Heterogeneity Testing in Internal Ratings-Based Credit Risk Models

Regulatory Requirements, Methods, Challenges, and Considerations

Andrija Djurovic*

19 December 2025

Abstract

One key aspect of analyzing the distribution and allocation of obligors and facilities in Internal Ratings-Based (IRB) credit risk models is heterogeneity. It refers to adequate differentiation in risk profiles across ratings, pools, or buckets. Heterogeneity is commonly tested in Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD) models, often using tests like the two-proportion test and Welch t-test. Credit risk practitioners closely examine the heterogeneity of the rating system as a regulatory requirement. However, challenges such as defining specific thresholds for the testing procedure and drawing overall conclusions about the level of heterogeneity remain insufficiently explored in practice. This paper summarizes regulatory requirements, standard practices, and challenges practitioners face in this exercise while also contributing to the field by proposing additional statistical methods to enhance heterogeneity assessment. Specifically, assuming a well-calibrated model, the paper demonstrates how statistical power analysis and extensions of the standard tests can help practitioners make more informed decisions about the heterogeneity of the rating system.

*External Consultant, Deloitte Sweden. The views expressed in this paper are the author's own and do not necessarily reflect those of the employer.

1 Regulatory Requirements for Model Heterogeneity

Assessing model heterogeneity is a critical aspect of credit risk modeling under the Internal Ratings-Based (IRB) approach. Regulatory requirements emphasize the need for meaningful differentiation across rating grades or pools to ensure accurate risk quantification. The Capital Requirements Regulation (CRR), the ECB Guide to Internal Models, the Commission Delegated Regulation (EU) 2022/439, and
Draft guidelines on Credit Conversion Factor estimation under Article 182(5) of Regulation (EU) No 575/2013 outline key requirements for model structure and validation in rating systems.

The CRR establishes the foundation for rating system requirements, ensuring that models effectively differentiate risk levels and maintain sufficient granularity. Specifically, Article 170(1) outlines the requirements for rating systems applied to corporate, institutional, central government, and central bank exposures, including:

- Obligor rating scales must reflect the risk of obligor default and include at least seven grades for non-defaulted obligors and one for defaulted obligors.
- Institutions must document the relationship between obligor grades and the corresponding level of default risk.
- Concentrations within a single grade must be justified by empirical evidence demonstrating that obligors within the grade share similar default risk characteristics.
- Institutions using their own estimates of Loss Given Default (LGD) must establish a distinct facility rating scale that captures LGD-specific transaction characteristics.

Furthermore, Article 170(3)(c) applies similar requirements to retail exposures, emphasizing that rating systems must:

- Ensure meaningful risk differentiation by grouping sufficiently homogeneous exposures.
- Avoid excessive concentrations within a single grade or pool unless justified by empirical evidence.
- Enable accurate and consistent estimation of loss characteristics at the grade or pool level.

The Commission Delegated Regulation (EU) 2022/439 provides further clarifications and assessment criteria for competent authorities to verify compliance with CRR requirements. Article 34(1) requires authorities to assess whether the number of grades and pools enables meaningful risk differentiation and accurate loss quantification. Specifically:

- The number of rating grades must meet the minimum requirements set out in CRR

Article 170.

- The distribution of obligors and exposures should avoid excessive concentration unless supported by empirical evidence of homogeneity.
- Retail exposure rating systems must ensure that grades or pools contain a sufficient number of exposures unless justified by direct risk parameter estimates.

Article 34(2) also requires authorities to evaluate the criteria institutions use to determine the overall number of grades or pools and the proportion of exposures assigned to each category.

The ECB Guide to Internal Models, in the Credit risk section, sets more detailed expectations for risk differentiation across grades or pools:

- Paragraph 211 requires institutions to ensure adequate differentiation of default risk across grades or pools, avoiding significant overlaps in risk distributions. This is especially important for institutions using highly granular rating scales.
- Paragraph 240 emphasizes that institutions must document their calibration approach and provide empirical evidence that PD estimates remain appropriate at both the grade and segment levels. Any method used to address data scarcity must not compromise risk differentiation, including heterogeneity across grades.
- Paragraph 284 outlines requirements for the appropriate distribution of facilities across grades and pools to ensure meaningful LGD quantification. This includes:
  - Providing empirical evidence to justify low or high concentrations within a grade.
  - Verifying homogeneity within each grade by demonstrating that facilities in the same grade exhibit similar LGD levels.
  - Ensuring heterogeneity across grades by proving that average realized LGD levels differ significantly between consecutive grades.

Although not explicitly stated, the same principles regarding heterogeneity apply to Exposure at Default (EAD) models. This is formally confirmed in paragraph 83(c) of section 6.1.2 Homogeneity and Heterogeneity in the Draft Guidelines on Credit Conversion Factor Estimation under Article 182(5) of Regulation (EU) No 575/2013, which indicates that institutions should avoid significant overlaps in the distribution of conversion risk between grades and pools.

The CRR, the Commission Delegated Regulation (EU) 2022/439, the ECB Guide to Internal Models, and Draft guidelines on Credit Conversion Factor estimation under Article 182(5) of Regulation (EU) No 575/2013 collectively establish stringent requirements for rating system heterogeneity. These regulations stress that rating systems must ensure risk differentiation,

prevent excessive concentration, and enable meaningful loss estimation at the grade or pool level. Compliance with these requirements is crucial for institutions to maintain regulatory approval for IRB models and ensure that risk parameters accurately reflect the creditworthiness of exposures.

2 Common Practices in Heterogeneity Testing

In practice, heterogeneity testing is performed for all three risk parameters - PD, LGD, and EAD - as a standard procedure for both initial and periodic model validation. For discrete model outputs, heterogeneity is tested at the level of rating grades or adjacent pool pairs. A common practice for continuous model outputs is to discretize the model estimates before conducting heterogeneity testing. A typical approach for defining the buckets follows the European Central Bank (2019) guidelines. Furthermore, standard practice suggests conducting heterogeneity testing based on the observed or realized values of specific risk parameters, typically using a one-sided hypothesis test. After sorting the rating scale, pools, or buckets in ascending order, the null hypothesis is usually defined to test whether the average observed risk parameter of the better grade, pool, or bucket is greater than or equal to that of the worse one. In this context, "better" and "worse" refer to adjacent pairs, where "better" indicates a lower calibrated risk parameter value. Therefore, the favorable outcome of heterogeneity testing is rejecting the null hypothesis.

Given that the Probability of Default (PD) is defined as the proportion of defaulted obligors or facilities to the total number of obligors or facilities, the test of two proportions is a natural choice for testing heterogeneity between adjacent rating grades or buckets. The following formula presents the test statistic for the two-proportion test:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where:

- p_1 is the observed proportion of the "better" rating grade;
- p_2 is the observed proportion of the "worse" rating grade;
- p is the pooled proportion calculated as $\frac{n_1\hat{p_1}+n_2\hat{p_2}}{n_1+n_2}$;
- n_1 and n_2 denote the sizes of the "better" and "worse" rating.

