The Traffic Light Effect in ESG Ratings

Florian Berg*, Jess Cornaggia†, Cristian Foroni‡, and Francesco Tripoli§✉

*Massachusetts Institute of Technology

†Pennsylvania State University

‡European University Institute

§Harvard Business School

Abstract

ESG ratings increasingly govern index construction and investment mandates, creating a conflict of interest for investor-paid raters: since rating changes trigger rebalancing costs for subscribers, raters face pressure to favor stability over timeliness. We document a *traffic light effect* in MSCI ESG ratings, the world's largest provider: companies systematically bunch just below or above letter-rating thresholds, as if queued behind a red light. By reverse-engineering MSCI's benchmarking methodology, we identify committee adjustments that create wedges between methodology-implied and published scores. These wedges grow with divergence from peer consensus, especially when peers agree closely, consistent with a model in which investor-paid raters trade methodological consistency for peer alignment. Exploiting a methodology change at Refinitiv, we show that MSCI conditions overrides on contemporaneously available peer ratings rather than on common fundamentals. Implemented downgrades predict negative returns of 2.8% over twelve months; suppressed downgrades do not.

Keywords: ESG ratings, information intermediaries, investor-pays model, herding, rating agencies, sustainable investing

JEL: G24, G14, M14, Q56

✉ Corresponding Author: Francesco Tripoli (ftripoli@hbs.edu)

We are grateful to Michaël Aklin, Shirley Lu, Roberto Rigobon and Charles Wang for helpful feedback on this paper. We also thank seminar participants at the Harvard Business School Accounting brownbag.

1 Introduction

Information intermediaries play a critical role in efficient capital allocation (Healy and Palepu, 2001; Bushman and Smith, 2001). By collecting, processing, and disseminating information about firms, they reduce information asymmetries between corporate insiders and outside investors (Akerlof, 1970). The quality of information produced by intermediaries depends critically on their compensation structure, which establishes the incentives governing their behavior. A central question in financial economics is whether intermediaries paid by information users (investors) produce different, and potentially higher quality, information than those paid by the subjects of their analysis (issuers).

The credit rating industry has served as the primary laboratory for studying how compensation structure affects information production. Major credit rating agencies such as Moody's, Standard & Poor's, and Fitch operate under an issuer-pays model, raising concerns about conflicts of interest that could lead to inflated ratings (Griffin and Tang, 2012; Griffin et al., 2013). In response, investor-paid agencies such as Egan-Jones and Rapid Ratings emerged as alternatives. A substantial body of research compares these two models, generally finding that investor-paid agencies produce more timely ratings, update ratings more frequently, and provide better ordinal rankings of credit risk (Cornaggia and Cornaggia, 2013; Bruno et al., 2016; Beaver et al., 2006). This evidence has led many to conclude that investor-paid compensation structures are superior for information quality. A caveat, however, is the trade-off between timeliness and stability: because credit ratings are embedded in investment guidelines and regulatory triggers, frequent revisions impose portfolio-rebalancing costs on the very investors who consume them, creating demand for stability even from subscribers (Bonsall et al., 2024).

We show that a closely related but distinct conflict of interest arises in ESG ratings, where investor-paid raters face even stronger pressure to favor stability over timeliness. ESG ratings now directly govern index construction, investment mandates, and exclusion screens (Amel-Zadeh and Serafeim, 2018). Subscribers who rely on these ratings to manage portfolios bear concrete costs when ratings change: index reconstitutions trigger forced trades, mandate-

driven rebalancing generates transaction costs, and rating volatility complicates long-horizon investment strategies. Because the rater's revenue depends on subscriber demand, it has a direct incentive to internalize these costs and limit the frequency and magnitude of rating changes, even when its own methodology would warrant an update. The result is a bias toward stability that operates through the very compensation structure that is often presumed to align rater and investor interests.

Unlike credit ratings, however, ESG attributes are multi-dimensional, slow-moving, and lack a clean, ex post benchmark. Different providers deliberately emphasize different dimensions of Environmental, Social, and Governance performance, apply distinct materiality frameworks, and weight underlying components in non-comparable ways (Berg et al., 2022; Azarmsa and Shapiro, 2025). There is no universally accepted notion of ESG "accuracy" analogous to the default outcomes that discipline credit ratings. In this setting, the relevant trade-off is not between accuracy and consensus, as in the credit ratings literature, but between *methodological consistency* and *peer alignment*. When investors cannot verify which rater is "right", they use cross-rater agreement as a heuristic for reliability and treat divergence as a signal of model risk. Rating alignment is interpreted as a proxy for quality, while standing apart from the market looks less like independent judgment and more like noise. This creates a second, reinforcing incentive for the rater: to adjust published scores toward peer consensus, sacrificing the internal consistency and methodological discipline that constitute its core analytical strength, in order to enhance the perceived informativeness and usability of its ratings.[1]

Using MSCI, the world's largest ESG rater, as our focal provider, we document a striking empirical pattern that we term the *traffic light effect*: companies systematically bunch just below or above the thresholds separating adjacent letter ratings, much like cars queuing behind a stop line while waiting for a green light. Since MSCI's letter ratings are determined by sharp cutoffs applied to a decimal score, even small numerical adjustments can have

[1] We use the term *peer alignment* rather than the conventional *herding* to reflect the ESG-specific context: since there is no universally agreed ground truth for sustainability performance, the relevant trade-off is between *methodological consistency* and *peer conformity*, not between accuracy and consensus. The formal mechanism is nonetheless isomorphic to the herding incentives formalized in the broader ratings literature.

discrete consequences. The traffic light effect therefore points to deliberate stability precisely where the incentives to manage rating transitions are strongest.

Institutional accounts shed light on how this mechanism operates in practice. MSCI's public materials indicate that rating stability was sufficiently important to be made an explicit topic of stakeholder consultation in 2024.[2] Moreover, a *Financial Times* piece describes an internal review process in which analysts proposing ESG rating changes may be required to justify them before a panel of senior colleagues, who ultimately decide whether a proposed update is acceptable. A former MSCI analyst, quoted anonymously, recalls both internal and external pressure to avoid substantial rating swings.[3] These testimonies point to an institutional setting in which large rating movements are subject to ex post moderation by a review committee, rather than being implemented mechanically. Committee governance is not, by itself, evidence of distortion: committees can improve consistency and discipline. However, in a setting where clients value both stability and peer alignment, and where there is no objective benchmark against which to evaluate rating changes, such discretion creates scope for a conflict of interest: the incentive to internalize subscribers' rebalancing costs by limiting rating changes and to avoid *looking* "wrong" relative to the rest of the market.

To interpret these patterns, we develop a deliberately simplified but illustrative model in which an investor-paid rater faces two coexisting incentive channels. The first is a *stability preference*: investors independently value continuity of published ratings, so any change in the published score reduces subscriber demand. The second is a *peer alignment preference*: when a provider's score diverges from other raters' evaluations, investors infer higher model risk and face higher integration costs, reducing demand especially when peers broadly agree. The model highlights two distinct forces shaping committee decisions: pressure to preserve rating stability and pressure to maintain alignment with peer assessments. Together, these forces can lead the committee to hold back an initial proposed revision ("red light"), while repeated proposals may become more likely to be implemented over time ("green light"). The traffic light effect thus emerges endogenously from the same profit-maximization problem,

[2] See MSCI 2024 ESG Ratings Model Consultation.

