Pitfalls and Remedies in Testing the Calibration Quality of Rating Systems

Wolfgang Aussenegg[a], Florian Resch[b], Gerhard Winkler[bc]

This Version: November 2009

[a] Vienna University of Technology, Favoritenstraße 9-11, A-1040 Vienna, Austria
[b] Oesterreichische Nationalbank, Otto Wagner Platz 3, A-1090 Vienna, Austria
[c] WU Wien - Vienna University of Economics and Business, Heiligenstädter Straße 46-48, A-1190 Wien, Austria

Abstract

Testing calibration quality by means of backtesting is an integral part in the validation of credit rating systems. Against this background this paper provides a comprehensive overview of existing testing procedures. We study the procedures'deficiencies theoretically and illustrate their impact empirically. Based on the insights gained thereof, we develop enhanced hybrid testing procedures which turn out to be superior to the commonly applied methods. We also propose computationally efficient algorithms for our calibration tests. Finally, we are able to demonstrate empirically that our method outperforms existing tests in a scenario analysis using rating data of Moody's.

Keywords: Rating System, Validation, Calibration Quality.

JEL classification: G20, C49.

Corresponding Author: Florian Resch, Oesterreichische Nationalbank,
Otto Wagner Platz 3, A-1090 Vienna, Austria. *E-mail address:* florian.resch@oenb.at
Any views expressed represent those of the authors only and not necessarily those of Oesterreichische Nationalbank. All remaining errors are those of the authors.

1 Introduction

Financial institutions as well as their regulators need means to assess the quality of rating systems. In the assessment, discriminatory power, calibration quality and stability are commonly regarded as the key attributes of a good rating system. For an introduction to currently employed methods see Basel Committee on Banking Supervision (2005). With the prime aim of a rating system being the estimation of probabilities of default, as argued by Krahnen and Weber (2001), calibration quality may be regarded as the most important aspect. In the most general sense, calibration quality refers to a rating system's ability to identify the true probability of default for a single borrower or a risk class of homogeneous borrowers (i.e., a rating grade).

In practice, it is not possible to precisely estimate the probability of default using a rating system as the true probability of default is latent. Yet a realization of the true latent probability of default is observable ex-post, which we call the realized default frequency. In essence, testing calibration quality thus means measuring the deviations between (ex-post) observable default frequencies and (ex-ante) estimated probabilities of default, and then judging whether these deviations are statistically significant. To this end, a variety of tests has been proposed for evaluating calibration quality. Most of the testing procedures have their origins in medical research. Partly due to this fact, their application in the assessment of a rating model's quality faces drawbacks and may yield biased results.

It is the purpose of this paper to provide a comprehensive overview of existing testing procedures, to study the procedures' deficiencies theoretically and to illustrate their impact empirically. Based on the insights gained thereof, we then develop enhanced hybrid testing procedures which turn out to be superior to the commonly applied methods for the evaluation of a rating system's calibration quality. We also propose computationally efficient algorithms for our procedures. Finally, we are able to demonstrate empirically that our method outperforms the others in a comprehensive scenario analysis using data on a portfolio rated by Moody's.

Section 2 provides an overview of existing testing procedures for single risk classes and of joint tests for complete rating systems. In section 3 we first derive an efficient algorithm for the calculation of the p-value of an exact joint test and then define a new class of hybrid tests. A comparative performance analysis employing multiple scenarios for all the testing procedures discussed in this paper is carried out in section 4. Section 5 concludes the paper.

2 A Revision of Common Tests

Dealing with the uncertainty whether an obligor will default over a specified time horizon constitutes a rater's fundamental problem. Thus it is necessary to consider this problem in an adequate probabilistic framework.

2.1 Probabilistic Framework

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Sample Space Ω contains the two possible outcomes $\Omega = \{\text{default}, \text{non} - \text{default}\}$. σ-algebra $\mathcal{F}$ shall be the power set of Ω, $\mathcal{F} = 2^{\Omega}$. Probability measure $\mathcal{P}$ assigns a unique probability to all the events contained in $\mathcal{F}$. Now consider a random variable D on this space defined as an $\mathcal{F}$-measurable mapping

$$D(\omega) := \begin{cases} (\Omega, \mathcal{F}) & \rightarrow & \{0,1\} \\ \omega & \rightarrow & \mathbf{1}_{\text{default}}(\omega) \end{cases}$$

where $\mathbf{1}_{\text{default}}$ is the indicator function

$$\mathbf{1}_{\text{default}} = \begin{cases} 1, & \text{if} \quad \omega = \text{default} \\ 0, & \text{if} \quad \omega = \text{non} - \text{default} \end{cases}$$

The probability mass function of the Bernoulli-distributed random variable D is given by

$$P(D = d) = b(d; p_{true}) = p_{true}^{d}(1 - p_{true})^{1-d} \tag{1}$$

where p_{true} is the true, latent probability of default. Note that ex-post only one realization of D is observable. From the result of one single experiment, i.e. a single obligor, we cannot infer information about what p_{true} might have been. Following the proposals of the Basel Committee on Banking Supervision (2005), obligors of identical credit quality are to be pooled into risk classes. Each risk class c comprises a certain number n_c of obligors, each with a probability of default $p_{c,true}$. Let S_c be a random variable counting the number of defaults in risk class c. If the n_c Bernoulli experiments D are independent, the measurable space $(\Omega_c, \mathcal{F}_c)$ on which random variable S_c is defined is then given by the product of the measurable spaces $(\Omega_c, \mathcal{F}_c) = (\Omega^{n_c}, \mathcal{F}^{n_c})$. Hence the random variable S_c is an $\mathcal{F}_c$-measurable mapping of the form

$$S_c := \begin{cases} (\Omega_c, \mathcal{F}_c) & \rightarrow \quad \{0, \ldots, n_c\} \\ (\omega_1, \ldots, \omega_{n_c}) & \rightarrow \quad \sum_{i=1}^{n_c} \mathbf{1}_{\text{default}}(\omega_i) \end{cases}$$

and follows a binomial distribution with probability mass function

$$P(S_c = s_c) = b(s_c; n_c, p_{c,true}) = \binom{n_c}{s_c} p_{c,true}^{s_c} (1 - p_{c,true})^{n_c - s_c}. \tag{2}$$

Further expanding on the ideas, one could also jointly consider an entire portfolio of obligors which are divided into C different risk classes. Such a segmentation we refer to as *rating system*. Our objective is the joint assessment of the calibration quality of the whole rating system (i.e., of all risk classes simultaneously). Hence we are again interested in the number of defaults in each risk class. To this end, we define the C-variate random variable $\mathbf{S} = (S_1, \ldots, S_C)^T$, where $S_c, c = 1, \ldots, C$, denotes the random variable counting the number of default in risk class c and is defined as in equation (2). Each dimension of $\mathbf{S}$ represents a risk class and we refer to the random variable $\mathbf{S}$ as *default pattern*.

Now assume again that defaults within each risk class and across risk classes are independent. The probability space $(\Omega_{\mathbf{S}}, \mathcal{F}_{\mathbf{S}}, \mathcal{P}_{\mathbf{S}})$ of $\mathbf{S}$ is then given by the product space of the probability spaces of the individual risk classes $(\Omega_{\mathbf{S}}, \mathcal{F}_{\mathbf{S}}, \mathcal{P}_{\mathbf{S}}) = (\times_{c=1}^{C} \Omega_c, \times_{c=1}^{C} \mathcal{F}_c, \times_{c=1}^{C} \mathcal{P}_c)$.

The random variable $\mathbf{S}$ is then defined as

$$\mathbf{S} := \begin{cases} (\Omega_{\mathbf{S}}, \mathcal{F}_{\mathbf{S}}) & \rightarrow \quad \mathbb{N}_0^C \\ (\omega_1, \ldots, \omega_C)^T & \rightarrow \quad (S_1(\omega_1), \ldots, S_C(\omega_C))^T, \end{cases}$$

and the probability measure is the product measure. Thus the probability to observe a specific default pattern is given by the product of the marginal distributions

$$P(\mathbf{S} = \mathbf{s}) = P(S_1 = s_1, \ldots, S_C = s_C) = \prod_{c=1}^{C} P(S_c = s_c) = \prod_{c=1}^{C} b(s_c; n_c, p_{c,true}) \qquad (3)$$

where n_c is the number of obligors in risk class c and $p_{c,true}$ is the probability of default for obligors in risk class c.

2.2 Tests for Single Risk Classes

The primary objective of a rating system is to provide precise probability of default estimates $\hat{p}_c$ for each rating class c (see Krahnen and Weber, 2001). In the assessment of calibration quality we are interested in the accuracy of the estimate $\hat{p}_c$ for the true latent $p_{c,true}$. Denote the *realized number of defaults* as m_c and the *realized default frequency* as $f_c = \frac{m_c}{n_c}$. A straightforward hypothesis test for the evaluation of the quality of the estimate relies on the null hypothesis that the (ex-ante) estimated probability of default equals the realized default frequency,

$$\mathrm{H}_0 : \ \hat{p}_c = f_c,$$

against the alternative hypothesis that the values differ at some specified level of significance

$$\mathrm{H}_1 : \ \hat{p}_c \neq f_c.$$

Under the null hypothesis the number of defaults S_c follows a binomial distribution

$$P(S_c = s_c) = b(s_c; n_c, \hat{p}_c) = \binom{n_c}{s_c} \hat{p}_c^{s_c} (1 - \hat{p}_c)^{n_c - s_c}. \tag{4}$$

A rejection of the null would indicate that $\hat{p}_c$ is a poor estimate for $p_{c,true}$.