The two-proportion test assumes that the test statistic Z follows a standard normal

distribution. Formally, the p-value is calculated as $pnorm(q = Z)$, where $pnorm$ is the cumulative standard normal distribution evaluated at the quantile Z. The final step in hypothesis testing is to compare the calculated p-value to the chosen significance level.

On the other hand, LGD and EAD are considered continuous variables, typically taking values within a specific range. A natural choice for testing heterogeneity is the Welch t-test. Practitioners sometimes use alternative statistical tests or complement the Welch t-test with other testing procedures, but this is beyond the scope of this paper. Therefore, the following sections will use the Welch t-test as the primary method for this purpose. The following formula represents the test statistic for the Welch t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{x}_1$ is the "better" pool/bucket observed mean;
- $\bar{x}_2$ is the "worse" pool/bucket observed mean;
- s_1^2 and s_2^2 are the sample variances;
- n_1 and n_2 are the sample sizes of the "better" and "worse" pool/bucket.

The test statistic above is assumed to follow a t-distribution, with degrees of freedom calculated using the Satterthwaite approximation, as given by the following formula:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

where:

- s_1^2 and $s_2^2$are the variances of the pools or buckets;
- n_1 and n_2 are the sizes of the respective pools or buckets.

After calculating the test statistic t, the conclusion is typically drawn by determining the p-value based on the t-distribution. For the test of heterogeneity, the p-value is calculated as $pt(q = t, df = df)$, where pt is the cumulative t-distribution evaluated at the quantile t with degrees of freedom df, both computed as described above. The final step in assessing heterogeneity involves comparing the observed p-value to a chosen significance level.

Finally, based on the rating grades, pools, or buckets and the statistical tests used for heterogeneity testing, practitioners define the testing level. Often, there is a discrepancy between initial and periodic model validation. For initial model validation, practitioners typically use the full development sample, which usually consists of multiple portfolio reference dates. This analysis is further supported by analyses performed on separate reference dates within the development sample. During periodic model validation, the heterogeneity test is often conducted at the application portfolio level, defined for a specific reference date, or at the aggregated portfolio level, covering the most recent three reference dates.

After running heterogeneity tests at different levels, practitioners draw an overall conclusion about the level of heterogeneity between adjacent rating grades, pools, or buckets. This conclusion is often supported by expert input on the expected number of pairs for which the heterogeneity test will fail.

Two commonly used thresholds are 20% and 30% for one or, at most, three reference dates, whereas for all reference dates, all pairs typically must pass the test.

3 Main Challenges in Heterogeneity Testing

Analyzing the common approach used in practice, a few questions arise:

- How should practitioners incorporate different levels of testing into the rating scale/pool design?
- Is having a unified threshold for each portfolio type or rating scale/pools sufficient?
- Is it sufficient to assess heterogeneity using only observed averages?
- If the model is well-calibrated, what is the probability of observing differences between adjacent grades, pools, or buckets?
- If the model is well-calibrated, what are the conclusions of heterogeneity testing?
- How is the monotonicity of risk parameters across the rating scale or pools connected to heterogeneity testing?
- What statistical methods can help practitioners support heterogeneity analysis?

The above questions summarize the main challenges practitioners face when testing for model heterogeneity. Focusing only on the observed averages per rating grade or pool, practitioners often overlook the broader aspects of model design and miss the connection to model calibration and final model use. Calibration is critical to model development and should not be separated from heterogeneity testing. In other words, since heterogeneity is tested using statistical hypothesis procedures, relying solely on observed averages can lead to

incorrect conclusions, as these averages incorporate variability that should be accounted for when designing rating scales or pools. In addition, expert-defined thresholds do not effectively capture the specific characteristics of the modeled portfolio.

Another aspect of heterogeneity that is not sufficiently incorporated into this analysis is monotonicity. Specifically, monotonicity is necessary for passing hypothesis testing, but practitioners may still observe disruptions based on the observed averages. When faced with such situations, practitioners often struggle to determine whether the issue stems from the design of the rating scale or pools or if it can be attributed to the variability of the observed averages. Performing heterogeneity testing over time further complicates the conclusion on potential issues with monotonicity disruptions.

This paper presents additional analyses to support current heterogeneity testing practices. Specifically, statistical power analysis can help design rating scales and pools and determine thresholds for overall decisions on heterogeneity levels when testing is conducted on a single or a few reference dates. Also, standard statistical hypothesis testing can be used to define thresholds for monotonicity disruption - another aspect of heterogeneity testing - when assessed on one or a few reference dates. The paper outlines analytical and simulation-based approaches and the underlying assumptions for both methods.

4 Statistical Power Analysis

Cohen (1988) defined the power of a statistical test as "the probability that it will correctly reject a false null hypothesis." Applied to heterogeneity testing in credit risk, a statistical power of 80% means there is an 80% chance of rejecting the null hypothesis - indicating, in other words, an 80% likelihood that heterogeneity exists between adjacent pairs of rating grades, pools, or buckets. Additionally, if the heterogeneity test were repeated 100 times, practitioners would be expected to reject the false null hypothesis in approximately 80 instances.

Power analysis is distinct from other statistical methods in several key ways. While most statistical analyses begin with existing data followed by analysis and interpretation, power analysis can be conducted during the planning phase. For example, statistical power can be used in heterogeneity testing as an input when designing the rating scale or defining pools. It also plays a unique role by helping to define thresholds for drawing overall conclusions about rating system heterogeneity - something not typically addressed by other statistical tools. Another key distinction lies in the interpretation of the output. Standard analyses often require interpreting complex results, whereas power analysis yields a straightforward outcome:

a single value representing the probability of correctly rejecting a false null hypothesis.

Power analysis requires the specification of four key parameters: statistical power level, significance level (α), effect size (ES), and sample size (n). When any three of these parameters are known, the fourth can be calculated. For any given statistical test, power is determined by the combination of sample size, effect size, and significance level.

The following subsections provide an overview of statistical power calculations for the test of two proportions and the Welch t-test using analytical solutions and Monte Carlo simulations. As mentioned in the Common Practices in Heterogeneity Testing section, the test of two proportions is a common choice for assessing the heterogeneity of PD models, while the Welch t-test is typically used for LGD and CCF models.

4.1 Test of Two Proportions

The statistical power of a one-sided test for the difference between two proportions can be calculated analytically using a closed-form expression. The formula below gives the power for the case where the null hypothesis assumes that p_1 is greater than or equal to p_2.

$$\text{Power} = \Phi\left(\frac{-z_{(1-\alpha)}\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})} - \delta}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\right)$$

where:

- Φ is standard normal cumulative distribution function;
- $z_{(1-\alpha)}$ denotes critical value of the standard normal distribution for a one-sided test at significance level α;
- p_1, p_2 are true proportions in grade 1 ("better") and 2 ("worse");
- n_1, n_2 are sample sizes in grade 1 ("better") and 2 ("worse");
- δ is true difference in proportions;
- p is the pooled proportion calculated under the null hypothesis.

In summary, the numerator reflects the difference between the critical threshold and the true effect under the null hypothesis, while the denominator represents the standard deviation under the alternative hypothesis.