[3] See ESG ratings: whose interests do they serve?, *Financial Times*, Oct 3, 2023.

generating both short-run stability and, if delayed, responsiveness to persistent signals.

Our empirical analysis proceeds in three steps. First, we reverse-engineer MSCI's benchmarking methodology to identify firms whose Industry Adjusted Scores (IAS) were likely modified by an internal committee review. MSCI's methodology translates raw Weighted Average Scores (WAS) into industry-adjusted scores using a linear transformation with industry-specific parameters. However, the published IAS sometimes deviates from what this formula would imply, particularly near rating thresholds. By optimizing over the benchmark parameters to maximize exact matches between implied and actual scores, we recover the methodology-implied score for each firm and identify wedges attributable to committee intervention, particularly concentrated at rating boundaries.

Second, we test whether these committee adjustments reflect both peer alignment and rating stability. We assemble firm-level ESG scores from eight major providers and construct a peer consensus measure along with a mean absolute dispersion statistic capturing peer agreement. We find that the wedge between MSCI's adjusted and methodology-implied score is strongly correlated both with the divergence between MSCI's model and the peer consensus and with the adjustment needed to preserve the prior rating. Critically, the peer-alignment relationship is moderated by peer agreement: when other raters broadly agree (low dispersion), divergence from the consensus triggers larger adjustments.

To sharpen identification, we further exploit a discrete methodology change in Thomson Reuters / Refinitiv ESG Ratings that occurred in 2020. Importantly for our purposes, this methodological overhaul was implemented not only prospectively but also retroactively: Refinitiv recomputed ESG scores for all historical fiscal years using the new framework, so that the series currently available to users are fully backfilled under the post-2020 methodology. This feature allows us to compare how MSCI's wedge responds to the legacy Thomson Reuters ratings that were actually available at the time versus the backfilled Refinitiv scores that did not exist when rating decisions were made. We find that before 2020, MSCI's adjustments co-move with divergence from the contemporaneously available Thomson Reuters ratings, while divergence from the retroactively computed Refinitiv scores has no explanatory

power. After 2020, the pattern reverses. This evidence rules out the hypothesis that wedges simply reflect common exposure to slowly moving fundamentals, since these are embedded in both the legacy and backfilled series. Instead, it points to an information channel: MSCI conditions its overrides on signals from contemporaneous peer assessments.

Third, we examine the financial consequences of the traffic light effect by studying stock returns following rating changes. We compare implemented and suppressed downgrades and find that the first predict significant negative buy-and-hold returns of approximately 2.8 percentage points over the subsequent twelve months. In contrast, suppressed downgrades show no systematic relationship with future returns. This asymmetry suggests that the suppressed downgrades contain information about ESG risk that the market does not fully incorporate, consistent with the view that committee adjustments impede the flow of value-relevant signals to investors.

Our findings contribute to several strands of literature. First, we add to the debate on optimal compensation structures for information intermediaries (Kashyap and Kovrijnykh, 2013; Lovo and Olivier, 2025). While prior work on credit ratings generally finds that investor-pay models produce more timely and informative signals (Cornaggia and Cornaggia, 2013; Bruno et al., 2016), we show that the same compensation structure can generate stability bias in the ESG context. The key difference is that ESG ratings are directly embedded in index products and portfolio mandates, so rating changes impose immediate, concrete costs on the very subscribers who fund the rater. Bonsall et al. (2024) document conflicts of interest at the subscriber-paid credit rating agency Egan-Jones, where the mechanism runs through favorable treatment of firms whose bonds are held by subscribing investors. Our paper identifies a distinct channel: the conflict in ESG ratings arises not from favorable treatment of specific rated entities, but from a structural bias toward stability and peer alignment driven by how subscribers *use* the ratings downstream. In credit markets, the stability–timeliness trade-off arises because frequent revisions impose rebalancing costs on investors subject to ratings-based regulations; in ESG markets, the same tension is amplified because there is no accepted notion of "accuracy" against which to benchmark rating changes, leaving peer consensus as the dominant signal of quality. Relatedly, Cornaggia and Cornaggia (2026) show

that corporations strategically manage their reported ESG performance in response to rating criteria, suggesting that the demand-side pressures we identify interact with supply-side gaming to further complicate the information environment.

Second, we speak to the literature on ESG rating divergence (Berg et al., 2022; Christensen et al., 2022; Brandon et al., 2021). Prior work documents substantial disagreement across ESG raters attributed to differences in scope, measurement, and weight. Our results instead suggest a source of *convergence*: rating agencies actively adjust their scores toward peer assessments, especially when these are tightly clustered. Chen et al. (2025) show that competition among ESG raters improves rating quality. Our findings suggest, however, that this disciplining effect operates in part through the peer alignment mechanism we identify, which need not enhance the underlying information content of ratings.

Third, we contribute to the literature on price formation and the capital-market effects of ESG information by showing that committee discretion affects whether value-relevant ESG signals are incorporated into prices (Derrien et al., 2025; Krüger, 2015).

The remainder of the paper proceeds as follows. Section 2 develops our theoretical framework. Section 3 describes the institutional background, data, and methodology for recovering benchmark parameters. Section 4 presents our main empirical results on peer alignment, the Refinitiv methodology change, and financial implications. Section 5 concludes.

2 A stylized model of investor–pay model incentives

Before turning to the empirical analysis, we develop a simple theoretical framework that clarifies how conflicts of interest can arise in an investor–pay ratings model. The goal is not to provide a full account of MSCI's methodology, but rather a stripped–down "toy model" that isolates two distinct and coexisting mechanisms: (i) an incentive for *peer alignment*, since investors use cross-rater agreement as a signal of rating quality; and (ii) a *stability bias*, since investors value continuity of published ratings.

Investors cannot directly observe firms' ESG fundamentals and therefore judge rating providers based on how informative, reliable and usable their signals appear to be. Under an investor–pay model, the provider's revenue depends on subscribers' willingness to pay for that information, which creates an incentive to manage not only methodological consistency, but also investors' *perceptions* of reliability and model risk (Holmström, 1979; Kreps, 1990; Kashyap and Kovrijnykh, 2013; Lovo and Olivier, 2025; Azarmsa and Shapiro, 2025). In this section we formalize this trade-off and provide a microfoundation for the peer-alignment and stability incentives emphasized in the remainder of the paper.

Let us consider a firm with a latent ESG state $\theta \in \mathbb{R}$ with prior $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$. MSCI's methodology-implied (industry-adjusted) score is given by:

$$y_M = \theta + \varepsilon_M, \qquad \varepsilon_M \sim \mathcal{N}(0, v_M).$$

Other raters $j = 1, \ldots, J-1$ report $y_j = \theta + \varepsilon_j$, $\varepsilon_j \sim \mathcal{N}(0, v_j)$. Let y_C be their average, and define the *mean absolute dispersion (MAD)*:

$$m_C := \frac{1}{J-1} \sum_{j=1}^{J-1} |y_j - y_C|.$$

m_C measures rater consensus: small m_C shows agreement, large m_C shows dispersion.