To obtain an exact test result we revert to Clopper and Pearson (1934). They provide exact confidence intervals for the probability estimate of a binomially distributed random variable. Their confidence intervals are often referred to as the 'gold standard' and are consequently recommended in the literature (e.g., Vollset, 1993). Inverting the procedure proposed by Clopper and Pearson for the computation of confidence intervals, we are able to derive an exact binomial test which handles both tails independently. The definition of Clopper and Pearson requires the test to be two-sided and 'central' in the sense that the critical regions for rejection for the number of defaults are the lower $\alpha/2$ quantile

$$\Big\{k | \sum_{i=0}^{k} b(i; n_c, \hat{p}_c) \leq \alpha/2\Big\}$$

and the upper $1 - \alpha/2$ quantile

$$\Big\{k | \sum_{i=k}^{n_c} b(i; n_c, \hat{p}_c) \leq \alpha/2\Big\}.$$

Hence the p-value $\alpha_{c,CP}$ of the Clopper-Pearson test is given by (see Vollset, 1993)

$$\alpha_{c,CP} = 2 \min \left\{ \sum_{s_c=0}^{m_c} b(s_c; n_c, \hat{p}_c), \sum_{s_c=m_c}^{n_c} b(s_c; n_c, \hat{p}_c) \right\}. \tag{5}$$

An alternative method for the computation of exact confidence intervals for a binomially distributed random variable was proposed by Sterne (1954). Reiczigel (2003) argued that the approach by Sterne outperforms the procedure suggested by Clopper and Pearson (1934) in terms of the mean coverage ratio and the minimum coverage ratio. Furthermore,

as the confidence intervals derived by Sterne are non-central, their application may be particularly useful when the binomial distribution is heavily skewed. This is very often the case in the validation of calibration quality, since both, the number of obligors and their corresponding default probabilities are very low in risk classes with good credit quality, which results in heavily skewed binomial distributions. In short, Sterne's method requires to rank the possible outcomes of the binomial random variable by their probability of occurrence in decreasing order. The acceptance region of a test relying on the method of Sterne comprises those outcomes of the ranked series for which the sum of the (ranked) probabilities of occurrence equals the desired level of confidence $1-\alpha$. Thus the Sterne p-value $\alpha_{c,Sterne}$ is defined as

$$\alpha_{c,Sterne} = Pr(b(S_c; n_c, \hat{p}_c) \leq b(m_c; n_c, \hat{p}_c)). \tag{6}$$

The method by Sterne can be linked to approximate tests for testing binomial proportions. To start with, let

$$h(z; \mu_c, \sigma_c) = \frac{1}{\sigma_c}\phi\left(\frac{z-\mu_c}{\sigma_c}\right) \tag{7}$$

be an approximation to the probability mass function of S_c by a normal density function,

$$b(\cdot; n_c, p_c) \approx h\left(\cdot; \mu_c, \sigma_c\right)$$

where μ_c and σ_c are the mean and the standard deviation of the normal distribution used in the approximation and $\phi(x)$ is the probability density function of a standard normal distribution

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

Replacing the binomial mass function in equation (6) by the normal density function $h(\cdot; \mu_c, \sigma_c)$ gives

$$\alpha_{c,approx} = Pr\left(\frac{1}{\sigma_c}\phi(\frac{S_c-\mu_c}{\sigma_c}) \leq \frac{1}{\sigma_c}\phi(\frac{m_c-\mu_c}{\sigma_c})\right). \tag{8}$$

Utilizing the definition of $\phi(\cdot)$, we obtain after some manipulation

$$\alpha_{c,approx} = Pr\Bigg(\underbrace{\left(\frac{S_c - \mu_c}{\sigma_c} \right)^2}_{X} \geq \left(\frac{m_c - \mu_c}{\sigma_c} \right)^2 \Bigg). \tag{9}$$

Thus an approximation of the probability mass function of S_c by means of h yields a χ^2_1 distribution for X.

Note that $m_c = f_c n_c$ and by setting $\mu_c = n_c \hat{p}_c$ and $\sigma_c = \sqrt{n_c \hat{p}_c (1 - \hat{p}_c)}$ we obtain the Score test (see Vollset, 1993). The p-value of the Score test is given by

$$\alpha_{c,Score} = Pr\Big(X \geq \left(\frac{m_c - n_c \hat{p}_c}{\sqrt{n_c \hat{p}_c (1 - \hat{p}_c)}} \right)^2 \Big), \quad X \sim \chi^2_1. \tag{10}$$

For small n_c and extreme values for p_c (close to 0 or 1) the binomial distribution is heavily skewed (see DeGroot and Schervish, 2002). Hence the approximation by the symmetric normal distribution used in the Score test may lead to biased results for the p-value $\alpha_{c,Score}$. Additionally, the approximation also suffers from the fact that the binomial distribution is discrete and the normal distribution is continuous especially when n_c is small. To mitigate the latter problem resulting from the discreetness of the binomial distribution, a continuity corrected version of the Score test has been proposed which yields a p-value $\alpha_{c,SCC}$ of

$$\alpha_{c,SCC} = Pr\left(X \geq \left(\frac{|m_c - n_c \hat{p}_c| - 0.5}{\sqrt{n_c \hat{p}_c (1 - \hat{p}_c)}} \right)^2 \right), \quad X \sim \chi^2_1. \tag{11}$$

When $\hat{p}_c$ is very close to zero the continuity correction may be too strong which would lead to biased results for the test. This is again typically the case for the best risk classes of any realistic portfolio's best risk classes containing obligors with very good credit quality.

An alternative to the Score test in the class of approximate tests is the Wald test. Here the realized frequency f_c is interpreted as the maximum likelihood estimate for $p_{c,true}$. Consequently, for the Wald test the maximum likelihood estimate f is used in the stan-

dard error $\sigma_c = \sqrt{n_c f_c(1-f_c)}$, while $m_c = f_c n_c$ and $\mu_c = n_c \hat{p}_c$ are defined as in the Score test. The p-value $\alpha_{c,Wald}$is then given by

$$\alpha_{c,Wald} = Pr\Big(X \geq \left(\frac{m_c - n_c\hat{p}_c}{\sqrt{n_c f_c(1-f_c)}}\right)^2\Big), \quad X \sim \chi^2_1. \tag{12}$$

Again the bias in the results may become very large when the probability of the binomial $\hat{p}_c$ is close to 0 or 1 and n_c is small due to the symmetry of the normal density (see Agresti and Coull, 1998). Furthermore, the p-value is undefined at the boundaries $m_c = 0$ or $m_c = n_c$. Vollset (1993) discussed Wald confidence intervals and suggested to use exact confidence intervals for boundary outcomes of m_c.

Addressing these deficiencies, Agresti and Coull (1998) suggested a modified Wald approximation. In essence, they recommend to add 2 successes and 2 failures to the realized outcome. Hence $\tilde{m}_c = m_c + 2$, $\tilde{n}_c = n_c + 4$ and $\tilde{f}_c = \frac{\tilde{m}_c}{\tilde{n}_c}$. This approach is motivated by the fact that $\tilde{n}_c = n_c + (\phi^{-1}(0.05/2))^2 = n_c + 1.96^2 \approx n_c + 4$. Therefore the confidence intervals proposed by Agresti and Coull are not 'central' around the maximum likelihood estimate f but close to Score intervals by Wilson (1927)) at a significance level 0.05. Inverting the procedure for obtaining Wald-Agresti-Coull confidence intervals, we derive the p-value $\alpha_{c,WAC}$ of a corresponding test which is given by

$$\alpha_{c,WAC} = Pr\left(X \geq \left(\frac{\tilde{m}_c - \tilde{n}_c\hat{p}_c}{\sqrt{\tilde{n}_c \tilde{f}_c(1-\tilde{f}_c)}}\right)^2\right), \quad X \sim \chi^2_1. \tag{13}$$

As an advantage of the modified version proposed by Wald-Agresti-Coull compared to the (standard) Wald test, the p-value based on Wald-Agresti-Coull is defined for all possible realizations of S_c including boundary outcomes. However, by definition results for the p-value become more and more biased the stronger the p-value deviates from 0.05.

2.3 Tests for Rating Systems

For a joint test of the calibration quality of all risk classes (i.e., the whole rating system) it seems advisable to extend the aforementioned testing procedures for single risk classes to the multivariate setting. When we examine a realization of all rating classes jointly we will refer to the vector $\mathbf{m} = (m_1, \ldots, m_C)^T$ as the *realized default pattern*. Accordingly we define the vector of realized default frequencies $\mathbf{f} = (f_1, \ldots, f_C)^T$ where $f_c = \frac{m_c}{n_c}$, $c = 1, \ldots, C$.