Monte Carlo simulation is an alternative approach used to estimate the power of a statistical test by mimicking repeated experiments under controlled conditions. In the context of heterogeneity testing for PD models, the simulation proceeds as follows:

1. Define the true proportions - p_1 and p_2 (calibrated PDs) - and the sample sizes, n_1 and n_2.
2. Generate two samples of sizes n_1 and n_2, consisting of independent binary outcomes from Bernoulli distributions with success probabilities p_1 and p_2.
3. Perform a one-sided test for the difference in proportions, testing whether the proportion in grade 1 is less than the proportion in grade 2 (alternative hypothsis).
4. Repeat steps 2 and 3 a total of N times.
5. Calculate the power as the proportion of simulations in which the null hypothesis is rejected.

4.2 Welch T-Test

The analytical approach for calculating the statistical power of the Welch t-test uses the noncentral t-distribution, with the noncentrality parameter (δ) and degrees of freedom (df) given by the formulas below:

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\left(\frac{\sigma_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\sigma_2^2}{n_2}\right)^2}{n_2 - 1}}$$

where:

- δ is the noncentrality parameter used in the noncentral t-distribution;
- df denotes the degrees of freedom based on the Satterthwaite approximation;
- μ_1, μ_2 are the calibrated estimates of adjacent pools or buckets, labeled "better" and "worse," respectively;
- σ_1^2, σ_2^2 are the variances of the adjacent pools or buckets, labeled "better" and "worse," respectively;
- n_1, n_2 denote the sample sizes of the adjacent pools or buckets.

Similarly to the Monte Carlo simulation for the test of two proportions, it is possible to mimic the data generation process for the Welch t-test as follows:

1. Define the true means (μ_1, μ_2), standard deviations (σ_1, σ_2), and sample sizes (n_1, n_2).
2. Select a distribution for sampling the target values.

3. Generate two samples of sizes n_1 and n_2 from the selected distribution in step 2, with means μ_1, μ_2 and standard deviations σ_1, σ_2, respectively.
4. Perform a one-sided Welch t-test for the difference in means, testing whether the mean in pool/bucket 1 is less than in pool/bucket 2.
5. Repeat steps 3 and 4 a total of N times.
6. Calculate the power as the proportion of simulations in which the null hypothesis is rejected.

Unlike the test of two proportions, where the variance is determined by the proportions and the number of observations, the Welch t-test requires additional assumptions for variance calculation. One possible approach is to estimate the variance based on a random sample of n observations of the target values, where n is the expected number of observations per pool or bucket at the reference date. Another approach is to calculate the variance as the average variance of the target across different reference dates. Both approaches assume that the final risk parameter estimates consist of the model estimates plus an add-on based on the margin of conservatism and any other regulatory adjustments. This assumption appears reasonable given the common practice of calibrating final risk parameter estimates. However, practitioners must decide on the most appropriate approach, as their choice may affect the overall statistical power. Similarly, practitioners should account for potential skewness in the target distribution when selecting a representative distribution for parametric Monte Carlo simulation. The effect of the chosen distribution may be negligible for larger sample sizes, but it is still an important consideration when designing the experiment.
In addition, practitioners can adjust the first three simulation steps to apply a bootstrap method and sample from the observed values of the analyzed risk parameter.

5 Monotonicity Disruption

Because heterogeneity testing is performed using a one-sided test where practitioners examine whether the average risk parameter of a "better" rating, pool, or bucket is smaller than that of a "worse" one (alternative hypothesis) - monotonicity is a minimally necessary condition. Beyond monotonicity, given the variability of the averages, a certain effect size (i.e., the difference between averages) is also required to achieve a statistically significant difference and fully meet the heterogeneity requirement. However, depending on whether the test is applied to the full modeling dataset (all reference dates) or one or a few reference dates, the heterogeneity test results are highly influenced by the number of observations per rating grade, pools, or buckets. Therefore, when testing heterogeneity as part of periodic model validation,

practitioners can combine the results of statistical tests with observed monotonicity, using the latter as a softer criterion for assessing this model aspect. A recurring challenge in this context is determining the threshold for the expected number of grades at which monotonicity disruption may occur. This section proposes a simple hypothesis and simulation testing framework for calculating the probability of monotonicity disruption when testing a given rating, pool, or bucket pair.

5.1 Test of Two Proportions

The probability of monotonicity disruption between two adjacent grades can be calculated by modifying the standard hypothesis testing framework. Unlike the two-proportion test for heterogeneity, which calculates the test statistic Z under the null hypothesis using pooled variance, this approach relies on unpooled variance. It tests the hypothesis that the PD of the "better" grade is greater than that of the "worse" grade. The following formula shows how the test statistic can be calculated under these assumptions:

$$Z = \frac{p_2 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

where:

- p_1 is the proportion of the "better" rating grade;
- p_2 is the proportion of the "worse" rating grade;
- n_1 and n_2 denote the sizes of the "better" and "worse" ratings.

The two-proportion test assumes that the test statistic follows a standard normal distribution. Consequently, the probability of monotonicity disruption is calculated as $1 - \Phi(Z)$, where Φ is the cumulative distribution function of the standard normal distribution.

An alternative method for calculating the probability of monotonicity disruption is based on Monte Carlo simulations. The following steps outline the simulation process:

1. Define the true default rates - p_1 and p_2 (calibrated PDs) - and the sample sizes, n_1 and n_2.
2. For rating grade 1, simulate the default rate as the average of default indicator values of size n_1 drawn from a binomial distribution with the probability of success equal to p_1.
3. For rating grade 2, simulate the default rate as the average of default indicator values of size n_2 drawn from a binomial distribution with the probability of success equal to p_2.
4. Repeat steps 2 and 3, N times, and collect the simulated default rates.

5. Estimate the probability of monotonicity disruption as the proportion of simulations where the simulated default rate for rating grade 1 exceeds that of rating grade 2.

One possible adjustment to the above simulation design concerns steps 2 and 3. Instead of calculating the average of simulated default indicators, practitioners can directly simulate the distribution of the true default rates (p_1 and p_2) from a normal distribution with means equal to p_1 and p_2, and standard deviations $\sqrt{\frac{p_1(1-p_1)}{n_1}}$ and $\sqrt{\frac{p_2(1-p_2)}{n_2}}$ for grades 1 and 2, respectively. However, practitioners should be mindful of potential issues with this approach, including small sample sizes and negative simulated default rates. Therefore, it is advisable to validate the results using the original simulation design described above.

5.2 Test for Mean Differences

Essentially, the probability of monotonicity disruption between two adjacent pools or buckets can be calculated using the same test statistic as Welch's t-test for heterogeneity testing, but with a different hypothesis and assumption about the distribution of the test statistic. Instead of the favorable hypothesis that the average value of the risk parameter in the "better" pool is lower than that in the "worse" grade, monotonicity disruption focuses on the opposite direction.

The test statistic for this test is given by:

$$Z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where:

- μ_1, μ_2 are the calibrated estimates of adjacent pools or buckets, labeled "better" and "worse," respectively;
- σ_1^2, σ_2^2 are the variances of the adjacent pools or buckets, labeled "better" and "worse," respectively;
- n_1, n_2 denote the sample sizes of the adjacent pools or buckets.