MSCI maps the continuous industry-adjusted score to one of $K = 7$ discrete letter categories (CCC, B, BB, BBB, A, AA, AAA) via a step function $L(\cdot)$ defined by fixed thresholds $\bar{y}_1 < \cdots < \bar{y}_{K-1}$: $L(y) = k$ if $\bar{y}_{k-1} \leq y < \bar{y}_k$. Since investors primarily act on the letter rating rather than the underlying continuous score, rating continuity is achieved by preventing the underlying score from crossing a threshold between two letter ratings. Let $\ell_{-1} := L(y_{A,-1})$ denote the letter rating published in the preceding period, with associated bucket $[\underline{y}_{\ell_{-1}}, \bar{y}_{\ell_{-1}})$. We define a *retention gap* as

$$g \;:=\; \max\Big(0,\; y_M - \bar{y}_{\ell_{-1}},\; \underline{y}_{\ell_{-1}} - y_M\Big) \;\geq\; 0. \tag{1}$$

Thus, the retention gap equals zero when the methodology-implied score already lies within last year's bucket, and is otherwise defined as the minimum distance from the previous year letter rating threshold.

Before releasing the updated scores, the rater may override y_M by blending with peers:

$$y_A \;=\; (1-\varphi)\, y_M + \varphi\, y_C, \qquad \varphi \in \mathbb{R}, \tag{2}$$

where $\varphi > 0$ moves the published score toward y_C and $\varphi < 0$ moves it away. This override incurs in a *wedge* $\Delta_A := y_A - y_M = -\varphi\Delta_0$, $|\Delta_A| = |\varphi||\Delta_0|$, with $\Delta_0 := y_M - y_C$. Now, since ESG attributes are *unobserved*, investors judge rating providers by *perceived* information value. Three observable drivers matter:

1. An ex-ante valuation of provider-specific data (granularity, coverage, methodology quality, documentation, etc.), summarized in a latent baseline value $B \geq 0$;
2. A *peer alignment preference*: a disagreement cost that rises when a provider's score diverges from other raters' evaluations, leading users to infer higher model risk and face higher integration costs. This cost is amplified when peers broadly agree (low m_C), since concordant providers are perceived as more likely to be correct;
3. A *rating stability preference*: investors relying on letter ratings for portfolio benchmarking, compliance mandates or passive tilts incur rebalancing costs whenever a letter rating changes, with disruption perceived as greater when the change results from a marginal threshold crossing.

Thus, we summarize the *subscriber demand* at fixed price p by

$$N(\varphi) \;=\; B \;-\; \rho\,\omega(m_C)\cdot(y_A - y_C)^2 - \;\frac{\sigma}{d_A(\varphi)}\cdot \mathbf{1}\{g>0,\, L(y_A)\neq \ell_{-1}\} \;-\; c\,\varphi^2, \tag{3}$$

with $\rho, \sigma, c > 0$, $\omega'(m_C) < 0$, $\omega(m_C) = 1/(1+m_C)$ as a canonical example, and $d_A(\varphi) := g - \varphi\Delta_0 > 0$. The first penalty captures peer alignment. The second captures stability demand: a cost incurred when the published letter rating is updated and whose magnitude

$\sigma/d_A(\varphi)$ is inversely proportional to the distance of the published score from the previous year letter rating threshold (a small d_A signals that ESG fundamentals barely moved, yet the letter rating changed). The third term captures subscribers' preference for transparency and methodological consistency: the larger $|\varphi|$, the more the published score departs from the methodology-implied one.

Let $F \geq 0$ be the minimum net gain in subscriber demand required to justify any override, capturing the fixed reputational cost of deviating from the stated methodology. Letting $h := 2pc$, the methodology committee faces the following Profit Maximization Problem:

$$\begin{aligned} \max_{\varphi} \quad & \Pi(\varphi) \;=\; p\,N(\varphi) \;-\; F \cdot \mathbf{1}\{\varphi \neq 0\} \\ \text{s.t.} \quad & \text{(C1) Score construction: } y_A(\varphi) - (1-\varphi)y_M - \varphi y_C = 0. \end{aligned} \tag{4}$$

Using (C1), $(y_A - y_C) = (1-\varphi)\Delta_0$. Defining the *smooth peer-alignment component*

$$\Pi_{PA,\mathrm{sm}}(\varphi) \;:=\; p\big(B - \rho\,\omega(m_C)\,(1-\varphi)^2\Delta_0^2\big) - \frac{h}{2}\varphi^2, \tag{5}$$

the full profit on the letter-changing region decomposes as

$$\Pi(\varphi) \;=\; \underbrace{\Pi_{PA,\mathrm{sm}}(\varphi) - F \cdot \mathbf{1}_{\varphi\neq 0}}_{\Pi_{PA}(\varphi)\text{ (peer-alignment component)}} \;-\; \underbrace{\frac{p\sigma}{d_A(\varphi)}}_{\text{stability penalty}}. \tag{6}$$

Since $\varphi \in \mathbb{R}$ is unconstrained, the first-order condition on each smooth region is $d\Pi_{PA,\mathrm{sm}}/d\varphi - d[p\sigma/d_A(\varphi)]/d\varphi = 0$, where $d[-p\sigma/d_A]/d\varphi = -p\sigma\Delta_0/d_A^2$: this term is negative when $\Delta_0 > 0$ (stability penalty worsens as y_A approaches the threshold, pulling φ below φ^{PA}) and positive when $\Delta_0 < 0$ (stability penalty improves as y_A moves further above the threshold, pushing φ above φ^{PA}). Letting $A := 2p\rho\,\omega(m_C)\,\Delta_0^2 \geq 0$, the peer-alignment-only ($\sigma = 0$) FOC is $A(1-\varphi) = h\varphi$, giving $\varphi^{PA} := A/(A+h)$.

Lemma 1 (Concavity and interior candidates)**.** *On the letter-preserving region, the smooth objective is $\Pi_{PA,\mathrm{sm}}(\varphi)$, strictly concave with unique global maximum $\varphi^{PA} := A/(A+h) \in [0,1)$. On the letter-changing region, the smooth objective $\Pi^{-}_{\mathrm{sm}}(\varphi) := \Pi_{PA,\mathrm{sm}}(\varphi) - p\sigma/d_A(\varphi)$*

is strictly concave with unique interior maximum φ^* satisfying: $\varphi^* < \varphi^{PA}$ when $\Delta_0 > 0$ (reinforcing channels); $\varphi^* > \varphi^{PA}$ when $\Delta_0 < 0$ (opposing channels). For $\sigma = 0$: $\varphi^* = \varphi^{PA}$ and $\Delta\Pi_{PA} := A^2/[2(A+h)] > 0$.

Proof. See Appendix. □

Since $h > 0$, the consistency penalty ensures $|\varphi^*|$ is finite on every smooth region.

Proposition 1 (Optimal committee adjustment)**.** *When $g = 0$ (stability channel inactive), the global maximizer of (4) is*

$$\varphi^\star \;=\; \begin{cases} \varphi^{PA} = \dfrac{A}{A+h} & \text{if } \dfrac{A^2}{2(A+h)} \;\geq\; F, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

When $g > 0$, $\Delta_0 > 0$ (reinforcing channels) and $\varphi^{PA} \geq \varphi^{RG}$, the stability penalty is avoided at φ^{PA}: the net gain from overriding rises to $\Delta\Pi_{PA} + p\sigma/g \geq \Delta\Pi_{PA}$, weakly lowering the effective fixed-cost hurdle.