The natural starting point for the development of a multivariate calibration quality test is the Clopper-Pearson test. Although it is considered the 'gold standard' in the univariate case, it cannot be transferred to a C-variate test. Per definition, the rejection region comprises the left tail with the lower $\frac{\alpha}{2}$ quantile and the right tail with the upper $1 - \frac{\alpha}{2}$ quantile. Thus to handle both tails independently a linear ranking has to be established to determine the p-value for a given realization. However, in a multivariate setting only inward to outward rankings may be defined but no linear ranking of the outcomes may be established. Consequently, the test cannot be applied to multiple dimensions. In contrast, we can easily extend the concept of the Sterne test to the multivariate case. We interpret the p-value in a multivariate Sterne test as the probability to observe the realized default pattern $\mathbf{m}$ or any other pattern with equal or less probability of occurrence than the realized default pattern. Thus we obtain

$$\alpha_{Sterne} = Pr\Big(\prod_{c=1}^{C} b(S_c; n_c, p_c) \leq \prod_{c=1}^{C} b(m_c; n_c, p_c) \Big). \tag{14}$$

Alternatively, we may again consider approximate solutions. To this end, in each class c the probability mass function of the binomial distribution $b(\cdot; n_c, p_c)$ is again approximated by a normal density function $h(\cdot; \mu_c, \sigma_c)$. The p-value α_{approx} of such an approximate test

can then be written as

$$\begin{aligned}\alpha_{approx} &= Pr\Big(\prod_{c=1}^{C}\frac{1}{\sigma_c}\phi\left(\frac{S_c-\mu_c}{\sigma_c}\right) \leq \prod_{c=1}^{C}\frac{1}{\sigma_c}\phi\left(\frac{m_c-\mu_c}{\sigma_c}\right)\Big) \\ &= Pr\Big(\underbrace{\sum_{c=1}^{C}\left(\frac{S_c-\mu_c}{\sigma_c}\right)^2}_{X} \geq \sum_{c=1}^{C}\left(\frac{m_c-\mu_c}{\sigma_c}\right)^2\Big). \end{aligned} \tag{15}$$

Hence X follows a χ^2_C distribution with C degrees of freedom.

By setting $\mu_c = n_c\hat{p}_c$ and $\sigma_c = \sqrt{n_c\hat{p}_c(1-\hat{p}_c)}$ for $c = 1,\ldots,C$ we obtain the test statistic of a multivariate Score test. The p-value α_{Score} of such a test is calculated as

$$\alpha_{Score} = Pr\left(X \geq \sum_{c=1}^{C}\left(\frac{m_c - n_c\hat{p}_c}{\sqrt{n_c\hat{p}_c(1-\hat{p}_c)}}\right)^2\right), \quad X \sim \chi^2_C \tag{16}$$

whereas for the corresponding continuity corrected version of the Score test we obtain

$$\alpha_{SCC} = Pr\left(X \geq \sum_{c=1}^{C}\left(\frac{|m_c - n_c\hat{p}_c| - 0.5}{\sqrt{n_c\hat{p}_c(1-\hat{p}_c)}}\right)^2\right), \quad X \sim \chi^2_C. \tag{17}$$

As an alternative we may also employ the procedure of Wald-Agresti-Coull. To this end, we define $\tilde{m}_c = m_c + 2$, $\tilde{n}_c = n_c + 4$ and $\tilde{f}_c = \frac{\tilde{m}_c}{\tilde{n}_c}$ for $c = 1,\ldots,C$ and thus obtain the p-value α_{WAC} of the corresponding test

$$\alpha_{WAC} = Pr\left(X \geq \sum_{c=1}^{C}\left(\frac{\tilde{m}_c - \tilde{n}_c\tilde{p}_c}{\sqrt{\tilde{n}_c\tilde{f}_c(1-\tilde{f}_c)}}\right)^2\right), \quad X \sim \chi^2_C. \tag{18}$$

Note that to the best of our knowledge, only the Score test is currently in use for jointly testing the calibration quality of a set of risk classes ex-post (see Basel Committee on Banking Supervision, 2005). A multivariate Score test has furthermore been proposed by Hosmer and Lemeshow (1980) to assess the calibration quality of a logistic regression model in-sample.

2.4 Intermediate Summary - Limitations and Shortcomings of Existing Tests

The literature on the derivation of confidence intervals for binomial proportions is extensive (see e.g. Newcombe, 1998). Based on this literature we have shown how to invert the procedures to obtain p-values for corresponding hypothesis tests. Table 1 summarizes the most important properties of the proposed tests. The abbreviations refer to the following tests respectively: "CP" to the Clopper-Pearson test, "SCC" to the Score test with continuity correction and "WAC" to the Wald-Agresti-Coull test.

[Insert Table 1 here]

To reflect on the univariate case first, the so-called 'gold standard' for testing calibration quality of a single risk class is the exact test by Clopper and Pearson (1934). However, Newcombe (1998) demonstrates that Clopper-Pearson confidence intervals might be too conservative. As a consequence, a shortcoming of the Clopper-Pearson test is its overconservatism in the rejection of the null hypothesis that the rating system is calibrated well. Furthermore, the p-value of the Clopper-Pearson test may be biased when the underlying distribution is heavily skewed. Reiczigel (2003) finds that the test proposed by Sterne (1954) may be a superior exact alternative to the Clopper-Pearson test. The rejection region of the Sterne test is not central which is advantageous when the test is applied to heavily skewed binomial distributions arising in risk classes with a small number of obligors n_c and estimated probabilities of default $\hat{p}_c$ close to the boundaries 0 or 1. Yet both exact tests, the Clopper-Pearson test and the Sterne test, are computationally more intensive than approximate tests since they require enumeration.

Thus concerning tests which are computationally faster but approximate, the Score test and the Score test with continuity correction are recommended by Vollset (1993) and Newcombe (1998) for testing binomial proportions. Note that both tests will yield strongly biased results when n_c is small and the probability of default $\hat{p}_c$ takes extreme values (i.e. close to 0 or 1).

An alternative approximate test is the Wald test. When the realized number of defaults is equal to zero or equal to the total number of obligors in this class then the Wald test is not defined. Thus Agresti and Coull (1998) suggest an alternative approximate test based on the Wald approximation whose corresponding confidence intervals are close to those of the Score test. Since this test is based on a normal approximation as well it will show biased results when n_c is small and $\hat{p}_c$ is close to 0 or 1. By definition, however, the Wald-Agresti-Coull test will perform best when the desired level of significance α in the test is set at or close to 0.05, and will yield (more or less strongly) biased results for other choices of α.

Regarding the class of joint tests for the assessment of the calibration quality of a rating system, we have shown that, by definition, the Clopper-Pearson test cannot be applied to a whole rating system. In the multivariate setting no linear ranking can be established and thus no test statistic can be defined. We have then extended the concept of the Sterne test to a multivariate setting. Albeit an exact test, the clear shortcoming of this approach is its potential computational expensiveness since it requires enumerative techniques. For a full enumeration, the probabilities of occurrence of $\prod_{c=1}^{C}(n_c+1)$ default patterns would have to be computed. Hence an exact calibration test based on full enumeration would require cluster computing for any realistic portfolio.

Thus resorting to approximations for our suggested multivariate Sterne test, we have shown how to transfer the concepts of the Score test, the Score test with continuity correction and the Wald-Agresti-Coull to the multivariate setting. However, these multivariate tests face the same shortcomings as described for univariate versions of these tests.

Hosmer and Lemeshow (1988) have empirically analyzed the performance of their multivariate Score test in an extensive simulation study. They find that the test statistic does not converge to a χ^2-distribution in case the probability of default estimate $\hat{p}_c$ or the number of obligors n_c are small. Yet this is the case for most realistic portfolios, where

the probabilities of default of the best rating classes are extremely small and the number of clients receiving the best ratings is also very limited. Thus the application of the Hosmer-Lemeshow test or any other approximate test in the assessment of the calibration quality of a rating system will very often yield extremely biased results.

To sum up, the efficient testing of the calibration quality of a rating system is constricted by the following trade-off: Whereas exact tests are computationally complex and thus resources intensive, approximate tests may yield strongly biased results for realistic portfolios.

3 Efficient Calculation of Exact Tests and Enhanced Hybrid Tests for Rating Systems

The insights gained in the previous section motivate the need for improved procedures to jointly test the calibration quality of a rating system. In light of the trade-off between speed and accuracy, we first propose a fast and efficient calculation method for the p-value of an exact multivariate Sterne test.

3.1 An efficient Algorithm for the Exact Calculation of the multivariate Sterne Test

At first we will give an intuitive description of the algorithm for the case with two risk classes and then proceed with presenting a mathematical derivation for a more general setup.

For any realistic rating system the probabilities of default will generally be very low (i.e., far below 50%, see for example table 2). This leads to heavily skewed binomial distributions for especially for the risk class with good credit quality. Whenever the realized default pattern does not deviate too strongly from the expected pattern, the number of

default patterns in the acceptance region will be much lower than the number of patterns contained in the rejection region. The idea of our approach is thus to calculate only the probability of occurrence of those default patterns which are contained in the acceptance region of the test for a given default pattern $\mathbf{m}$ instead of fully enumerating all possible default patterns.

Having two risk classes we would carry out a full enumeration of all possible default patterns by employing an outer loop with index $i \,\epsilon\, \{0, \ldots, n_1\}$ for class one and an inner loop with index $j \;\epsilon\, \{0, \ldots, n_2\}$ for risk class two. Let $P(\mathbf{m}) = b(m_1; n_1, p_1)b(m_2; n_2, p_2)$ be the probability of occurrence of the realized pattern and let v_2 be the mode of the binomial distribution of S_2. Then the minimum probability for the number of defaults in risk class 1 that is allowed so that the default pattern contributes to the acceptance region is given by $p_1^\star = P(\mathbf{m})/b(v_2; n_2, p_2)$. Knowing that the binomial distribution is unimodal we may find a smaller range for the index $i \,\varepsilon\, \{l_1^\star, \ldots, u_1^\star\}$ by calculating $l_1^\star = \min\{s_1 | b(s_1; n_1, p_1) \geq p_1^\star\}$ and $u_1^\star = \max\{s_1 | b(s_1; n_1, p_1) \geq p_1^\star\}$.

