EBA STAFF PAPER SERIES

N.24 – 4/2026

SYSTEMATIC BACKTESTING OF PROBABILITY OF DEFAULT MODELS WITH REGULATORY DATA

METHODOLOGICAL ADVANCES AND EMPIRICAL INSIGHTS FROM EUROPEAN REGULATORY DATA

by Simone Casellina, Gaetano Chionsini, Raphael M. Kopp, and Maroua Riabi

ABSTRACT

Internal ratings-based models play a central role in bank risk management and regulatory capital determination, yet their validation remains methodologically challenging and operationally resource-intensive. In this paper, we contribute to the quantitative validation of probability of default models through a systematic backtesting exercise using a new proprietary dataset collected by the European Banking Authority between 2017 and 2024. We propose a generalised correction to the canonical binomial test that simultaneously accounts for both asset and serial correlation and is supported by extensive simulations. Acknowledging the iterative nature of model validation, we use order statistics to identify persistent miscalibrations over time. We present an approach to aggregate the results of backtesting procedures, which are typically designed for bank-level evaluation, whereas our focus is to provide evidence on the performance of the models across EU banks. Empirically, we find that the share of miscalibrated exposures of the small and medium-sized enterprises corporates asset class ranges from around 3.0% under realistic assumptions to a conservative upper bound of 16.7% implied by the canonical binomial test. We also quantify the impact on capital requirements and show that prudent model recalibrations would reduce system-wide Tier 1 capital ratios by 4 to 10 basis points. By offering scalable backtesting tools and enhancing transparency, we support more effective supervisory oversight and contribute to restoring market confidence in internal models.

KEYWORDS
Credit Risk; PD; Backtesting; Model Validation; Asset Correlation; Serial Correlation

JEL CODES
C15; C18; G21; G32

1. Introduction

Amid ongoing discussions on enhancing the European Union's (EU) competitiveness, policymakers have increasingly emphasised the importance of streamlining the regulatory framework without compromising financial stability (Berg et al. 2025; Buch 2025; Restoy 2025). Recent landmark reports such as those by Draghi (2024) and Letta (2024) highlight regulatory simplification as an important lever to promote innovation and economic growth in the EU. Importantly, simplification in this context is not synonymous with deregulation but rather aims to achieve regulatory objectives more efficiently (Villeroy de Galhau 2025). This perspective underlines the challenge of striking a balance between rigorous oversight and the need to enhance competitiveness. Complementing these efforts, regulators are increasingly promoting a data-centric culture and the use of advanced technologies to strengthen analytical capabilities and improve supervisory efficiency, risk identification, and responsiveness (Elderson 2024).

One area in which this balance is particularly important is model validation as a critical, albeit resource-intensive, component of every credit institution's internal risk management framework and supervisory processes. According to the Basel Committee on Banking Supervision (BCBS 2005b), model validation examines whether the internal rating systems effectively differentiate risk and whether the estimates of risk parameters—such as probability of default (PD), loss given default (LGD), and exposure at default (EAD)—accurately capture the relevant risk characteristics. Article 185 of the (EU) No 575/2013 Capital Requirements Regulation (CRR) requires regular validation of internal ratings-based (IRB) models, which includes both qualitative (e.g., model design, documentation, governance, and management oversight) and quantitative (e.g., backtesting and benchmarking) aspects.[1] While qualitative validation focuses on 'how a model works', quantitative validation addresses 'whether a model works'.

Traditional supervisory validation practices often rely on resource-intensive, institution-specific onsite inspections that are supplemented by quantitative backtesting.[2,3] In this context, backtesting refers to the comparison of a model-based forecast or expectation (e.g., PD) with the ex-post realisation of the outcome (e.g., default rate (DR)). While these practices are thorough and essential from a supervisory perspective, they also present challenges: they are time-consuming, lack scalability, and their results are rarely publicly disclosed.[4] This opacity limits the ability of a wider audience—including market participants, policymakers, and academics—to understand whether internal models are working as intended. In addition, backtesting is typically conducted at the individual model and bank level, which is essential for risk modelling and banking supervision, but does not easily allow for a sector-wide perspective. An aggregated view is necessary to inform

[1] For a detailed description of backtesting and benchmarking, we refer to Castermans et al. (2010).
[2] We refer to European Central Bank (ECB 2024) for a detailed guide to internal models.
[3] Recent supervisory opinions increasingly emphasise the need for more efficient validation processes, including risk-based approaches and streamlined assessments for limited model changes (Comfort and Arons 2023; Noonan and Comfort 2025).
[4] It is worth noting that the current Pillar 3 framework requires certain disclosures comparing predicted PDs with observed DRs. In particular, 'Template CR9 –IRB approach – Back-testing of PD per exposure class (fixed PD scale)' provides a structured basis that can be used to replicate the analyses presented in this study (see Final Draft Annex I of the Implementing Technical Standards on Institutions' Public Disclosures of the Information Referred to in Titles II and III of Part Eight of Regulation (EU) No 575/2013)).

macroprudential oversight and financial stability discussions. Additional aggregation and synthesis are therefore needed to bridge the gap between micro-level supervision and broader systemic risk considerations.

Against this background, our paper examines whether systematic, data-driven backtesting using regulatory datasets can serve as a reliable and cost-effective complement to traditional supervisory practices, with the aim of informing a broader audience about the reliability of EU banks' internal models.[5] Our paper also responds directly to recent calls for higher reliance on backtesting. For example, Cannata and Serafini (2025, p. 30) suggest that assessing the "effectiveness of models based on actual results" is a valuable complement to resource-intensive on-site inspections, which are inherently constrained by supervisory capacity and pose the risk of promoting a mere tick-the-box approach. The scope of the paper is specifically focused on credit risk model validation, with a particular emphasis on PD models. This distinction is important because banks use a variety of other models that are also subject to backtesting, including models for market risk, counterparty credit risk (IMM), Repo-VaR, and derivative pricing.

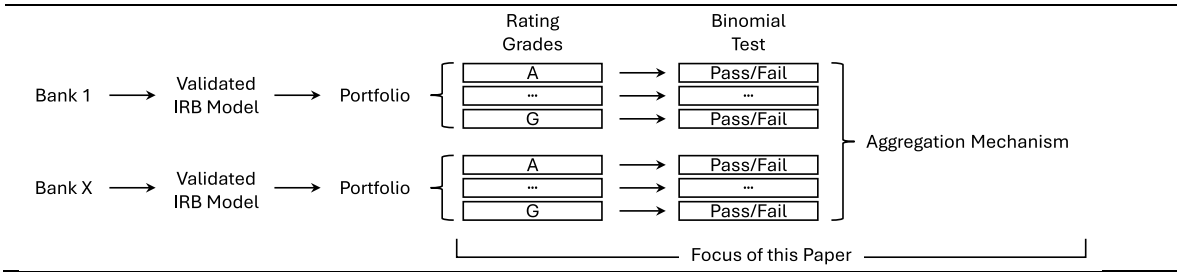
To operationalise this idea, we use variations of the binomial test at the rating grade level for each bank in our sample. This is consistent with the way banks typically assign PDs—at the obligor and rating grade levels—to ensure homogeneity of risk profiles within each grade.

Figure 1 shows a stylised example of our setting. Our approach focuses on potential miscalibrations at the rating grade level rather than at the portfolio level. Specifically, we perform separate binomial tests for each rating grade within a bank's portfolio to detect potential miscalibrations. To complement this with an aggregated view, we introduce an aggregation methodology to quantify the miscalibrated exposures across EU banks. Although global (model) calibration tests (e.g., Hosmer-Lemeshow) can assess model calibration across all rating grades simultaneously, our approach targets weaknesses at the rating grade level to determine whether the internal models are working as intended. Our approach can be considered more conservative as it identifies potential miscalibrations in rating grades, even if the models themselves are prudently calibrated overall. Therefore, we do not automatically conclude that a model is imprudently calibrated, as this requires additional (qualitative) considerations. Our methodology offers scalability, comparability across banks and time periods, and the ability to flag cases that warrant further supervisory scrutiny. By improving transparency and providing insights at both bank and aggregate level, we aim to contribute to effective supervisory oversight and restore market confidence in the use of internal models.

Our setting contributes to the ongoing discussion on supervisory efficiency by offering novel insights into model validation through backtesting from three perspectives: (A) from a data perspective through the use of a new, proprietary regulatory dataset across banks; (B) from an empirical perspective through the application of the framework to a sample of corporate small and medium-sized enterprises (SME) exposures from EU IRB banks; and (C) from a methodological perspective through extensions of the canonical binomial test.

[5] From a practical perspective, the idea of systematic backtesting is used, e.g., in the context of the Eurosystem Credit Assessment Framework (ECAF) to mitigate the credit risk of collateral used in monetary policy operations. On an annual basis, all credit assessment systems accepted for ECAF purposes (e.g., IRB systems) are subject to backtesting. For an overview of the ECAF, we refer to Guideline (EU) 2015/510 and for a description of the financial risk management of the Eurosystem's monetary policy operations to ECB (2015).

Figure 1: Simple Stylised Example of Our Setting



(A) Data perspective. Under the supervisory reporting framework, banks that use validated IRB models are required to submit detailed information to their supervisory authorities via template C 08.05 of the Common Reporting Framework (COREP). This includes data on the number of obligors, PDs, and observed DRs, disaggregated by asset class and rating grade, and mapped to a common master scale with fixed PD ranges. This reporting requirement introduced by the European Banking Authority (EBA) aims to facilitate the supervisory monitoring of institutions' compliance with the CRR. The EBA has been collecting this proprietary regulatory data since 2017, providing a solid basis for large-scale cross-sectional and longitudinal model validation analyses, including backtesting of PD models under the IRB approach applied by EU banks. We utilise this novel proprietary dataset in our empirical exercise.

(B) Empirical perspective. Our empirical contributions are twofold: (i) we develop a consistent framework for model validation and aggregation across institutions, and (ii) we assess the economic impact of model miscalibrations in terms of capital adequacy:

(i) We conduct comprehensive bottom-up backtesting, analysing model calibrations both at the individual bank level and at the aggregate level across EU banks. The aggregate view is essential from a macroprudential and financial stability perspective, while the bank level perspective is essential from a risk modelling and banking supervision perspective. To bring these perspectives together, we propose a robust aggregation method to transform granular backtesting results into an informative, system-wide representation of model calibration performance. We estimate that the proportion of miscalibrated corporate SME exposures in EU banks is between an upper bound of 16.7% using the standard binomial test and between 2.9% and 3.5% using a more realistic scenario that takes into account both asset and serial correlation. Importantly, we observe a decreasing trend in the share of miscalibrated exposures in recent years, which is likely due to improvements in data quality and the inclusion of higher margins of conservatism in the model calibrations.

(ii) We quantify the economic impact of miscalibrated corporate SME PD models by assessing their impact on risk-weighted assets (RWAs) and Tier 1 capital ratios. This extension is particularly important for regulatory purposes, as capital adequacy ratios must reflect banks' actual risk exposure. We introduce metrics that link miscalibrated corporate SME exposures to changes in RWAs and Tier 1 capital ratios by capturing the difference between reported PDs and the minimum PDs required for prudent model calibration. Prudently re-calibrating PD models that fail backtesting would imply a reduction in the Tier 1 capital ratios of 3.8 to 10.3 basis points at the system level, with the upper bound again based on the standard binomial test. Over time, a generally stable to slightly decreasing trend in these effects can be observed, which is consistent with the observed decrease in miscalibrated exposures and the simultaneous increase in the Tier 1 capital ratios. Taken together, these developments are consistent with a broad-based increase in the conservatism of the calibration of credit risk models.

(C) Methodological perspective. We contribute a generalised correction to the canonical binomial test. We propose a theoretically based generalised correction of the well-established binomial test that simultaneously corrects for asset and serial correlation.[6] The classical binomial test, which is widely used to validate PD models, works under the assumption of independent defaults (Blümke 2022b). Consequently, it tends to over-reject the null hypothesis of prudent model calibration, leading to an inflated type I error rate. Although this conservative bias may be less concerning from a supervisory standpoint than under-rejection, excessive over-rejection can trigger unnecessary model re-calibrations, potentially undermining model stability and increasing costs for both banks and supervisors. As the estimated risk parameters have a direct impact on RWAs and therefore capital requirements, it is essential to assess the extent to which the lack of independence of DRs affects backtesting results. While the existing literature and regulatory discussions focus predominately on cross-sectional correlation (i.e., asset correlation), DRs also exhibit notable temporal dependence (i.e., serial correlation) (Blümke 2022a)—the existing literature is largely silent on the impact of serial correlation. Our generalised correction fills this gap by accounting for both types of correlation. We underpin our approach with an extensive simulation study to demonstrate its effectiveness.

Related Work

Our study builds on the broader model validation literature with a particular emphasis on quantitative model validation through backtesting with one of the most commonly employed tests for validating PDs—the binomial test.[7,8] The binomial test works under the restrictive assumption of independent defaults, which leads to an inflated type I error rate, resulting in an overestimation of the significance of deviations between the realised DRs and estimated PDs (Tasche 2005; Miu and Ozdemir 2008; Blümke 2013). For this reason, several tests have been proposed in the literature to overcome this limitation.

One of the earliest modifications was proposed by Tasche (2003), who introduced a traffic light approach in conjunction with a methodology that adjusts the confidence boundaries of PDs taking into account a correlation parameter. In addition, a correlated binomial test that corrects for asset correlation was introduced by Balthazar (2004) and further elaborated by Tasche (2005). Other modifications include the integration of the one-factor model from the Basel framework proposed by Blümke (2012) and the novel statistical approach of Schechtman (2017), which focuses primarily on the risk of accepting incorrectly calibrated credit risk models. The proposed approach addresses the risk of validating misspecified credit risk models and presents an asymptotic analytical framework for joint testing of multiple PDs under the assumption of default correlation.[9]

[6] We refer to Conover (1999) and to BCBS (2005a) for a detailed discussion of the standard binomial test.