The test statistic Z follows a standard normal distribution. Consequently, the probability of monotonicity disruption is calculated as $1 - \Phi(Z)$, where Φ is the cumulative distribution function of the standard normal distribution.

An alternative method for calculating the probability of monotonicity disruption is based on Monte Carlo simulations. The following steps outline the simulation process:

1. Define the true means (μ_1, μ_2), standard deviations (σ_1, σ_2), and sample sizes (n_1, n_2).
2. Select a distribution for sampling the target values per pool/bucket.
3. Generate two samples of sizes n_1 and n_2 from the selected distribution in step 2, with means μ_1, μ_2 and standard deviations σ_1, σ_2, respectively.
4. Calculate the average for each sample from step 3.
5. Repeat steps 3 and 4 a total of N times.
6. Estimate the probability of monotonicity disruption as the proportion of simulations where the average values from pool/bucket 1 exceeds the value from pool/bucket 2.

Similar to statistical power calculation, practitioners should consider the most appropriate method for calculating the standard deviations (σ_1, σ_2) used in the simulation.
Another aspect of the above simulation framework that can be modified relates to step 3 and, consequently, specific subsequent steps. Specifically, instead of simulating target values and then calculating their averages for further use in the simulation, it is possible to simulate the average target values directly from a normal distribution with means equal to μ_1 and μ_2, and standard deviations of the mean equal to $\sqrt{\frac{\sigma_1^2}{n_1}}$ and $\sqrt{\frac{\sigma_2^2}{n_2}}$ for pools/buckets 1 and 2, respectively. However, practitioners should be mindful of potential issues with this approach, including small sample sizes, negative averages, and skewed distributions of the target variable. Therefore, it is advisable to validate the results using the original simulation design described above.
In addition, practitioners can adjust the first three simulation steps to apply a bootstrap method and sample from the observed values of the analyzed risk parameter.

6 Simulation Study

As mentioned in previous sections, practitioners sometimes face challenges in drawing overall conclusions from heterogeneity testing due to different possible testing levels. Therefore, analyses that provide deeper insights and distinguish between more and less likely events prove helpful. This is particularly convenient during model development, when designing the rating scale or pools, as well as in initial and periodic model validation. Through a practical example, this section demonstrates how statistical power analysis and hypothesis testing results for monotonicity disruption analysis can enhance heterogeneity testing.

6.1 Statistical Power Analysis

One straightforward extension of the statistical analysis for heterogeneity testing is defining thresholds for the expected number of pairs that fail the testing procedure. Assume prac-

titioners design the rating scale or pools such that the expected probability of observing heterogeneity between all adjacent pairs is constant at 80%. Given this input and the number of rating grades or pools, it is possible to derive the probability that n pairs will fail heterogeneity testing. The table below provides these probabilities.

Table 1: Probability of heterogeneity failure by number of pairs and different sizes of rating scales, pools, or buckets.

Number of pairs	Maximum number of rating, pool, or bucket pairs						
	5	7	10	12	14	16	20
0	32.77%	20.97%	10.74%	6.87%	4.40%	2.81%	1.15%
1	40.96%	36.70%	26.84%	20.62%	15.39%	11.26%	5.76%
2	20.48%	27.53%	30.20%	28.35%	25.01%	21.11%	13.69%
3	5.12%	11.47%	20.13%	23.62%	25.01%	24.63%	20.54%
4	0.64%	2.87%	8.81%	13.29%	17.20%	20.01%	21.82%
5	0.03%	0.43%	2.64%	5.32%	8.60%	12.01%	17.46%
6		0.04%	0.55%	1.55%	3.22%	5.50%	10.91%
7		0.00%	0.08%	0.33%	0.92%	1.97%	5.45%
8			0.01%	0.05%	0.20%	0.55%	2.22%
9			0.00%	0.01%	0.03%	0.12%	0.74%
10			0.00%	0.00%	0.00%	0.02%	0.20%
11				0.00%	0.00%	0.00%	0.05%
12				0.00%	0.00%	0.00%	0.01%
13					0.00%	0.00%	0.00%
14					0.00%	0.00%	0.00%
15						0.00%	0.00%
16						0.00%	0.00%

Table 1: Probability of heterogeneity failure by number of pairs and different sizes of rating scales, pools, or buckets.

Number of pairs	Maximum number of rating, pool, or bucket pairs						
	5	7	10	12	14	16	20
17							0.00%
18							0.00%
19							0.00%
20							0.00%

The probabilities from the table above can be interpreted as follows. For example, consider practitioners having a PD rating scale with eight rating grades, resulting in seven possible adjacent pairs to be tested. Given a constant statistical power of 80%, the probability that none of the adjacent pairs will fail heterogeneity testing is 20.97%. Extending this analysis, practitioners can determine the maximum number of pairs likely to fail heterogeneity testing with a 5% or less probability. This threshold is three pairs for the same design with 80% power and eight rating grades. According to the table, the likelihood of having four or more pairs fail heterogeneity testing is the sum of 2.87%, 0.43%, 0.04%, and 0.00%, totaling 3.33%. Similarly, the results can be interpreted for other sizes of the rating scale, pools, or buckets.

Reversing the previous example, practitioners can also calculate the required power of the test for a given number of rating grades or pools and an expected number of pairs that will fail heterogeneity. The presented analysis is convenient for model development, can be performed during the rating scale or pool design planning phase, and is equally applicable to all risk parameters: PD, LGD, and EAD.

While the previous example is general and applicable to all risk parameters, the following subsections demonstrate the power analysis applied specifically to PD models on the one hand, and to LGD and EAD models on the other. As already described, since PD models primarily involve a binary variable, a test of two proportions is used, while for LGD and EAD models, the Welch t-test is applied.

6.1.1 Probability of Default Models

When validating credit risk models, practitioners already have rating scales or pools designed and delivered, making them the subject of validation. However, the same analysis and approach can still be used to enhance heterogeneity testing. Unlike the previous example, in the following one, for a given rating scale or pools, practitioners first need to calculate the power of the heterogeneity test and then apply the same approach to support different testing levels of the heterogeneity aspect.
The following example demonstrates how this can be performed for a PD model.

Assume the following rating scale of the PD model delivered to the validation team, where the distribution of obligors is defined as the average number over the period covered by the calibration dataset, and the column PD represents the final model estimates.

Table 2: PD model rating scale.

Rating	Number of obligors	PD
RG1	1,500	0.57%
RG2	1,920	1.05%
RG3	2,925	1.69%
RG4	4,515	3.10%
RG5	2,535	5.30%
RG6	1,365	7.93%
RG7	91	14.51%
RG8	148	25.90%

To utilize statistical power analysis, validators can first investigate and ultimately accept that the model under validation is well-calibrated. Under this assumption, the typical first step is to test for heterogeneity within a given distribution of obligors and calibrated PD values. The following table presents the results of applying the test of two proportions described in the section Common Practices in Heterogeneity Testing.

Table 3: PD model heterogeneity testing results.