Assume that $i_{cur} \,\epsilon\, \{l_1^\star, \ldots, u_1^\star\}$ is the current index for the outer loop for class one. Then the range of indices for the inner loop given on the current index i_{cur} of the outer loop is denoted by $j \,\epsilon\, \{l_2^\star, \ldots, u_2^\star\}$ where $\{l_2^\star = \min\{s_2 | b(s_2; n_2, p_2) \geq P(\mathbf{m})/b(i_s; n_1, p_1)\}$ and $\{u_2^\star = \max\{s_2 | b(s_2; n_2, p_2) \geq P(\mathbf{m})/b(i_s; n_1, p_1)\}$. Hence the range for index j for risk class two depends on the current index i_{cur} for risk class one. This leads to a recursive algorithm which can be easily transfered to multiple dimensions. A mathematical derivation is given below.

As argued, it will generally be computationally advantageous to define α_{Sterne} using the complementary probability

$$\alpha_{Sterne} = 1 - Pr\Big(\prod_{c=1}^{C} b(S_c; n_c, p_c) > \prod_{c=1}^{C} b(m_c; n_c, p_c)\Big). \tag{19}$$

Using the law of total probability, we may write α_{Sterne} as

$$\begin{aligned}\alpha_{Sterne} &= 1 - \sum_{s_1=0}^{n_1} \cdots \sum_{s_{a-1}=0}^{n_{a-1}} b(s_1; n_1, p_1) \ldots b(s_{a-1}; n_{a-1}, p_{a-1}) \cdot \\ &\underbrace{Pr\Big(\prod_{i=1}^{a-1} b(s_i; n_i, p_i) \prod_{j=a}^{C} b(S_j; n_j, p_j) > \prod_{c=1}^{C} b(m_c; n_c, p_c) \Big| s_1, \ldots, s_{a-1} \Big)}_{A}. \end{aligned} \tag{20}$$

The conditional probability A in equation (20) may then be written as

$$\begin{aligned} A &= Pr\Big(\prod_{j=a}^{C} b(S_j; n_j, p_j) > \\ &\quad \prod_{c=1}^{C} b(m_c; n_c, p_c) \prod_{i=1}^{a-1} b(s_i; n_i, p_i)^{-1} \Big| s_1, \ldots, s_{a-1} \Big). \end{aligned} \tag{21}$$

We now want to know the minimum value $p_a^\star$ which $b(S_a; n_a, p_a)$ may take so that the inequality in (21) holds. Conditional on $s_1, \ldots, s_{a-1}$ this value can be calculated when considering the most likely outcomes for $S_{a+1}, \ldots, S_C$. Let $v_c \, \epsilon \, \{0, \ldots, n_c\}$ be the mode of S_c,

$$v_c = \arg\max_{s_c} b(s_c; n_c, p_c), \quad c = 1, \ldots, C. \tag{22}$$

Hence the minimum value $p_a^\star$ for the probability $b(S_a; n_a; p_a)$ for $a = 2, \ldots, C$ so that the inequality in equation (21) holds is given by

$$p_a^\star(s_1, \ldots, s_{a-1}) = \prod_{c=1}^{C} b(m_c; n_c, p_c) \prod_{i=1}^{a-1} b(s_i; n_i, p_i)^{-1} \prod_{j=a+1}^{C} b(v_j; n_j, p_j)^{-1}. \tag{23}$$

and $p_1^\star$ is denoted by

$$p_1^\star = \prod_{c=1}^{C} b(m_c; n_c, p_c) \prod_{j=2}^{C} b(v_j; n_j, p_j)^{-1}. \tag{24}$$

Using these results and the property that the binomial distribution is unimodal, an upper limit $u_a^\star$ and a lower limit $l_a^\star$ for the number of defaults in class $b = 1, \ldots, C$, so that the inequality in equation (21) is satisfied can easily be derived:

$$l_a^\star(p_a^\star(s_1, \ldots, s_{a-1})) = \min\{s_a | b(s_a; n_a, p_a) \geq p_a^\star\}, \tag{25}$$

$$u_a^\star(p_a^\star(s_1, \ldots, s_{a-1})) = \max\{s_a | b(s_a; n_a, p_a) \geq p_a^\star\}. \tag{26}$$

Hence the exact p-value α_{Sterne} may be calculated as

$$\begin{aligned} \alpha_{Sterne} = 1 - \sum_{s_1=l_1^\star(p_1^\star)}^{u_1^\star(p_1^\star)} b(s_1; n_1, p_1) \Big(\sum_{s_2=l_2^\star(p_2^\star)}^{u_2^\star(p_2^\star)} b(s_2; n_2, p_2) \Big(\ldots \\ \Big(\sum_{s_C=l_C^\star(p_C^\star)}^{u_C^\star(p_C^\star)} b(s_C; n_C, p_C) \Big) \ldots \Big) \Big). \end{aligned} \tag{27}$$

where the boundaries $l_c^\star(p_c^\star(s_1, \ldots, s_{c-1}))$ and $u_c^\star(p_c^\star(s_1, \ldots, s_{c-1}))$, $c = 1, \ldots, C$ are defined as in (25) and (26).

The implementation of equation (27) can be achieved by looping recursively over each dimension. By definition, only the probabilities of occurrence of default patterns which are contained in the acceptance region are calculated, summed up and then subtracted from 1. This yields a fast and efficient calculation of the exact p-value α_{Sterne}. Note that the computational effort of our procedures approaches the effort of a full enumeration when the deviation between the realized default pattern $\mathbf{m}$ and the default pattern with the highest probability of occurrence increases. To limit the calculation time a threshold p_{min} on the p-value may be introduced. Then the algorithm aborts when the actual p-value falls below p_{min} during the calculation. As a consequence, unlikely patterns with little contribution to the acceptance region need not to be computed. If such a threshold is implemented and the algorithm aborts then it is only known that the p-value is less than p_{min}. In case the p-value is greater than p_{min} the exact value is calculated.

3.2 A new Class of Hybrid Tests

Having dealt with efficient enumeration techniques applicable in exact testing we return to approximate tests as introduced in section (2.3). Remember that the possible bias of approximate tests arises from those risk classes where the approximation of the binomial distribution by a normal distribution cannot be deemed appropriate. Against this background we thus suggest an alternative calculation of the p-value α combining enumerative and approximate testing procedures. Without loss of generality, we assume that for the first a classes the approximation of the binomial distribution by a normal distribution does not hold but that for the remaining $C-a$ classes the binomial distributions can be well approximated by normal distributions. This leads to a joint distribution with binomial marginal distributions for the first a classes and normal marginal distributions for the remaining $C-a$ classes. We refer to the calibration quality tests carried out on this joint distribution as *hybrid tests*.

Using the law of total probability α_{Sterne} may be written as in equation (20). For the last $C-a$ classes the binomial probability mass function $b(\cdot; n_j, p_j)$ shall be approximated by a normal probability density function $h(\cdot; \mu_j, \sigma_j)$ as defined in (7), $j = a+1, \ldots, C$. Focusing on the term A in equation (20) and employing the approximations for classes $a+1, \ldots, C$, yields

$$\begin{aligned} A &= Pr\Big(\prod_{i=1}^{a} b(s_i; n_i, p_i) \prod_{j=a+1}^{C} \frac{1}{\sigma_j} \phi\left(\frac{S_j - \mu_j}{\sigma_j} \right) \\ &\qquad > \prod_{i=1}^{a} b(m_i; n_i, p_i) \prod_{j=a+1}^{C} \frac{1}{\sigma_j} \phi\left(\frac{m_j - \mu_j}{\sigma_c} \right) \Big) \end{aligned} \tag{28}$$

Using the definition of $\phi(\cdot)$, we obtain after some manipulations

$$= Pr\Big(\underbrace{\sum_{j=a+1}^{C} \left(\frac{S_j - \mu_j}{\sigma_j} \right)^2}_{X_a} < \underbrace{\sum_{j=a+1}^{C} \left(\frac{m_j - \mu_j}{\sigma_j} \right)^2 - 2 \sum_{i=1}^{a} (\ln b(m_i; n_i, p_i) - \ln b(s_i; n_i, p_i))}_{r_a} \Big). \quad (29)$$

Thus X_a follows a χ^2_{C-a} distribution with $C-a$ degrees of freedom and the value of A is given by

$$A = Pr(X_a < r_a | s_i,\ i = 1, \ldots, a), \quad X_a \sim \chi^2_{C-a}. \quad (30)$$

Therefore, the p-value α_{hybrid} of our hybrid test can be calculated as

$$\alpha_{hybrid} = 1 - \sum_{s_1=0}^{n_1} \cdots \sum_{s_a=0}^{n_a} b(s_1; n_i, p_i) \ldots b(s_a; n_i, p_i) \cdot$$

$$Pr(X_a < r_a | s_i,\ i = 1, \ldots, a) \quad (31)$$

where

$$X_a \sim \chi^2_{C-a} \quad (32)$$

and

$$r_a = \sum_{j=a+1}^{C} \left(\frac{m_j - \mu_j}{\sigma_j} \right)^2 - 2 \sum_{i=1}^{a} (\ln b(m_i; n_i, p_i) - \ln b(s_i; n_i, p_i)) \quad (33)$$

When calculating α_{hybrid} as defined in equation (31), we may utilize different approximations which may be combined with enumerative techniques. Hybrid p-values utilizing Score tests, Score tests with continuity correction and Wald-Agresti-Coull tests may be derived by choosing appropriate values for μ_j and σ_j, $j = a+1, \ldots, C$, as defined in section 2.2. To enumerate the first a classes efficiently, the algorithm proposed for the calculation of the multivariate Sterne test may be employed. The combination of enumerative and approximate techniques is computationally superior to pure enumerative procedures since the number of dimensions where an exact calculation is required is reduced from C to a.

