, Vol. 28, special issue.

Jafry, Y. and Schuermann, T. (2004), Measurement, Estimation and Comparison of Credit Migration Matrices, Journal of Banking and Finance.

Johnson, N. L. (1969), Discrete distributions, Houghton Mifflin Company, Boston.

Krahnen, J. P., and Weber, M. (2001), Generally accepted rating principles: a primer, Journal of Banking and Finance, Vol. 25.

Lando, D. and Skødeberg, T. M. (2002), Analyzing rating transitions and rating drift with continuous observations, Journal of Banking and Finance, Vol. 26.

Moody's (2005), Moody's Default Report 2005. Annual Default Study.

Moore, D. S., and McCabe, G. P. (1999), Introduction to the practice of statistics, W. H. Freemand and Company, New York.

Newcombe, R. G. (1998), Two-sided confidence intervals for the single proportion: comparison of seven methods, Statistics in Medicine, Vol. 17.

Nickel, P., Perraudin, W. and Varotto, S. (2000), Stability of Rating Transitions, Journal of Banking and Finance, Vol. 24..

Rauhmeier, R. (2006), PD-validation, experience from banking practice, in Engelman, B. and Rauhmeier, R. (eds.), The Basel II risk parameter, estimation, validation and stress testing, Springer, Berlin, pp. 307-343.

Rauhmeier, R. and Scheule, H. (2005), Rating properties and their implications for Basel II capital, Risk, 18(3), pp. 78-81.

Reiczigel, J. (2004), Confidence intervals for the binomial parameter: some new considerations, Working Paper, Szent István University, Budapest.

Rohatgi, V. K. (1984), Statistical Inference, Wiley Series in Probability and mathematical statistics, John Wiley & Sons.

Spanos, A. (1986), Statistical foundations of econometric modelling, Cambridge University Press.

Spiegelhalter, D. (1986), Probabilistic prediction in patient management and clinical trails, Statistics in Medicine, Vol. 5, pp. 421-433.

Standard & Poor's (2005), Annual global corporate default study: corporate defaults poised to rise in 2005. Global fixed income research.

Tasche, D. (2003), A traffic lights approach to PD validation, Working Paper, Deutsche Bundesbank, Frankfurt.

Tasche, D. (2006), Validation of internal rating systems and PD estimates, Working paper, Deutsche Bundesbank, Frankfurt.

Tiomo, A. (2004), Credit risk and variability of default rates: an empirical analysis using simulations, Working Paper, Banque de France, Paris.

Vollset, S. E. (1993), Confidence intervals for a binomial proportion, Statistics in Medicine, Vol. 12.