Proof. See Appendix. □

Proposition 2 (Comparative statics)**.** *If $\varphi^\star = \varphi^{PA}$ (i.e. $\sigma = 0$), then*

$$\frac{\partial \varphi^{PA}}{\partial \Delta_0^2} \;=\; \frac{2p\rho\,\omega(m_C)\,h}{(A+h)^2} \;>\; 0, \qquad \frac{\partial \varphi^{PA}}{\partial m_C} \;=\; \frac{2p\rho\,\Delta_0^2\,h}{(A+h)^2}\,\omega'(m_C) \;<\; 0.$$

The realized wedge $|\Delta_A| = \varphi^{PA}|\Delta_0|$ satisfies $\partial|\Delta_A|/\partial|\Delta_0| > 0$ and $\partial|\Delta_A|/\partial m_C < 0$. The event $\{\varphi^\star \neq 0\}$ is (weakly) increasing in $|\Delta_0|$ and (weakly) decreasing in m_C.

Proof. See Appendix. □

Proposition 3 (Peer-alignment activation threshold)**.** *A peer-alignment override ($g = 0$, $\varphi^\star = \varphi^{PA} > 0$) occurs if and only if*

$$\omega(m_C) \;\geq\; \bar{\omega}(\Delta_0) \;:=\; \frac{F + \sqrt{F^2 + 2Fh}}{2p\rho\,\Delta_0^2}. \tag{8}$$

For $\omega(m_C) = 1/(1+m_C)$, this is equivalent to $m_C \leq \bar{m}_C(\Delta_0) := 1/\bar{\omega}(\Delta_0) - 1$. $\bar{\omega}(\Delta_0)$ is strictly decreasing in $|\Delta_0|$ and strictly increasing in F and h.

Proof. See Appendix. □

Proposition 4 (Stability bias)**.** *Suppose $g > 0$. Define the* retention gain

$$\Gamma \;:=\; \frac{p\sigma}{g} \;-\; \frac{A+h}{2}(\varphi^{RG} - \varphi^{PA})^2, \tag{9}$$

and let $\tilde{\Delta}(\varphi^{RG}) := \Pi_{PA,\mathrm{sm}}(\varphi^{RG}) - \Pi_{PA,\mathrm{sm}}(0)$. Then:

1. Reinforcing channels ($\Delta_0 > 0$, $\varphi^{PA} < \varphi^{RG}$):

$$\varphi^\star = \begin{cases} \varphi^{RG} & \text{if } \Gamma \geq 0 \text{ and } \tilde{\Delta}(\varphi^{RG}) + p\sigma/g \geq F, \\ \varphi^* & \text{if } \Gamma < 0 \text{ and } \Pi^-_{\mathrm{sm}}(\varphi^*) - \Pi^-_{\mathrm{sm}}(0) \geq F, \\ 0 & \text{otherwise.} \end{cases}$$

   When $\varphi^\star = \varphi^{RG}$: $|\Delta_A| = g$. Since $p\sigma/g$ is decreasing in g, the retention incentive is strongest for marginal threshold crossings.

2. Opposing channels ($\Delta_0 < 0$): *$\varphi^* > \varphi^{PA}$ on the letter-changing region. The stability override $\varphi^{RG} = g/\Delta_0 < 0$ eliminates the penalty but sets $|\Delta_A| = g$ at alignment cost.*

Proof. See Appendix. □

In short, the stability channel interacts with peer alignment through $d_A(\varphi) = g - \varphi\Delta_0$, and its role depends on the sign of Δ_0. When $\Delta_0 > 0$ (reinforcing channels) and $\varphi^{PA} < \varphi^{RG}$, peer alignment alone does not retain last year's bucket and the committee faces a genuine tradeoff between overriding to φ^{RG} (eliminating the stability penalty) and stopping at the letter-changing interior solution $\varphi^* < \varphi^{PA}$. When $\Delta_0 < 0$ (opposing channels), both channels call for positive φ within the letter-changing region since increasing d_A reduces the stability penalty, but eliminating it entirely requires a stability-escape override $\varphi \leq g/\Delta_0 < 0$, which

incurs large peer-alignment and consistency costs: the outcome is a tug of war between a "green team" pushing for the threshold crossing in the direction of peer alignment, and the "red team", which pulls the score towards stability, incurring in the traffic light effect.

Since the rater revenues under an investor-pay model scale with $N(\varphi)$ and N penalizes both divergence from peers when consensus is high *and* threshold crossings in proportion to their marginal severity, the rater is incentivized to align with peers relative to a planner that only minimizes statistical error. The latter minimizes $\mathrm{Var}(y_A - \theta) = (1-\varphi)^2 v_M + \varphi^2 v_C$, yielding the classical MSE shrinkage (Holmström, 1979; Kreps, 1990):

$$\varphi^{\mathrm{pl}} \;=\; \frac{v_M}{v_M + v_C}. \tag{10}$$

Since there is no universally agreed ESG ground truth, φ^{pl} represents a statistical benchmark rather than a normative ideal. The investor-pay model introduces both demand channels in (3), incentivizing peer alignment beyond this benchmark.

Proposition 5 (Excess peer alignment condition)**.** *An investor-pay model exhibits excess peer alignment relative to the planner benchmark when*

$$\varphi^{PA} \;\geq\; \varphi^{\mathrm{pl}} \quad\Longleftrightarrow\quad 2p\,\rho\,\omega(m_C)\,\Delta_0^2 \;\geq\; \frac{v_M}{v_C}\,h.$$

The stability channel ($g > 0$) makes this condition more likely to hold: when $\Delta_0 > 0$, by raising the net gain from any positive override through $p\sigma/g$; when $\Delta_0 < 0$, by amplifying the peer-alignment override above φ^{PA} within the letter-changing region.

Proof. See Appendix. □

Thus, when peer consensus is strong (low m_C, large ω) and/or the methodology-implied score has crossed a rating threshold, the investor-pay objective tilts toward *conformity* and *stability* to preserve N, pushing $|\varphi^\star|$ above the statistical benchmark φ^{pl}.

3 Empirical setting

We now turn to the empirical setting in which we test the mechanisms modeled in the previous Section. Our analysis focuses on MSCI ESG Ratings, which is considered the most influential commercial ESG rater in capital markets [4]. Throughout this paper, we document how score adjustments decided at specific points in time and attributed to the intervention of an internal "methodology committee", lead firms to bunch just below or above rating thresholds and thereby prevent upgrades or downgrades. This pattern, which we refer to as the *traffic light effect*, is visible in the sharp peaks of the rating distribution close to letter-grade cutoffs. As Figure 1 shows, companies line up just before the thresholds separating two letter ratings, much like cars queuing behind a stop line while waiting for the green light.

To interpret these descriptive patterns, and to understand why such bunching is unlikely to be a purely mechanical implication of the scoring model, we provide a brief overview of MSCI's ESG ratings methodology, focusing on the elements that matter for our analysis: the construction of the underlying scores, the industry benchmarking step, and the role of committee review.