At the same time, the result will not have a bias as strong as the one of a pure approximate test since for those classes where an approximation is not appropriate, the exact binomial distribution is used.

Note that our suggested hybrid procedure requires the tester to make one important decision. She has to define the classes for which an approximation is to be applied depending on her preferences in terms of accuracy versus speed. We propose to follow an often cited rule-of-thumb for the appropriateness of an approximation of a binomial distribution by a normal distribution and thus decide to approximate if $n_c p_c \geq 5$ and $n_c(1-p_c) \geq 5$ (i.e., see DeGroot and Schervish, 2002).

4 Performance Analysis

4.1 Scenario Analysis Utilizing Moody's data

In this section we employ a scenario analysis to analyze the performance of the proposed tests in terms of accuracy and computational effort. To allow for a realistic parametrization of our tests, we utilize data reported by Moody's Investor Service (2009). In our base case scenario, the distribution of obligors across risk classes is given by Moody's static pools as of 2008. The default pattern to be tested in the base case scenario is given by the expected default pattern. To determine the expected default pattern, we set the probability of default estimates for all risk classes equal to the average annual default rates reported for Moody's rating grades for the years 1970 to 2007. As the average default frequency for Aaa is 0% we correct the probability of default estimate for this risk class to 0.01%. The expected number of defaults e_c is obtained by rounding the expected values $n_c\hat{p}_c$ to the next smallest integer. Table 2 shows the sizes of the static pools (i.e., the number of obligors n_c), the probability of default estimates $\hat{p}_c$ and the expected number of defaults e_c for all risk classes of Moody's.

[Insert Table 2 here]

To study accuracy and computational effort for different settings, we modify the base case scenario by varying the observed number of defaults in each risk class. The different default patterns constituting the scenarios are presented in the form of deviations from the expected default pattern and are shown in table 3. Two factors (F1 and F2) are varied independently on three levels which yields nine scenarios plus the actually realized default pattern of Moody's static pools 2008 as an extra scenario. The first factor (F1) introduces no (0 defaults), moderate (+1 default) and strong (+2 defaults) deviations from the expected number of defaults in class Aa. The second factor (F2) introduces no (0 defaults), moderate (+10 defaults) and strong (+20 defaults) deviations from the expected number of defaults in classes B and C. For all hybrid tests the suggested rule-of-thumb for the approximation of a binomial distribution by a normal distribution that $n_c p_c \geq 5$ and $n_c(1-p_c) \geq 5$ is used. The application of this rule leads to normal approximations for the classes BB, B and C. Hence the first (second) factor represents the influence of deviations in those classes where the binomial distribution cannot (can) be approximated by a normal distribution.

[Insert Table 3 here]

The computational effort of the p-value of the exact Sterne and the proposed hybrid approaches are solely dependent on the endurance of the enumeration. To illustrate the computational expensiveness the number of runs through the innermost loop in the recursive algorithm is presented. For the exact test this is equal to the number of default patterns that lie within the acceptance region of the test for the calculated p-value.

The results of the scenario analysis are presented in tables 4 and 5. The abbreviations refer to the following tests respectively: "SCC" to the Score test with continuity correction and "WAC" to the Wald-Agresti-Coull test. The same abbreviations are used to address the hybrid versions of these tests.

[Insert Table 4 and 5 here]

For the interpretation of the results we assume that a significance level of 0.05 is desired by the statistician. We address the scenarios in the form of (F1/F2). In table 4 the

p-values of the Sterne test will be used as a benchmark for all other tests. Looking at the Score test with continuity corrected (SCC) we find that the p-values are equal to zero for all scenarios. As expected, this is due to the fact that the continuity correction is too strong when the probability of default is very close to 0 or 1 for any class, as it is the case for example in class Aaa with a probability of default of 0.01%. Hence the test yields extremely biased results. For the hybrid Score test with continuity correction (Hybrid SCC) no normal approximations are applied in the classes with low probability of default and thus the p-values for the hybrid Score test with continuity correction are in line with the p-values obtained by the Sterne and the hybrid Score test. Turning to the Wald-Agresti-Coull (WAC) test we find that the p-values deviate heavily from those of the exact Sterne test. By definition, this test is supposed to perform well only if the p-value is close to 0.05. The p-values of the hybrid Wald-Agresti-Coull (Hybrid WAC) test are closer to those of the Sterne test than those derived by the pure approximate version of the Wald-Agresti-Coull (WAC) test but they deviate stronger than the p-values of the other hybrid tests.

In the scenario which represents the actually realized number of defaults in 2008, the null hypothesis that the average historical probability of default does not differ significantly from the realized default frequency in 2008 is rejected by all tests. The influence of deviations in classes B and C with no or moderate deviations in the first class can be examined in scenarios (0/0), (0/1), (0/2), (1/0), (1/1) and (1/2). For these scenarios the test decisions of the Sterne, Score and the hybrid tests are consistent in each scenario; only the p-value of the hybrid Wald-Agresti-Coull test (Hybrid WAC) is slightly below 0.05 in scenario (0/2). The results show that the difference in p-values between the Sterne test and the Score test is greatest for small deviations from the expected pattern. In scenario (2/0) the Score test clearly rejects the null hypothesis while the Sterne, the hybrid Score and the hybrid Score tests with continuity correction (Hybrid SCC) yield p-values of around 15%. Note that in this scenario only strong deviations in the Aa class are introduced and that the Aa class cannot be approximated well by a normal distribution. This may bias the results of the Score test. A similar result is obtained for scenario (2/1) where the Score

test leads to a different test decision than the Sterne, the hybrid Score and the hybrid Score test with continuity correction (Hybrid SCC).

The number of runs through the through the innermost loop are presented in table 5. Pure approximate tests such as the Score test, Score test with continuity correction (SCC) and the Wald-Agresti-Coull test (WAC) require only the evaluation of the quantile of a χ^2 distribution. Thus for these tests the number of runs through the innermost loop is 0 for all default patterns. For the hybrid tests the results indicate that they require far less runs through the innermost loop of the recursion to calculate the p-value than the Sterne test. Hence hybrid tests are computationally superior to the exact Sterne test. When the probability of occurrence of the tested default pattern is low as for example in the scenario with the realized default frequency in 2008 the computation time of the exact Sterne test increases significantly. Note again the calculation time could have been limited by introducing a minimum threshold for the p-value (i.e., 1%) at which the recursive algorithm would abort. The interpretation of the result of the exact test would then be that the actual p-value is below the specified threshold but it would not be known exactly.

Summing up the results, we find that the Wald-Agresti-Coull (WAC) test yields the same test decisions as the Sterne test for most scenarios but the differences in p-values compared to the exact test may be very large. Furthermore, the Score test yields biased results, especially if strong deviations from the expected pattern are observed in risk classes with a very low probability of default. In practice, the Score test will thus be particularly problematic in the assessment of portfolios with a high proportion of obligors with good credit quality. The Score test with continuity correction (SCC) may not be applied to rating systems where the probability of default for any class is very close to zero. In the class of hybrid tests, the hybrid Wald-Agresti-Coull test (Hybrid WAC) shows the greatest deviations from the exact p-values. To conclude, the performance analysis provides strong indications that the hybrid Score test and the hybrid Score test with continuity correction (Hybrid SCC) outperform their competing procedures: they yield p-values differing only slightly from the Sterne p-values while their computational effort is much lower.

4.2 Further Robustness Checks Allowing for Different Portfolio Qualities and Sizes

In our base case portfolio, the distribution of obligors across risk classes is given by Moody's static pools as of 2008. To demonstrate the performance under more general conditions the size and the credit quality of the underlying portfolio are varied independently while a set of default scenarios, i.e. different default patterns, is applied to each of the portfolios. To assess the sensitivity of the proposed tests on the size of the underlying portfolio the number of obligors from the base case portfolio is manipulated. One out of three different levels for the total number of obligors n, $n/2$, $n/4$ is chosen which leads to a realistic portfolio size for commercial banks. To modify the distribution of obligors across risk classes obligors are shifted from good to bad classes or vice versa. To create an up shift in credit quality the distribution of obligors across classes is changed to

$$n_c^{up} = \begin{cases} n_c + \frac{n_{c+1}}{2}, & c = 1, \\ \frac{n_c}{2} + \frac{n_{c+1}}{2}, & c = 2, \ldots, C-1, \\ \frac{n_c}{2}, & c = C. \end{cases}$$

To obtain a down shift in credit quality we applied

$$n_c^{down} = \begin{cases} \frac{n_c}{2}, & c = 1, \\ \frac{n_c}{2} + \frac{n_{c-1}}{2}, & c = 2, \ldots, C-1, \\ n_c + \frac{n_{c-1}}{2}, & c = C. \end{cases}$$

Combining the three choices for both, portfolio sizes and shifts in credit quality, leads to nine unique portfolios on which the performance of the proposed tests is assessed. An overview of the numbers of obligors per risk class for each portfolio is given in table 6.