NATIONAL BANK OF BELGIUM - WORKING PAPERS SERIES

1. "Model-based inflation forecasts and monetary policy rules" by M. Dombrecht and R. Wouters, *Research Series*, February 2000.
2. "The use of robust estimators as measures of core inflation" by L. Aucremanne, *Research Series*, February 2000.
3. "Performances économiques des Etats-Unis dans les années nonante" by A. Nyssens, P. Butzen, P. Bisciari, *Document Series*, March 2000.
4. "A model with explicit expectations for Belgium" by P. Jeanfils, *Research Series*, March 2000.
5. "Growth in an open economy: some recent developments" by S. Turnovsky, *Research Series*, May 2000.
6. "Knowledge, technology and economic growth: an OECD perspective" by I. Visco, A. Bassanini, S. Scarpetta, *Research Series*, May 2000.
7. "Fiscal policy and growth in the context of European integration" by P. Masson, *Research Series*, May 2000.
8. "Economic growth and the labour market: Europe's challenge" by C. Wyplosz, *Research Series*, May 2000.
9. "The role of the exchange rate in economic growth: a euro-zone perspective" by R. MacDonald, *Research Series*, May 2000.
10. "Monetary union and economic growth" by J. Vickers, *Research Series*, May 2000.
11. "Politique monétaire et prix des actifs: le cas des Etats-Unis" by Q. Wibaut, *Document Series*, August 2000.
12. "The Belgian industrial confidence indicator: leading indicator of economic activity in the euro area?" by J.-J. Vanhaelen, L. Dresse, J. De Mulder, *Document Series*, November 2000.
13. "Le financement des entreprises par capital-risque" by C. Rigo, *Document Series,* February 2001.
14. "La nouvelle économie" by P. Bisciari, *Document Series*, March 2001.
15. "De kostprijs van bankkredieten" by A. Bruggeman and R. Wouters, *Document Series*, April 2001.
16. "A guided tour of the world of rational expectations models and optimal policies" by Ph. Jeanfils, *Research Series*, May 2001.
17. "Attractive Prices and Euro - Rounding effects on inflation" by L. Aucremanne and D. Cornille, *Documents Series*, November 2001.
18. "The interest rate and credit channels in Belgium: an investigation with micro-level firm data" by P. Butzen, C. Fuss and Ph. Vermeulen, *Research series*, December 2001.
19. "Openness, imperfect exchange rate pass-through and monetary policy" by F. Smets and R. Wouters, *Research series*, March 2002.
20. "Inflation, relative prices and nominal rigidities" by L. Aucremanne, G. Brys, M. Hubert, P. J. Rousseeuw and A. Struyf, *Research series*, April 2002.
21. "Lifting the burden: fundamental tax reform and economic growth" by D. Jorgenson, *Research series*, May 2002.
22. "What do we know about investment under uncertainty?" by L. Trigeorgis, *Research series*, May 2002.
23. "Investment, uncertainty and irreversibility: evidence from Belgian accounting data" by D. Cassimon, P.-J. Engelen, H. Meersman, M. Van Wouwe, *Research series*, May 2002.
24. "The impact of uncertainty on investment plans" by P. Butzen, C. Fuss, Ph. Vermeulen, *Research series*, May 2002.
25. "Investment, protection, ownership, and the cost of capital" by Ch. P. Himmelberg, R. G. Hubbard, I. Love, *Research series*, May 2002.
26. "Finance, uncertainty and investment: assessing the gains and losses of a generalised non-linear structural approach using Belgian panel data", by M. Gérard, F. Verschueren, *Research series*, May 2002.
27. "Capital structure, firm liquidity and growth" by R. Anderson, *Research series*, May 2002.
28. "Structural modelling of investment and financial constraints: where do we stand?" by J.- B. Chatelain, *Research series*, May 2002.
29. "Financing and investment interdependencies in unquoted Belgian companies: the role of venture capital" by S. Manigart, K. Baeyens, I. Verschueren, *Research series*, May 2002.
30. "Development path and capital structure of Belgian biotechnology firms" by V. Bastin, A. Corhay, G. Hübner, P.-A. Michel, *Research series*, May 2002.
31. "Governance as a source of managerial discipline" by J. Franks, *Research series*, May 2002.

