

Designing Stable ESG Grade Bands: Cut-Point Optimisation and a Quantisation Perspective

Hira Mahmood

`hira.mahmood@iusspavia.it`

IUSS Pavia

Minh N. Dao

RMIT University

Claudio Conversano

University of Cagliari

Research Article

Keywords: ESG ratings, grade bands, cut-points, rating stability, scalar quantisation, Refinitiv/LSEG

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11277119/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Designing Stable ESG Grade Bands: Cut-Point Optimisation and a Quantisation Perspective

Hira Mahmood^{1*}, Minh N. Dao² and Claudio Conversano³

^{1*}School of Advanced Sciences, IUSS Pavia, Pavia, Italy.

²School of Science, STEM College, RMIT University, Melbourne, Australia.

³Department of Business and Economics, University of Cagliari, Cagliari, Italy.

*Corresponding author(s). E-mail(s): hira.mahmood@iusspavia.it;
Contributing authors: minh.dao@rmit.edu.au; conversa@unica.it;

Abstract

Environmental, social, and governance (ESG) ratings are often communicated as letter grades obtained by applying cut-points to an underlying numerical score. The location of these boundaries is therefore a discrete design problem with direct implications for rating stability. We formulate the legacy Refinitiv/LSEG ESG grade scale as a constrained scalar-quantisation problem and evaluate alternative internal cut-points while retaining the main A/B/C/D boundaries at 25, 50, and 75. Using 177 large-capitalisation firms observed from 2018 to 2023, we compare the incumbent equal-width structure with six non-uniform alternatives. Published scores are highly concentrated at incumbent grade midpoints, creating a quantisation limit that prevents most alternative scales from being distinguished on the static data. We therefore evaluate stability through Monte Carlo score perturbations, boundary-safety measures, midpoint distortion, and four numerical optimisation procedures. The equal-width scale produces the lowest or near-lowest grade migration under small and moderate disturbances, maximises the observed minimum distance to a boundary, and is selected or closely approximated by all four optimisers. Under the largest disturbance considered, a wider central band becomes more stable. The results provide a computational framework for designing discrete ESG grade boundaries and show that the preferred partition depends jointly on the score distribution and the magnitude of expected score revisions.

Keywords: ESG ratings, grade bands, cut-points, rating stability, scalar quantisation, Refinitiv/LSEG

1 Introduction

ESG rating systems convert multidimensional information into numerical scores that are often communicated to investors as discrete categories or letter grades. Once a continuous score is mapped into ordered grades, the placement of the cut-points becomes a mathematical design choice: movements within a band leave the reported grade unchanged, whereas an otherwise small score revision can trigger reclassification when it crosses a boundary. For financial users who screen securities, construct portfolios, or monitor rating transitions, the stability of these thresholds is therefore distinct from the construction of the underlying ESG score itself.

In the majority of existing research regarding ESG ratings, the focus is on the differences between the various rating providers. Different companies receive different ratings from the various providers due to the differences in the indicators that are used to calculate ESG ratings of those companies (Chatterji et al. 2016; Christensen et al. 2022; Berg et al. 2022; Billio et al. 2021; Escrig-Olmedo et al. 2019; Dimson et al. 2020; Larcker et al. 2022). The ratings provided by those companies have an impact on the investment and financial return decisions of investors (Gibson Brandon et al. 2021; Avramov et al. 2022; Pástor et al. 2021; Pedersen et al. 2021; Krüger 2015; Serafeim and Yoon 2023), alongside the rating histories of those companies, which can introduce difficulties in the replication of those studies (Berg et al. 2021). However, less research is performed on the step of assigning a letter grade to those companies based on their calculated score.

Beyond being a presentation of the ESG rating of a company, those letter grades have an impact on investors and the companies' portfolios; investors may have an interest in certain categories of sustainability within those companies, for instance (Hartzmark and Sussman 2019). In addition, changes in those grades have the potential to affect the companies' portfolios and the decisions of their investors. Recent evidence also suggests that ESG rating providers may manage scores around grade thresholds to reduce unnecessary changes (Berg et al. 2026). This means that the position of the cut-points should be treated as an important methodological decision.

This study evaluates the grade-band structure of the legacy Refinitiv/LSEG ESG scale. Under this historical methodology firms received a score between 0 and 100 and were classified into twelve grades from D− to A+ using equal-width intervals (LSEG 2022). The data cover 2018–2023 and therefore relate to the legacy system. In March 2026, LSEG introduced a redesigned ESG scoring system based on a 0–5 scale, while the earlier scores were moved to legacy status (LSEG 2026). The study therefore examines the historical Refinitiv/LSEG scale rather than the current methodology.

The analysis keeps the main A/B/C/D boundaries fixed at 25, 50, and 75 and modifies only the three sub-grades within each quartile. Let e denote the width of each edge sub-grade; the width of the centre sub-grade is then $c = 25 - 2e$. Under the incumbent uniform scale all three sub-grades have equal width, $e = c = 25/3 \approx 8.33$.

Six alternative scales are considered by making the centre band wider and the edge bands narrower. This lets us test whether a non-uniform scale would give greater grade stability than the uniform scale without changing the existing A/B/C/D structure.

The major limitation of the available data on the ESG rating is its quantisation. Out of 1,048 observations of firms and their non-missing ESG scores, there are only eleven distinct scores. Furthermore, 99.9% of firms have an ESG score that corresponds to the midpoint of one of the A/B/C/D grade bands. As a result, five of the six alternative grade scales produce the same grades as the uniform scale on the static data. This does not settle which design is better; it shows that alternative cut-points cannot be properly compared when the available scores have already been rounded according to the incumbent scale.

The analysis therefore focuses on stability under score changes. Observed scores are repeatedly perturbed using mean-zero Gaussian noise, and grades are reassigned under each candidate scale. The probability of a grade change is then calculated. The study also measures boundary safety, defined as the minimum distance between an observed score and the nearest cut-point, and optimises the edge width using grid search, golden-section search, simulated annealing, and a genetic algorithm.