[Insert Table 6 here]

To each of the nine portfolios the nine default scenarios are applied. Note that the results for the first portfolio constituting the base case with no change in size and no shift in credit quality have already been discussed in detail in section 4.1 and that the realized

default pattern of Moody's 2008 is applied to this portfolio only. The resulting p-values for all other portfolios, scenarios and tests are presented in the appendix in tables 9 to 16. The number of runs through the innermost loop required for each calculation are given in tables 17 to 23.

To summarize the results we investigated how often the considered tests would lead to differing test decisions compared to the Sterne test for a given confidence level α in the 81 portfolio-default scenario combinations (3 portfolio sizes times 3 levels of credit quality times 9 default scenarios resulting from the variations of F1 and F2 = 81). The results of this analysis are summarized in table 7.

[Insert Table 7 here]

Additionally we study the absolute percentage deviations (APD) between the p-value of the Sterne test and the p-values derived by an alternative definition. The APD for an alternative test is defined as

$$APD_{alternative} = \frac{|\alpha_{Sterne} - \alpha_{alternative}|}{\alpha_{Sterne}}$$

The range of p-values below 10% will be of highest importance to practioners. Hence for the summary calculation we consider only those default scenarios where the p-value of the Sterne test is below 10% average the absolute percentage deviations for these combinations. The results are presented in table 8.

[Insert Table 8 here]

The key findings of the additional scenario analyses allowing for differences in portfolio size and credit quality are the following: Depending on the portfolio and the default scenario, the Score test may over- or underestimate the p-value compared to the exact Sterne test. This may lead to test decisions which do not coincide with those of the Sterne test.

The smaller the portfolio size, the more classes may not be approximated by normal distributions. As a consequence, the bias of the pure approximate tests increases as the number of obligors decreases. At the same time the results of the hybrid tests remain close

to the results of the exact Sterne test while their computational effort reduces further due to the smaller portfolio size.

The distribution of the obligors across risk classes also affects the performance of the tests. A high concentration of obligors in classes with good rating quality results in a strong bias of the pure approximate tests. Again, this effect is more pronounced for small portfolios. Both, the hybrid Score test and the hybrid Score test with continuity correction show p-values which are close to those of the Sterne test. The summary of differing test decisions presented in table 7 shows that the hybrid Score test has at most 1.23% differing test decisions while the hybrid Score test with continuity correction has at most 2.47% for a given p-value α. Looking at the average absolute percentage deviations summarized in table 8 we find that the deviation is smallest for the Hybrid Score test with 1.09%. Our additional analyses thus underline the robustness of the results already obtained in the previous section. Again the hybrid Score test outperforms the other procedures. Summing up, we thus recommend the hybrid Score test as it offers an excellent performance both, in terms of speed and accuracy.

5 Conclusion

In this paper we provide an overview of existing tests for calibration quality in an appropriate probabilistic framework. We discuss their advantages and drawbacks in terms of accuracy and computational effort. Based on the insights gained thereof an efficient algorithm for the calculation of the exact multivariate Sterne test is derived. Furthermore a new class of hybrid tests is developed which combines approximate and enumerative techniques. In a scenario analysis utilizing Moody's data we study the performance of the different testing procedures. Hybrid approaches outperform other existing and proposed procedures in terms of accuracy versus speed. Based on the results of an extensive scenario analysis we recommend using a hybrid Score test. It yields p-values closest to the ones of the exact Sterne test while at the same time being computationally much faster.

The testing procedures discussed in this paper together with the results obtained in an

extensive performance analysis are meant to provide guidelines for practical applications by risk managers. Furthermore, they may be of relevance for future policy decisions as regulators are given detailed information about advantages and disadvantages of individual testing procedures. Finally, the results of this paper should inspire researchers to further study the problem of testing calibration quality, e.g. by allowing for dependence in default events.

References

Agresti, A., Coull, B., 1998. Approximation is better than 'exact' for interval estimation of binomial proportions. The American Statistician 52, 119–126.

Basel Committee on Banking Supervision, 2005. Studies on the validation of internal rating systems. Working Paper 14.

Clopper, C., Pearson, S., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26, 404–413.

DeGroot, M., Schervish, M., 2002. Probability and Statistics, 3rd Edition. Addison-Wesley, Boston, San Francisco, New York.

Hosmer, D., Lemeshow, S., 1980. A goodness-of-fit test for the multiple logistic regression model. Communication in Statisitics - Theory and Methods 10, 1043–1069.

Hosmer, D., Lemeshow, S., 1988. Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. Biometrical Journal 8, 911–924.

Krahnen, J., Weber, M., 2001. Generally accepted rating principals: a primer. Journal of Banking and Finance 25, 3–23.

Moody's Investor Service, 2009. Corporate default and recovery rates, 1920-2008 (Excel data). Available from www.moodys.com.

Newcombe, R., 1998. Two-sided confidence intervals for the single proportion: a comparison of seven methods. Statistics in Medicine 12, 857–872.

Reiczigel, J., 2003. Confidence intervals for the binomial parameter: Some new considerations. Statistical Methods in Medical Research 22, 611–621.

Sterne, T., 1954. Some remarks on confidence or fiducial limits. Biometrika 41, 275–278.

Vollset, S., 1993. Confidence intervals for a binomial proportion. Statistics in Medicine 12, 809–824.

Wilson, E. B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22, 209–212.

Test	Exact	Central	Applicable to boundary outcomes	Multivariate extension possible
CP	Yes	Yes	Yes	No
Sterne	Yes	No	Yes	Yes
Score	No	Yes	Yes	Yes
SCC	No	Yes	Yes	Yes
Wald	No	Yes	No	Yes
WAC	No	Yes	Yes	Yes

Table 1: Overview of currently employed tests for a single risk class

Exact refers to the property that no approximation to the binomial distribution is used, *central* describes if the rejection regions of the test are symmetric around the median, *applicable to boundary outcomes* states if the test is well defined for cases when there are no defaults or all obligors in a risk class default, and *multivariate extension possible* shows if the definition of the test can be transfered to a setup with several dimensions. The abbreviations refer to the following tests: "CP" to the Clopper-Pearson test, "SCC" to the Score test with continuity correction, and "WAC" to the Wald-Agresti-Coull test.

	Rating classes						
	Aaa	Aa	A	Baa	Ba	B	C
n_c	182	795	1240	1138	590	1210	425
$\hat{p}_c$	0.01%	0.02%	0.02%	0.16%	1.03%	5.02%	21.41%
e_c	0	0	0	1	6	60	90

Table 2: Portfolio of the base case scenario based on Moody's data

Here n_c denotes the number of obligors per rating class of Moody's static pools as of 2008, $\hat{p}_c$ is the ex-ante estimated probability of default using the average annual default frequencies of Moody's static pools from 1970 to 2007 and e_c is the expected number of defaults derived by rounding $n_c p_c$ to the next smallest integer.

Scenario		Rating classes							Probability
F 1	F 2	Aaa	Aa	A	Baa	Ba	B	C	of occurrence
Realized		0	4	4	4	0	-36	-28	2.56-23
0	0	0	0	0	0	0	0	0	7.80e-05
0	1	0	0	0	0	0	10	10	2.00e-05
0	2	0	0	0	0	0	20	20	3.06e-07
1	0	0	1	0	0	0	0	0	1.24e-05
1	1	0	1	0	0	0	10	10	3.18e-06
1	2	0	1	0	0	0	20	20	4.87e-08
2	0	0	2	0	0	0	0	0	9.85e-07
2	1	0	2	0	0	0	10	10	2.52e-07
2	2	0	2	0	0	0	20	20	3.87e-09

Table 3: Default patterns employed in the scenario analysis

The default scenarios for validating the testing procedures are defined as deviations from the expected default pattern. The realized default pattern of Moody's static pools of 2008 is used as first scenario and additionally nine other scenarios are defined by varying two factors. Factor 1 (F1) and Factor 2 (F2) introduce deviations from the expected pattern in the AA as well as the B and C risk classes respectively. The last column represents the probability of occurence of the default pattern.

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
Realized		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0	0	1.0000	0.9974	0.0000	0.5711	0.9998	0.9997	0.8578
0	1	0.7204	0.8456	0.0000	0.2639	0.7366	0.7667	0.4910
0	2	0.0679	0.0923	0.0000	0.0215	0.0609	0.0736	0.0488
1	0	0.5996	0.6485	0.0000	0.4569	0.6033	0.6021	0.4061
1	1	0.2984	0.3606	0.0000	0.1974	0.3122	0.3346	0.1722
1	2	0.0187	0.0205	0.0000	0.0149	0.0167	0.0205	0.0128
2	0	0.1475	0.0026	0.0000	0.3559	0.1495	0.1491	0.0872
2	1	0.0596	0.0009	0.0000	0.1451	0.0634	0.0689	0.0306
2	2	0.0028	0.0000	0.0000	0.0102	0.0025	0.0031	0.0019

Table 4: P-values of Moody's static pools portfolio 2008 using different default scenarios and varying tests

The abbreviations refer to the following tests: "SCC" to the Score test with continuity correction and "WAC" to the Wald-Agresti-Coull test. The same abbreviations are used to address the hybrid versions of these tests.