3.1 Institutional background on MSCI ESG Ratings

MSCI ESG Ratings are designed to help investors identify financially material environmental, social, and governance risks and opportunities at the company level. Ratings are expressed on a seven-point letter scale, from CCC (worst-in-class) to AAA (best-in-class), and are defined relative to a firm's industry peers rather than in absolute terms.[5] Within each sector, firms that receive the highest letter ratings are intended to be those judged most effective at managing their exposure to material ESG risks and opportunities.

The raw inputs to this process are exclusively public sources, including financial and sustainability reports, specialized government and academic datasets, media and NGO reports,

[4] See "MSCI, the largest ESG rating company [...]", Bloomberg News.
[5] See MSCI (2022) for the methodological guide and MSCI (2024) for the materiality mapping framework.

and other third-party information. The core building blocks are 35 "Key Issues" that capture specific environmental, social, or governance risks. For each of the 158 Global Industry Classification Standard (GICS) sub-industries (S&P, 2023), MSCI selects a subset of two to seven Environmental and Social Key Issues that are deemed financially material, while Governance is assessed using a common set of six issues applied to all sectors, consistent with the rater's view that *"governance is universally important and should be evaluated in an integrated way, regardless of industry"* (MSCI, 2022).

For a given firm, each relevant Key Issue is scored based on its *Risk Exposure* and *Risk Management*. These issue-level scores are then combined using analyst-assigned weights to obtain several aggregate metrics. The most important for our purposes is the *Weighted Average Score* (WAS), a 0–10 index that averages (with weights) all industry-specific Environmental and Social Key Issues and the Governance Pillar Score. MSCI also calculates Environmental and Social Pillar Scores and a set of thematic scores, again on a 0–10 scale, but these are not used directly to determine the letter rating.

The methodology allows for deviations from the standard set of Key Issues. If a firm is exposed to a material risk that is not common in its industry, for example because of a unique business model, analysts may add a company-specific Key Issue and adjust the weights on existing issues. Conversely, if a firm is not exposed to a risk that is otherwise considered material for its peers, that issue may be dropped and the remaining weights rescaled (MSCI, 2022). Such company-specific changes require approval by an *ESG Methodology Committee*, which might therefore introduce a layer of discretion in the final ratings.

Finally, before being mapped into letter ratings, the WAS is transformed into an *Industry Adjusted Score* (IAS) through a benchmarking procedure. Conceptually, this step simply rescales each firm's WAS using the cross-sectional distribution of WAS among its industry peers. At the beginning of each year, MSCI identifies an industry-specific minimum score m and maximum score M based on the distribution of WAS across all rated firms in that sector. Typically, m is chosen between the 0th and 5th percentiles and M between the 95th and 100th percentiles, though the percentile ranges can be widened to the 10th and 90th

percentiles if needed to stabilize ratings over time.

Given these bounds, the IAS for firm i is defined as

$$\text{IAS}_i = 10 \times \frac{\text{WAS}_i - m}{M - m}, \tag{11}$$

with scores truncated to lie between 0 and 10 and rounded to one decimal place. Values at or below m receive an IAS of 0, while those at or above M receive an IAS of 10. The IAS is then translated into a letter rating using seven equally sized intervals on the $0-10$ scale (roughly, 0-1.43 corresponds to CCC, and so on until the AAA range being $8.67-10$). In industries with a very narrow distribution of the WAS, these ranges can be truncated so that, for example, no firms receive CCC or AAA ratings. Specifically, if the bottom benchmark value exceeds 4, the industry minimum is capped at this number, while the industry maximum is truncated at 6 in cases when M is lower(MSCI, 2023).

Before releasing the updated scores, following the benchmarking procedure, MSCI relies on committee review to vet and, if necessary, override model-implied outcomes. According to the methodology hand book, a committee review is triggered when (i) upgrades to AAA or downgrades to CCC are proposed; (ii) changes of two or more notches are suggested; (iii) the analyst believes the model-determined rating does not adequately reflect the firm's operational context or ESG performance; or (iv) there is substantial uncertainty about data quality or coverage. In such cases, the committee may approve, modify, or reject the proposed change, including by directly adjusting the IAS. Information about how frequently these committees intervene, and on what basis, is not available. Subscribers can observe the date of each reassessment but not whether a committee review took place or how the final score was reached.

3.2 Descriptive evidence of the traffic light effect

We now present descriptive evidence that connects these institutional features to the traffic light effect mentioned earlier. The key idea is that, if committee review is used to prevent

ratings from crossing certain thresholds (from either sides), we should observe an excess mass of firms in the thresholds' neighborhoods. Figure 1 shows that, rather than a smooth density, the distribution of IAS over time has sharp spikes just below and above the cutoffs between letter ratings (in dashed lines). In particular, firms tend to accumulate right below the threshold for higher letter ratings or just above the cutoffs for lower ones. This bunching, absent in the un-adjusted scores (see Appendix Figure 10), is difficult to reconcile with a purely benchmarking procedure, which would tend to spread firms relatively evenly within each interval. Instead, the pattern is suggestive of an additional force, the committee, limiting the extent to which scores cross rating boundaries.

In Figure 2, we show changes in IAS between two consecutive years. If rating changes followed only from the model-based updating of underlying Key Issues, we would expect a relatively diffuse cloud around the 45-degree line, with some firms upgrading and others downgrading. What we observe instead is a striking concentration of changes within bands with thresholds between letter ratings appearing to constraint firms from crossing boundaries. Upgrades to AAA and downgrades to CCC are particularly scarce, with scores often capped just below the AAA threshold or just above the CCC threshold. Overall, only 24.3% of the 129,702 firm-year observations that undergo a full reassessment experience any letter rating change; among these, 20,588 are upgrades and 10,944 downgrades, indicating a pronounced upward skew.

These dynamic patterns echo the cross-sectional bunching in Figure 1. They indicate that rating mobility is strongly constrained: two-notch moves are rare, upgrades to best-in-class and downgrades to worst-in-class are exceptional events, and a non-negligible share of firms appears to hit the ceiling or floor just before these extreme ratings. Coupled with the absence of sharp kinks in the underlying WAS distribution, these stylized facts illustrate the active intervention of the committees.

3.3 Benchmarking recovery

Having laid out the institutional background and descriptive evidence for the traffic light effect, we now describe the empirical strategy that allows us to reverse–engineer MSCI's benchmarking step and identify where committee intervention is most likely to occur.

As detailed in Section 3.1, the benchmarking step and subsequent committee review, which can results in scores being overwritten, are plausible candidates to explain the traffic light effect. Thus, we reverse-engineer this process by recovering the benchmarking weights, m and M, through a non-linear optimization problem. Note that a simple OLS regression of IAS on WAS would be ill-suited since, as shown in Figure 3, the score adjustments we wish to detect introduce systematic deviations from the methodology-implied linear relationship. Our objective is therefore not to minimize the average distance between observed scores and a fitted line, but to maximize the number of firms whose Industry Adjusted Score exactly coincides with what the benchmarking formula would predict absent any intervention. In doing so, we assume that any score adjustment is an exception rather than a rule.