Rating	Number of obligors	PD	Estimated number of defaults	p-value
RG1	1,500	0.57%	8.5500	
RG2	1,920	1.05%	20.1600	6.34%
RG3	2,925	1.69%	49.4325	3.35%
RG4	4,515	3.10%	139.9650	0.01%
RG5	2,535	5.30%	134.3550	0.00%
RG6	1,365	7.93%	108.2445	0.06%
RG7	91	14.51%	13.2041	1.40%
RG8	148	25.90%	38.3320	1.88%

Typically, the p-value from the applied test is compared to a selected significance level of 1%, 5%, or 10% and is applied consistently throughout the testing process. Validation teams also often use the so-called traffic light approach, applying different significance levels. In the example above, for instance, if the selected significance level is 5%, practitioners would conclude that only the rating grade pair `RG1-RG2` shows no statistical evidence of sufficient heterogeneity. However, having just one pair of grades that fails the heterogeneity test at a specific reference date in a well-calibrated model does not necessarily imply that the model as a whole fails the heterogeneity assessment. Additionally, since heterogeneity is typically tested on observed default rates, assuming a well-calibrated model allows practitioners to easily derive the expected number of grade pairs that will fail the test at a given confidence level. In addition to its use in the validation exercise, model developers can apply the same approach when designing the rating scale.

The following table present the results of the statistical power analysis, following the approaches from subsection Test of Two Proportions in section Statistical Power Analysis.

Table 4: Statistical power of the test of two proportions.

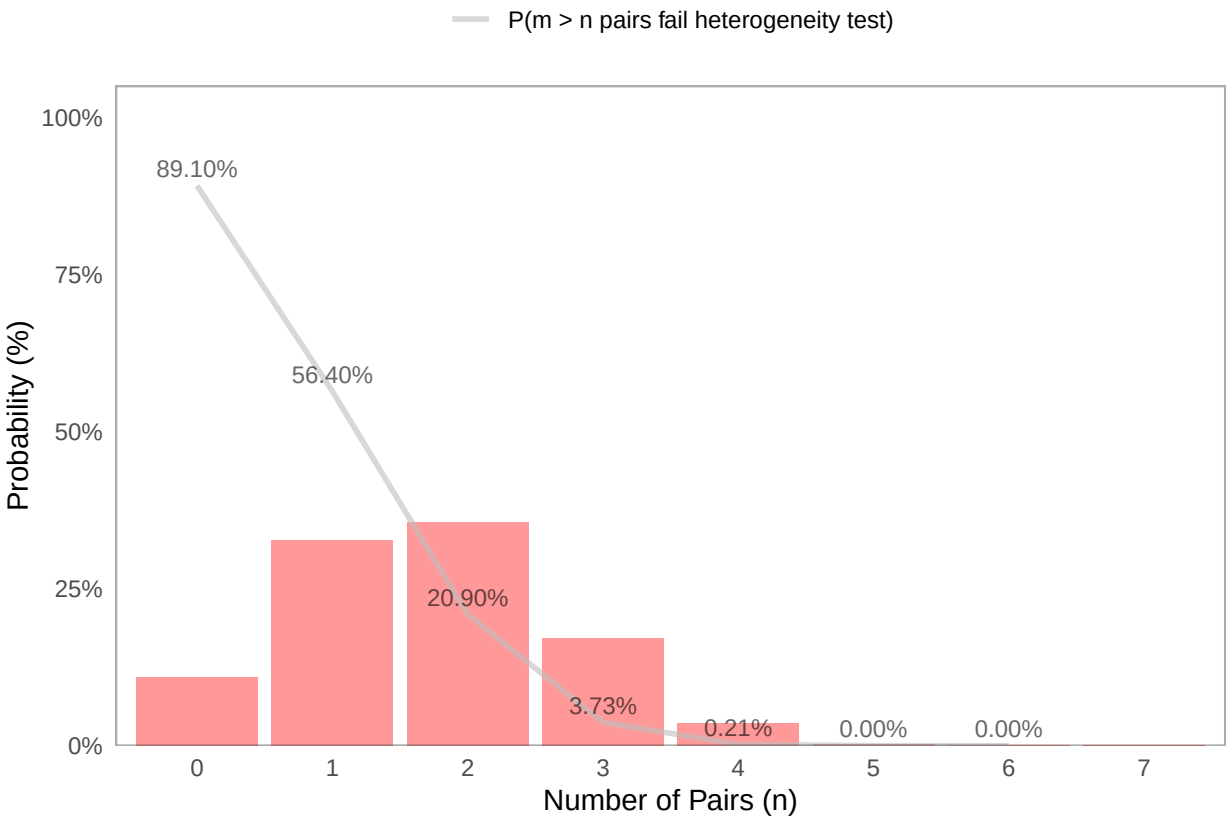
Rating	Number of obligors	PD	Estimated number of defaults	p-value	Power Analytical Approach	Power Simulation Approach
RG1	1,500	0.57%	8.5500			
RG2	1,920	1.05%	20.1600	6.34%	45.12%	45.44%
RG3	2,925	1.69%	49.4325	3.35%	57.75%	58.30%
RG4	4,515	3.10%	139.9650	0.01%	98.82%	98.87%
RG5	2,535	5.30%	134.3550	0.00%	99.70%	99.69%
RG6	1,365	7.93%	108.2445	0.06%	93.50%	93.58%
RG7	91	14.51%	13.2041	1.40%	67.00%	66.24%
RG8	148	25.90%	38.3320	1.88%	67.77%	68.82%

The above results show that the power to identify heterogeneity between rating grades ranges from slightly above 45% for the rating pair `RG1-RG2` to over 99% for the pair `RG4-RG5`, for both analytical and simulation-based power calculations.

Extending this analysis, practitioners can calculate the probabilities for each number of rating grade pairs expected to fail the heterogeneity test. This can be useful for setting a threshold for the maximum number of failing pairs, beyond which the failure may indicate an issue with the rating scale design rather than being attributable to randomness. A typical choice for the confidence level is 90% or 95%.

The following figure shows the results of this calculation based on the power derived from the analytical approach.

Figure 1: PD model: Probability of number of pairs that fail the heterogeneity for significance level of 5%



6.1.2 Loss Given Default and Exposure at Default Models

Similar to the analysis performed for the PD model, practitioners can apply statistical power analysis to LGD and EAD models as well. For example, suppose the validation team receives the LGD model pools shown in the following table, along with the expected number of facilities, calibrated LGD values, and estimated standard deviations of the realized LGD values for each pool.

Table 5: LGD model pools.

Pool	Number of facilities	LGD	Standard deviation
P1	166	10.43%	0.1687
P2	129	17.72%	0.2594

Table 5: LGD model pools.

Pool	Number of facilities	LGD	Standard deviation
P3	131	17.79%	0.2568
P4	162	32.50%	0.3147
P5	198	41.80%	0.3423
P6	238	49.73%	0.3879
P7	176	66.44%	0.3437

Given the data above and using the Welch t-test described in the section Common Practices in Heterogeneity Testing practitioners can calculate the p-value to analyze the heterogeneity aspect of the LGD model.

Table 6: LGD model heterogeneity testing results.