[7] A non-exhaustive list of other commonly used PD validation methods includes the level test (Blöchlinger and Leippold 2011, 2018), maximum likelihood estimators (Miu and Ozdemir 2008), the Brier score (Brier 1950), and the Hosmer-Lemeshow test (Hosmer and Lemeshow 2000).

[8] Another commonly used approach in the literature for PD backtesting is the Jeffreys test (ECB 2019). The Jeffreys test, like the binomial test, compares forecasted defaults with observed defaults within a binomial framework, either at portfolio level or at rating grade level. The p-value of the Jeffreys test is given by the cumulative distribution function (CDF) of the Beta distribution: $F_{Beta}(PD; \alpha, \beta)$, where the shape parameters are $\alpha = D + 0.5$ and $\beta = N - D + 0.5$, with *PD* being the probability of default, *D* the number of observed defaults, and *N* the total number of exposures. The Jeffreys test and binomial test generally yield similar results, particularly for large *N*. In Appendix Part B, we show a comparison between the binomial test and the Jeffreys test, which leads to identical results, by replicating our Monte Carlo simulations with known parameters, which we show in Section 3.3, Table 9.

[9] For a detailed overview of the general impact of asset correlations in credit risk, we refer to Global Credit Data (2025). For a review of different rating philosophies, we refer to Carlehed and Petrov (2012).

Another sub-strand of the literature extends the model validation based on one-year observations to multi-period tests with data from several years. These tests are used to determine whether the observed DRs align with the long-run average PDs. Contributions that address the issue of multi-period model validation include the normal test (Tasche 2005), the extension of the traffic light approach to the normal approximation of the binomial test (Blochwitz et al. 2006), and the inclusion of the one-factor model in an order statistic (Blümke 2012).

All of the one-year and multiyear extensions described above focus primarily on solving the problem of asset correlation, but do not account for serial correlation. It is well-known from the literature that DRs do not change rapidly but exhibit a high degree of persistence (Blümke 2022a). A first attempt to tackle the problem of serial correlation was recently proposed by Blümke (2025); the proposed approach deals with the modelling of DRs considering serial correlations by using a Vasicek distribution.

2. Real-world Systematic Backtesting Exercise

In this section, we present a backtesting exercise of PD models for a sample of EU banks building on the most commonly employed tests for validating PDs—the binomial test. Thereby, we show the number of miscalibrated rating grades and miscalibrated exposures, as well as well as the economic impact on the banks' Tier 1 capital ratios. The results obtained with the binomial test can be interpreted as a prudent upper bound, which is preferable from a prudential point of view compared to overly permissive results. However, in Section 3.5 we extend our analysis to more realistic conditions by considering both asset and serial correlation through our proposed generalised correction to the binomial test.

2.1 Modelling Framework

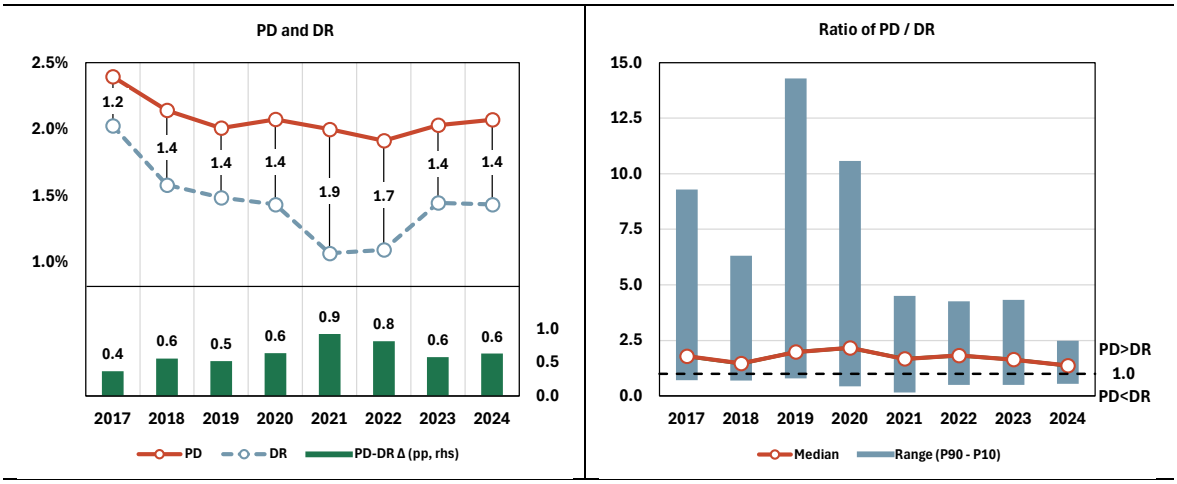
In what follows, we briefly formalise the need for formal backtesting, describe the well-established binomial test and aggregation mechanism, and outline our approach to assessing the economic impact of miscalibrations.

Need for Formal Backtesting

Backtesting procedures are used to assess the reliability of the calibration of statistical models, e.g., credit risk models used to estimate PDs. A simple starting point is to compare the EAD weighted average PDs reported by different banks for a given asset class with the corresponding EAD weighted average annual DRs. Figure 2 shows this comparison over time for all banks included in our empirical exercise in Section 2.3. The results (left panel) show that the reported PDs at rating grade level are generally significantly higher than the observed defaults rates—on average by a factor of 1.5, with a range between 1.2 and 1.9. The right panel complements this view by showing the median and dispersion (P90 – P10) of the ratio between PD and DR at bank level. The average lower bound (P10) is around 0.5, the average upper bound (P90) is around 7.0, and the average median is around 1.7. This more nuanced picture highlights that the ratio falls below one for some banks, which suggests possible issues with undercalibrations. Note that miscalibrations can occur

in both directions. In general, a well-calibrated model should exhibit neither upward nor downward bias. We focus on identifying cases of undercalibrations, i.e., we conduct one-sided tests, because this is of primary interest to supervisors and it facilitates the aggregation of results across banks. However, an unduly conservative PD may adversely affect bank's profitability and competitiveness and impair their ability to provide funding efficiently to the real economy. In what follows, we use the term miscalibration interchangeably to refer to undercalibrations.

Figure 2: Simple Comparison of the Weighted Average PD and the Weighted Average DR



Note. This figure shows a simple comparison of the EAD weighted average PD and the EAD weighted average DR over time for all banks in our sample. The values on the vertical segments (left panel) indicate the factor by which the PD exceeds the DR.

However, this simple comparison has a decisive limitation from a supervisors' perspective. On the one hand, if the ratio were to consistently exceed one over a prolonged period (e.g., for several consecutive years), it would provide an unambiguous signal that PD estimates are, on average, overly conservative. On the other hand, by averaging across banks, conservative PD estimates are implicitly offset against less conservative ones. Such offsetting is problematic, as the capital requirements for one bank cannot be used to offset the risks of another bank. Additionally, the PD estimates reported for regulatory purposes are not intended to predict DRs in a given year, but rather to reflect long-run average PDs. Consequently, a direct year-to-year comparison between reported PDs and observed DRs is insufficient for a sound assessment of model calibration. To address these shortcomings, in the following section we introduce the binomial test as a formal backtesting procedure.

Binomial test

A natural approach to assessing the reliability of the calibration of a statistical model is to reduce this assessment to a binary decision: (i) accepting the null hypothesis H_0, which states that a given model is prudently calibrated, or (ii) rejecting the null hypothesis H_0, which states that the model is not prudently calibrated (BCBS 2005a). In the context of credit risk models, this decision is made by comparing the observed DRs for a given rating grade at a specific point in time with a pre-specified threshold. If the observed DR exceeds this threshold, the PD is underestimated and H_0 is rejected.

For this reason, the ex-ante specification of the threshold is crucial and depends on three components: (i) a statistical model, (ii) the estimated parameter (e.g., PD), and (iii) a chosen confidence

level. According to the binomial test, H_0 is rejected for a given confidence level q and a given number of obligors n if the DR for a rating grade is greater than or equal to the critical value (i.e., threshold), which is defined as (BCBS 2005a):

$$p^*_{\text{binomial}} \approx \Phi^{-1}(q)\sqrt{\frac{PD(1-PD)}{n}} + PD, \tag{1}$$

where Φ^{-1} denotes the inverse standard normal distribution function. The binomial test is based on the assumption of independent defaults—an assumption that is rarely true in practice. This implies that the test is known to inflate the type I error rate and is therefore overly conservative, as the null hypothesis H_0 is rejected more frequently than the nominal significance level α would suggest. In this paper, we set $\alpha = 5\%$ meaning that it is acceptable to reject H_0 with a probability of 5% if the model is correctly calibrated. Therefore, our results can be interpreted as an upper bound for potential model miscalibrations.

From a supervisory perspective, over-rejection of well-calibrated models (i.e., being overly conservative), although safer than under-rejection, can still pose practical challenges. For example, frequent re-calibrations triggered by false positives could affect model stability over time and lead to more frequent interactions between banks and supervisors. In Section 3, we propose a generalised correction of the binomial test that simultaneously corrects for asset and serial correlation and provides a more reliable framework for backtesting under realistic conditions.

Aggregation Mechanism

As our analysis includes model validation both at the individual bank level and in aggregate across EU banks, an aggregation mechanism is required to synthesise the results. The aggregate view is particularly important from a macroprudential and financial stability perspective, where a system-wide view is essential. This section describes the methodology used to aggregate the results at bank level into a coherent and informative representation at the system level.

The binomial test is applied at the rating grade level, where a binary decision is made: (i) accepting the null hypothesis H_0, which states that the risk model is prudently calibrated, or (ii) rejecting H_0, which states that the risk model is not prudently calibrated (BCBS 2005a). To formalise this decision process, we define an indicator variable $\mathcal{M}_g$ that indicates whether a particular rating grade g is prudently calibrated:

$$\mathcal{M}_g = \begin{cases} 0, & \text{if } PD_g \text{ is prudently calibrated,} \\ 1, & \text{if } PD_g \text{ is not prudently calibrated.} \end{cases} \tag{2}$$

This decision rule is applied to each rating grade of all banks in our sample. To draw an overall conclusion, we aggregate these binary decisions by computing the weighted average percentage of miscalibrations using EAD. For a given bank b, this is specified as:

$$\text{MCB}_b = \frac{\sum_{g=1}^{G} \mathcal{M}_{g,b} \cdot \text{EAD}_{g,b}}{\sum_{g=1}^{G} \text{EAD}_{g,b}}, \tag{3}$$

where g indicates a rating grade and b indicates a bank.[10]

[10] Other weighting and aggregation schemes are also conceivable, for example, an obligor weighting. However, exposure weighting seems to be a more natural candidate with a clear economic interpretation, which

It is important to recognise that the aggregation rule we propose in (3) introduces an element of prudential conservatism that may be viewed as a conservative bias. This aspect should be borne in mind when interpreting the results. However, as discussed in relation to Figure 2, it is not possible to assume that capital from more conservative banks can be reallocated to offset underestimation by less conservative ones. However, within a single bank, it is plausible that an overestimation of risk in one rating grade could offset an underestimation in another. In other words, even if the PD is underestimated for one or more rating grades, the overall portfolio risk may still be appropriately estimated (see the example in Table 1 below). Nevertheless, under the aggregation rule specified in (3), we consider only those rating grades identified as underestimated, without accounting for potential offsetting effects at the portfolio level.

Table 1: Example of Aggregating Backtesting Results at the Bank Level ($\alpha = 5\%$)

RG	Exposure	n	PD	DR	$p^*_{binomial}$	MCB	MCB %
1	4.56	1,000	0.83	0.95	1.26	No	-
2	4.75	700	2.77	2.25	3.71	No	-
3	0.55	500	8.92	12.00	10.99	Yes	5.63
All	**9.81**	**2,200**	**3.29**	**3.29**	**3.87**	**No**	**5.63**

Note. This table presents an example of the aggregation mechanism of backtesting results at the bank level. Exposures are expressed in EUR millions; PD, DR, $p^*_{binomial}$, and Miscalibrated (MCB) % are expressed in percentages; and the number of obligors n is expressed in frequencies. In this example, only one rating grade (RG) fails the binomial test. While the model appears to be correctly calibrated at the portfolio level ('All'), the aggregation rule in (3) attributes the exposures in the third rating class to the pool of miscalibrated exposures.

Risk-weighted Assets Impact

A less commonly discussed limitation of the binomial test—and, more broadly, of any test yielding a binary (Y/N) outcome—arises when the number of observations (n) becomes large. In such cases, the threshold $p^*_{binomial}$ approaches the estimated PD, reducing the test's practical discriminative power. Consider the following example in Table 2: as n increases, $p^*_{binomial}$ converges toward the PD. In the last rating grade, the test fails, even though the observed DR exceeds the PD by only 9 basis points. Under the aggregation rule in (3), such marginal deviations are treated equivalently to cases where the DR substantially exceeds the PD. To mitigate this issue, one may instead assess the impact on RWAs, and consequently on capital requirements, that a prudent model re-calibration would entail.

Table 2: Example of the Functioning of the Binomial Test with Different Levels of n

Class	n	PD	DR	$p^*_{binomial}$	MCB
1	500	0.83	0.92	1.50	No
2	3,000	0.83	0.92	1.10	No
3	50,000	0.83	0.92	0.90	Yes

Note. This table presents an example of the functioning of the binomial test with different levels of obligors n. PD, DR, and $p^*_{binomial}$ are expressed in percentages and n is expressed in frequencies.