Let $m_{i,t}$ denote the bottom benchmark value for the WAS in industry i and year t, and $M_{i,t}$ the corresponding top benchmark defined in Section 3.1. Given the pair $(\text{WAS}_{j,i,t}, \text{IAS}_{j,i,t})$ for each firm j, we search over $(m_{i,t}, M_{i,t})$ to maximize the number of exact matches between the IAS implied by Equation (11) and the IAS reported by MSCI. Formally, we solve

$$\max_{m_{i,t}, M_{i,t}} \sum_{j \in J} \mathbb{1} \left(\lfloor \text{IAS}_{j,i,t} \rceil_{0.1} = \begin{cases} 0 & \text{if } \text{WAS}_{i,j,t} < m_{i,t} \\ 10 & \text{if } \text{WAS}_{i,j,t} > M_{i,t} \\ \left\lfloor 10 \cdot \dfrac{\text{WAS}_{i,j,t} - m_{i,t}}{M_{i,t} - m_{i,t}} \right\rceil_{0.1} & \text{otherwise} \end{cases} \right)$$

subject to

$$0 \leq m_{i,t} \leq 4 \quad \forall\, i \in I,\ t \in T, \qquad 6 \leq M_{i,t} \leq 10 \quad \forall\, i \in I,\ t \in T,$$

where $\lfloor \cdot \rceil_{0.1}$ denotes rounding to one decimal place.

The problem's constraints mirror the bounds imposed by MSCI's methodology, as discussed in Section 3.1. Firms with WAS below $m_{i,t}$ or above $M_{i,t}$ are assigned an IAS of 0 or 10, respectively; firms in between receive the linearly interpolated score given by Equation (11), rounded to one decimal. The piecewise function in the objective implements exactly this rule, while the indicator function counts only the firms for which the methodology-implied IAS equals the firm's actual IAS. The sets I, J and T denote industries, firms and years (2014–2022), respectively. Because the rounding operator renders the problem non-linear and non-differentiable, standard gradient-based methods are inappropriate. We therefore rely on a dual annealing algorithm (Tsallis and Stariolo, 1996), which efficiently explores the parameter space and is well-suited for global optimization with discontinuities.

Solving this problem yields, for each industry–year cell, the pair $(m_{i,t}, M_{i,t})$ that best rationalizes the observed IAS under the published benchmarking rule. We refer to the score implied by these optimized weights as the *model-implied* or *implied* IAS, denoted $\widehat{\text{IAS}}$, and we continue to use "*actual*" when referring to the scores released by MSCI after any committee review. Whenever implied and actual scores differ, we interpret the discrepancy as the outcome of committee intervention. Importantly, these gaps are often large enough to alter the letter rating, either by preventing an upgrade or a downgrade. We refer to such cases as *extensive adjustments*; score changes that leave the letter grade unchanged are instead called *intensive adjustments*.

Figure 3 plots the IAS against WAS for a representative industry and year. In the absence of any discretionary adjustments, all firms should lie on a straight line between the industry minimum and maximum WAS. In practice, we observe systematic deviations: some firms with higher WAS might receive lower IAS than peers, or vice versa, and firms with identical WAS occasionally receive different IAS. Figure 4 summarizes how frequently these adjustments occur across the IAS distribution. For each score on the 0–10 scale, the figure plots the share of firms that achieved that score due to a modification. We observe that discrepancies between implied and actual scores are most pronounced precisely where a one-tenth change in IAS would trigger a letter rating change. For leaders, this typically means lowering the implied score to avoid an upgrade; for laggards, it means raising the implied score to avoid a

downgrade. Consistent with this interpretation, adjustment frequencies are relatively modest across most of the distribution, but spike sharply in narrow neighborhoods around rating cutoffs, where extensive adjustment rates rise by several multiples relative to adjacent IAS.

These rating discontinuities matter since MSCI's commercial ESG indexes are built directly on the same ratings. Best-in-class indexes such as MSCI ESG Leaders and MSCI SRI explicitly construct portfolios by selecting, within each sector and region, the top 50% or 25% of firms by ESG rating, while other indexes tilt or optimize benchmark weights as a function of the same ratings and their changes. A one-notch move at these thresholds can therefore entail a discrete change in index eligibility or in portfolio weight, even when the underlying $\widehat{\text{IAS}}$ moves by only 0.1 points. Moreover, because selection into the top 50% or 25% is defined within sector–subregion cells, even *intensive* adjustments that leave the letter rating unchanged can, at least in principle, move a firm just inside or outside the relevant percentile cutoff and swap one index constituent for another; at this stage, however, this remains a speculative interpretation of the intensive adjustments.

Viewed through this lens, the pattern in Figure 4 might suggest that committee intervention internalizes the consequences of rating changes for MSCI's own ESG index family. Upgrades to AA or AAA expand the pool of "ESG leaders" that populate the ESG Leaders and SRI indexes, while downgrades to B or CCC increase the mass of "ESG laggards" that are down-weighted or excluded from those same products. Since MSCI markets these indexes as simultaneously improving ESG quality and delivering attractive, stable risk-return characteristics, large and frequent shifts in the size of the leader and laggard buckets would translate into higher index turnover and larger tracking error versus the parent benchmark. MSCI itself emphasizes tracking-error constraints in its index design precisely to reassure benchmark-oriented investors MSCI (2019b).

The optimization procedure carries two important insights. First, it provides a transparent way to identify firms whose scores must have been modified during committee review, even though no information on committee decisions is disclosed to subscribers. Second, these findings warn against the use of IAS as a running variable in Regression Discontinuity Designs

(RDD). Indeed, the core identifying assumption in RDD is that the running variable behaves as a quasi-random "assignment variable" around the cutoff (Thistlewaite and Campbell, 1960). Our reconstructed scores show that the model-implied $\widehat{\text{IAS}}$ might differ from the actual IAS, especially near thresholds between letter ratings. Since the committee can (and does) precisely nudge scores away from the cutoff to avoid treatment (for instance, an upgrade to best-in-class or a downgrade into the laggard bucket), the non-manipulation assumption of RDD is violated: the realized IAS is an endogenous variable rather than the mechanical output of the model (McCrary, 2008).

3.4 Data

We source firm-level data on MSCI ESG Ratings from 2014 to 2022 at a monthly frequency. The starting year is dictated by a major revision of MSCI's industry benchmarking process in 2014, when the benchmark universe was shifted from the MSCI World Index to the broader MSCI ACWI Index in order to better reflect the global investment landscape and incorporate more large-cap emerging-market firms in the peer set (MSCI, 2019a). Because much of our analysis focuses on the benchmarking step, we restrict attention to the post-2014 regime to avoid mixing observations generated under different benchmark definitions. The impact of this change is visible in Figure 1: the 2013 distribution of Industry Adjusted Scores (IAS) exhibits more mass around the B/BB threshold than in later years, underscoring the lack of comparability with the subsequent period.

MSCI updates company data each month, but, as discussed in Section 3.1, a full reassessment of the IAS and the associated letter rating typically occurs only about once per year. In months without a full review, Key Issue scores may move while the IAS and rating are left unchanged. Since our empirical strategy focuses on comparing methodology-implied IAS with the actual score assigned at the time of a rating update, we drop months in which no reassessment takes place. Including those months would mechanically generate spurious gaps between implied and actual scores simply because the underlying Weighted Average Score (WAS) changed but the IAS was not updated. After this filtering, we obtain an unbalanced

panel of firm-year observations: some firms start being rated after 2014, others cease to be rated due to delistings or coverage changes, and a small subset of companies is not reassessed in a given year because of the exceptions described in Section 3.1. The number of rated firms nonetheless expands markedly over time, from 11,271 in 2014 to 16,025 in 2022, mirroring the growth of MSCI's coverage universe.