The results show that the uniform scale has the lowest or nearly lowest migration rate for small and moderate score changes of up to about four points. It also provides the largest distance between the observed scores and the nearest boundary. All four optimisation methods return an edge width close to the uniform value of 25/3. However, at the largest disturbance considered, $\sigma = 5$, the most extreme wide-centre scale produces a lower migration rate. This helps to explain how the preferred scale relates to whether the designer is protecting against small or large changes in measurement.

Using the theory of scalar quantisation, grades can be thought of as dividing the continuous score into intervals that are assigned to the same grade. The Lloyd–Max conditions suggest that the boundaries between these grade bands should be established between the representative scores of the groups within the percentile ranks (Max 1960; Lloyd 1982; Gray and Neuhoﬀ 1998). If the score uses percentile ranks, the distribution of scores within a given group is approximately uniform, so equal-width grade bands are one way to accomplish this grade assignment. Grade bands of different widths may instead fit the distribution better when scores are not uniformly distributed within the groups.

This paper makes three main contributions. First, it treats the boundaries of ESG grades as a separate methodological choice. Second, it shows that alternative ESG grading rules may not be estimable because of the rounding of published ESG scores. Third, it relates ESG grading rules to work on credit-rating stability and quantisation (Altman and Rijken 2004; Löffler 2004). This study is purely methodological and does not examine the effect of ESG grades on company returns, nor does it claim that the existing ESG score is a complete measure of a company’s sustainability in relation to environmental, social, and governance issues.

1.1 Research questions

Although ESG rating differences across providers have been widely studied, much less attention has been given to how the boundaries between ESG grades should be

designed. Based on the objectives of this study, the following four research questions are examined:

RQ1: Can alternative grade-band designs be distinguished from the existing uniform scale using the published ESG scores?

RQ2: Which grade-band design is most stable when ESG scores change by small or moderate amounts, and which design keeps scores furthest from the grade boundaries?

RQ3: Do different optimisation methods identify the same or a similar preferred grade-band structure?

RQ4: Under what conditions are uniform grade bands preferable to non-uniform grade bands?

The remainder of this paper is organised as follows. Section 2 reviews the literature on rating divergence, the discrete grade layer, threshold design in credit ratings, and scalar quantisation. Section 3 sets out the methodology: the family of candidate scales, the static identification test, the perturbation and boundary-safety measures, the optimisation criterion, and the quantisation interpretation. Section 5 reports the results and is organised around the research questions: RQ1 is answered in Subsection 5.1; RQ2 is answered through Subsections 5.2 and 5.3; RQ3 is answered in Subsection 5.4; and RQ4 is addressed jointly in Subsections 5.2 5.4, which identify the conditions under which uniform and non-uniform grade bands are preferable. Section 6 concludes.

2 Literature Review

Work on ESG ratings has grown quickly, but it has clustered around one question: why do providers disagree, and what follows from that disagreement? The step that turns a continuous score into a discrete grade has stayed in the background. This review organises the field around four threads — rating divergence and its consequences, the discrete layer that investors actually see, stability and threshold design in credit ratings, and scalar quantisation theory — and shows that the placement of grade cut-points falls in the space none of them occupies.

2.1 Rating divergence and its consequences

The finding in most studies is that the rating providers tend to disagree. [Chatterji et al. \(2016\)](#) found that the ratings of corporate social responsibility converged weakly between providers, and that the reasons for the gaps in the ratings were differences in the theories of and methods for measuring corporate social responsibility. [Berg et al. \(2022\)](#) performed a more detailed investigation into the gaps between social responsibility ratings, decomposing the differences according to the scope of the ratings, the method for scoring different indicators of social responsibility, and the way that the scores were combined to arrive at a final social responsibility score, finding that the correlations between the scores provided by different companies were low on average. [Christensen et al. \(2022\)](#) found a slightly different result: that higher levels of environmental, social, and governance (ESG) disclosures by companies were associated with more disagreement between providers regarding the social responsibility of that company. [Billio et al. \(2021\)](#) found a similar result in the context of analysing portfolios of companies, while [Kotsantonis and Serafeim \(2019\)](#) documented some of the issues

with the raw data from companies that may explain some of the gaps between the scores provided by the different rating companies.

The consensus that divergence exists in the industry is matched by evidence that such divergence has an effect on the market. [Gibson Brandon et al. \(2021\)](#) have shown that the difference in ratings between the individuals that provide those ratings correlates with the returns provided by the stocks. [Avramov et al. \(2022\)](#) have shown that the uncertainty within such ratings has an effect on the portfolios that are created with those stocks, leading to a weakening of the relationship between ESG ratings and stock returns. [Berg et al. \(2021\)](#) have discovered that many of the ratings published by these companies are often changed after the fact, indicating a potential issue in the reliability of those ratings. Where the literature becomes somewhat inconclusive is in the interpretation of these divergent results. Some interpret the differences as noise that must be explained, but others, based on the findings of [Berg et al. \(2022\)](#) and [Christensen et al. \(2022\)](#), suggest that many of the differences are actually differences of definition among the various organisations. Each of these interpretations, however, stops at the continuous score; the rule that defines which grade a company receives is largely treated as a matter of presentation alone.

2.2 The discrete layer: grades, thresholds, and investor response

A separate thread of research shows that the discrete label is anything but cosmetic. [Hartzmark and Sussman \(2019\)](#) found that funds with the worst ratings in a new Morningstar sustainability category saw investment outflows, while those with the best ratings in that category saw investment inflows. In each case, the investor reacted to the category that the fund received, not the continuous attribute that determined that category. Furthermore, the discrete category has implications for eligibility for benchmarks and for compliance with various regulations, so a small change in a grade can lead to movement of investments. While the literature on discrete ratings establishes that the thresholds have weight, it does not examine whether the existing discrete categories are placed well relative to one another. In other words, while the placement of the lines has been established to have weight, the question of where those lines should sit relative to the continuous score has not been directly examined.