To analyse the impact on RWAs, we compute the following metric for rating grade g:[11]

is not necessarily the case for obligors, as a smaller number of miscalibrated obligors does not necessarily imply a less miscalibrated monetary value.

[11] For brevity, we omit the rating grade indicator if it is clear from the exposition. The impact can subsequently be easily aggregated to bank level as: $\sum_{g=1}^{G} \text{numerator} / \sum_{g=1}^{G} \text{denominator}$.

$$\mathrm{RWA}_{\mathrm{impact}} = \frac{\mathrm{RWA}^{+} - \mathrm{RWA}}{\mathrm{RWA}}, \tag{4}$$

where RWA^{+} represents the hypothetical RWAs resulting from applying p^{+} values instead of the bank's actual PD values. For rating grades classified as miscalibrated, $(\mathrm{RWA}^{+} - \mathrm{RWA}) \leq 0$; for all other rating grades, this difference is set equal to zero. Specifically, p^{+} reflect the lowest value for which the risk model remains prudently calibrated using the binomial test. For each rating grade *g*, p^{+} is the largest positive value that satisfies the equation:

$$p^{+} - \Phi^{-1}(q)\sqrt{\frac{p^{+}(1-p^{+})}{n}} \leq DR. \tag{5}$$

In other terms, p^{+} is the smallest grade PD for which the grade's DR would pass the binomial test. The main difference to (1) is that we use p^{+} instead of the actual PD. To compute the hypothetical RWAs, we follow the approach outlined by BCBS (2020) for RWAs under the IRB approach for non-defaulted corporate exposures, which is the regulatory standard. The final computation of RWAs is given by:

$$\mathrm{RWA} = \mathrm{K} \cdot 12.5 \cdot \mathrm{EAD}, \tag{6}$$

where K is the capital requirement. For the computation of RWAs, we rely on the data reported by each bank at the reference date. Specifically, we use the EAD, maturity, and LGD reported by each individual bank at the rating-grade level.[12]

Tier 1 Capital Ratio Impact

To analyse the impact on Tier 1 capital ratios (expressed in basis points), we compute the following metric at the bank level b:

$$\mathrm{T1CR}_{\mathrm{impact}} = \left(\frac{\mathrm{T1C}_b}{\mathrm{RWA}_{\mathrm{b}}^{adj}} - \frac{\mathrm{T1C}_b}{\mathrm{RWA}_b}\right) \cdot 10^4, \tag{7}$$

where $\mathrm{T1C}_b$ and RWA_b are the Tier 1 capital and RWAs at the bank level, respectively. $\mathrm{RWA}_{\mathrm{b}}^{adj}$ is an adjusted RWA figure that we augment by the miscalibrated RWAs at the rating grade level—which is equivalent to the sum of the numerator of (4). Formally, RWA_b^{adj} is defined as:

$$\mathrm{RWA}_b^{adj} = \mathrm{RWA}_b - \sum_{g \in G_b^{MCB}} \left(\mathrm{RWA}_{\mathrm{g}}^{+} - \mathrm{RWA}_{\mathrm{g}}\right), \tag{8}$$

where G_b^{MCB} denotes the set of miscalibrated rating grades for bank *b*.

[12] Because information on borrower size is not reported, we use the formulation set out in Article 153(1) of Regulation (EU) No 575/2013, rather than the alternative specified in Article 154(4). As a consequence, our RWA calculations do not necessarily exactly replicate the figures reported by banks. We therefore computed the RWA along with both the reported *PD* and the adjusted p^{+}, and use the difference between the two to assess the impact in terms of capital ratios.

2.2 Data

Our study is based on proprietary regulatory data collected by the EBA from 2017 to 2024 for corporate SME exposures.[13] The main datasets utilised in this analysis are (I) template C 08.05 COREP and (II) data collected as part of the supervisory benchmarking exercise.

Regarding (I), this data is collected explicitly for the backtesting of PDs of IRB systems of EU banks and contains information on the number of obligors, the PDs, the number of defaults, and the annual and long-run DRs per asset class and rating grade of the bank. The exposures are assigned to a common master scale with fixed PD ranges based on the PDs estimated for each obligor at the beginning of the reporting period. A few additional clarifications are warranted. The data are defined with respect to a standardized master scale to which banks are required to map their internal rating scales, which may differ in the number of rating grades. Even when reported at the asset-class level, the data may therefore reflect the output of multiple underlying models. In practice, a given asset class can be covered by several models (for example, differentiated by counterparty country), and banks typically conduct backtesting at the model level rather than at the asset-class level. We do not have access to model-level data, which represents a limitation of our analysis. Our objective, however, is not to identify individual miscalibrated models, this is the responsibility of banks' internal validation functions, but rather to provide aggregate evidence on the performance of IRB models in the EU. For this purpose, we believe the available data are representative. However, this template for COREP reporting is relatively new and information is only available from 2021. To expand our sample, we rely on (II) to obtain data related to the supervisory benchmarking exercise for the period from 2017 to 2020. The information we obtain from this dataset is the same as in the new COREP template and is used for the annual assessment of internal models in accordance with Article 78 of the 2013/36/EU Capital Requirements Directive (CRD). The only difference in content is that, for example, the number of obligors in (I) refers to the number of obligors at the beginning of the period (e.g., 1 January 2024), while in (II) it refers to the number obligors at the end of the period (e.g., 31 December 2024). We therefore assume that the number of obligors at the beginning of the year (t) is the same as the number of obligors at the end of the previous year ($t-1$).

To provide additional perspective, we supplement our dataset with information from COREP templates C 01.00 (Tier 1 capital) and C 02.00 (RWAs) at bank level. In addition, we obtain information at the rating-grade level on RWAs, EAD, LGD, and weighted-average maturities from C 08.02 in the same granularity as the data from C 08.05. In our analysis, we consider a constant sample of banks that report data for all COREP years from 2021 to 2024 and have continuous annual reporting for all years in which they appear in the dataset in order to exclude mechanical effects in the results. In addition, we restrict the sample to all observations where the obligors and EAD are greater than 1 and the PD is between 0 and 1;[14] we obtain 2,250 rating grade observations from 45 banks.

In Table 3, we present the descriptive statistics of our data. The average bank has RWAs of EUR 184bn and a Tier 1 capital ratio of 16.5%. The average rating grade comprises 3,670 obligors, ranging from 34 obligors (P5) to more than 14,700 obligors (P95). The simple average PD is 8.0% and

[13] For brevity, we restrict the empirical exercise to this asset class as an important use case in Europe, but our proposed approach can equally be applied and extended to any other asset class.

[14] This filter is intended to eliminate errors in the reporting. In some cases, banks report rating grades with an assigned PD but zero or missing numbers of borrowers or exposures. In others, the reported PD exceeds the admissible range $[0,1]$. It should be noted, however, that such reporting issues are more prevalent in the earlier reporting periods and have become increasingly rare over time.

the simple average DR is 4.8%. In terms of RWAs, the average rating grade is in the order of less than 0.7% of total average RWAs. Note that, while the backtesting analysis is conducted solely with reference to the SME corporates asset class, the computation of capital ratios uses each bank's total RWA, including contributions from other asset classes as well as other sources of RWA (e.g., market risk, operational risk, and others).

Table 3: Summary Statistics of Bank Level and Rating Grade Level Variables

	Mean	SD	P5	P25	P50	P75	P95	RG
A. Bank Level								
RWA	183,613	185,430	3,413	52,556	99,071	314,149	605,244	267
Tier 1 Ratio	16.5	2.4	13.2	14.8	16.0	18.0	21.1	267
B. Rating Grade Level								
PD	8.0	10.9	0.1	0.5	3.4	12.0	33.4	2,250
DR	4.8	8.5	0.0	0.2	1.5	5.7	20.5	2,250
Defaults	63	121	0	2	16	71	274	2,250
Obligors	3,670	6,990	34	260	1,140	4,314	14,738	2,250
EAD	2,582	4,733	16	155	734	2,927	11,402	2,250
RWA	1,204	1,885	18	116	489	1,431	5,149	2,250
Maturity	2.6	0.7	1.7	2.5	2.5	2.8	3.6	2,250

Note. This table presents summary statistics. We provide all variable definitions in the Appendix Part A. RWA and EAD are expressed in EUR millions; Tier 1 Ratio, PD, and DR are expressed in percentages; Defaults and Obligors are expressed in frequencies; Maturity is expressed in years; and RG indicates the number of observed rating grades.

2.3 Empirical Results

In this section, we provide insights into the miscalibrations of EU banks' internal PD models, the miscalibrations over time, and the economic effects of miscalibrations.

Baseline Aggregate Results

Table 4 shows the distributions of weighted, miscalibrated exposures of EU banks. In our prudent baseline specification using the binomial test, we find that on aggregate 13.4% (301/2,250) of the rating grades are miscalibrated, which translates into 16.7% of the exposures in our sample being miscalibrated. The distribution of miscalibrations shows a strongly right-skewed distribution with a fat tail, with an average and median miscalibrated exposure in the order of EUR 3,214mn and EUR 1,394mn, respectively.

To provide additional perspective, we decompose the total miscalibrations into investment grade (IG) and non-investment grade (Non-IG) subsets. We define IG and non-IG rating grades based on the estimated PD, using a threshold of 1%.[15] It can be observed that although the IG subset

[15] For reference, the Moody's master scale ranges from Aaa to C and comprises 21 notches grouped into two broad categories: investment grade (from Aaa to Baa) and non-investment grade, or speculative grade (from Ba1 to C). In terms of PD, the IG boundary in the Moody's master scale corresponds to a PD below 1%. However, it is important to note that Moody's ratings typically apply to large corporations, while our analysis focuses on SMEs, which are generally associated with higher default rates. We verify that adopting a 1% PD

accounts for less than half of the miscalibrated rating grades, it covers almost 60% of the miscalibrated exposures. The main reason for this observation is that the composition of the IG and non-IG rating grades differs considerably; the average number of obligors and exposures in the IG rating grades is around two times higher than in the non-IG rating grades, respectively. From a regulatory perspective, the fact that more non-IG rating grades are miscalibrated may be considered of greater concern, as obligors in these rating classes by definition not only have higher PDs on average, but the actual observed DRs are also significantly higher.

Table 4: Miscalibrations of EU Banks Weighted by the Exposure at Default

Sample	MCB	%	Mean	SD	P5	P25	P50	P75	P95	RG
Total	967,444	16.7	3,214	4,716	31	372	1,394	3,895	12,071	301
IG	573,966	20.3	4,381	4,751	125	1,020	2,859	6,181	13,280	131
Non-IG	393,478	13.2	2,315	4,499	19	174	594	2,139	11,584	170

Note. This table presents the number of miscalibrated rating grades and miscalibrated exposures based on the binomial test. RG indicates the number of miscalibrated rating grades, % indicates the miscalibrated exposures in percentages, and the remaining values are in terms of the miscalibrated exposures in EUR millions.

In Table 5, we show the number of miscalibrated exposures over time. In the last four years of the sample—aligned with the data available through COREP—the number of miscalibrated exposures is substantially lower than in the pre-COREP sample. The trend likely reflects improved data quality resulting from well-established supervisory reporting channels, which include numerous validation rules to ensure high-quality data. A plausible explanation for the slightly higher values in 2023 and 2024 is the phase-out of COVID-19-related moratoria and public guarantees.

Table 5: Miscalibrations of EU Banks Weighted by the Exposure at Default Over Time

Total	2017	2018	2019	2020	2021	2022	2023	2024
16.7	23.2	26.0	20.9	17.3	10.0	10.2	17.2	12.1

Note. This table presents the percentage of miscalibrated exposures based on the binomial test over time.

However, the 2023 value is well-above what can reasonably be attributed solely to the withdrawal of COVID-19-related measures. We argue that this is due to the sensitivity of the test, where even a single additional default can result in an entire rating grade being flagged as miscalibrated.[16] Consider the following simple example of a rating grade in our sample:

EAD	N	PD	Defaults	DR
~ 29,000	9,493	1.54%	166	1.75%

A miscalibration in this rating grade implies miscalibrated exposures totalling around EUR 29bn. Yet, a slightly higher PD of 3.6 basis points or a slightly lower DR of 4.2 basis points would have been sufficient to achieve a positive result. Under these adjusted parameters, the share of miscalibrated exposures for 2023 would fall by 3.7 percentage points to 13.5%. This underlines why the analysis of miscalibrated exposures alone is not sufficient. In particular, it highlights the limitations of a

threshold yields a balanced partition of the sample, with each category accounting for approximately 50% of total exposures.

[16] As outlined before, from a supervisory perspective, over-rejection of well-calibrated models (i.e., being overly conservative) is generally less critical than under-rejection.

binary decision framework, such as that underlying the binomial test and related procedures, which fails to account for the magnitude by which the test criterion is violated. In the next section, we therefore extend the analysis to assess the economic impact of miscalibrations, in particular their impact on Tier 1 capital ratios.

Economic Effects of Miscalibrations

As the banks' internal risk models are ultimately used, for example, to derive regulatory capital requirements, we examine in Table 6 how the identified miscalibrations affect Tier 1 capital ratios. For completeness, we also report the levels and shares of miscalibrated exposures. Based on the hypothetical RWAs from (4), a prudent calibration of the risk models would result in a reduction (Δ) of the Tier 1 capital ratio by 10.3 basis points at the system level across all banks.[17] This figure is close to the computation of a Tier 1 impact weighted by Tier 1 capital, which would be 10.5 basis points. In contrast, the simple average Tier 1 impact is slightly higher at 21.0 basis points, with a maximum impact exceeding 300 basis points at the bank-year level. While our preferred measure in this section is the Tier 1 capital ratio impact calculated at the system level across all banks, the following section also presents the impact at the individual bank level.