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
Realized		3.45e+08	0.00e+00	0.00e+00	0.00e+00	3.30e+03	3.30e+03	9.03e+03
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	4.00e+00
0	1	8.26e+03	0.00e+00	0.00e+00	0.00e+00	4.00e+00	4.00e+00	7.00e+00
0	2	2.94e+05	0.00e+00	0.00e+00	0.00e+00	3.90e+01	3.90e+01	3.20e+01
1	0	1.60e+04	0.00e+00	0.00e+00	0.00e+00	8.00e+00	8.00e+00	1.40e+01
1	1	6.55e+04	0.00e+00	0.00e+00	0.00e+00	1.50e+01	1.50e+01	1.80e+01
1	2	6.80e+05	0.00e+00	0.00e+00	0.00e+00	6.60e+01	6.60e+01	5.80e+01
2	0	1.51e+05	0.00e+00	0.00e+00	0.00e+00	2.80e+01	2.80e+01	3.50e+01
2	1	3.24e+05	0.00e+00	0.00e+00	0.00e+00	3.90e+01	3.90e+01	4.60e+01
2	2	1.71e+06	0.00e+00	0.00e+00	0.00e+00	1.25e+02	1.25e+02	1.07e+02

Table 5: Number of runs through the innermost loop needed for the calculation of the p-value for different default scenarios and varying tests

The abbreviations refer to the following tests: "SCC" to the Score test with continuity correction and "WAC" to the Wald-Agresti-Coull test. The same abbreviations are used to address the hybrid versions of these tests.

Portfolio		Risk class						
Size	Shift	Aaa	Aa	A	Baa	Ba	B	C
full	no	182	795	1240	1138	590	1210	425
full	down	91	489	1017	1189	864	900	1030
full	up	580	1017	1189	864	900	818	212
1/2	no	91	398	620	569	295	605	212
1/2	down	45	244	509	595	432	450	515
1/2	up	290	509	595	432	450	408	106
1/4	no	46	199	310	284	148	302	106
1/4	down	23	123	254	297	216	225	257
1/4	up	146	254	297	216	225	204	53

Table 6: Portfolios for further robustness checks based on Moody's static pools 2008

The portfolio with full size and no shift in credit quality consists of Moody's static pools of 2008. The smaller portfolios where derived by reducing the number of obligors in each rating class to the half respectively the fourth of the original number of obligors in the portfolio. Then a shift function was employed to change the distribution of obligors across risk classes. This leads to an upshift or downshift in credit quality.

α	Method						
	Sterne	S	SCC	WAC	Hybrid S	Hybrid SCC	Hybrid WAC
1.0 %	0.00%	23.46%	58.02%	8.64%	1.23%	0.00%	3.70%
2.5 %	0.00%	16.05%	49.38%	13.58%	0.00%	1.23%	3.70%
5.0 %	0.00%	13.58%	43.21%	14.81%	1.23%	2.47%	6.17%
7.5 %	0.00%	14.81%	35.80%	14.81%	0.00%	0.00%	6.17%
10.0 %	0.00%	11.11%	35.80%	17.28%	1.23%	0.00%	6.17%

Table 7: Further robustness checks: Differing test decisions compared to the Sterne test for various confidence levels

The table shows the percentage of 81 portfolio-default scenario combinations where the test leads to a differing test decision than the Sterne test for the given confidence level α.

α	Method						
	Sterne	S	SCC	WAC	Hybrid S	Hybrid SCC	Hybrid WAC
Average APD	0.00%	42.20%	100.00%	51.73%	1.09%	4.64%	16.19%

Table 8: Further robustness checks: average absolute percentage deviations from the Sterne p-value

The table shows the average absolute percentage deviations (APD) for all portfolio-default scenario combinations where the Sterne p-value is less than 10%.

Appendix

Table 9: P-values for portfolio with full size and down shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	1.0000	0.9980	0.0000	0.5188	0.9984	0.9998	0.9957
0	1	0.6513	0.8329	0.0000	0.2329	0.6527	0.7017	0.6052
0	2	0.0687	0.0967	0.0000	0.0254	0.0532	0.0664	0.0939
1	0	0.3955	0.1272	0.0000	0.4094	0.3897	0.3970	0.3809
1	1	0.1700	0.0503	0.0000	0.1726	0.1713	0.1911	0.1511
1	2	0.0114	0.0020	0.0000	0.0176	0.0087	0.0110	0.0156
2	0	0.0517	0.0000	0.0000	0.3150	0.0509	0.0522	0.0494
2	1	0.0188	0.0000	0.0000	0.1258	0.0190	0.0216	0.0166
2	2	0.0010	0.0000	0.0000	0.0120	0.0007	0.0009	0.0013

Table 10: P-values for portfolio with full size and up shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	1.0000	0.9996	0.0000	0.4751	0.9999	0.9999	0.9875
0	1	0.4177	0.5871	0.0000	0.0998	0.4249	0.4865	0.4043
0	2	0.0036	0.0031	0.0000	0.0012	0.0020	0.0030	0.0081
1	0	0.6045	0.7215	0.0000	0.3721	0.6058	0.6057	0.5665
1	1	0.1490	0.2109	0.0000	0.0715	0.1525	0.1779	0.1333
1	2	0.0009	0.0006	0.0000	0.0008	0.0005	0.0007	0.0020
2	0	0.1585	0.0045	0.0000	0.2848	0.1596	0.1596	0.1438
2	1	0.0288	0.0006	0.0000	0.0507	0.0297	0.0353	0.0248
2	2	0.0001	0.0000	0.0000	0.0005	0.0001	0.0001	0.0003

Table 11: P-values for portfolio with $n/2$ size and no shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	1.0000	0.9936	0.0000	0.4109	0.9997	1.0000	0.9947
0	1	0.2819	0.4412	0.0000	0.0674	0.2824	0.3329	0.3154
0	2	0.0009	0.0007	0.0000	0.0005	0.0004	0.0006	0.0040
1	0	0.3121	0.0435	0.0000	0.3164	0.3116	0.3123	0.3051
1	1	0.0473	0.0050	0.0000	0.0476	0.0476	0.0582	0.0490
1	2	0.0001	0.0000	0.0000	0.0003	0.0000	0.0001	0.0005
2	0	0.0329	0.0000	0.0000	0.2384	0.0330	0.0331	0.0321
2	1	0.0039	0.0000	0.0000	0.0332	0.0039	0.0049	0.0039
2	2	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000

Table 12: P-values for portfolio with $n/2$ size and up shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	1.0000	0.9950	0.0000	0.4361	0.9985	0.9999	0.9963
0	1	0.1089	0.1631	0.0000	0.0267	0.0975	0.1343	0.1693
0	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
1	0	0.3810	0.1376	0.0000	0.4361	0.3804	0.3836	0.9963
1	1	0.0204	0.0046	0.0000	0.0267	0.0181	0.0258	0.1693
1	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
2	0	0.0543	0.0000	0.0000	0.3380	0.0542	0.0548	0.3753
2	1	0.0019	0.0000	0.0000	0.0185	0.0017	0.0025	0.0313
2	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 13: P-values for portfolio with $n/2$ size and down shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	1.0000	0.9926	0.0000	0.4406	0.9995	1.0000	0.9976
0	1	0.3191	0.5003	0.0000	0.0977	0.3131	0.3590	0.3345
0	2	0.0023	0.0016	0.0000	0.0020	0.0008	0.0012	0.0112
1	0	0.2242	0.0010	0.0000	0.3407	0.2237	0.2247	0.2201
1	1	0.0397	0.0001	0.0000	0.0696	0.0388	0.0466	0.0412
1	2	0.0002	0.0000	0.0000	0.0013	0.0001	0.0001	0.0009
2	0	0.0151	0.0000	0.0000	0.2572	0.0151	0.0152	0.0148
2	1	0.0021	0.0000	0.0000	0.0488	0.0021	0.0026	0.0023
2	2	0.0000	0.0000	0.0000	0.0009	0.0000	0.0000	0.0000

Table 14: P-values for portfolio with $n/4$ size and no shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	1.0000	0.9980	0.0000	0.3421	0.9972	0.9989	0.9963
0	1	0.0499	0.0909	0.0000	0.0119	0.0386	0.0587	0.1041
0	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
1	0	0.1893	0.0001	0.0000	0.2586	0.1899	0.1909	0.1893
1	1	0.0042	0.0000	0.0000	0.0081	0.0032	0.0050	0.0083
1	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0	0.0111	0.0000	0.0000	0.1913	0.0112	0.0112	0.0111
2	1	0.0002	0.0000	0.0000	0.0054	0.0001	0.0002	0.0003
2	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 15: P-values for portfolio with $n/4$ size and up shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	1.0000	0.9995	0.0000	0.3135	0.9983	0.9991	0.9774
0	1	0.0044	0.0042	0.0000	0.0016	0.0017	0.0037	0.0234
0	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0	0.2261	0.0020	0.0000	0.2359	0.2271	0.2277	0.2108
1	1	0.0004	0.0000	0.0000	0.0010	0.0001	0.0003	0.0021
1	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0	0.0167	0.0000	0.0000	0.1739	0.0169	0.0170	0.0154
2	1	0.0000	0.0000	0.0000	0.0007	0.0000	0.0000	0.0001
2	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 16: P-values for portfolio with $n/4$ size and down shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	1.0000	0.9992	0.0000	0.3121	0.9994	0.9993	0.9839
0	1	0.0639	0.1126	0.0000	0.0180	0.0432	0.0650	0.1301
0	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004
1	0	0.1334	0.0000	0.0000	0.2334	0.1341	0.1341	0.1267
1	1	0.0036	0.0000	0.0000	0.0122	0.0024	0.0037	0.0073
1	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0	0.0048	0.0000	0.0000	0.1704	0.0049	0.0049	0.0045
2	1	0.0001	0.0000	0.0000	0.0082	0.0001	0.0001	0.0002
2	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 17: # runs for portfolio with full size and down shift in credit quality