2.3 Rating stability and threshold design

The credit-rating literature is where the idea of a threshold as a design choice has been explored. [Altman and Rijken \(2004\)](#) examined how credit rating agencies balance the responsiveness of their ratings with the desire for those ratings to be stable, and [Löffler \(2004\)](#) provided a formalisation ([Cantor and Mann 2007](#)) of that notion of “through-the-cycle” behaviour by the agencies. The role of the threshold as an instrument to control how often a company crosses it is visible in current ESG practice. For instance, the redesigned LSEG methodology includes devices to provide stability to its ratings on the firm’s 0–5 scale, such as a buffer that places a company on the “rewarded” side of a threshold if its distance from that threshold is within ten percentage points, and a cap on a company’s rating on a particular ESG theme at three

unless certain gateway metrics are fulfilled (London Stock Exchange Group 2026). These are boundary-management choices of exactly the kind the credit-rating literature studied, now applied to ESG. What is missing is a framework that evaluates such choices for an ESG scale rather than assuming them.

2.4 Grade bands as scalar quantisation

A grading rule is a scalar quantiser: it partitions a continuous variable into cells and represents everything in a cell by one label. The distortion this introduces, and the partition that minimises it, are classical results. Max (1960) and Lloyd (1982) derived the classical optimality conditions: each representative is located at its cell centroid and each boundary lies midway between neighbouring representatives — and Gray and Neuhoff (1998) consolidated the theory (Bennett 1948; Panter and Dite 1951; Cover and Thomas 2006). A consequence rarely stated in the rating context is that for a source that is uniform on an interval, these conditions place the boundaries at equal spacing, and non-uniform cells reduce distortion only where the density is uneven. A percentile-rank score, of the kind the legacy Refinitiv/LSEG scale produced, is close to uniform within its peer group by construction, which is precisely the case where equal-width bands are distortion-optimal. This theory has not been connected to ESG grade design, yet it supplies both a design rule and a reason for it.

3 Methodology

This study examines whether changing the internal boundaries of the legacy Refinitiv/LSEG ESG scale can improve grade stability. The analysis has five stages. First, a family of alternative grade scales is constructed while keeping the main A/B/C/D categories unchanged. Second, the grade assignments under the alternative scales are compared with those of the incumbent scale. Third, stability is evaluated by perturbing the observed scores and measuring grade migration. Fourth, boundary safety and midpoint distortion are calculated and combined in an optimisation criterion. Finally, an auxiliary indicator anchor and scalar quantisation theory are used to interpret the results.

3.1 Candidate grade scales

The legacy ESG scale divides the 0–100 score into four main letter bands:

$$D = [0, 25], \quad C = (25, 50], \quad B = (50, 75], \quad A = (75, 100]. \quad (1)$$

The boundaries at 25, 50, and 75 are retained in every candidate scale. Each 25-point band is divided into three sub-grades with widths (e, c, e) , where e is the width of each edge sub-grade and c is the width of the centre sub-grade. Since the three widths must cover the full band,

$$e + c + e = 25, \quad \text{and therefore} \quad c = 25 - 2e. \quad (2)$$

The incumbent uniform scale has three equally sized sub-grades:

$$e = c = \frac{25}{3} \approx 8.33. \quad (3)$$

The alternative scales reduce e and increase c , while preserving the number and order of the twelve grades. The centre sub-grade is required to be at least as wide as either edge sub-grade. The feasible range is therefore

$$0 < e \leq \frac{25}{3}. \quad (4)$$

The incumbent and the six alternative scales used in the analysis are listed in Table 1. Each alternative reduces the edge width e and widens the centre band $c = 25 - 2e$, while keeping the twelve grades and their order. The labels S1–S6 defined in Table 1 are used throughout the results. For each main band beginning at $25i$, where $i \in \{0, 1, 2, 3\}$,

Table 1 Incumbent and candidate grade scales

Scale	Edge width e	Centre width c	Sub-grade widths ($e:c:e$)
Uniform	8.33	8.33	8.33:8.33:8.33
S1	5.0	15.0	5:15:5
S2	7.0	11.0	7:11:7
S3	8.0	9.0	8:9:8
S4	6.0	13.0	6:13:6
S5	7.5	10.0	7.5:10:7.5
S6	4.0	17.0	4:17:4

Notes: Edge and centre widths are in score points and satisfy $c = 25 - 2e$. The uniform scale has $e = c = 25/3 \approx 8.33$. Scales S1–S6 keep the A/B/C/D boundaries at 25, 50, and 75 and reallocate width within each quartile.

the two internal boundaries are located at $25i + e$ and $25i + 25 - e$. The complete boundary set is

$$\mathcal{B}(e) = \{25i + e, 25i + 25 - e : i \in \{0, 1, 2, 3\}\} \cup \{25, 50, 75\}. \quad (5)$$

After ordering these eleven cut-points as $b = (b_1, \dots, b_{11})$, a score s is assigned to a grade using

$$g(s; b) = 1 + \sum_{k=1}^{11} \mathbf{1}\{s > b_k\}, \quad (6)$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

3.2 Static identification of alternative scales

Each firm-year score is assigned to a grade under the incumbent and each alternative scale. Agreement with the incumbent is calculated as

$$\text{Agreement}(b) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{g(s_i; b) = g(s_i; b^0)\}, \quad (7)$$

where b^0 is the incumbent boundary vector and N is the number of observed scores.

Proposition 1 *Let b^0 and b be two ordered cut-point vectors with the same number of boundaries and the same assignment rule (6). If, for every k , no observed score lies between b_k^0 and b_k , then $g(s_i; b) = g(s_i; b^0)$ for every observed score s_i .*

Proof By (6),

$$g(s; b) - g(s; b^0) = \sum_{k=1}^{11} \left(\mathbf{1}\{s > b_k\} - \mathbf{1}\{s > b_k^0\} \right).$$

The k -th term is non-zero only if s lies between b_k and b_k^0 , that is, above one of the two boundaries and below or equal to the other. If no observed score s_i lies between b_k^0 and b_k for any k , every term vanishes, so $g(s_i; b) = g(s_i; b^0)$. $\square$

Proposition 1 gives a simple test for when two scales cannot be told apart on static data: it is enough that no observed score falls in the gap between any pair of corresponding boundaries. An alternative scale can therefore produce a different grade only when an observed score lies between a candidate boundary and the corresponding incumbent boundary. If no score lies within any such interval, the two scales produce the same grade assignments and cannot be distinguished using the static data. This is particularly relevant here because nearly all observed scores are located at the midpoints of the incumbent grade bands.