Table 6: Real Effects of Miscalibrations

Subset	MCB	%	Tier 1	Tier 1 Δ
A. Overall Sample				
Total	967,444	16.7	15.7	-10.3
B. Across Rating Categories				
IG	573,966	20.3	15.6	-13.8
Non-IG	393,478	13.2	15.8	-3.0
C. Over Time				
2017	130,085	23.2	14.8	-6.2
2018	165,164	26.0	15.2	-9.2
2019	156,476	20.9	15.7	-7.0
2020	130,552	17.3	16.5	-8.4
2021	77,119	10.0	15.4	-17.9
2022	80,157	10.2	15.5	-17.8
2023	134,970	17.2	16.3	-18.2
2024	92,921	12.1	16.3	-4.6

Note. This table presents the real effects of miscalibrations based on the binomial test. MCB and % indicates the miscalibrated exposures in EUR millions and percentages, respectively. Tier 1 is expressed in percentages and the corresponding economic impact Tier 1 Δ is expressed in basis points.

Looking at the impact on the individual rating categories (Panel B), the effect is more pronounced in the IG category, with a corresponding decline in the Tier 1 capital ratio of 13.8 basis points. Looking at the time dimension (Panel C), we observe the smallest impact in 2024 at -4.6 basis points.

[17] The difference in the Tier 1 capital ratio between Table 6 and the summary statistics in Table 3 arises from the aggregation method: Table 3 presents simple averages across banks, whereas Table 6 reports a system level measure calculated as the ratio of the sum of Tier 1 capital to the sum of RWAs across all observations, i.e., $\sum \mathrm{T1C}_{b,t} / \sum \mathrm{RWA}_{b,t}$.

The slightly higher impact between 2021 and 2023 is primarily due to a small number of bank rating grades that have been severely miscalibrated. If these outliers are excluded, there is a generally stable to slightly declining trend over time, which is accompanied by both a reduction in miscalibrated exposures and an increase in Tier 1 capital ratios. Overall, this pattern is plausibly explained by a higher margin of conservatism applied to the estimated PDs.

Taken together, miscalibrated risk models are not only an important aspect from a theoretical perspective but also have a significant real effect on the Tier 1 capital ratios. This emphasises in particular the importance of the annual assessment of internal models in accordance with Article 78 of the CRD, which contributes to ensuring the robustness of the EU banking sector and promoting financial stability.

Bank-by-Bank Results

In the previous sections, we focused primarily on aggregate developments across banks. However, risk modelling and banking supervision occur at the individual bank level. Hence, Table 7 shows the percentage of miscalibrated exposures over time for each bank in our sample and the Tier 1 capital ratio impact. Since we use the binomial test, the results can again be interpreted as an upper bound for miscalibrated exposures.

In our sample, 20% of banks have no miscalibrations and around 31% have miscalibrations of less than 10% at the aggregate level. Only three banks have miscalibrations of more than 50%, and of these, two have no miscalibrations in the last year of our sample (2024). Most banks with up to four miscalibrated years, around 71% of banks, tend to show miscalibrations in earlier years, with only five banks having a miscalibration in 2024 and only one bank having miscalibrations in 2023 and 2024. In contrast, of the banks with at least five miscalibrated years, almost 85% also have a miscalibration in the last year of the sample.

The simple average (median) impact on the Tier 1 capital ratio is -17.4 (-7.2) basis points, with values ranging from 0 to -137.7 basis points. This figure is derived by first computing the Tier 1 impact at the bank level and then taking the simple average (median) across banks. In contrast, the figure presented in the previous section was based on a simple average across all bank-year observations. For almost two-thirds of the banks, the impact is less than 10 basis points and only in three cases more than 50 basis points. This measure provides additional insights that highlight the importance of not relying solely on the share of miscalibrated exposures. Miscalibrated exposures do not translate one-to-one into economic effects in terms of Tier 1 capital, which emphasises the need to consider both dimensions in the analysis.

This section presented real-world evidence from a systematic backtesting exercise conducted with a sample of EU banks using the binomial test. Since the binomial test assumes independent defaults (Blümke 2022b), it is well-documented in the literature that the test is overly conservative and too often rejects the null hypothesis of prudently calibrated models, thereby inflating the type I error rate. It is therefore essential to assess, at least approximately, the extent to which the assumption of independent DRs affects backtesting results (Blochwitz et al. 2006). While discussions in banking regulation and credit risk backtesting often emphasise asset correlation, in practice DRs also exhibit significant persistence and thus serial correlation (Blümke 2022a).

In what follows, we introduce a generalised version of a corrected binomial test that simultaneously corrects for asset and serial correlation. Previous literature has addressed the breakdown of the independence assumption through a correlated binomial test that extends the standard binomial

distribution to account for asset correlation (Balthazar 2004; Tasche 2005). For brevity, we do not discuss this intermediate correction, as the generalisation we propose includes the asset correlation correction as a special case within a broader correction that also accounts for serial correlation.

Table 7: Miscalibrations at EU Bank Level Weighted by the Exposure at Default and Impact on Tier 1 Capital Ratios

Bank	2017	2018	2019	2020	2021	2022	2023	2024	MCB	Tier 1 Δ
1	0.7	54.9	73.2	74.9	66.6	82.7	81.4	85.2	**61.9**	**-29.8**
2	0.0	39.6	79.5	65.6	98.2	99.0	90.3	0.0	**59.5**	**-15.3**
3	21.8	16.6	2.6	12.2	100.0	100.0	100.0	0.0	**50.3**	**-137.7**
4	6.5	9.3	20.7	89.6	5.7	96.5	81.9	90.2	**45.3**	**-17.8**
5	17.6	24.4	65.5	47.3	38.2	32.9	37.2	41.1	**37.9**	**-7.2**
6	0.0	90.8	54.1	92.4	0.0	0.0	3.7	2.0	**36.4**	**-11.7**
7	100.0	73.7	55.0	0.0	0.0	0.0	58.7	6.4	**35.8**	**-22.9**
8	100.0	43.9	0.0	0.0	38.1	0.0	0.0	52.4	**29.5**	**-86.3**
9	42.1	91.1	98.7	19.6	0.0	0.0	0.0	0.0	**28.3**	**-31.2**
10	16.9	30.3	0.0	58.6	53.0	0.0	18.8	25.5	**24.8**	**-5.7**
11	1.9	0.0	0.0	0.0	100.0	64.5	61.5	33.1	**23.2**	**-35.6**
12	0.0	0.0	15.8	100.0	15.7	58.3	0.0	0.0	**22.1**	**-25.3**
13	74.0	63.9	42.3	0.0	0.0	1.9	0.0	3.0	**21.5**	**-33.2**
14	-	-	0.3	0.0	0.0	0.0	52.5	59.7	**20.8**	**-25.0**
15	3.3	0.0	77.0	0.0	0.0	0.0	77.2	0.0	**18.8**	**-58.8**
16	1.0	51.0	16.6	90.2	0.0	0.0	0.0	5.8	**18.4**	**-5.3**
17	0.0	57.8	0.0	75.6	0.0	0.0	0.0	0.0	**16.6**	**-2.5**
18	38.2	45.3	6.2	0.0	10.6	5.1	7.5	7.8	**14.9**	**-2.5**
19	0.0	25.3	22.6	50.3	0.0	14.4	0.0	0.0	**14.1**	**-10.6**
20	0.0	0.0	37.6	1.8	0.0	27.4	65.0	0.0	**13.9**	**-0.2**
21	27.7	0.0	62.8	51.8	0.0	0.0	0.0	0.0	**11.7**	**-1.2**
22	-	-	0.0	100.0	0.0	0.0	0.0	0.0	**10.6**	**-19.1**
23	0.0	17.0	0.0	46.0	0.0	0.0	2.8	2.7	**8.8**	**-10.8**
24	0.0	30.0	0.0	25.8	0.0	0.0	0.0	0.0	**7.0**	**-5.6**
25	14.0	0.0	0.0	6.5	0.0	31.7	0.0	0.0	**6.9**	**-0.7**
26	0.0	0.0	9.9	21.8	23.7	0.0	0.0	0.0	**6.9**	**-2.3**
27	0.0	0.0	2.5	0.0	16.7	0.0	0.0	35.3	**6.8**	**-2.1**
28	2.9	0.0	0.0	0.0	0.0	0.0	0.0	40.9	**4.9**	**-7.5**
29	27.8	0.0	4.5	0.0	0.0	0.0	0.0	0.0	**4.6**	**-0.3**
30	0.0	0.0	0.0	0.0	0.0	0.0	3.2	60.8	**4.4**	**-0.9**
31	25.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**4.0**	**-19.4**
32	20.8	3.7	0.0	6.4	0.0	0.0	0.0	0.0	**3.0**	**-3.2**
33	-	0.0	3.2	0.0	0.0	0.0	0.0	0.0	**0.5**	**-1.7**

34	0.0	0.0	2.2	0.0	0.0	0.0	0.0	0.0	**0.3**	**-1.8**
35	-	0.3	0.9	0.4	0.0	0.0	0.0	0.0	**0.2**	**-0.5**
36	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	**0.1**	**0.0**
37	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	**0.0**	**-0.2**
38	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**
39	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**
40	-	-	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**
41	-	-	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**
42	-	-	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**
43	-	-	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**
44	-	-	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**
45	-	-	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**

Note. This table presents the percentage of miscalibrated exposures based on the binomial test and is sorted according to the miscalibrated exposures on the total level (MCB). White indicates no miscalibrated exposures, and the darker the grey, the higher the percentage of miscalibrated exposures. Tier 1 Δ shows the impact on the Tier 1 capital ratios in basis points across all years.

3. Generalised Binomial Test Correction for Asset and Serial Correlation

In this section, we introduce a generalised version of a corrected binomial test that simultaneously corrects for asset and serial correlation. We also explore its distributional properties, present a simulation study to evaluate its performance, and use the order statistic for multiyear backtesting. Finally, we apply the corrected test to replicate the real-world systematic backtesting exercise shown in the previous section, thereby demonstrating its practical implications and potential improvements over the unadjusted approach.

3.1 Model Specification

Blümke (2012) proposes an alternative to the binomial test based on a maximum likelihood approach. This test is of interest because it explicitly accounts for default correlation. In what follows, however, we take a different approach and propose a correction to the binomial test itself.[18] Our objective is to show that, once the assumption of no correlation, both cross-sectional and serial, is relaxed, the distribution of default rates becomes right-skewed, with a heavier right tail. This shift increases the threshold required to reject the null hypothesis of correct calibration. In doing so, we

[18] The methodological structure employed in this paper is closely related to the framework used by Benjamin (2006) for low-default portfolios. In both cases, the starting point is a one-factor Vasicek credit risk model, in which default rates are driven by asset correlation and a serially correlated systemic factor. The key difference lies in the conceptual orientation. In Benjamin (2006), the objective is to derive a conservative PD estimate that is consistent with observed default outcomes under given correlation assumptions. In contrast, in the present paper the PD is taken as given, and the analysis instead focuses on deriving critical default-rate thresholds for backtesting purposes.

aim to clarify why the binomial test can be viewed as a particularly conservative, or prudential, testing approach.

Let us assume that an obligor $i \in (1, \dots, n)$ defaults at time t if her creditworthiness $z_{i,t}$ falls below a certain threshold c. Defaults are recorded by an indicator variable $\mathcal{D}_{i,t}$ as follows:

$$\mathcal{D}_{i,t} = \begin{cases} 0, & \text{if } z_{i,t} \geq c, \\ 1, & \text{if } z_{i,t} < c. \end{cases} \tag{9}$$

Creditworthiness $z_{i,t}$ is a random variable that is unique for each obligor i at each time t:

$$\begin{aligned} z_{i,t} &= \sqrt{\rho} y_t + \sqrt{1-\rho} u_{i,t}, \\ y_t &= \psi y_{t-1} + e_t, \end{aligned} \tag{10}$$

where $u_{i,t} \sim \mathcal{N}(0,1)$ and $e_t \sim \mathcal{N}(0,1)$. y_t represents a systemic factor that is common to all obligors n (e.g., the macroeconomic environment) and $u_{i,t}$ represents an idiosyncratic, risk-specific factor that is unique for each obligor i. The parameter ρ is assumed to be fixed for all obligors and captures the underlying asset correlation: setting $\rho = 0$ implies independent defaults, while setting $\rho = 1 - \epsilon$ with $\epsilon \geq 0$ and $\epsilon < 1$ implies correlated defaults at time t. Our representation of the systemic factor in (10) differs slightly from the usual representations in the literature, as the systemic factor follows an autoregressive process of order one (AR(1)), where ψ denotes the persistence parameter. This framework allows us to consider the potential presence of serial dependence in default rates. Although the available time series of eight years is too short to support a formal assessment of such dynamics, the data nevertheless suggest an average serial correlation of approximately 40%.

Blümke (2022b, 2025) also introduces serial dependence by modelling the systemic factor as an AR(1) process. In contrast to our approach, Blümke (2025) assumes that the variance of the innovation e_t is equal to $1 - \psi^2$ rather than one, while Blümke (2022b) scales the innovation e_t by $\sqrt{1-\psi^2}$. Both formulations ensure that the variance of the systemic factor y_t remains equal to one. This is an elegant modelling choice that allows the introduction of serial correlation in defaults without deviating substantially from the baseline framework. Our objective differs. We aim to illustrate how departures from the assumptions of the binomial test can generate more dispersed and asymmetric default rate distributions with heavier right tails, which in turn lead to over rejection when the binomial test is used. While we do not claim that our specification is more realistic, we argue that is better suited to our purpose of inflating the right tail of the default rate distribution and thereby highlighting the prudential nature of the binomial test.