To control for firm fundamentals, we merge the MSCI data with financial information from Compustat Capital IQ. Specifically, we control for the following annual variables: *Size* is the logarithm of total assets; *CAPEX/Assets*, defined as the ratio of capital expenditures to total assets; *Cash/Assets*, constructed as cash plus short-term investments over total assets; *Debt/Assets* is total debt (current plus long-term) over total assets; *EBIT/Assets*, defined as earnings before interest and taxes over total assets; *PP&E/Assets* being property, plant and equipment over total assets; and *Sales growth* given by the growth rate of sales between two consecutive fiscal years. To mitigate any outliers, all financial variables are winsorized at the 1% and 99% levels. Because Compustat coverage is narrower than MSCI's, merging the two sources reduces the sample size by roughly 60%; the resulting merged dataset contains 25,111 firm-year observations.

Table 1 provides descriptive statistics for the main variables in this merged sample. Financial characteristics are comparable to those reported in earlier work on large-firm samples (e.g. Opler et al., 1999; Lewellen, 2015), while the distribution of IAS, WAS and pillar scores matches previous analyses of MSCI ESG data (Berg et al., 2022). This reassures that the filtered panel is representative of the broader universe covered by MSCI.

4 Results

4.1 Stability bias and peer alignment in MSCI's ESG Ratings

In Section 2 we argued that MSCI's adjustment process may reflect two distinct but co-existing forces. On the one hand, the committe may display a *stability bias*, i.e. a tendency

to preserve the previous year rating bucket unless there is a strong reason to let the score move across a threshold. On the other hand, they may also respond to *peer alignment* concerns, adjusting the published score when the methodology-implied value diverges substantially from the consensus of other major ESG providers. We now test these two channels jointly.

We begin by assembling firm-level ESG scores from the main commercial providers that cover a broad cross-section of publicly listed firms. Specifically, we use data from Refinitiv, Sustainalytics/Morningstar, S&P Global, TrueValue, RepRisk, ISS, OWL ESG, and MSCI.[6] Since these raters employ heterogeneous scales and aggregation conventions, we first map all raw scores to a common $[0, 10]$ interval using a monotone linear transformation. This preserves within-provider rankings while making cross-rater dispersion directly comparable.

Sustainalytics' scores are defined on a risk scale on which lower values correspond to better ESG performance. To align direction across all sources, we first map Sustainalytics scores to the common $[0, 10]$ interval and then invert the scale by defining $\widetilde{y}_{it}^{\text{Sus}} = 10 - y_{it}^{\text{Sus}}$, so that higher values of $\widetilde{y}_{it}^{\text{Sus}}$ indicate better ESG performance, as for the other providers. All remaining raters' scores are simply rescaled to the 0–10 interval with higher values corresponding to better ESG performance.

Let y_{it}^{k} denote the harmonized ESG score of firm i in year t from provider k. For each firm-year we define the peer average score as the simple mean across all other raters:

$$y_{it}^{C} = \frac{1}{J_{it} - 1} \sum_{k \neq \text{MSCI}} y_{it}^{k},$$

where J_{it} is the number of raters covering firm i in year t. This y_{it}^{C} is the empirical analogue of the peer consensus in Section 2. Because coverage differs across firms and years, we restrict the sample to firm-years rated by at least two providers other than MSCI. To measure the

[6] Decision-makers might have different channels to access third-party ratings, including (i) public issuer-lookup pages that display the current rating—for example, Refinitiv/LSEG ESG scores, Sustainalytics via Morningstar company pages, and S&P Global ESG scores—and/or (ii) issuer "badges" or labels that rated companies may feature on their website—for example, MSCI ESG Ratings badges, Sustainalytics Top-Rated badges, S&P Global Sustainability Yearbook emblems, and ISS ESG Prime badge. For the remaining providers in our dataset, ESG assessments may still be observable through other channels (e.g., issuer sustainability communications) or via licensed access to the underlying data products.

strength of peer agreement, we compute the mean absolute dispersion (MAD) of peer scores around their average,

$$\mathrm{MAD}_{it} = \frac{1}{J_{it} - 1} \sum_{k \neq \mathrm{MSCI}} \left| y_{it}^{k} - y_{it}^{C} \right|.$$

Small values of MAD_{it} indicate a tight peer consensus, while large values reflect disagreement.

For MSCI, we work with three firm-year objects. First, we reconstruct the methodology-implied Industry Adjusted Score $\widehat{\mathrm{IAS}}_{it}$ using the reverse-engineering procedure described in Section 3.3. Second, we observe the actual adjusted score IAS_{it} released to subscribers. Third, we define the realized wedge between the published score and the methodology-implied score as

$$\mathrm{Wedge}_{it} = \mathrm{IAS}_{it} - \widehat{\mathrm{IAS}}_{it}.$$

A positive wedge corresponds to an upward adjustment relative to the model, while a negative wedge corresponds to a downward adjustment. To capture peer misalignment, we define the divergence between MSCI's model and the peer consensus as

$$\Delta_{it} = \widehat{\mathrm{IAS}}_{it} - y_{it}^{C}.$$

Thus, $\Delta_{it} > 0$ means that the methodology-implied score is more optimistic than the peer average, whereas $\Delta_{it} < 0$ means it is more conservative. We will estimate both absolute-wedge specification using $|\Delta_{it}|$ to measure the magnitude of disagreement irrespective of direction, and signed specifications that retain the sign of Δ_{it} in order to distinguish cases in which the model lies above or below the peer consensus.

To isolate the stability channel, we construct a measure of the extent of adjustment needed to keep the firm's ESG score in last year's actual MSCI letter rating: the *retention gap*. Let last year's actual letter rating determine an interval of continuous IAS values. The variable *retention gap* is then the minimum absolute adjustment required to move the methodology-implied score back into that interval. By construction, it equals 0 whenever the methodology-implied score already lies inside last year's bucket, and it becomes larger the further the model would otherwise move the firm across a rating threshold. For the signed regressions we use a

directional version, whose sign indicates whether preserving last year's bucket would require an upward or downward correction.

We also decompose the wedge into one-sided outcomes. Let

$$\text{Wedge}_{it}^{+} = \text{Wedge}_{it} \cdot \mathbf{1}\{\text{IAS}_{it} \geq \widehat{\text{IAS}}_{it}\}, \qquad \text{Wedge}_{it}^{-} = \text{Wedge}_{it} \cdot \mathbf{1}\{\widehat{\text{IAS}}_{it} \geq \text{IAS}_{it}\}.$$

Hence, $\text{Wedge}_{it}^{+} \geq 0$ captures upward adjustments relative to the model, while $\text{Wedge}_{it}^{-} \leq 0$ captures downward adjustments. This decomposition allows us to study whether committee intervention actually moves scores in the direction of peer alignment and stability.