Pool	Number of facilities	LGD	Standard deviation	p-value
P1	166	10.43%	0.1687	
P2	129	17.72%	0.2594	0.31%
P3	131	17.79%	0.2568	49.13%
P4	162	32.50%	0.3147	0.00%
P5	198	41.80%	0.3423	0.38%
P6	238	49.73%	0.3879	1.20%
P7	176	66.44%	0.3437	0.00%

As shown above, the p-values for heterogeneity testing on adjacent pools range from nearly 0% for the pairs `P3-P4` and `P6-P7` to as high as 49.13% for the pair `P2-P3`. Comparing these

p-values to commonly used significance levels of 1%, 5%, and 10%, practitioners conclude that one pair of adjacent pools failed the hypothesis test.

Formal heterogeneity testing can be taken a step further by calculating the probability that more than `m` out of a total of `n` pool pairs are expected to fail the heterogeneity test under the assumption of a well-calibrated model. As mentioned earlier, this analysis can help practitioners set a threshold for the number of failing pairs above which heterogeneity issues should be investigated as a potential problem with the pool design, rather than attributed to the variability of the realized LGD values. This analysis can be performed using the approaches described in subsection Welch T-Test within section Statistical Power Analysis. Unlike the PD model, where in a simulation-based approach the default indicator can be drawn from a binomial distribution with known parameters, for LGD and EAD models practitioners need to assume a specific distribution for the realized values, with the expected mean and standard deviation as given in the table at the beginning of this section. For this purpose, let's assume the realized LGD values are drawn from a beta distribution with parameters shown in the following table for each pool. Practitioners should note that drawing a sufficiently large number of samples from these beta distributions should result in average LGD values and standard deviations that match those reported at the start of this section.

Table 7: Beta distribution parameters.

Pool	Shape 1	Shape 2
P1	0.2381	2.0445
P2	0.2068	0.9600
P3	0.2166	1.0011
P4	0.3949	0.8202
P5	0.4499	0.6264
P6	0.3289	0.3325
P7	0.5897	0.2979

Given the above inputs, the following table shows the analytical and simulation-based results of the power analysis for each pair of adjacent pools.

Table 8: Statistical power of the Welch t-test.

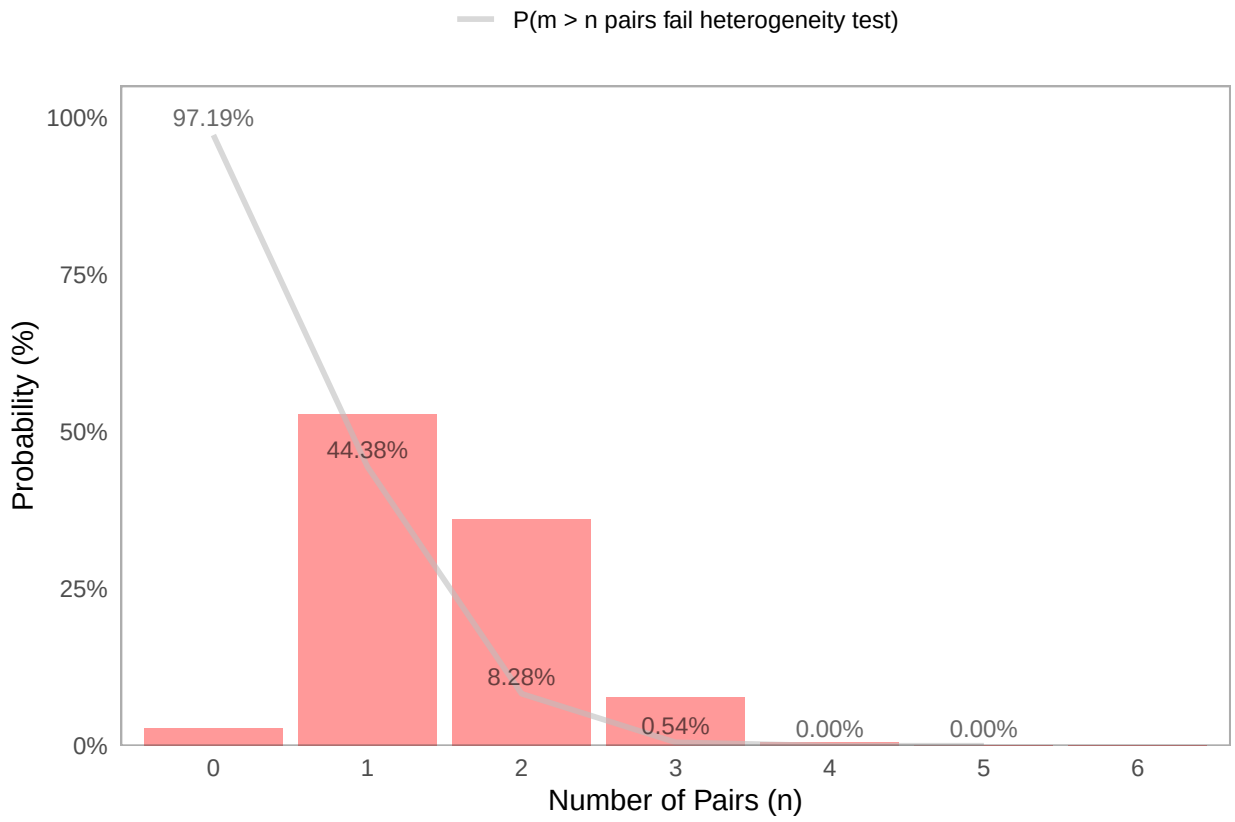
Pool	Number of facilities	LGD	Standard deviation	p-value	Power Analytical Approach	Power Simulation Approach
P1	166	10.43%	0.1687			
P2	129	17.72%	0.2594	0.31%	86.76%	88.13%
P3	131	17.79%	0.2568	49.13%	5.23%	5.18%
P4	162	32.50%	0.3147	0.00%	99.70%	99.53%
P5	198	41.80%	0.3423	0.38%	84.88%	84.88%
P6	238	49.73%	0.3879	1.20%	73.18%	73.57%
P7	176	66.44%	0.3437	0.00%	99.85%	99.81%

The results above show that the power to detect heterogeneity between rating grades ranges from just over 5% for the pool pair `P2-P3` to more than 99% for the pairs `P3-P4` and `P6-P7`, in both the analytical and simulation-based calculations.

Building on this, practitioners can also calculate the probabilities for how many pool pairs are expected to fail the heterogeneity test. This helps establish a threshold for the maximum number of failing pairs, beyond which the failures may indicate issues with the pool design rather than just random variation in the realised LGD values. A typical confidence level for this is 90% or 95%.

The following figure presents these results, based on the power calculated using the analytical approach.

Figure 2: LGD model: Probability of number of pairs that fail the heterogeneity for significance level of 5%



6.2 Monotonicity Disruption Analysis

Whether heterogeneity testing is applied to the whole modeling dataset (across all reference dates) or just one or a few reference dates, the results are strongly affected by the number of observations per rating grade, pool, or bucket. For this reason, when performing heterogeneity testing as part of periodic model validation, practitioners can complement statistical test results with observed monotonicity, using the latter as a softer criterion for evaluating this aspect of the model.