Using this specification, we obtain the following distributions of the creditworthiness $z_{i,t}$ and the systemic factor y_t:[19]

$$z_{i,t} \sim \mathcal{N}\left(0, \frac{1-\psi^2(1-\rho)}{1-\psi^2}\right), \tag{11}$$

$$y_t \sim \mathcal{N}\left(0, \frac{1}{1-\psi^2}\right). \tag{12}$$

As is common in the literature, we define the unconditional probability of default as $PD = \mathbb{P}(\mathcal{D}_{i,t} = 1)$, which allows us to determine the threshold c in (9) such that $c = \Phi^{-1}(PD)\sqrt{\mathbb{V}(z_{i,t})}$, where

[19] It is straightforward to compute the variance of $z_{i,t}$ from (10) such that $\mathbb{V}(z_t) = \frac{\rho}{1-\psi^2} + 1 - \rho = \frac{1-\psi^2(1-\rho)}{1-\psi^2}$.

$\mathbb{V}(z_{i,t})$ is the variance of the creditworthiness.[20] Setting $c = \Phi^{-1}(PD)\sqrt{\mathbb{V}(z_{i,t})}$, we obtain from (9) $\mathbb{P}(\mathcal{D}_{i,t} = 1) = \mathbb{P}(z_{i,t} < c) = \mathbb{P}(z_{i,t} < \Phi^{-1}(PD)\sqrt{\mathbb{V}(z_{i,t})})$, and since $z_{i,t}/\sqrt{\mathbb{V}(z_{i,t})}$ is a standard normal random variable, we have $\mathbb{P}(\mathcal{D}_{i,t} = 1) = \Phi(\Phi^{-1}(PD)) = PD$.

Now, we define the conditional probability of default as $\mathbb{P}(\mathcal{D}_{i,t} = 1|y_t = y)$. Consequently, if the common systemic factor is fixed, the conditional probability for obligor i is:[21]

$$\mathbb{P}(z_{i,t} < c|y_t = y) = \mathbb{P}(\sqrt{\rho}y + \sqrt{1-\rho}u_{i,t} < c) = \mathbb{P}\left(u_{i,t} < \frac{c-\sqrt{\rho}y}{\sqrt{1-\rho}}\right) = \Phi\left(\frac{c-\sqrt{\rho}y}{\sqrt{1-\rho}}\right) = f(y). \quad (13)$$

Equipped with this, we can determine the probability that the number of defaults in the portfolio is greater than or equal to a critical value k. To start with, we define $\mathcal{D}_{n,t} = \sum_{i=1}^{n} \mathcal{D}_{i,t}$ as the number of defaults observed in a portfolio with n obligors. Intuitively, the random variable $\mathcal{D}_{i,t}$ for a given t and consequently fixed $y_t = y$ is a Bernoullian random variable with parameter $f(y)$. As a result, $\mathcal{D}_{n,t}$ is then a binomial random variable with parameters $f(y)$ and n, such that we can write:

$$\mathbb{P}(\mathcal{D}_{n,t} \leq k|y_t = y) = \sum_{j=0}^{k} \binom{n}{j} f(y)^j (1 - f(y))^{n-j}. \quad (14)$$

The evaluation of (14) for each value of y_t, weighted with its probability, enables us to derive the cumulative distribution function (CDF) of the random variable $\mathcal{D}_{n,t}$:

$$\mathbb{P}(\mathcal{D}_{n,t} \leq k) = \int_{-\infty}^{\infty} \sum_{j=0}^{k} \binom{n}{j} f(y)^j (1 - f(y))^{n-j} \phi\left(\frac{y}{\sqrt{1-\psi^2}}\right) dy, \quad (15)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ denotes the standard normal density function. Therefore, the exact critical value k, depending on asset and serial correlation, can be determined by iterative computations.

Approximation

Since solving (15) iteratively is computationally expensive, we now provide a simpler approximation to obtain the CDF when n is large. In order to achieve this, we must obtain a pre-specified threshold. Let us define $\mathcal{D}R_t = \mathcal{D}_{n,t}/n$ as the observed DR. Since $\mathcal{D}_{n,t}$ is a binomial random variable with parameters $f(y)$ and n for a given $y_t = y$, we have by the law of large numbers $\lim_{n\to\infty} \mathcal{D}R_t = f(y) = \Phi\left(\frac{c-\sqrt{\rho}y}{\sqrt{1-\rho}}\right)$.

Hence:

$$\mathbb{P}(\mathcal{D}_{n,t} \leq k) = \mathbb{P}(\mathcal{D}R_t \leq k/n) \to \mathbb{P}\left(\Phi\left(\frac{c-\sqrt{\rho}y}{\sqrt{1-\rho}}\right) \leq k/n\right) = \mathbb{P}\left(\frac{c-\sqrt{\rho}y}{\sqrt{1-\rho}} \leq \Phi^{-1}(k/n)\right). \quad (16)$$

Solving this expression for y, we obtain:

[20] Omitting the serial correlation results in $\mathbb{V}(z_{i,t}) = 1$ and $c = PD$, which brings us back to the threshold defined in classical models such as Vasicek (2002).

[21] Notice that this is also referred to as the Asymptotic Single Risk Factor (ASRF) model, which is included in the IRB supervisory formula for deriving capital requirements for credit institutions (BCBS 2004, 2011, 2020).

$$\mathbb{P}(\mathcal{D}R_t \leq k/n) \to \mathbb{P}\left(y \leq \frac{\sqrt{1-\rho}\Phi^{-1}(k/n)-c}{\sqrt{\rho}}\right). \quad (17)$$

Now, remember from (12) that $y_t \sim \mathcal{N}\left(0, \frac{1}{1-\psi^2}\right)$, such that we ultimately have:

$$\mathbb{P}(\mathcal{D}R_t \leq k/n) \to \mathbb{P}\left(\frac{y}{\sqrt{1/(1-\psi^2)}} \leq \frac{\sqrt{1-\rho}\Phi^{-1}(k/n)-c}{\sqrt{\rho/(1-\psi^2)}}\right) = \Phi\left(\frac{\sqrt{1-\rho}\Phi^{-1}(k/n)-c}{\sqrt{\rho(1-\psi^2)}}\right). \quad (18)$$

This enables us to compute an approximate value k by solving this expression numerically for k such that (18) equals a certain confidence level (e.g., 95%). Therefore, with a reasonable calibration of the required parameters, it is always possible to set the threshold for the test so that the probability of rejection of the model (i.e., type I error rate) is equal to a desired level.

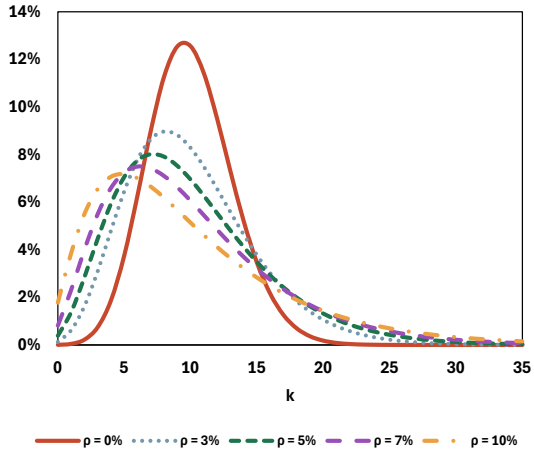
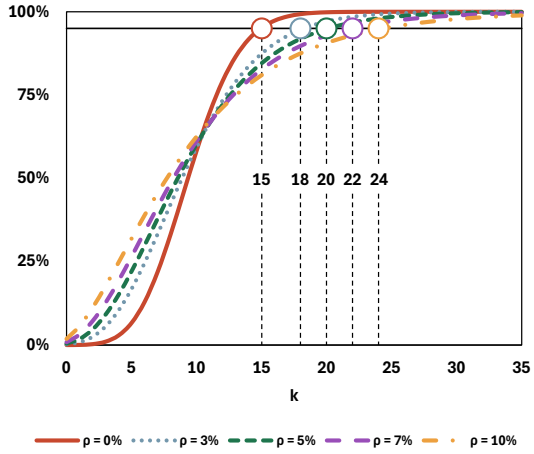
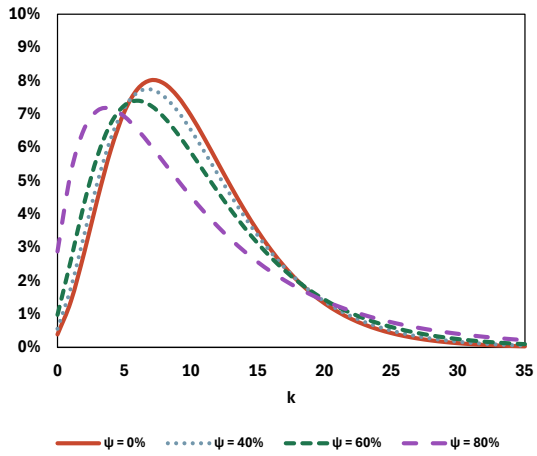
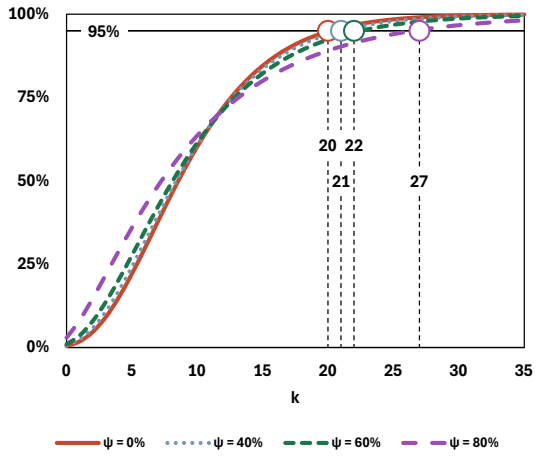
3.2 Distributional Properties

Following the theoretical model specification, we now examine the distributional properties in Figure 3 to explore how the threshold is affected when the assumption of independent and serially uncorrelated defaults does not hold. We also analyse how different levels of asset and serial correlation affect both the probability distribution function (PDF; i.e., $\mathbb{P}(\mathcal{D}_n(t) = k)$) and the cumulative distribution function (CDF; i.e., $\mathbb{P}(\mathcal{D}_n(t) \leq k)$). Panel A shows how the presence of asset correlation affects the maximum number of expected defaults at a given confidence level, assuming a serial correlation of $\psi = 0\%$. Panel B shows how the presence of serial correlation affects the maximum number of expected defaults, assuming a constant asset correlation of $\rho = 5\%$. To illustrate and without loss of generality, we consider $n = 200$ obligors and $PD = 5\%$. Generally, the presence of asset and serial correlation clearly widens the dispersion of the distributions (left panels); with the effect being somewhat more modest for serial correlation up to a certain level. Given the importance of the threshold in the backtesting exercise, we now examine the effects on the quantiles of the distribution (right panels).

In Panel A, where we analyse the impact of asset correlation, the 95%-quantile is 15 when $\rho = 0\%$. This implies that at a 95% confidence level, the DR is not expected to exceed 7.5% (15/200), or put differently, the DR will only exceed 7.5% in one out of 20 cases. This threshold would be used in the backtesting exercise under the assumption that the defaults are independent and serially uncorrelated. In practice, with an estimated PD of 5%, we would accept the hypothesis that the model is prudently calibrated if the observed DR is below 7.5%. However, as noted above, this assumption is overly restrictive and is generally not met in credit risk practice. Assuming an asset correlation of $\rho = 5\%$, the 95%-quantile increases to 20, which corresponds to a DR of 10% (20/200). This corresponds to an increase of 2.5 percentage points compared to the scenario without asset correlation.

In Panel B, where we analyse the impact of serial correlation, the 95%-quantile is 20 when $\psi = 0\%$. As already noted for the PDF, the effect of persistence appears to be modest up to certain levels. The 95%-quantile increases only slightly from 20 to 22 when persistence increases from 0% to 60%. However, as persistence approaches 80%, the quantile rises significantly from 20 to 27. This implies that assuming a constant asset correlation of $\rho = 5\%$ and a serial correlation of $\psi = 80\%$, the threshold for the DR rises to 13.5% (27/200), which is 3.5 percentage points higher than in the case without serial correlation and 6.0 percentage points higher than in the case without both asset and serial correlation. It is therefore clear that if we assume $\rho = 0\%$ and $\psi = 0\%$, but in reality $\rho > 0\%$ or $\psi > 0\%$, the 7.5% threshold used at a confidence level of 95% is too low, which leads to the rejection of the hypothesis that the model is prudently calibrated too often.

Figure 3: Probability Distribution Function (PDF) and Cumulative Distribution Function (CDF) with Asset and Serial Correlation

$\mathbb{P}(\mathcal{D}_n(t) = k)$	$\mathbb{P}(\mathcal{D}_n(t) \leq k)$
Panel A. Asset Correlation (ρ)	
	
Panel B. Serial Correlation (ψ)	
	

Note. This figure shows the probability distribution function (PDF) in the left panel and the cumulative distribution function (CDF) in the right panel. For illustrative purposes and without loss of generality, we consider $n = 200$ obligors and $PD = 5\%$. In Panel A, only the distributional impacts of asset correlation are considered and therefore a serial correlation of $\psi = 0\%$ is assumed. In Panel B, the distributional impacts of serial correlation are considered using a constant asset correlation of $\rho = 5\%$. In the right panels, we show in the body of the charts the numerical quantities k^* = number of defaults such that $\mathbb{P}(\mathcal{D}R_t \leq k^*/n) = 95\%$.