32. "Financing constraints, fixed capital and R&D investment decisions of Belgian firms" by M. Cincera, *Research series*, May 2002.
33. "Investment, R&D and liquidity constraints: a corporate governance approach to the Belgian evidence" by P. Van Cayseele, *Research series*, May 2002.
34. "On the Origins of the Franco-German EMU Controversies" by I. Maes, *Research series*, July 2002.
35. "An estimated dynamic stochastic general equilibrium model of the Euro Area", by F. Smets and R. Wouters, *Research series*, October 2002.
36. "The labour market and fiscal impact of labour tax reductions: The case of reduction of employers' social security contributions under a wage norm regime with automatic price indexing of wages", by K. Burggraeve and Ph. Du Caju, *Research series*, March 2003.
37. "Scope of asymmetries in the Euro Area", by S. Ide and Ph. Moës, *Document series*, March 2003.
38. "De autonijverheid in België: Het belang van het toeleveringsnetwerk rond de assemblage van personenauto's", by F. Coppens and G. van Gastel, *Document series*, June 2003.
39. "La consommation privée en Belgique", by B. Eugène, Ph. Jeanfils and B. Robert, *Document series*, June 2003.
40. "The process of European monetary integration: a comparison of the Belgian and Italian approaches", by I. Maes and L. Quaglia, *Research series*, August 2003.
41. "Stock market valuation in the United States", by P. Bisciari, A. Durré and A. Nyssens, *Document series*, November 2003.
42. "Modeling the Term Structure of Interest Rates: Where Do We Stand?, by K. Maes, *Research series*, February 2004.
43. Interbank Exposures: An Empirical Examination of System Risk in the Belgian Banking System, by H. Degryse and G. Nguyen, *Research series*, March 2004.
44. "How Frequently do Prices change? Evidence Based on the Micro Data Underlying the Belgian CPI", by L. Aucremanne and E. Dhyne, *Research series*, April 2004.
45. "Firms' investment decisions in response to demand and price uncertainty", by C. Fuss and Ph. Vermeulen, *Research series*, April 2004.
46. "SMEs and Bank Lending Relationships: the Impact of Mergers", by H. Degryse, N. Masschelein and J. Mitchell, *Research series*, May 2004.
47. "The Determinants of Pass-Through of Market Conditions to Bank Retail Interest Rates in Belgium", by F. De Graeve, O. De Jonghe and R. Vander Vennet, *Research series*, May 2004.
48. "Sectoral vs. country diversification benefits and downside risk", by M. Emiris, *Research series*, May 2004.
49. "How does liquidity react to stress periods in a limit order market?", by H. Beltran, A. Durré and P. Giot, *Research serie*s, May 2004.
50. "Financial consolidation and liquidity: prudential regulation and/or competition policy?", by P. Van Cayseele, *Research series*, May 2004.
51. "Basel II and Operational Risk: Implications for risk measurement and management in the financial sector", by A. Chapelle, Y. Crama, G. Hübner and J.-P. Peters, *Research series,* May 2004.
52. "The Efficiency and Stability of Banks and Markets", by F. Allen, *Research series*, May 2004.
53. "Does Financial Liberalization Spur Growth?" by G. Bekaert, C.R. Harvey and C. Lundblad, *Research series*, May 2004.
54. "Regulating Financial Conglomerates", by X. Freixas, G. Lóránth, A.D. Morrison and H.S. Shin, *Research series*, May 2004.
55. "Liquidity and Financial Market Stability", by M. O'Hara, *Research series*, May 2004.
56. "Economisch belang van de Vlaamse zeehavens: verslag 2002", by F. Lagneaux, *Document series*, June 2004.
57. "Determinants of Euro Term Structure of Credit Spreads", by A. Van Landschoot, *Research series*, July 2004.
58. "Macroeconomic and Monetary Policy-Making at the European Commission, from the Rome Treaties to the Hague Summit", by I. Maes, *Research series*, July 2004.
59. "Liberalisation of Network Industries: Is Electricity an Exception to the Rule?", by F. Coppens and D. Vivet, *Document series*, September 2004.
60. "Forecasting with a Bayesian DSGE model: an application to the euro area", by F. Smets and R. Wouters, *Research series*, September 2004.
61. "Comparing shocks and frictions in US and Euro Area Business Cycle: a Bayesian DSGE approach", by F. Smets and R. Wouters, *Research series*, October 2004.