Scenario		Method						
F1	F2	Sterne	S	SCC	WAC	Hybrid S	Hybrid SCC	Hybrid WAC
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	1.00e+00
0	1	1.50e+04	0.00e+00	0.00e+00	0.00e+00	5.00e+00	5.00e+00	4.00e+00
0	2	3.40e+05	0.00e+00	0.00e+00	0.00e+00	3.10e+01	3.10e+01	1.70e+01
1	0	4.88e+04	0.00e+00	0.00e+00	0.00e+00	1.10e+01	1.10e+01	1.10e+01
1	1	1.53e+05	0.00e+00	0.00e+00	0.00e+00	1.70e+01	1.70e+01	1.60e+01
1	2	1.03e+06	0.00e+00	0.00e+00	0.00e+00	6.40e+01	6.40e+01	3.80e+01
2	0	4.19e+05	0.00e+00	0.00e+00	0.00e+00	3.20e+01	3.20e+01	3.20e+01
2	1	7.84e+05	0.00e+00	0.00e+00	0.00e+00	4.50e+01	4.50e+01	4.00e+01
2	2	3.02e+06	0.00e+00	0.00e+00	0.00e+00	1.23e+02	1.23e+02	8.50e+01

Table 18: # runs for portfolio with full size and up shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	1.00e+00
0	1	2.13e+04	0.00e+00	0.00e+00	0.00e+00	1.10e+01	1.10e+01	6.00e+00
0	2	9.64e+05	0.00e+00	0.00e+00	0.00e+00	1.19e+02	1.19e+02	5.50e+01
1	0	8.64e+03	0.00e+00	0.00e+00	0.00e+00	6.00e+00	6.00e+00	6.00e+00
1	1	8.53e+04	0.00e+00	0.00e+00	0.00e+00	2.20e+01	2.20e+01	1.50e+01
1	2	1.71e+06	0.00e+00	0.00e+00	0.00e+00	1.72e+02	1.72e+02	8.40e+01
2	0	8.00e+04	0.00e+00	0.00e+00	0.00e+00	2.10e+01	2.10e+01	2.10e+01
2	1	3.16e+05	0.00e+00	0.00e+00	0.00e+00	4.90e+01	4.90e+01	3.80e+01
2	2	3.38e+06	0.00e+00	0.00e+00	0.00e+00	2.63e+02	2.63e+02	1.44e+02

Table 19: # runs for portfolio with $n/2$ size and no shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	2.00e+00
0	1	9.51e+03	0.00e+00	0.00e+00	0.00e+00	4.00e+01	4.00e+01	1.90e+01
0	2	3.55e+05	0.00e+00	0.00e+00	0.00e+00	6.64e+02	6.64e+02	2.34e+02
1	0	8.20e+03	0.00e+00	0.00e+00	0.00e+00	3.40e+01	3.40e+01	3.50e+01
1	1	5.05e+04	0.00e+00	0.00e+00	0.00e+00	1.21e+02	1.21e+02	8.60e+01
1	2	7.40e+05	0.00e+00	0.00e+00	0.00e+00	1.13e+03	1.13e+03	4.76e+02
2	0	6.41e+04	0.00e+00	0.00e+00	0.00e+00	1.42e+02	1.42e+02	1.43e+02
2	1	1.96e+05	0.00e+00	0.00e+00	0.00e+00	3.29e+02	3.29e+02	2.61e+02
2	2	1.60e+06	0.00e+00	0.00e+00	0.00e+00	1.99e+03	1.99e+03	9.72e+02

Table 20: # runs for portfolio with $n/2$ size and up shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	1.00e+00
0	1	1.79e+04	0.00e+00	0.00e+00	0.00e+00	9.20e+01	9.20e+01	4.20e+01
0	2	1.16e+06	0.00e+00	0.00e+00	0.00e+00	3.39e+03	3.39e+03	6.87e+02
1	0	3.84e+03	0.00e+00	0.00e+00	0.00e+00	2.70e+01	2.70e+01	1.00e+00
1	1	6.00e+04	0.00e+00	0.00e+00	0.00e+00	2.34e+02	2.34e+02	4.20e+01
1	2	1.91e+06	0.00e+00	0.00e+00	0.00e+00	4.77e+03	4.77e+03	6.87e+02
2	0	3.14e+04	0.00e+00	0.00e+00	0.00e+00	1.33e+02	1.33e+02	2.80e+01
2	1	1.93e+05	0.00e+00	0.00e+00	0.00e+00	5.60e+02	5.60e+02	1.20e+02
2	2	3.34e+06	0.00e+00	0.00e+00	0.00e+00	7.04e+03	7.04e+03	1.17e+03

Table 21: # runs for portfolio with $n/2$ size and down shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	1.00e+00
0	1	1.16e+04	0.00e+00	0.00e+00	0.00e+00	3.30e+01	3.30e+01	2.20e+01
0	2	3.45e+05	0.00e+00	0.00e+00	0.00e+00	5.74e+02	5.74e+02	1.74e+02
1	0	1.87e+04	0.00e+00	0.00e+00	0.00e+00	4.80e+01	4.80e+01	4.80e+01
1	1	8.25e+04	0.00e+00	0.00e+00	0.00e+00	1.38e+02	1.38e+02	1.08e+02
1	2	8.66e+05	0.00e+00	0.00e+00	0.00e+00	1.09e+03	1.09e+03	4.14e+02
2	0	1.46e+05	0.00e+00	0.00e+00	0.00e+00	2.12e+02	2.12e+02	2.12e+02
2	1	3.58e+05	0.00e+00	0.00e+00	0.00e+00	4.21e+02	4.21e+02	3.36e+02
2	2	2.13e+06	0.00e+00	0.00e+00	0.00e+00	2.07e+03	2.07e+03	9.52e+02

Table 22: # runs for portfolio with $n/4$ size and no shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	1.00e+00
0	1	8.64e+03	0.00e+00	0.00e+00	0.00e+00	4.70e+01	4.70e+01	1.70e+01
0	2	4.86e+05	0.00e+00	0.00e+00	0.00e+00	1.70e+03	1.70e+03	2.67e+02
1	0	2.72e+03	0.00e+00	0.00e+00	0.00e+00	1.80e+01	1.80e+01	1.90e+01
1	1	3.35e+04	0.00e+00	0.00e+00	0.00e+00	1.27e+02	1.27e+02	6.10e+01
1	2	8.82e+05	0.00e+00	0.00e+00	0.00e+00	2.52e+03	2.52e+03	4.99e+02
2	0	2.08e+04	0.00e+00	0.00e+00	0.00e+00	8.00e+01	8.00e+01	8.30e+01
2	1	1.13e+05	0.00e+00	0.00e+00	0.00e+00	3.19e+02	3.19e+02	1.89e+02
2	2	1.65e+06	0.00e+00	0.00e+00	0.00e+00	3.92e+03	3.92e+03	9.60e+02

Table 23: # runs for portfolio with $n/4$ size and up shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	1.00e+00
0	1	2.31e+04	0.00e+00	0.00e+00	0.00e+00	1.90e+02	1.90e+02	4.40e+01
0	2	2.00e+06	0.00e+00	0.00e+00	0.00e+00	1.22e+04	1.22e+04	1.18e+03
1	0	1.49e+03	0.00e+00	0.00e+00	0.00e+00	1.80e+01	1.80e+01	1.90e+01
1	1	6.16e+04	0.00e+00	0.00e+00	0.00e+00	3.85e+02	3.85e+02	1.23e+02
1	2	2.93e+06	0.00e+00	0.00e+00	0.00e+00	1.56e+04	1.56e+04	1.83e+03
2	0	1.17e+04	0.00e+00	0.00e+00	0.00e+00	8.40e+01	8.40e+01	8.40e+01
2	1	1.60e+05	0.00e+00	0.00e+00	0.00e+00	7.87e+02	7.87e+02	3.09e+02
2	2	4.48e+06	0.00e+00	0.00e+00	0.00e+00	2.07e+04	2.07e+04	2.96e+03

Table 24: # runs for portfolio with $n/4$ size and down shift in credit quality

Scenario		Method						
F1	F2	Sterne	Score	SCC	WAC	Hybrid Score	Hybrid SCC	Hybrid WAC
0	0	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.00e+00	1.00e+00	1.00e+00
0	1	1.11e+04	0.00e+00	0.00e+00	0.00e+00	5.00e+01	5.00e+01	1.60e+01
0	2	4.52e+05	0.00e+00	0.00e+00	0.00e+00	1.49e+03	1.49e+03	1.74e+02
1	0	5.95e+03	0.00e+00	0.00e+00	0.00e+00	2.40e+01	2.40e+01	2.40e+01
1	1	5.24e+04	0.00e+00	0.00e+00	0.00e+00	1.54e+02	1.54e+02	7.20e+01
1	2	9.32e+05	0.00e+00	0.00e+00	0.00e+00	2.37e+03	2.37e+03	3.95e+02
2	0	4.59e+04	0.00e+00	0.00e+00	0.00e+00	1.16e+02	1.16e+02	1.16e+02
2	1	1.93e+05	0.00e+00	0.00e+00	0.00e+00	3.98e+02	3.98e+02	2.36e+02
2	2	1.93e+06	0.00e+00	0.00e+00	0.00e+00	3.89e+03	3.89e+03	8.46e+02