To provide additional perspective, we now examine in more detail in Table 8 how the threshold k^* (i.e., the number of defaults such that $\mathbb{P}(\mathcal{D}R_t \leq k^*/n) = 95\%$) changes with the introduction of asset and serial correlation at the 95% confidence level. In line with our previous examples, we consider $n = 200$ obligors and $PD = 5\%$. Note that (15) represents the binomial distribution when both $\rho = 0\%$ and $\psi = 0\%$, and that it corresponds to the Vasicek distribution when only $\psi = 0\%$. By varying the parameters ρ and ψ, we can clearly observe how the threshold changes. For example, if $\rho = 0\%$ and $\psi = 0\%$, the highest observed DR at a 95% confidence level is expected to be 7.5%. However, when the 'true' $\rho > 0\%$, the expected highest observed DR ranges from 9% ($\rho = 3\%$) to 12% ($\rho = 10\%$) assuming no serial correlation, and from 9.5% ($\rho = 3\%$ and $\psi = 40\%$) to

17% ($\rho = 10\%$ and $\psi = 80\%$) assuming both asset and serial correlation. Consequently, depending on the values of asset and serial correlation, the expected highest observed DR can be more than twice as high as that under the assumption of independent and serially uncorrelated defaults.

Table 8: 95%-quantiles of the Distribution of the DRs

		ψ			
		0.0	**40.0**	**60.0**	**80.0**
ρ	**0.0**	7.5	-	-	-
	3.0	9.0	9.5	10.0	11.5
	5.0	10.0	10.5	11.0	13.5
	7.0	11.0	11.5	12.5	15.0
	10.0	12.0	13.0	14.0	17.0

Note. This table shows the numerical quantities k^* = number of defaults divided by *n* such that $\mathbb{P}(\mathcal{D}R_t \leq k^*/n) = 95\%$ (i.e., the threshold). For example, consider from Figure 3 the value of 15 (Panel A, right-hand side) divided by 200, which equals 7.5%. For illustrative purposes and without loss of generality, we consider $n = 200$ obligors and $PD = 5\%$. White indicates the expected highest observed DR under the assumption of independent and serially uncorrelated defaults, and the darker the grey, the higher the expected highest observed DR, suggesting that the assumption of independent and serially uncorrelated defaults does not hold.

3.3 Simulation Study

In what follows, we will provide evidence for the functioning of our proposed generalised binomial test correction that corrects for asset and serial correlation using Monte Carlo simulations. We assume that all parameters are known, which is consistent with simulation studies in the literature. The simulation procedure is repeated for a wide range of parameter combinations with 10^6 replications per run; we iteratively set the parameters to: portfolio size $n \in \{500; 1{,}000; 5{,}000; 10{,}000\}$, asset correlation $\rho \in \{0\%; 5\%\}$, serial correlation $\psi \in \{0\%; 40\%; 60\%; 80\%\}$, and probability of default $PD \in \{0.5\%; 5.0\%\}$. In Table 9, we report the results of our simulation study for the binomial test (BT), the binomial test correction proposed in the literature that is correcting only for asset correlation (BT-C ρ), and our proposed binomial test correction that corrects for both asset and serial correlation (BT-C ρ & ψ).

The left panel shows the results at a $PD = 0.5\%$, which can be considered as a PD in the order of an investment grade rating class, and the right panel shows the results at a $PD = 5.0\%$, which can be considered as a PD in the order of a non-investment grade rating class. In all specifications, we use a confidence level of 95%, which implies that it is acceptable to reject H_0 with a probability of 5% if the risk model is actually prudently calibrated. Thus, if the test works as expected, we expect a probability of rejection of the model of 5%.

Starting with the classical binomial test (BT), we find that the test under the independence assumption ($\rho = 0\%$ and $\psi = 0\%$) rejects the model at around 5% based on the 95% confidence level, in line with expectations. As soon as the independence assumption no longer holds (i.e., $\rho > 0\%$), it can be observed that the probability of rejection for the binomial test increases substantially and ranges between 10.3% ($n = 500$) and 27.5% ($n = 10{,}000$) for a $PD = 0.5\%$ and increases even further for a $PD = 5.0\%$ and ranges between 21.9% ($n = 500$) and 36.8% ($n = 10{,}000$). The increase in the probability of rejection between the different levels of PD is not surprising, as the

threshold depends on the respective parameter (i.e., PD).[22] The introduction of serial correlation (i.e., $\psi > 0\%$) has a relatively modest effect and does not substantially increase the probability of rejection.

Table 9: Probability of Rejecting the Model with Known Parameters

			PD = 0.5%			PD = 5.0%		
n	ρ	ψ	BT	BT-C ρ	BT-C ρ & ψ	BT	BT-C ρ	BT-C ρ & ψ
500	0.0	0.0	**4.2**	-	-	**4.1**	-	-
500	5.0	0.0	10.3	**4.1**	-	21.9	**4.8**	-
500	5.0	40.0	10.9	4.6	**4.6**	22.7	5.9	**4.9**
500	5.0	60.0	11.6	5.6	**5.6**	23.7	7.6	**4.8**
500	5.0	80.0	12.9	7.6	**4.6**	25.0	11.2	**4.7**
1,000	0.0	0.0	**6.8**	-	-	**4.7**	-	-
1,000	5.0	0.0	16.4	**5.6**	-	27.0	**5.1**	-
1,000	5.0	40.0	16.9	6.4	**5.0**	27.6	6.2	**5.0**
1,000	5.0	60.0	17.5	7.6	**5.1**	28.2	8.0	**4.9**
1,000	5.0	80.0	17.9	9.7	**5.0**	28.6	11.7	**4.9**
5,000	0.0	0.0	**4.4**	-	-	**4.9**	-	-
5,000	5.0	0.0	23.7	**4.9**	-	34.7	**5.0**	-
5,000	5.0	40.0	23.9	5.8	**5.0**	34.8	6.2	**5.0**
5,000	5.0	60.0	24.0	7.1	**4.9**	34.7	8.0	**5.0**
5,000	5.0	80.0	23.2	9.7	**4.9**	33.5	11.9	**4.9**
10,000	0.0	0.0	**5.1**	-	-	**4.7**	-	-
10,000	5.0	0.0	27.5	**5.0**	-	36.8	**5.0**	-
10,000	5.0	40.0	27.5	6.0	**5.0**	36.7	6.2	**5.0**
10,000	5.0	60.0	27.1	7.4	**5.0**	36.2	8.1	**5.0**
10,000	5.0	80.0	25.4	9.9	**4.9**	34.6	11.9	**4.9**

Note. This table shows the probability of rejecting the model for the binomial test (BT), the binomial test corrected for asset correlation (BT-C ρ), and the generalised binomial test simultaneously correcting for asset and serial correlation (BT-C ρ & ψ). Light grey indicates that the probability of rejection is close to the expected value of 5% at a confidence level of 95%, and the darker the grey, the further away the probability of rejection is from the expected value.

Next, we examine the binomial test corrected for asset correlation proposed in the literature (BT-C ρ). The main correction that this specification achieves applies to the specification with $\rho > 0\%$ and $\psi = 0\%$. We find that by applying this correction, the probability of rejection is close to the desired level of 5%. For example, instead of the probability of rejection of 36.8% ($n = 10{,}000$ and $PD = 5.0\%$) observed for the standard binomial test, a probability of rejection of 5% is now observed. However, unlike the previous observation, the introduction of serial correlation (i.e., $\psi > 0\%$) is no longer negligible, and the probability of rejection increases substantially up to 11.9% ($n = 10{,}000$, $PD = 5.0\%$, and $\psi = 80\%$). Consequently, the binomial test correction currently applied in the literature, which only corrects for asset correlation, does not provide a sufficient correction in practice, as DRs also exhibit a high degree of persistence.

Finally, our generalised binomial test correction (BT-C ρ & ψ), which simultaneously corrects for asset and serial correlation, brings the probability of rejection close to the expected value of 5%,

[22] More specifically, we know from (18) that if $PD_1 > PD_2$, then $k(PD_1) > k(PD_2)$ and it follows that $\mathbb{P}\big(\mathcal{D}_{\mathrm{n,t}} \leq k(PD_1)\big) \geq \mathbb{P}\big(\mathcal{D}_{\mathrm{n,t}} \leq k(PD_2)\big)$.

regardless of the persistence level specifications. As a result, it becomes clear that the generalised correction we propose can be considered as a sufficient correction of the binomial test, allowing the threshold to be set so that the probability of rejection of the model corresponds to a desired level.

3.4 Multiyear Backtesting

Article 185 of the (EU) No 575/2013 CRR requires credit institutions to regularly validate IRB models to ensure their appropriateness, robustness, and reliability. Accordingly, model validation should not be seen as a one-off exercise, but rather as a continuous and iterative process (BCBS 2005b). This aspect raises key supervisory questions for backtesting: When should an internal model be considered problematic or imprudently calibrated? Is a single year of miscalibration sufficient cause for concern, or should multiple occurrences be observed before such a conclusion is reached?

Blümke (2012) proposes a backtesting procedure for the PD of a given rating grade that considers default data observed over multiple years. The approach relies on order statistics derived from multiyear default rate observations to assess whether the observed default frequency is consistent with the PD assigned to that grade, thereby providing a long-horizon test of calibration at the rating grade level.

The proposed test builds on a general result from probability theory. Given *T* observations of a random variable, the r-th order statistics, denoted $X_{(r)}$ (i.e., the r-th largest observation), has a known cumulative distribution function (CDF):

$$\mathbb{P}\left[X_{(r)} < x\right] = \sum_{i=k}^{T} \binom{T}{i} [F_X(x)]^k [1 - F_X(x)]^{k-1}. \tag{19}$$

In the context of PD model validation, this result applies by interpreting $X_{(r)}$ as the r-th largest observed default rate. Following Blümke (2012),[23] we focus on the second-highest observed default rate, so that $X_{(r)}$=$\mathcal{D}R_t^{(2)}$. The underlying CDF $F_X(x)$ is given by $\mathbb{P}(\mathcal{D}R_t \leq x)$, which allows the Merton-Vasicek model to be incorporated into the derivation of the test. In particular, Blümke (2012) makes use of (17) to account explicitly for asset correlation, yielding:

$$\mathbb{P}\left[\mathcal{D}R_t^{(2)} < x\right] = \sum_{i=k}^{T} \binom{T}{i} \left[\mathbb{P}\left(\frac{\sqrt{1-\rho}\Phi^{-1}(x)-c}{\sqrt{\rho}}\right)\right]^k \left[1 - \mathbb{P}\left(y \leq \frac{\sqrt{1-\rho}\Phi^{-1}(x)-c}{\sqrt{\rho}}\right)\right]^{k-1}. \tag{20}$$

Given a confidence level α (in our case $\alpha = 5\%$), it is necessary to determine numerically the value of x^* such that:

$$\mathbb{P}\left[\mathcal{D}R_t^{(2)} < x^*\right] = \alpha = 5\%. \tag{21}$$

The resulting quantity x^*, referred to by Blümke (2012) as the critical PD, serves as the threshold against which $\mathcal{D}R_t^{(2)}$ is compared. If the observed second-highest default rate exceeds x^*, the null hypothesis of correct model calibration is rejected.

The following example illustrates the behaviour of the critical threshold x^*. Consider a time series of eight annual default rates for a given rating grade, with an assigned PD of 1% and an estimated

[23] Selecting the second-largest value avoids reliance on a single exceptional year, which could otherwise disproportionately influence the results if the maximum were used.

asset correlation of 5%. In this case, the critical threshold is $x^* = 2.27\%$. If the second-highest observed default rate exceeds this value, the null hypothesis of correct model calibration is rejected. If instead the sample consist of 15 annual observations, with all other parameters unchanged, the critical threshold increases to $x^* = 2.66\%$. The increase in x^* with the length of the sample is consistent with the intuition that, as more observations are available, the likelihood of observing more extreme realizations increases.

Table 10: Critical Values for the Order Test in Case the Second-highest Default Rate is Considered

PD	T	ρ					
		0.25%	**0.50%**	**1.00%**	**2.50%**	**5.00%**	**10.00%**
1%	**8**	1.24%	1.35%	1.50%	1.85%	2.27%	2.93%
	15	1.29%	1.42%	1.62%	2.07%	2.66%	3.64%
5%	**8**	5.91%	6.30%	6.89%	8.13%	9.65%	12.03%
	15	6.08%	6.56%	7.29%	8.85%	10.84%	14.10%

As proposed, the order test does not account for potential serial correlation in default rates. Table 11 reports the results of a Monte Carlo simulation in which five million times series of length 8 and 15 are generated using the model in (9) and (10), and the order test is applied to the second-highest observed default rate. The critical value is computed under the assumption that the PD is known and that asset correlation equals 5%.

As expected, when the asset corelation used to generate default rates matches the assumed value (5%) and there is no serial correlation, the empirical rejection frequency is close to the nominal confidence level. When default rates are uncorrelated both cross-sectionally ($\rho = 0$) and serially ($\psi = 0$), the order test never rejects the null hypothesis, as the critical value is overly conservative in this case. Finally, when serial correlation is present but not accounted for in the order test, the rejection frequency exceeds the confidence level.

Table 11: Probability of Rejecting the Hypothesis that the Model is Correctly Calibrated with the Order Test Knowing the PD and with Asset Correlation $\rho = 5\%$

PD	T	$\rho = 0\%$	$\rho = 5\%$	$\rho = 5\%$
		$\psi = 0\%$	$\psi = 0\%$	$\psi = 40\%$
1%	**8**	0.0%	5.29%	11.22%
	15	0.0%	5.28%	12.01%
5%	**8**	0.0%	5.10%	11.03%
	15	0.0%	5.11%	11.81%

Section 4 also presents aggregate results from applying the order test to the bank-level data. In that application, the critical value is computed using the PD associated with each rating grade and assuming an asset correlation of 5%.