The following subsections use a practical example to show how monotonicity disruption analysis can enhance heterogeneity testing.

6.2.1 Probability of Default Models

To incorporate monotonicity analysis into heterogeneity testing for the PD model, practitioners can use the same inputs as in the previous example - rating grade, number of obligors (average over the calibration period), and calibrated PDs (final model estimates).

Following the approaches described in subsection Test of Two Proportions within section Monotonicity Disruption the table below presents the probability of monotonicity disruption for each pair of adjacent rating grades.

Table 9: PD model: Probability of monotonicity disruption.

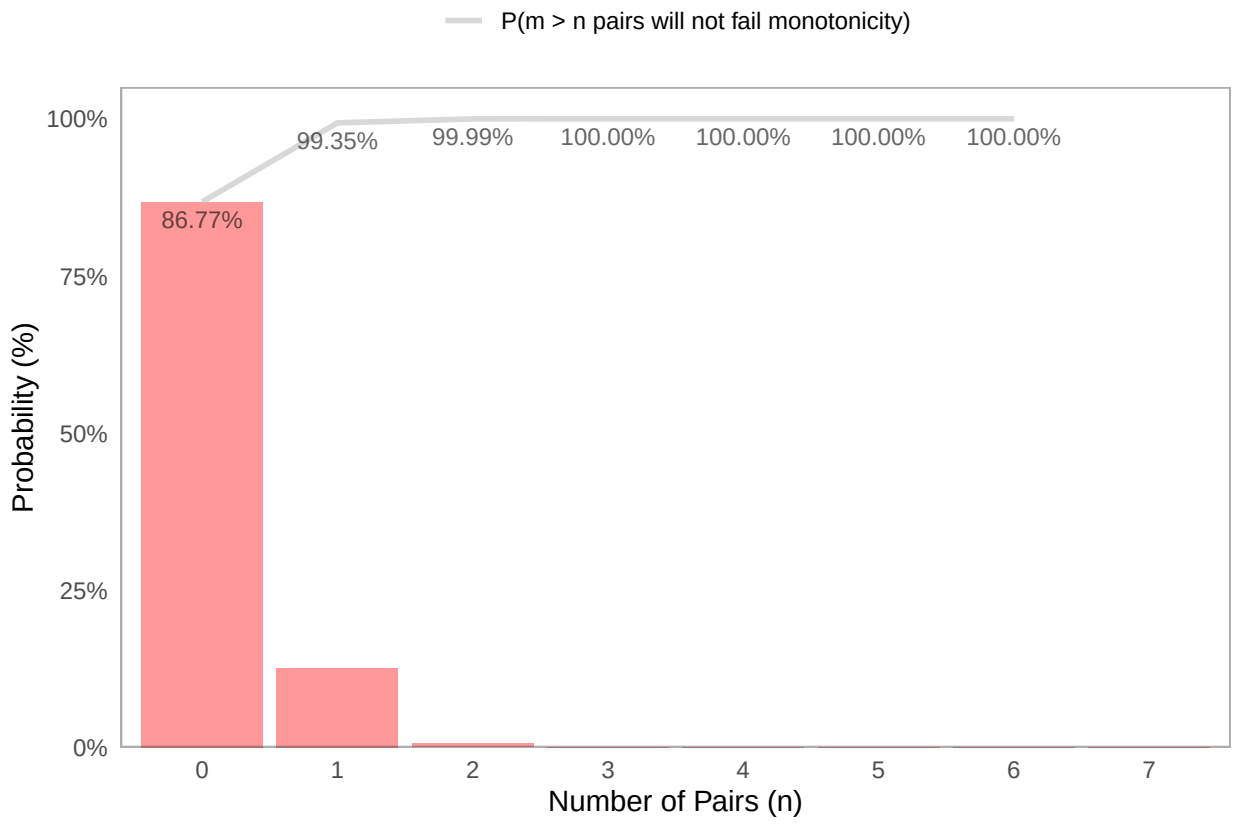
Rating	Number of obligors	PD	Mono. Disr. Analytical Approach	Mono. Disr. Simulation Approach
RG1	1,500	0.57%		
RG2	1,920	1.05%	5.67%	5.78%
RG3	2,925	1.69%	2.73%	2.77%
RG4	4,515	3.10%	0.00%	0.01%
RG5	2,535	5.30%	0.00%	0.00%
RG6	1,365	7.93%	0.11%	0.08%
RG7	91	14.51%	4.02%	3.25%
RG8	148	25.90%	1.36%	1.43%

The results above show that both approaches—statistical hypothesis testing and the simulation method - lead to the same overall conclusion, with only slight differences in the final values. The largest misalignment occurs for rating pairs with smaller sample sizes (`RG6 - RG7` and `RG7 - RG8`), which require special caution when interpreting the results in relation to sample size. Aside from the comparison of methods, the results indicate that the highest probability of monotonicity disruption, for both approaches, is observed for the rating pairs `RG1 - RG2` and `RG6 - RG7`.

Building on this analysis, practitioners can calculate the probabilities for each possible number of rating grade pairs expected to show monotonicity disruption. This can help in setting a threshold for the maximum number of failing pairs beyond which the issue is more likely due to problems with the rating scale design rather than random variation. A typical confidence level for this purpose is 90% or 95%.

The figure below shows the results of this calculation, based on the power derived from the analytical approach.

Figure 3: PD model: Probability of number of pairs that will not fail the monotonicity



Practitioners should note that the probability of observing monotonicity disruption for a properly defined rating scale is lower than the probability of passing heterogeneity testing using the standard methods described in section Common Practices in Heterogeneity Testing.

6.2.2 Loss Given Default and Exposure at Default Models

The following simulation study applies equally to LGD and EAD models, although the inputs and process will be demonstrated for the LGD model. Similar to the statistical power analysis for the LGD model, the necessary inputs for this analysis include pools or buckets, the (expected) number of facilities per reference date, calibrated LGD values, and estimated standard deviations of the realized LGD values for each pool. For the simulation-based approach, we additionally assume that the realized LGDs are drawn from a beta distribution with the parameters reported in the same table (`Beta Shape 1` and `Beta Shape 2`).
Given these inputs, and following the approaches described in subsection Test for Mean

Differences within section Monotonicity Disruption, the table below shows the probability of monotonicity disruption for each pair of adjacent pools.