3.3 Stability under score perturbations

Grade stability is evaluated using a Monte Carlo perturbation procedure. Firm-year scores are sampled with replacement from the observed score distribution. Each sampled score is then changed by independent Gaussian noise:

$$\tilde{s}_j = \min\{100, \max\{0, s_j + \varepsilon_j\}\}, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2). \quad (8)$$

The minimum and maximum functions restrict the perturbed scores to the valid interval $[0, 100]$. For a given scale, the migration rate is the proportion of simulated scores that move to another grade:

$$\widehat{\text{Mig}}(b, \sigma) = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{g(s_j; b) \neq g(\tilde{s}_j; b)\}. \quad (9)$$

The parameter σ represents the size of the score disturbance. Values from 0.5 to 5 score points are considered. The reported migration rates are based on 20,000 simulation draws for each scale and each value of σ . The same random seed and perturbation draws are used across the scales to ensure a consistent comparison.

3.4 Boundary safety and midpoint distortion

Boundary safety measures how close the observed scores are to the grade cut-points. It is defined as the smallest distance between any distinct observed score and any boundary:

$$\text{Buffer}(b) = \min_{s \in M} \min_{1 \leq k \leq 11} |s - b_k|, \quad (10)$$

where M is the set of distinct observed scores. A larger buffer indicates that scores must move further before changing grade. A small buffer indicates that at least one observed score is close to a boundary. Midpoint distortion measures how closely the midpoint of an assigned grade represents its underlying score. We use the mean absolute deviation:

$$\text{Delta}(b) = \frac{1}{N} \sum_{i=1}^N |s_i - m(g(s_i; b))|, \quad (11)$$

where $m(g)$ is the midpoint of grade g . A lower value indicates that the grade midpoints represent the observed scores more closely. The absolute deviation is used, rather than the squared error, so that Delta stays in score points; once divided by the uniform sub-grade width in (12) it becomes a unit-free quantity that is directly comparable with the migration probability. The squared analogue is the mean-squared distortion D of Section 3.7, which we use only for the Lloyd–Max argument. The two measures describe different aspects of the scale. Boundary safety measures resistance to grade changes, while midpoint distortion measures the information lost when a continuous score is represented by a discrete grade.

3.5 Optimisation of the band structure

The preferred edge width is selected by combining migration and midpoint distortion in a single loss function:

$$L(e) = \omega_1 \overline{\text{Mig}}(e) + \omega_2 \frac{\text{Delta}(e)}{25/3}, \quad (12)$$

where average migration is calculated as

$$\overline{\text{Mig}}(e) = \frac{1}{3} \sum_{\sigma \in \{1, 2, 3\}} \widehat{\text{Mig}}(\mathcal{B}(e), \sigma). \quad (13)$$

Migration is averaged over score disturbances of one, two, and three points because the optimisation focuses on small and moderate revisions. Midpoint distortion is divided

by $25/3$, the width of a uniform sub-grade, so that the two components are more comparable. The weights are

$$\omega_1 = 0.6, \quad \omega_2 = 0.4. \quad (14)$$

These weights give greater importance to grade stability. They are design choices rather than estimates of transaction costs or investor preferences. The optimisation problem is

$$e^* = \arg \min_{e \in [0.5, 25/3]} L(e). \quad (15)$$

The lower bound of 0.5 rules out degenerate designs whose edge sub-grades would be narrower than the smallest perturbation considered and finer than the rounding of the published scores; below this width an edge sub-grade would no longer correspond to a distinguishable grade. All six candidate scales in Table 1 have $e \geq 4$, so the bound binds only for the optimiser and not for the tabulated scales. Four optimisation procedures are used: grid search, golden-section search (Kiefer 1953), simulated annealing (Kirkpatrick et al. 1983), and a genetic algorithm (Holland 1975; Goldberg 1989). Grid search evaluates the loss at fixed intervals. Golden-section search provides a more precise solution when the loss function is approximately unimodal. Simulated annealing and the genetic algorithm provide additional global searches that make fewer assumptions about the shape of the loss function. Similar results across the four methods indicate that the selected value is not determined by one particular algorithm. The optimisation procedure uses 4,000 simulation draws for each noise level because the loss function must be evaluated repeatedly. The final migration tables and figures use 20,000 draws.

3.6 Auxiliary indicator anchor

A partial sustainability anchor is constructed from the ten available environmental, social, and governance indicators. This anchor is used only as an additional validity check and is not intended to reconstruct the full legacy Refinitiv/LSEG score. Missing indicator values are replaced by their within-year medians. Each indicator is then standardised within year:

$$z_{i,t}^{(k)} = \frac{x_{i,t}^{(k)} - \bar{x}_t^{(k)}}{s_t^{(k)}}, \quad (16)$$

where $\bar{x}_t^{(k)}$ and $s_t^{(k)}$ are the annual mean and standard deviation of indicator k . The indicators are polarity-adjusted so that higher values represent better performance. Carbon emissions, waste, water use, energy use, lost working days, and anti-takeover provisions enter negatively. Women employees, employee training, trade-union representation, and ESG reporting enter positively. The resulting anchor is

$$A_{i,t} = \sum_{k \in \mathcal{P}^+} z_{i,t}^{(k)} - \sum_{k \in \mathcal{P}^-} z_{i,t}^{(k)}. \quad (17)$$

where $\mathcal{P}^+$ and $\mathcal{P}^-$ contain the positively and negatively oriented indicators, respectively. Spearman's rank correlation is used to compare the anchor with the published

ESG score. Because the anchor excludes the provider’s complete indicator set, materiality weights, peer-group adjustments, and controversy rules, it is interpreted as a limited cross-check rather than an independent measure of true ESG performance.

3.7 Quantisation interpretation

Converting a continuous ESG score into a fixed number of grades is a scalar quantisation problem. The score range is divided into cells, and all scores within a cell are represented by the same grade. The quality of the grading system therefore depends on the positions of its boundaries and representative values. For a score distribution with density $f(x)$, boundaries b_i , and grade representatives r_i , mean-squared quantisation distortion is

$$D = \sum_i \int_{b_{i-1}}^{b_i} (x - r_i)^2 f(x) dx. \quad (18)$$

The Lloyd–Max conditions place each representative at the conditional mean of the scores within its cell and each boundary midway between neighbouring representatives (Max 1960; Lloyd 1982; Gray and Neuhoﬀ 1998). For an approximately uniform score distribution, these conditions support equal-width grade bands. If the score distribution is clustered or skewed, narrower bands may instead be appropriate in regions where observations are concentrated.