62. "Voting on Pensions: A Survey", by G. de Walque, *Research series*, October 2004.
63. "Asymmetric Growth and Inflation Developments in the Acceding Countries: A New Assessment", by S. Ide and P. Moës, *Research series*, October 2004.
64. "Importance économique du Port Autonome de Liège: rapport 2002", by F. Lagneaux, *Document series*, November 2004.
65. "Price-setting behaviour in Belgium: what can be learned from an ad hoc survey", by L. Aucremanne and M. Druant, *Research series,* March 2005.
66. "Time-dependent versus State-dependent Pricing: A Panel Data Approach to the Determinants of Belgian Consumer Price Changes", by L. Aucremanne and E. Dhyne, *Research series*, April 2005.
67. "Indirect effects – A formal definition and degrees of dependency as an alternative to technical coefficients", by F. Coppens, *Research series,* May 2005.
68. "Noname – A new quarterly model for Belgium", by Ph. Jeanfils and K. Burggraeve, *Research series*, May 2005.
69. "Economic importance of the Flemish maritime ports: report 2003", F. Lagneaux, *Document series*, May 2005.
70. "Measuring inflation persistence: a structural time series approach", M. Dossche and G. Everaert, *Research series*, June 2005.
71. "Financial intermediation theory and implications for the sources of value in structured finance markets", J. Mitchell, *Document series*, July 2005.
72. "Liquidity risk in securities settlement", J. Devriese and J. Mitchell, *Research series*, July 2005.
73. "An international analysis of earnings, stock prices and bond yields", A. Durré and P. Giot, *Research series*, September 2005.
74. "Price setting in the euro area: Some stylized facts from Individual Consumer Price Data", E. Dhyne, L. J. Álvarez, H. Le Bihan, G. Veronese, D. Dias, J. Hoffmann, N. Jonker, P. Lünnemann, F. Rumler and J. Vilmunen, *Research series*, September 2005.
75. "Importance économique du Port Autonome de Liège: rapport 2003", by F. Lagneaux, *Document series,* October 2005.
76. "The pricing behaviour of firms in the euro area: new survey evidence, by S. Fabiani, M. Druant, I. Hernando, C. Kwapil, B. Landau, C. Loupias, F. Martins, T. Mathä, R. Sabbatini, H. Stahl and A. Stokman, *Research series*, November 2005.
77. "Income uncertainty and aggregate consumption, by L. Pozzi, *Research series*, November 2005.
78. "Crédits aux particuliers - Analyse des données de la Centrale des Crédits aux Particuliers", by H. De Doncker, *Document series*, January 2006.
79. "Is there a difference between solicited and unsolicited bank ratings and, if so, why?" by P. Van Roy, *Research series*, February 2006.
80. "A generalised dynamic factor model for the Belgian economy - Useful business cycle indicators and GDP growth forecasts", by Ch. Van Nieuwenhuyze, *Research series*, February 2006.
81. "Réduction linéaire de cotisations patronales à la sécurité sociale et financement alternatif" by Ph. Jeanfils, L. Van Meensel, Ph. Du Caju, Y. Saks, K. Buysse and K. Van Cauter, *Document series*, March 2006.
82. "The patterns and determinants of price setting in the Belgian industry" by D. Cornille and M. Dossche, *Research series*, May 2006.
83. "A multi-factor model for the valuation and risk management of demand deposits" by H. Dewachter, M. Lyrio and K. Maes, *Research series*, May 2006.
84. "The single European electricity market: A long road to convergence", by F. Coppens and D. Vivet, *Document series*, May 2006.
85. "Firm-specific production factors in a DSGE model with Taylor price setting", by G. de Walque, F. Smets and R. Wouters, *Research series*, June 2006.
86. "Economic importance of the Belgian ports: Flemish maritime ports and Liège port complex - report 2004", by F. Lagneaux, *Document series*, June 2006.
87. "The response of firms' investment and financing to adverse cash flow shocks: the role of bank relationships", by C. Fuss and Ph. Vermeulen, *Research series*, July 2006.
88. "The term structure of interest rates in a DSGE model", by M. Emiris, Research series, July 2006.
89. "The production function approach to the Belgian output gap, Estimation of a Multivariate Structural Time Series Model", by Ph. Moës, Research series, September 2006.
90. "Industry Wage Differentials, Unobserved Ability, and Rent-Sharing: Evidence from Matched Worker-Firm Data, 1995-2002", by R. Plasman, F. Rycx and I. Tojerow, Research series, October 2006.