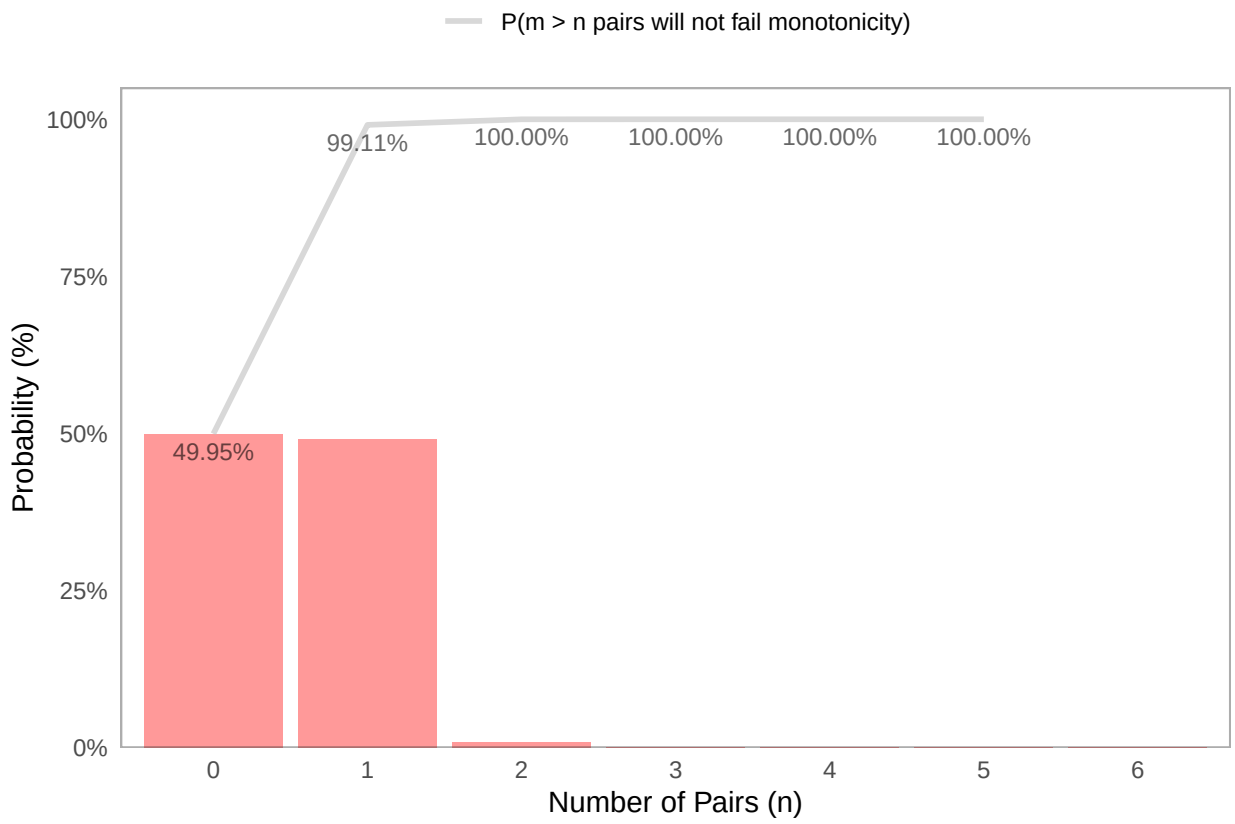
Table 10: LGD model: Probability of monotonicity disruption.

Pool	Number of facilities	LGD	Standard deviation	Beta Shape 1	Beta Shape 2	Mono. Disr. Analytical Approach	Mono. Disr. Simulation Approach
P1	166	10.43%	0.1687	0.2381	2.0445		
P2	129	17.72%	0.2594	0.2068	0.9600	0.28%	0.30%
P3	131	17.79%	0.2568	0.2166	1.0011	49.13%	49.22%
P4	162	32.50%	0.3147	0.3949	0.8202	0.00%	0.00%
P5	198	41.80%	0.3423	0.4499	0.6264	0.37%	0.33%
P6	238	49.73%	0.3879	0.3289	0.3325	1.17%	1.09%
P7	176	66.44%	0.3437	0.5897	0.2979	0.00%	0.00%

As shown by the reported results, both approaches lead to the same conclusion, identifying the pool pair `P2 - P3` as having the highest probability of monotonicity disruption. This outcome is expected, given that their average realized LGD values are very close to each other.

Building on this analysis, practitioners can calculate the probabilities for each possible number of pool pairs expected to exhibit monotonicity disruption. This can help in setting a threshold for the maximum number of failing pairs beyond which the issue is more likely due to problems with the rating scale design rather than random variation. A typical confidence level for this purpose is 90% or 95%. As with PD models, it should be noted that the probability of observing monotonicity disruption for properly defined pools is lower than the probability of passing heterogeneity testing using the standard methods described in section Common Practices in Heterogeneity Testing.

Figure 4: LGD model: Probability of number of pairs that will not fail the monotonicity



7 Conclusions

This paper examined the regulatory requirements, standard practices, and practical challenges associated with testing heterogeneity in Internal Ratings-Based (IRB) credit risk models. Regulatory frameworks such as the CRR, Commission Delegated Regulation (EU) 2022/439, the ECB Guide to Internal , and Draft Guidelines on Credit Conversion Factor Estimation emphasize the need for meaningful differentiation of risk across rating grades, pools, or buckets in PD, LGD, and EAD models.

While standard approaches - most often the two-proportion test for PD and the Welch t-test for LGD and EAD - are widely applied, practitioners continue to face difficulties in setting appropriate thresholds, interpreting results in the context of model calibration, and addressing the interaction between heterogeneity and monotonicity.

To enhance current practice, the paper proposed two complementary analyses. First, statistical power analysis can be applied both in model development and validation to quantify the probability of detecting heterogeneity, set thresholds for acceptable numbers of failed pairs,

and account for sampling variability in observed results. Second, monotonicity disruption analysis provides a structured way to assess the likelihood of breaks in the expected ordering of risk parameters, either as a standalone diagnostic or as a softer criterion alongside formal hypothesis testing. Both methods can be implemented analytically or through simulation, allowing flexibility for different data structures and sample sizes.

Incorporating these analyses into heterogeneity testing offers tangible benefits. Power analysis helps ensure that conclusions are not unduly influenced by low detection capability, leading to more robust threshold setting and a clearer distinction between genuine design issues and random variation. Monotonicity disruption analysis, in turn, strengthens the interpretation of heterogeneity results by highlighting whether observed irregularities are statistically meaningful or likely due to noise. Together, these tools support more informed, transparent, and defensible decisions in both the development and periodic validation of IRB credit risk models.

References

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

European Banking Authority. 2025. "Draft Guidelines on Credit Conversion Factor Estimation Under Article 182(5) of Regulation (EU) No 575/2013." https://www.eba.europa.eu/sites/default/files/2025-07/b3f9af47-ab61-4e89-94f2-7910c39c372f/Consultation%20paper%20Guidelines%20CCF.pdf.

European Central Bank. 2019. "Instructions for Reporting the Validation Results of Internal Models: IRB Pillar i Models for Credit Risk." https://www.bankingsupervision.europa.eu/activities/internal_models/shared/pdf/instructions_validation_reporting_credit_risk.en.pdf.

———. 2025. "ECB Guide to Internal Models." https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide202507.en.pdf.

European Commission. 2021. "Commission Delegated Regulation (EU) 2022/439 of 20 October 2021." https://eur-lex.europa.eu/eli/reg_del/2022/439/oj.

European Parliament and Council. 2013. "Regulation (EU) No575/2013 of the European Parliament and of the Council of 26 June 2013 on Prudential Requirements for Credit Institutions and Investment Firms and Amending Regulation (EU) No648/2012." https://eur-lex.europa.eu/eli/reg/2013/575/oj.