4 Data and preprocessing

The analytical panel includes 177 large-capitalisation issuers tracked annually from 2018 to 2023, for a total of 1,062 firm-year records. The legacy ESG score is available for 1,048 of these observations, and missing scores are not imputed. For each firm-year, the dataset includes ten indicators covering environmental, social, and governance dimensions: carbon emissions, waste, water use, energy use, women employees, average training, trade-union representation, lost working days, ESG reporting, and anti-takeover provisions. These variables are used only as an auxiliary validation anchor and are not intended to reproduce the full legacy Refinitiv/LSEG methodology.

Percentage values are converted to numeric form, textual missing codes are recoded as missing, and indicator gaps are replaced by within-year medians for the auxiliary analysis. The indicators are then standardised within year and polarity-adjusted so that higher values represent better performance. Carbon emissions, waste, water use, energy use, lost working days, and anti-takeover provisions enter negatively, while the remaining indicators enter positively.

The observed ESG score is highly quantised. Nearly all non-missing scores (99.9%) correspond to the midpoints of the existing grade bands, and the sample contains only eleven distinct score values. This allows the study to examine the placement and stability of grade boundaries, but it does not capture within-grade variation, the original unrounded scores, or the causal effects of grade changes on financial markets. The findings should therefore be interpreted as evidence from a historical sample of large-capitalisation firms rather than as results that apply to all companies (Drempetic et al. 2020).

5 Results

5.1 Static non-identification

This subsection addresses RQ1: whether alternative grade-band designs can be distinguished from the existing uniform scale using the published ESG scores.

Table 2 reports agreement with the incumbent. S1 through S5 assign exactly the same grade as the uniform scale to every observed firm-year. Only S6 differs materially, agreeing on 29.3% of observations. Geometrically, the observed scores sit close to the existing midpoints, so small shifts of the internal boundaries do not cross any of the mass points. The extreme case of S6 shifts the boundaries far enough to cross some of the mass points.

Table 2 Static agreement with the incumbent grade

Scale	Grades identical to uniform (%)	Interpretation
S1 (5:15:5)	100.0	Not identified on static data
S2 (7:11:7)	100.0	Not identified on static data
S3 (8:9:8)	100.0	Not identified on static data
S4 (6:13:6)	100.0	Not identified on static data
S5 (7.5:10:7.5)	100.0	Not identified on static data
S6 (4:17:4)	29.3	Static reclassification occurs

Notes: The scale definitions are given in Table 1. “Not identified” means the scale assigns the same grade as the uniform scale to every observed firm-year, so the two cannot be distinguished on the static data.

The weak score anchor correlation ($\rho_s = 0.098$) does not resolve this problem. Because S1–S5 generate the same grade vector, every grade-based indicator-separation statistic is algebraically identical across them. More importantly, the low correlation indicates that the ten-variable anchor captures only a limited portion of the information or transformation embedded in the legacy score. It would therefore be inappropriate to select a grade scale on the basis of this anchor alone. Accordingly, RQ1 is answered negatively for S1–S5: these alternatives cannot be distinguished from the uniform scale using the published static scores, whereas S6 can be distinguished because it produces substantial reclassification.

5.2 Migration under score perturbations

This subsection addresses the grade-stability component of RQ2 and provides evidence for RQ4 by showing how the preferred design changes with the size of the score disturbance.

Table 3 reports the estimated migration probabilities. The incumbent is exceptionally stable at small perturbations, because its mass points lie close to the centres of their bands. At $\sigma = 0.5$, there is no observed migration for the incumbent in 20,000 draws. The same is true of S3 and several other near-uniform scales. In contrast, S1 and S6 show migration even at $\sigma = 0.5$: 3.65% and 26.16%, respectively. At $\sigma = 2$, the

incumbent migration rate is 3.90%, compared with 4.14% for S3, 5.37% for S5, 7.47% for S2, 14.74% for S4, 25.40% for S1, and 32.96% for S6.

Table 3 Grade migration rate under Gaussian score perturbation

σ	Uniform	S1	S2	S3	S4	S5	S6
0.5	0.00	3.65	0.00	0.00	0.01	0.00	26.16
1.0	0.00	15.06	0.22	0.00	2.59	0.04	30.73
1.5	0.62	21.22	2.44	0.74	8.50	1.24	32.23
2.0	3.90	25.40	7.47	4.14	14.74	5.37	32.96
2.5	9.77	29.60	13.69	10.22	20.48	11.48	33.47
3.0	16.93	33.68	20.56	17.23	26.22	18.48	33.92
4.0	30.28	41.85	33.03	30.70	36.78	31.80	35.10
5.0	41.39	49.01	43.16	41.60	45.58	42.22	37.19

Notes: Based on 20,000 resampled firm-years with scores perturbed by $\mathcal{N}(0, \sigma^2)$ and clipped to $[0, 100]$. The Monte Carlo seed is held fixed across scales; all values are in percentages. Scale definitions are given in Table 1.