3.5 Replication of Real-world Systematic Backtesting Exercise

In this section, we replicate our backtesting exercise for a sample of EU banks, incorporating different levels of asset and serial correlation to provide a more nuanced perspective. Table 12 shows the distribution of weighted, miscalibrated exposures, where Panel A is identical to Table 4.

Table 12: Replication of Miscalibrations of EU Banks Weighted by the Exposure at Default (EAD) from Table 4

A. Binomial Test										
Sample		**MCB**	**%**	**Mean**	**SD**	**P5**	**P25**	**P50**	**P75**	**P95**	**RG**
Total		967,444	16.7	3,214	4,716	31	372	1,394	3,895	12,071	301
B. Binomial Test Corrected (ρ & ψ)											
ρ	ψ	**MCB**	**%**	**Mean**	**SD**	**P5**	**P25**	**P50**	**P75**	**P95**	**RG**
1	40	647,671	11.1	2,699	4,023	20	271	1,018	3,388	10,984	240
3	40	398,218	6.9	2,505	3,739	18	263	952	3,396	10,561	159
5	40	317,281	5.5	2,498	3,957	17	184	809	3,205	10,561	127
7	40	251,612	4.3	2,396	3,851	17	186	806	3,203	10,242	105
10	40	205,501	3.5	2,418	4,065	15	154	772	3,205	10,242	85
1	60	613,927	10.6	2,729	4,112	22	279	983	3,396	11,111	225
3	60	383,287	6.6	2,625	3,870	19	186	998	3,471	10,561	146
5	60	256,890	4.4	2,357	3,807	15	174	806	3,203	10,242	109
7	60	214,320	3.7	2,381	3,966	15	154	789	3,205	10,242	90
10	60	174,777	3.0	2,330	4,220	13	118	597	2,375	10,756	75
1	80	443,935	7.6	2,466	3,619	20	235	966	3,300	9,986	180
3	80	256,681	4.4	2,399	3,830	15	184	806	3,205	10,242	107
5	80	204,006	3.5	2,458	4,104	15	154	772	3,396	10,242	83
7	80	171,566	3.0	2,486	4,362	15	184	772	2,859	10,756	69
10	80	167,916	2.9	2,624	4,498	17	169	789	3,300	10,756	64

Note. This table presents the number and miscalibrated exposures based on the standard binomial test and the binomial test corrected for asset and serial correlation under the assumption of different 'known' parameter realisations. The literature suggests that asset correlations are between 5% and 10% depending on the segment (Blümke 2022b), and serial correlations are often estimated to be much higher, at around 80% (Blümke 2022a). ρ and ψ are expressed in percentages, RG indicates the number of miscalibrated rating grades, % indicates the miscalibrated exposures in percentages, and the remaining values are in terms of the miscalibrated exposures in EUR millions.

Even under modest assumptions, namely an asset correlation of $\rho = 1\%$ and a serial correlation of $\psi = 40\%$, the share of miscalibrated exposures falls by 5.6 percentage points, which corresponds to a reduction of miscalibrated exposures of around one third. Previous literature suggests that asset correlations tend to be between 5% and 10% depending on the segment (Blümke 2022b), while the regulatory parameters are between 4% and 24% (BCBS 2020 and Article 153(4) of Regulation (EU) No 575/2013). In addition, the serial correlation is often estimated to be much higher, at around 80% (Blümke 2022a). Under these more realistic assumptions (i.e., $\rho = 5 - 10\%$ and $\psi = 80\%$), the reduction in miscalibrated exposures is between 13.2 and 13.8 percentage points. As a result, the miscalibrated exposures fall between 3.5% and 2.9% of the total exposures, which corresponds to a reduction in the identified miscalibrations of around 80% compared to the standard binomial test. The data available to us, particularly the limited time dimension, do not allow us to estimate the parameters governing asset and serial correlation. Our objective here, however, is not parameter estimation, but rather to demonstrate how substantially the results can differ once the binomial test is corrected to account for cross-sectional and/or serial correlation. Put

differently, we aim to provide an illustrative range of outcomes (see Figure 4), with one extreme corresponding to the classic binomial test and the other to a correlation-adjusted version. For the serial correlation parameter, one could adopt the value observed in our data, approximately 40%. Instead, we prefer a higher value at 80%, in order to construct a range that spans two extreme scenarios and thereby highlights the sensitivity of the results to alternative dependence assumptions.

Table 13: Replication of Real Effects of Miscalibrations ($\rho = 5\% \ \& \ \psi = 80\%$)

Subset	MCB	%	Tier 1	Δ
A. Overall Sample				
Total	204,006	3.5	15.7	-3.8
B. Across Rating Categories				
IG	192,706	6.8	15.6	-6.1
Non-IG	11,300	0.4	15.8	-0.2
C. Over Time				
2017	22,132	3.9	14.8	-1.7
2018	42,289	6.7	15.2	-2.5
2019	22,195	3.0	15.7	-1.3
2020	37,189	4.9	16.5	-2.1
2021	42,715	5.5	15.4	-9.5
2022	18,455	2.3	15.5	-8.9
2023	19,033	2.4	16.3	-8.9
2024	0	0.0	16.3	0.0

Note. This table presents the real effects of miscalibrations based on the binomial test corrected for asset and serial correlation under the assumption of 'known' realisations of ρ and ψ. The literature suggests that asset correlations are between 5% and 10% depending on the segment (Blümke 2022b), and serial correlations are often estimated to be much higher, at around 80% (Blümke 2022a). MCB and % indicates the miscalibrated exposures in EUR millions and percentages, respectively. Tier 1 is expressed in percentages and the corresponding economic impact Δ is expressed in basis points.

Table 13 shows the miscalibrated exposures and their corresponding impact on the Tier 1 capital ratios. In all dimensions, the values are substantially lower compared to those shown in Table 6. Under more realistic assumptions (i.e., $\rho = 5\%$ and $\psi = 80\%$) and using the hypothetical RWAs from (4), a prudent re-calibration of the risk models would lead to a reduction (Δ) in the Tier 1 capital ratios of 3.8 basis points at the system level across all banks. This corresponds to a reduction of 6.5 basis points compared to the standard binomial test. Disaggregated by rating category (Panel B), the effects are again more pronounced in the IG category, where the Tier 1 capital ratio falls by 6.9 basis points compared to 13.8 basis points with the standard binomial test. Similar patterns emerge over time (Panel C), with no miscalibrated exposures and therefore no impact on Tier 1 capital ratios in 2024. This indicates a relatively high margin of conservatism in the banks' internal risk models.

4. Backtesting Dashboard

Although backtesting of PD models used by IRB banks is common practice among both banks and supervisors, publicly available information on the results of these exercises remains scarce. This is probably due to two main factors: First, backtesting techniques are typically designed to assess a single model in a single institution, rather than providing insights across multiple banks and models; and second, none of the currently available backtesting methods are entirely free of limitations. These factors may raise concerns that the publication of such results could lead a non-expert audience to draw premature or misleading conclusions.

Nonetheless, we argue that the regular publication of backtesting results for PD models across the EU banking sector could increase transparency, restore confidence in internal models, and facilitate the identification of systematic problems. To this end, we suggest using regulatory reporting data to publish a dashboard summarising these results, as shown in Table 14.[24] Comparing the standard binomial test with alternative specifications allows for a more nuanced interpretation and ensures the necessary caution in drawing conclusions, as each approach emphasises different aspects. Recall, the binomial test relies on the assumption of independent defaults, which makes the test a conservative benchmark; the resulting miscalibration rates can therefore be interpreted as an upper bound. We also introduce a novel aggregation mechanism from the rating grade level to the system level. While the rating grade perspective remains valuable for risk modelling and supervisory purposes, the system level view is particularly relevant from a macroprudential and financial stability perspective.

Panel A of the proposed dashboard shows the aggregate input factors by comparing the EAD weighted average PD reported by the different banks with the corresponding EAD weighted average annual DRs. The results show a comfortable average PD to DR ratio of 1.5, with a range between 1.2 and 1.9. However, as the capital requirements of one bank cannot offset the risks of another, more formal and systematic backtesting procedures are needed to assess the calibration of the PD models of EU banks.

Panel B applies the standard binomial test, which leads to a conservative estimate of the upper bound of around 12.1% miscalibrated exposures in 2024 at the aggregate level. This miscalibration corresponds to a decrease in the average Tier 1 capital ratio of 4.6 basis points assuming a prudent re-calibration of risk models. These results underline the importance of regular validation of internal models to ensure the resilience of the EU banking sector and promote financial stability. Previous literature suggests that asset correlations are typically between 5% and 10% and that serial correlation is around 80%. When these more realistic parameter values and our corrected binomial test are taken into account (Panel C), the miscalibrated exposures and their impact on Tier 1 capital ratios are significantly lower. Figure 4 provides further insight by illustrating the likely range of miscalibrated exposures (left panel) and the corresponding Tier 1 capital impact (right panel). The estimates range from the conservative upper bound derived from the standard binomial test to the more realistic values derived using the corrected binomial test.

[24] Comparative dashboards are widely used in other domains, such as used-car brand reliability, business school rankings, and electronic product evaluations.

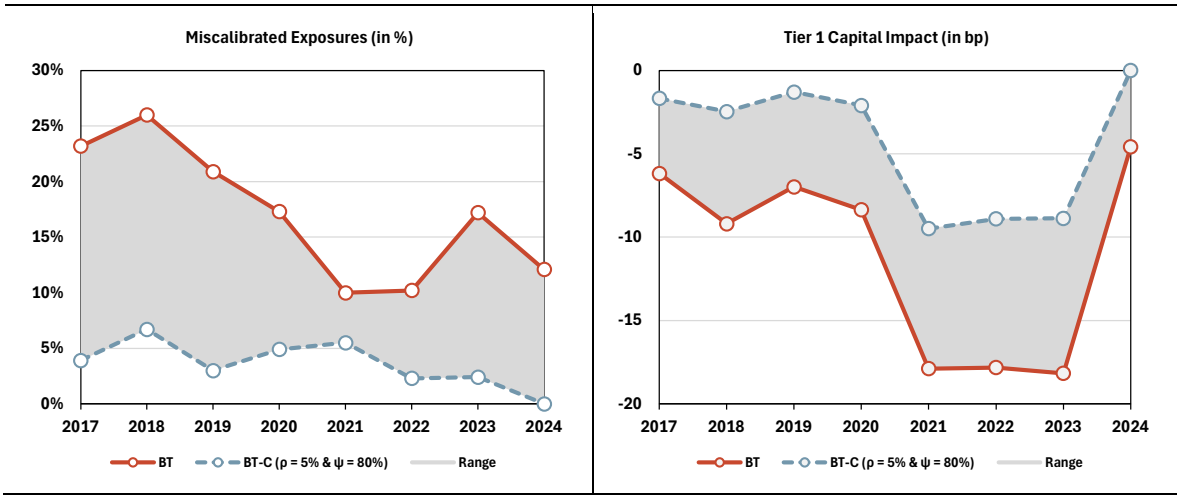
Table 14: Dashboard on Miscalibrations of EU Banks Over Time

	2017	2018	2019	2020	2021	2022	2023	2024
A. Inputs								
Weighted Avg. PD	2.4	2.1	2.0	2.1	2.0	1.9	2.0	2.1
Weighted Avg. DR	2.0	1.6	1.5	1.4	1.1	1.1	1.4	1.4
PD / DR	1.2	1.4	1.4	1.4	1.9	1.7	1.4	1.4
B. Binomial Test								
Weighted Avg. PD (MCB)	1.7	1.6	1.5	1.2	0.9	1.1	1.7	1.6
Weighted Avg. DR (MCB)	3.7	2.7	2.4	2.7	2.4	2.4	3.3	2.3
PD / DR	0.5	0.6	0.6	0.4	0.4	0.5	0.5	0.7
EAD MCB	130.1	165.2	156.5	130.6	77.1	80.2	135.0	92.9
% MCB	23.2	26.0	20.9	17.3	10.0	10.2	17.2	12.1
Impact RWA %	6.5	7.2	6.0	7.3	10.9	10.2	10.6	3.2
Impact Tier 1 bp	-6.2	-9.2	-7.0	-8.4	-17.9	-17.8	-18.2	-4.6
C. Binomial Test Corrected ($\rho = 5\%$ & $\psi = 80\%$)								
Weighted Avg. PD (MCB)	0.4	0.2	0.5	1.2	0.2	1.0	1.6	-
Weighted Avg. DR (MCB)	3.9	1.3	1.9	4.6	1.9	3.7	7.8	-
PD / DR	0.1	0.1	0.3	0.3	0.1	0.3	0.2	-
EAD MCB	22.1	42.3	22.2	37.2	42.7	18.5	19.0	-
% MCB	3.9	6.7	3.0	4.9	5.5	2.3	2.4	-
Impact RWA %	1.8	1.9	1.1	1.8	5.8	5.1	5.1	-
Impact Tier 1 bp	-1.7	-2.5	-1.3	-2.1	-9.5	-8.9	-8.9	-
D. Multiyear Test ($\rho = 5\%$ & $\psi = 0\%$)								
%EAD MCB								16.1
Weighted Avg. Critical PD MCB								1.41
Weighted Avg. 2nd Highest DR MCB								1.84

Note. This table presents an exemplary dashboard of miscalibrations that show a number of metrics at the system level to provide insights in the backtesting results of PD models across EU banks over time. The distributions reported in Panel A are EAD weighted and computed using all available rating grades, whereas Panels B and C consider miscalibrated rating grades only. We use the binomial test corrected for asset and serial correlation under the assumption of 'known' realisations of ρ and ψ. The literature suggests that asset correlations are between 5% and 10% depending on the segment (Blümke 2022b), and serial correlations are often estimated to be much higher, at around 80% (Blümke 2022a). The miscalibrations are expressed in EUR billions, the Tier 1 impact in basis points, the PD/DR ratio in percentage points, and the remaining variables are expressed in percentages.