91. "The dynamics of trade and competition", by N. Chen, J. Imbs and A. Scott, Research series, October 2006.
92. "A New Keynesian Model with Unemployment", by O. Blanchard and J. Gali, Research series, October 2006.
93. "Price and Wage Setting in an Integrating Europe: Firm Level Evidence", by F. Abraham, J. Konings and S. Vanormelingen, Research series, October 2006.
94. "Simulation, estimation and welfare implications of monetary policies in a 3-country NOEM model", by J. Plasmans, T. Michalak and J. Fornero, Research series, October 2006.
95. "Inflation persistence and price-setting behaviour in the euro area: a summary of the Inflation Persistence Network evidence ", by F. Altissimo, M. Ehrmann and F. Smets, Research series, October 2006.
96. "How Wages Change: Micro Evidence from the International Wage Flexibility Project", by W.T. Dickens, L. Goette, E.L. Groshen, S. Holden, J. Messina, M.E. Schweitzer, J. Turunen and M. Ward, Research series, October 2006.
97. "Nominal wage rigidities in a new Keynesian model with frictional unemployment", by V. Bodart, G. de Walque, O. Pierrard, H.R. Sneessens and R. Wouters, Research series, October 2006.
98. "Dynamics on monetary policy in a fair wage model of the business cycle", by D. De la Croix, G. de Walque and R. Wouters, Research series, October 2006.
99. "The kinked demand curve and price rigidity: evidence from scanner data", by M. Dossche, F. Heylen and D. Van den Poel, Research series, October 2006.
100. "Lumpy price adjustments: a microeconometric analysis", by E. Dhyne, C. Fuss, H. Peseran and P. Sevestre, Research series, October 2006.
101. "Reasons for wage rigidity in Germany", by W. Franz and F. Pfeiffer, Research series, October 2006.
102. "Fiscal sustainability indicators and policy design in the face of ageing", by G. Langenus, Research series, October 2006.
103. "Macroeconomic fluctuations and firm entry: theory and evidence", by V. Lewis, Research series, October 2006.
104. "Exploring the CDS-Bond Basis" by J. De Wit, Research series, November 2006.
105. "Sector Concentration in Loan Portfolios and Economic Capital", by K. Düllmann and N. Masschelein, Research series, November 2006.
106. "R&D in the Belgian Pharmaceutical Sector", by H. De Doncker, Document series, December 2006.
107. "Importance et évolution des investissements directs en Belgique", by Ch. Piette, Document series, January 2007.
108. "Investment-Specific Technology Shocks and Labor Market Frictions", by R. De Bock, Research series, February 2007.
109. "Shocks and frictions in US Business cycles: a Bayesian DSGE Approach", by F. Smets and R. Wouters, Research series, February 2007.
110. "Economic impact of port activity: a disaggregate analysis. The case of Antwerp", by F. Coppens, F. Lagneaux, H. Meersman, N. Sellekaerts, E. Van de Voorde, G. van Gastel, Th. Vanelslander, A. Verhetsel, Document series, February 2007.
111. "Price setting in the euro area: some stylised facts from individual producer price data", by Ph Vermeulen, D. Dias, M. Dossche, E. Gautier, I. Hernando, R. Sabbatini, H. Stahl, Research series, March 2007.
112. "Assessing the Gap between Observed and Perceived Inflation in the Euro Area: Is the Credibility of the HICP at Stake?", by L. Aucremanne, M. Collin, Th. Stragier, Research series, April 2007.
113. "The spread of Keynesian economics: a comparison of the Belgian and Italian experiences", by I. Maes, Research series, April 2007.
114. "Imports and Exports at the Level of the Firm: Evidence from Belgium", by M. Muûls and M. Pisu, Research series, May 2007.
115. "Economic importance of the Belgian ports: Flemish maritime ports and Liège port complex - report 2005", by F. Lagneaux, Document series, May 2007.
116. "Temporal Distribution of Price Changes: Staggering in the Large and Synchronization in the Small", by E. Dhyne and J. Konieczny, Research series, June 2007.
117. "Can excess liquidity signal an asset price boom?", by A. Bruggeman, Research series, August 2007.
118. "The performance of credit rating systems in the assessment of collateral used in Eurosystem monerary policy operations", by Francois Coppens, Fernando González, Gerhard Winkler, Research series, August 2007.