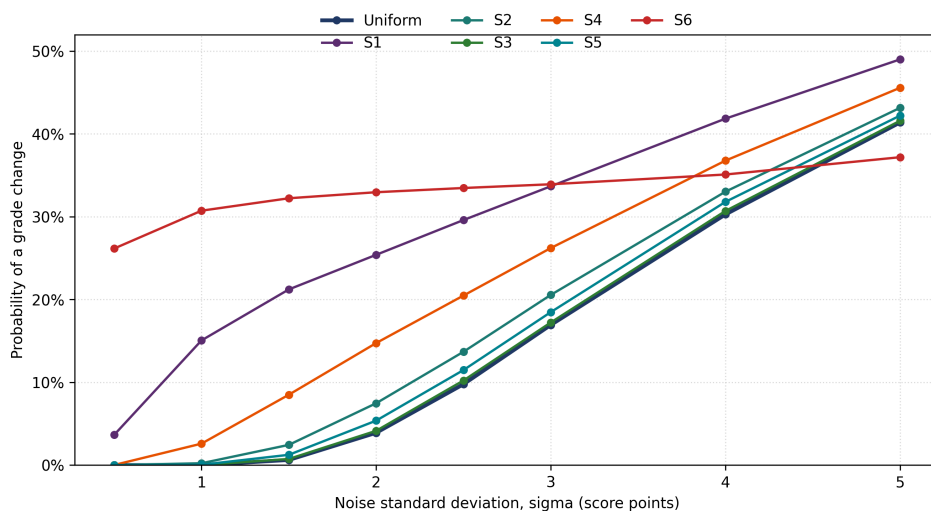


Fig. 1 Grade-migration probabilities under Gaussian score perturbations. The uniform and S3 curves are essentially identical for small and moderate noise. S6 is highly unstable to score perturbations, and its curve crosses the uniform curve at $\sigma = 5$

Figure 1 visualises the migration patterns reported in Table 3 and highlights the crossover between the uniform and S6 designs at the largest disturbance.

Two features of these curves are worth noting. First, S3 closely follows the incumbent curve, reflecting its relatively wide edge of 8 rather than the incumbent’s 8.33. The conclusion is therefore not that every departure from uniformity is disastrous; it is that the data provide no gain from a near-uniform redesign. Second, the ranking changes at the largest disturbance. At $\sigma = 5$, S6 has a migration rate of 37.19%, below

the incumbent’s 41.39%. The wide centre bands that make S6 highly vulnerable to tiny changes around its closely placed boundaries also retain more observations after sufficiently large movements, once scores enter those broad cells. This crossover shows that “stability” is not a scale-free property. The preferred partition depends on the disturbance distribution and on whether the design problem prioritises local measurement noise or large shocks. For score revisions of one to three points—the range used in the optimisation criterion—the incumbent is the lowest or effectively tied with S3 and consistently outperforms the wider centre alternatives. The result is thus strongest as a local-robustness claim. For the stability part of RQ2, the uniform scale is therefore the most stable or effectively tied for small and moderate score changes. For RQ4, these results show that uniform or near-uniform bands are preferable when the design objective is robustness to small and moderate revisions, whereas a wide-centre non-uniform design can become more stable under very large disturbances.

5.3 Boundary safety

This subsection addresses the boundary-safety component of RQ2: which design keeps observed scores furthest from the grade boundaries.

The boundary-safety results in Table 4 support the perturbation evidence. The incumbent has an empirical minimum buffer of 4.14 score points. S3 retains 3.81 points and S5 retains 3.31. The broader-centre designs place at least one cut-point much closer to an observed mass point: S1 leaves 0.81 points and S6 only 0.17. A score revision smaller than one fifth of a point can therefore trigger a migration under S6 for at least one observed value, whereas the incumbent requires a movement of just over four points from its nearest rounded midpoint.

Table 4 Minimum observed distance to the nearest grade boundary

Scale	Minimum buffer (score points)	Relative safety
Uniform	4.14	Highest
S1	0.81	Low
S2	2.81	Moderate
S3	3.81	Near incumbent
S4	1.81	Low–moderate
S5	3.31	High
S6	0.17	Very low

Notes: Buffers use the rounded score values from the worksheet. The theoretical half-width of a uniform band is $25/6 = 4.17$ points, while the empirical value is 4.14 owing to rounding at the recorded midpoints.

Figure 2 shows the same boundary-safety comparison graphically and confirms the ranking reported in Table 4.

This measure also clarifies why a centre-widening heuristic fails in this sample. Widening a centre band does not move every boundary away from the data. Some boundaries are shifted towards the current midpoints, where the observations are concentrated. The first-order effect is therefore to manufacture threshold proximity

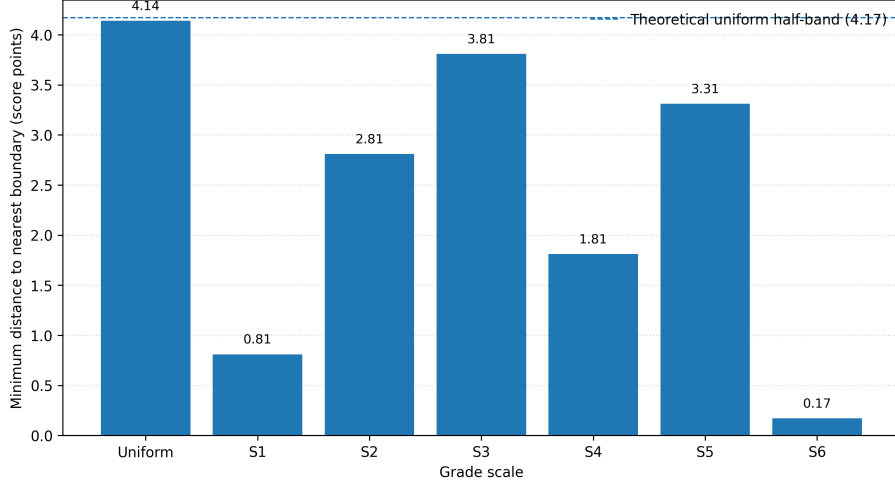


Fig. 2 Minimum observed distance from each score mass point to the nearest grade boundary. The uniform scale provides the largest empirical buffer, whereas S6 places a cut-point only 0.17 score points from an observed value. The dashed reference is the theoretical half-width of a uniform band, $25/6 = 4.17$

rather than remove it. Thus, the boundary-safety part of RQ2 is answered in favour of the uniform scale, which has the largest observed minimum buffer.

5.4 Optimisation of the edge width

This subsection addresses RQ3 and, together with the migration and boundary-safety results, completes the empirical answer to RQ4.

Table 5 presents the numerical solutions to [equation \(12\)](#). Golden-section search and simulated annealing return $e = 8.328$, effectively the upper feasible bound of $25/3$. Grid search returns 8.25 because the grid advances in 0.25-point increments. The genetic algorithm returns 8.125 and a slightly higher objective value. All four methods therefore select the incumbent or a near-uniform design.