Model validation is not a one-off exercise, but a continuous and iterative process. To support this, Table 14, Panel D presents the results of a novel heuristic for multiyear backtesting. First, note that the share of EAD associated with miscalibrated rating grades is higher than that reported for the binomial test in 2024 (16.1% versus 12.1%). At first glance, this may appear counterintuitive, since the order test assumes a non-zero asset corelation and, as shown in the previous section, this assumption widens the acceptance range of the null hypothesis of correct model calibration. However, it is important to recall that the order test is applied over the entire time span of observed years, whereas the binomial test is conducted on a single-year basis. It is also noteworthy that the average PD of the miscalibrated rating grades identified by the order test is lower than the average PD of the rating grades that did not pass the binomial test. This indicates that the order test is able to detect calibration issues in rating grades with relatively low PDs that are not flagged by the binomial test.

Figure 4: Range of Miscalibrated Exposures and Tier 1 Capital Impact of EU Banks Over Time



Note. This figure shows the range of miscalibrated exposures in the left panel and the impact on Tier 1 capital in the right panel based on the standard binomial test and the binomial test corrected for asset and serial correlation. We use the binomial test corrected for asset and serial correlation under the assumption of 'known' realisations of ρ and ψ. The literature suggests that asset correlations are between 5% and 10% depending on the segment (Blümke 2022b), and serial correlations are often estimated to be much higher, at around 80% (Blümke 2022a).

5. Conclusions

This paper expands the current understanding of backtesting practices for PD models by mitigating the main limitations of conventional validation tools. Leveraging a standardised, cross-institutional dataset reported by banks to their supervisors, we provide a comprehensive framework that combines theoretical refinements, simulation-based analysis, and empirical validation, applied to the IRB models of EU banks. Our generalised binomial test correction, which accounts for both asset and serial correlation, improves the reliability of statistical validation. Furthermore, we use order statistics for the evaluation of multi-period settings and provide a methodology for the aggregation of validation results for different institutions. Finally, we highlight the potential economic impact of incorrect PD model calibrations on capital adequacy.

Taken together, our results provide theoretical, policy-relevant, and practical insights for model validation, prudential risk assessment, and banking supervision. Building on these contributions, the paper offers a comprehensive framework to assist banks and supervisors in the evaluation and monitoring of PD models, ensuring their robustness, accuracy, and compliance with regulatory standards. By utilising the methodologies and insights presented, institutions can enhance risk management practices and support the resilience and stability of the financial system.

References

Balthazar, L. (2004). PD Estimates for Basel II. Risk, 17, 84–85.

Basel Committee on Banking Supervision (2004). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework.

Basel Committee on Banking Supervision (2005a). Studies on the Validation of Internal Rating Systems. Working Paper.

Basel Committee on Banking Supervision (2005b). Update on Work of the Accord Implementation Group Related to Validation Under the Basel II Framework, Newsletter No. 4.

Basel Committee on Banking Supervision (2011). Final Elements of the Reforms to Raise the Quality of Regulatory Capital Issued by the Basel Committee.

Basel Committee on Banking Supervision (2020). CRE 31 IRB Approach: Risk Weight Functions.

Benjamin, N., Cathcart, A., & Ryan, K. (2006). Low Default Portfolios: A Proposal for Conservative Estimation of Default Probabilities.

Berg, J., Boivin, N., & Geeroms, H. (2025). The European Union Should Not Water Down Bank Capital Standards. Bruegel.

Blochwitz, S., Martin, M. R. W., & Wehn, C. S. (2006). Statistical Approaches to PD Validation. In: The Basel II Risk Parameters: Estimation, Validation, and Stress Testing – with Applications to Loan Risk Management, Berlin, Heidelberg: Springer Berlin Heidelberg.

Blöchlinger, A., & Leippold, M. (2011). A New Goodness-of-Fit Test for Event Forecasting and Its Application to Credit Defaults. Management Science, 57, 487–505.

Blöchlinger, A., & Leippold, M. (2018). Are Ratings the Worst Form of Credit Assessment Except for All the Others? Journal of Financial and Quantitative Analysis, 53, 299–334.

Blümke, O. (2012). Probability of Default Validation: A Single-year and Multiyear Methodology for the Basel Framework. Journal of Risk Model Validation, 6, 47–79.

Blümke, O. (2013). Probability of Default Validation: Introducing the Likelihood-ratio Test and Power Considerations. Journal of Risk Model Validation, 7, 29–59.

Blümke, O. (2022a). A Structural Hidden Markov Model for Forecasting Scenario Probabilities for Portfolio Loan Loss Provisions. Knowledge-Based Systems, 249, 108934.

Blümke, O. (2022b). Testing the Predictive Power: A Comparative Study of Current Default Probability Validation Tests. Expert Systems With Applications, 203, 117312.

Blümke, O. (2025). The Vasicek Distribution Autoregressive Time-Series Model for Aggregated Data of Default and Delinquency Rates. Journal of the Royal Statistical Society Series A: Statistics in Society, forthcoming.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review, 78, 1–3.

Buch, C. (2025). Introductory Statement: Hearing of the Committee on Economic and Monetary Affairs of the European Parliament, Speech.

Cannata, F., & Serafini, L. (2025). A Pragmatic Approach to Simplification: The Case of Banking Regulation in the EU. Bank of Italy Occasional Papers, 955, 1–34.

Carlehed, M., & Petrov, A. (2012). A Methodology for Point-in-time—Through-the-cycle Probability of Default Decomposition in Risk Classification Systems. Journal of Risk Model Validation, 6, 3–25.

Castermans, G., Martens, D., Van Gestel, T., Hamers, B., & Baesens, B. (2010). An Overview and Framework for PD Backtesting and Benchmarking. Journal of the Operational Research Society, 61, 359–373.

Comfort, N., & Arons, S. (2023). Global Banks' Post-Brexit Shift to Europe Almost Done, ECB Says. Bloomberg.

Conover, W. J. (1999). Practical Nonparametric Statistics, Hoboken: Wiley.

Draghi, M. (2024). The Future of European Competitiveness: In-depth Analysis and Recommendations. European Commission.

Elderson, F. (2024). Introductory Remarks: Embedding a Strong Data Culture in Supervision – Another Stepping Stone Towards Effective Supervision. European Central Bank, Speech.

European Central Bank (2015). The Financial Risk Management of the Eurosystem's Monetary Policy Operations.

European Central Bank (2019). Instructions for Reporting the Validation Results of Internal Models: IRB Pillar I Models for Credit Risk.

European Central Bank (2024). ECB Guide to Internal Models.

EU (2013a). Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms.

EU (2013b). Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms.

EU (2015). Guideline (EU) 2015/510 of the European Central Bank of 19 December 2014 on the implementation of the Eurosystem monetary policy framework.

Global Credit Data (2025). Asset Correlations in Credit Risk: An Empirical Study with GCD Data. Report.

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression, Hoboken: Wiley.

Letta, E. (2024). Much More Than a Market – Speed, Security, Solidarity: Empowering the Single Market to Deliver a Sustainable Future and Prosperity for all EU Citizens.

Miu, P., & Ozdemir, B. (2008). Estimating and Validating Long-run Probability of Default with Respect to Basel II Requirements. Journal of Risk Model Validation, 2, 3–41.

Noonan, L., & Comfort, N. (2025). ECB Seeks to Speed Up Bank Capital Model Approvals, Maijoor Says. Bloomberg.

Restoy, F. (2025). Financial Regulation and Growth: What Should be the European Policy Priorities? Speech.

Schechtman, R. (2017). Joint Validation of Credit Rating PDs under Default Correlation. International Journal of Central Banking, 49, 235–282.

Tasche, D. (2003). A Traffic Lights Approach to PD Validation. Working Paper.

Tasche, D. (2005). Rating and Probability of Default Validation. BCBS Working Paper No. 14 on Studies on the Validation of Internal Rating Systems.

Vasicek, O. (2002). The Distribution of Loan Portfolio Value. Risk, 15, 160–162.

Villeroy de Galhau, F. (2025). A European Approach to Simplification: Avoiding Three Misconceptions and Suggesting Concrete Milestones, Speech.

Appendix

Equation, Table, and Figure numbers not preceded by a letter refer to the main article.

A. Variable Description

Variable	Description
A. Bank Level	
RWA	Total risk weighted exposure amount calculated according to Article 92(3) and Articles 95, 96 and 98 CRR (in EUR millions)
Tier 1 Ratio	Tier 1 capital ratio is the Tier 1 capital of the institution expressed as a percentage of the total risk exposure amount (in decimal percentage)
B. Rating Grade Level	
Probability of Default (PD)	Arithmetic average of PD at the beginning of the reporting period of the obligors that fall within the bucket of the fixed PD range (in decimal percentage)
Default Rate (DR)	One-year DR referred to in point (78) Article 4(1) CRR (in decimal percentage)
Defaults	Number of obligors which defaulted during the year (i.e. the observation period of the DR calculation). Defaults are determined in accordance with Article 178 CRR. Each defaulted obligor is counted only once in the numerator and denominator of the one-year DR calculation, even if the obligor defaulted more than once during the relevant one-year period (frequency)
Obligors	All obligors carrying a credit obligation at the relevant point in time shall be included (frequency)
EAD	Exposure value calculated in accordance with Article 166(1) to (7) CRR without taking into account any credit risk adjustments (in EUR millions)
RWA	Total risk weighted exposure amount for credit risk calculated under the IRB approach for a given rating grade (in EUR millions)
Maturity	For all exposures included in each bucket of the fixed PD range, the average maturity of each exposure, weighted by the exposure value post-CCF; the maturity is determined in accordance with Article 162 CRR (in years)

B. Binomial Test vs. Jeffreys Test

Table B.1: Probability of Rejecting the Model with Known Parameters

			PD = 0.5%			PD = 5.0%		
n	ρ	ψ	**BT**	**JT**	Δ	**BT**	**JT**	Δ
500	0.0	0.0	**4.2**	**4.2**	**0.0**	**4.1**	**4.1**	**0.0**
500	5.0	0.0	10.3	10.3	0.0	21.9	21.9	0.0
500	5.0	40.0	10.9	10.9	0.0	22.7	22.7	0.0
500	5.0	60.0	11.6	11.6	0.0	23.7	23.7	0.0
500	5.0	80.0	12.9	12.9	0.0	25.0	25.0	0.0
1,000	0.0	0.0	**6.8**	**6.8**	**0.0**	**4.7**	**4.7**	**0.0**
1,000	5.0	0.0	16.4	16.4	0.0	27.0	27.0	0.0
1,000	5.0	40.0	16.9	16.9	0.0	27.6	27.6	0.0
1,000	5.0	60.0	17.5	17.5	0.0	28.2	28.2	0.0
1,000	5.0	80.0	17.9	17.9	0.0	28.6	28.6	0.0
5,000	0.0	0.0	**4.4**	**4.4**	**0.0**	**4.9**	**4.9**	**0.0**
5,000	5.0	0.0	23.7	23.7	0.0	34.7	34.7	0.0
5,000	5.0	40.0	23.9	23.9	0.0	34.8	34.8	0.0
5,000	5.0	60.0	24.0	24.0	0.0	34.7	34.7	0.0
5,000	5.0	80.0	23.2	23.2	0.0	33.5	33.5	0.0
10,000	0.0	0.0	**5.1**	**5.1**	**0.0**	**4.7**	**4.7**	**0.0**
10,000	5.0	0.0	27.5	27.5	0.0	36.8	36.8	0.0
10,000	5.0	40.0	27.5	27.5	0.0	36.7	36.7	0.0
10,000	5.0	60.0	27.1	27.1	0.0	36.2	36.2	0.0
10,000	5.0	80.0	25.4	25.4	0.0	34.6	34.6	0.0

Note. This table shows the probability of rejecting the model for the standard binomial test (BT) and the Jeffreys test (JT). Light grey indicates that the probability of rejection is close to the expected value of 5% at a confidence level of 95%, and the darker the grey, the further away the probability of rejection is from the expected value. Δ is the difference between the BT and the JT.

ACKNOWLEDGEMENTS

We thank Roberto Mosca, Werner Osterkamp, Giovanni Rinna, Franco Varetto, and three anonymous referees for useful comments and suggestions. All remaining errors are ours.

Simone Casellina
Economic Analysis and Impact Assessment Unit, Economic and Risk Analysis Department, European Banking Authority

Gaetano Chionsini
Statistics Unit, Data Analytics, Reporting and Transparency Department, European Banking Authority

Raphael M. Kopp
Economic Analysis and Impact Assessment Unit, Economic and Risk Analysis Department, European Banking Authority & University of Duisburg-Essen

Maroua Riabi*
* At the time of developing the methodology described in this paper, Ms. Riabi was part of the Statistics Unit, Data Analytics, Reporting and Transparency Department, European Banking Authority

Tour Europlaza, 20 avenue André Prothin CS 30154
92927 Paris La Défense CEDEX, FRANCE
Tel. +33 1 86 52 70 00

E-mail: info@eba.europa.eu

https://eba.europa.eu

ISBN 978-92-9407-288-7
ISSN 2599-7831
doi:10.2853/4618175
DZ-01-26-002-EN-N