Table 5 Convergence of independent optimisation procedures

Method	Optimal e	Objective value	Deviation from $25/3$
Grid search	8.250	0.0427	1.0%
Golden-section search	8.328	0.0414	0.1%
Simulated annealing	8.328	0.0414	0.1%
Genetic algorithm	8.125	0.0453	2.5%

Figure 3 displays the composite loss over the feasible edge-width range and visually confirms the optimiser results reported in Table 5.

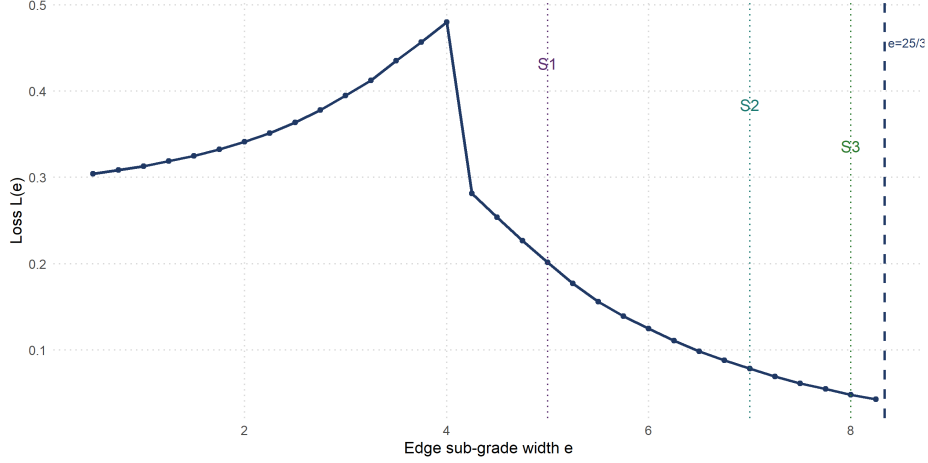


Fig. 3 Composite loss over the feasible edge-width interval. The candidate positions S1, S2, and S3 are marked, and the dashed vertical line denotes the uniform upper bound $e = 25/3$. Under the stated 0.6/0.4 weighting of migration and midpoint distortion, the loss is minimised at the uniform or near-uniform boundary

The optimisation result should be interpreted conditionally. It shows that, for the observed quantised score distribution and a loss function that places 60% weight on average migration at one-, two-, and three-point perturbations, no centre-widening design improves on the incumbent. It does not establish a universal welfare optimum, because the loss weights are not estimated from transaction costs, mandate breaches, or investor preferences. The strongest evidence is the concordance between the optimiser, the migration curves over the local range, the boundary buffer, and the Lloyd–Max argument. RQ3 is therefore answered affirmatively: all four optimisation procedures identify the same or a very similar near-uniform preferred structure. Taken together, the evidence answers RQ4 conditionally: uniform or near-uniform bands are preferable for the observed quantised distribution when small and moderate revisions and the stated loss function are prioritised, while non-uniform wide-centre bands may become preferable when protection against very large score disturbances is the dominant objective.

6 Conclusion

Grade cut-points decide when a change in an ESG score becomes a change in an ESG grade, and they are worth designing rather than inheriting. Studying the legacy Refinitiv/LSEG scale, we reach three qualified conclusions.

First, the design question is hard to answer from published data. Because the released scores are rounded onto the incumbent grade grid, five of six alternative scales label every firm exactly as the incumbent does, so the alternatives are not identified on the static data. We state this as a quantisation-invariance result (Proposition 1), with a simple test for when it applies.

Second, within this sample and for the perturbations we examine, the uniform scale is among the most stable for small and moderate score revisions of up to about four points, and it keeps observed scores furthest from every cut-point; four independent optimisers agree on an edge width at or near the uniform value. This is a local-robustness finding, conditional on the observed score distribution and the chosen loss weights, not a universal welfare optimum. The single crossover at the largest shock shows that the preferred scale depends on the distribution of shocks a designer is protecting against.

Third, scalar quantisation theory offers an interpretation: equal-width bands minimise distortion for the approximately uniform, percentile-rank score that the legacy system produced, and non-uniform bands help only where a score is concentrated.

These conclusions rest on a historical sample of large-capitalisation firms and on quantised published scores; they do not speak to within-grade variation, unrounded scores, or the effect of grades on returns. The same design questions could be asked of the new 0–5 scales now replacing the 0–100 system, but answering them would require access to the unrounded scores.

Statements and Declarations

Funding

No funding was received for this study.

Competing interests

The authors declare no competing interests.

Author contributions

Hira Mahmood performed the data analysis and drafted the manuscript. Minh N. Dao contributed to the methodology, mathematical formulation, and manuscript review. Claudio Conversano contributed to the research idea, data collection, and manuscript review. All authors approved the final manuscript.

Data availability

The data were obtained from Refinitiv/LSEG. Data may be available from the corresponding author upon reasonable request, subject to Refinitiv/LSEG licensing restrictions.

References

Avramov, D., Cheng, S., Lioui, A., Tarelli, A.: Sustainable investing with ESG rating uncertainty. *Journal of Financial Economics* **145**(2), 642–664 (2022) <https://doi.org/10.1016/j.jfineco.2021.09.009>

- Altman, E.I., Rijken, H.A.: How rating agencies achieve rating stability. *Journal of Banking & Finance* **28**(11), 2679–2714 (2004) <https://doi.org/10.1016/j.jbankfin.2004.06.006>
- Berg, F., Cornaggia, J., Foroni, C., Tripoli, F.: The traffic light effect in ESG ratings. Working Paper 26-072, Harvard Business School (April 2026). <https://doi.org/10.2139/ssrn.6630858>
- Billio, M., Costola, M., Hristova, I., Latino, C., Pelizzon, L.: Inside the ESG ratings: (dis)agreement and performance. *Corporate Social Responsibility and Environmental Management* **28**(5), 1426–1445 (2021) <https://doi.org/10.1002/csr.2177>
- Bennett, W.R.: Spectra of quantized signals. *Bell System Technical Journal* **27**(3), 446–472 (1948) <https://doi.org/10.1002/j.1538-7305.1948.tb01340.x>
- Berg, F., Fabisik, K., Sautner, Z.: Is history repeating itself? the (un)predictable past of ESG ratings. Finance Working Paper 708/2020, European Corporate Governance Institute (2021). <https://doi.org/10.2139/ssrn.3722087>
- Berg, F., Kölbel, J.F., Rigobon, R.: Aggregate confusion: The divergence of ESG ratings. *Review of Finance* **26**(6), 1315–1344 (2022) <https://doi.org/10.1093/rof/rfac033>
- Chatterji, A.K., Durand, R., Levine, D.I., Touboul, S.: Do ratings of firms converge? implications for managers, investors and strategy researchers. *Strategic Management Journal* **37**(8), 1597–1614 (2016) <https://doi.org/10.1002/smj.2407>
- Cantor, R., Mann, C.: Analyzing the tradeoff between ratings accuracy and stability. *The Journal of Fixed Income* **16**(4), 60–68 (2007) <https://doi.org/10.3905/jfi.2007.683317>
- Christensen, D.M., Serafeim, G., Sikochi, A.: Why is corporate virtue in the eye of the beholder? the case of ESG ratings. *The Accounting Review* **97**(1), 147–175 (2022) <https://doi.org/10.2308/TAR-2019-0506>
- Cover, T.M., Thomas, J.A.: *Elements of Information Theory*, 2nd edn. Wiley-Interscience, Hoboken, NJ (2006)
- Drempetic, S., Klein, C., Zwergel, B.: The influence of firm size on the ESG score: Corporate sustainability ratings under review. *Journal of Business Ethics* **167**(2), 333–360 (2020) <https://doi.org/10.1007/s10551-019-04164-1>
- Dimson, E., Marsh, P., Staunton, M.: Divergent ESG ratings. *The Journal of Portfolio Management* **47**(1), 75–87 (2020) <https://doi.org/10.3905/jpm.2020.1.175>
- Escrig-Olmedo, E., Fernández-Izquierdo, M.Á., Ferrero-Ferrero, I., Rivera-Lirio, J.M.,

- Muñoz-Torres, M.J.: Rating the raters: Evaluating how ESG rating agencies integrate sustainability principles. *Sustainability* **11**(3), 915 (2019) <https://doi.org/10.3390/su11030915>
- Gibson Brandon, R., Krueger, P., Schmidt, P.S.: ESG rating disagreement and stock returns. *Financial Analysts Journal* **77**(4), 104–127 (2021) <https://doi.org/10.1080/0015198X.2021.1963186>
- Gray, R.M., Neuhoﬀ, D.L.: Quantization. *IEEE Transactions on Information Theory* **44**(6), 2325–2383 (1998) <https://doi.org/10.1109/18.720541>
- Goldberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA (1989)
- Holland, J.H.: *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor (1975)
- Hartzmark, S.M., Sussman, A.B.: Do investors value sustainability? a natural experiment examining ranking and fund flows. *The Journal of Finance* **74**(6), 2789–2837 (2019) <https://doi.org/10.1111/jofi.12841>
- Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. *Science* **220**(4598), 671–680 (1983)
- Kiefer, J.: Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society* **4**(3), 502–506 (1953)
- Krüger, P.: Corporate goodness and shareholder wealth. *Journal of Financial Economics* **115**(2), 304–329 (2015) <https://doi.org/10.1016/j.jfineco.2014.09.008>
- Kotsantonis, S., Serafeim, G.: Four things no one will tell you about ESG data. *Journal of Applied Corporate Finance* **31**(2), 50–58 (2019) <https://doi.org/10.1111/jacf.12346>
- Lloyd, S.P.: Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**(2), 129–137 (1982) <https://doi.org/10.1109/TIT.1982.1056489>
- Löffler, G.: An anatomy of rating through the cycle. *Journal of Banking & Finance* **28**(3), 695–720 (2004) [https://doi.org/10.1016/S0378-4266\(03\)00041-4](https://doi.org/10.1016/S0378-4266(03)00041-4)
- London Stock Exchange Group: *LSEG Sustainability Ratings and Data: ESG Scores and ESG Scores Plus Scoring Methodology, v1.0*. LSEG Data & Analytics (2026)
- Larcker, D.F., Pomorski, L., Tayan, B., Watts, E.M.: ESG ratings: A compass without direction. Working paper, Rock Center for Corporate Governance, Stanford University (2022). Stanford Closer Look Series
- LSEG: *Environmental, Social and Governance Scores from LSEG: Methodology*.

- Methodology guide for the legacy Refinitiv/LSEG ESG score. Accessed 30 July 2026 (2022)
- LSEG: LSEG Launches New Suite of ESG Scores and Sustainability Analytics to Enhance Transparency for Global Markets. Press release. Introduces a 0–5 threshold-based scoring scale; legacy 0–100 scores move to legacy status (2026)
- Max, J.: Quantizing for minimum distortion. *IRE Transactions on Information Theory* **6**(1), 7–12 (1960) <https://doi.org/10.1109/TIT.1960.1057548>
- Panter, P.F., Dite, W.: Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proceedings of the IRE* **39**(1), 44–48 (1951) <https://doi.org/10.1109/JRPROC.1951.230419>
- Pedersen, L.H., Fitzgibbons, S., Pomorski, L.: Responsible investing: The ESG-efficient frontier. *Journal of Financial Economics* **142**(2), 572–597 (2021) <https://doi.org/10.1016/j.jfineco.2020.11.001>
- Pástor, Ľ., Stambaugh, R.F., Taylor, L.A.: Sustainable investing in equilibrium. *Journal of Financial Economics* **142**(2), 550–571 (2021) <https://doi.org/10.1016/j.jfineco.2020.12.011>
- Serafeim, G., Yoon, A.: Stock price reactions to ESG news: The role of ESG ratings and disagreement. *Review of Accounting Studies* **28**(3), 1500–1530 (2023) <https://doi.org/10.1007/s11142-022-09675-3>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [7scaleesgscore.rmd](#)
- [ESGindicatorandscore.xlsx](#)
- [7scaleesgscore.rmd](#)
- [ESGsevenscalesanalysis.xlsx](#)