We proceed in three steps. First, we estimate a *stability* specification. Second, we estimate a separate specification relevant to the peer alignment channel. Finally, we estimate the *joint* specification that includes both channels:

$$|\text{Wedge}_{it}| = \alpha_1 \text{RetainGap}_{it} + \mathbf{X}_{it}'\delta + \gamma_i + \lambda_{ct} + \varepsilon_{it}. \tag{12}$$

$$|\text{Wedge}_{it}| = \beta_1 |\Delta_{it}| + \beta_2 \text{MAD}_{it} + \beta_3 \big(|\Delta_{it}| \times \text{MAD}_{it}\big) + \mathbf{X}_{it}'\eta + \gamma_i + \lambda_{ct} + \nu_{it}. \tag{13}$$

$$|\text{Wedge}_{it}| = \theta_1 \text{RetainGap}_{it} + \theta_2 |\Delta_{it}| + \theta_3 \text{MAD}_{it} + \theta_4 \big(|\Delta_{it}| \times \text{MAD}_{it}\big) + \mathbf{X}_{it}'\rho + \gamma_i + \lambda_{ct} + \xi_{it}, \tag{14}$$

In all regressions, $\mathbf{X}_{it}$ contains the firm-level controls introduced in Section 3.4, γ_i are firm fixed effects, and λ_{ct} are country-year fixed effects. Standard errors are clustered at the firm level. To make coefficients comparable across specifications, within each outcome family we impose a common support.

Table 2 reports the results. The first finding is that the stability channel is strong and remarkably robust. In the absolute-wedge regression, the coefficient on the retention gap is positive and highly significant, implying that the larger the adjustment needed to keep the firm in last year's rating bucket, the larger the wedge actually introduced by MSCI, suggesting that the committee actively leans against score changes that would otherwise move a firm across a letter-rating threshold. The same logic is visible in the one-sided regressions.

For upward adjustments, the coefficient is positive and significant, indicating that when preserving last year's bucket requires an upward correction, the realized wedge becomes more positive. For downward adjustments, the corresponding coefficient is also significant and consistent with the committee moving the score in the direction needed to preserve the previous bucket *(more negative gap, more negative wedge, thus positive association)*. Taken together, these estimates point to a clear stability bias in the committee review process.

The second finding is that peer alignment matters as well. In the absolute-wedge regressions, the coefficient on $|\Delta_{it}|$ is positive and significant, meaning that larger deviations between the methodology-implied score and the peer average are associated with larger wedges. Since $\Delta_{it} = \widehat{\text{IAS}}_{it} - y_{it}^{C}$, this implies that the committee intervenes more when MSCI's model stands further away from the consensus of other providers. At the same time, the interaction between $|\Delta_{it}|$ and MAD_{it} is negative, which means that this effect is attenuated when peers disagree more and strengthened when peer consensus is tight. In other words, distance from peers is most consequential precisely when the peer benchmark is more informative.

Third, the joint regressions show that the two mechanisms coexist rather than substitute for each other. Once we include both channels simultaneously, the retain-gap coefficient remains large and precisely estimated, while the divergence terms continue to carry explanatory power. The empirical picture that emerges is therefore consistent with an adjustment process shaped jointly by rating stability and external benchmarking.

We next zoom in on the subset of cases in which these two forces point in opposite directions, in order to study how decisively the committee implements a rating change when peer alignment prevails over stability. Table 3 shows regression results focusing on the subset of firm-years in which the stability channel and the peer-alignment channel point in opposite directions and the committee nevertheless implements a rating change. The estimates show that when the realized adjustment follows the peer-alignment direction, the score is placed deeper inside the new rating bucket: the post-cross buffer is larger by about 0.13 points in the baseline specification and by about 0.15 points once controls are included, while the probability of a "bare cross", a rating change leaving the score within 0.20 points of the

crossed threshold, falls by roughly 11 to 14 percentage points. This pattern matches Proposition 4: if the committee is willing to override stability and accept a letter-rating change, it tends not to do so in a knife-edge way as a threshold crossing that leaves the score just barely inside the new bucket would still look fragile and could be perceived by stability-sensitive investors as an avoidable disruption. By contrast, moving the score further into the new bucket makes the change appear more deliberate and less likely to be quickly reversed.

Finally, to complement the regression evidence above, we provide three visual diagnostics that focus more closely on the peer-alignment channel. This additional attention is warranted because, while the existence of a stability bias is intuitive in a rating setting, the fact that MSCI's realized adjustments appear to respond to model-implied position relative to a cross-rater benchmark is far less obvious and therefore deserves more scrutiny.

Figure 5 offers a first non-parametric view of this mechanism. It displays the average absolute wedge as a function of two objects: the magnitude of the divergence between the methodology-implied score and the peer consensus, $|\Delta_{it}|$, on the horizontal axis, and the degree of peer disagreement, MAD_{it}, on the vertical axis. Two patterns stand out clearly. First, for any given level of peer dispersion, the average wedge rises as the methodology-implied score moves further away from the peer average. Second, for any given level of divergence, the wedge tends to be larger when peer dispersion is low. Thus, the visual pattern lines up closely with the estimates in Table 2: committee adjustments are not only larger when MSCI's model disagrees more with peers, but they are especially pronounced when that disagreement occurs against a stronger peer consensus.

Figure 6 complements this analysis with a simpler binned representation. The figure plots the average absolute wedge $|\text{Wedge}_{it}|$ against quantiles of $|\Delta_{it}|$, splitting the sample into two groups based on the median of MAD_{it} and displaying separate lines for low-dispersion and high-dispersion firm-years. Within each dispersion group, the average absolute wedge is increasing in the divergence quantile: firms in the upper deciles of $|\Delta_{it}|$ experience substantially larger adjustments than firms in the lower deciles. Moreover, for every divergence quantile, the wedge is systematically larger in the low-dispersion subsample than in the high-

dispersion one. We acknowledge that, visually, the relationship between $|\text{Wedge}_{it}|$ and $|\Delta_{it}|$ also appears mildly convex, especially when peer disagreement is low, suggesting that the marginal effect of peer misalignment becomes stronger at higher levels of disagreement. For ease of interpretation, we retain a linear baseline specification in the main regressions, but the figures indicate that the underlying relationship may steepen in the tails.

Finally, Figure 7 speaks to the economic significance of these effects. At first glance, the regression coefficients in Table 2 may appear small relative to the 0–10 score range. Yet, what matters for market-relevant outcomes is not the absolute size of the continuous wedge per se, but whether it is large enough to push a firm across one of MSCI's discrete rating thresholds. Because the mapping from the continuous Industry Adjusted Score to letter ratings relies on sharp cutoffs, even modest numerical adjustments can be sufficient to prevent an upgrade or a downgrade. To illustrate this point, we conduct a simple counterfactual exercise in which the methodology-implied score is assumed never to diverge from the peer consensus. Operationally, we set $|\Delta_{it}| = 0$ for all firm-years, keep all other covariates at their observed values, and use the relevant specification from Table 2 to compute counterfactual wedges. We then compare the resulting distribution with the observed one. The figure shows that, under this no-divergence scenario, predicted wedges generate very few threshold-crossing adjustments. By contrast, the actual data exhibit substantial mass exactly where wedges are large enough to block rating changes. This comparison makes clear that the peer-alignment channel is economically meaningful: numerically modest deviations from the methodology-implied score can nevertheless be large enough to systematically prevent upgrades and downgrades.

4.2 Evidence from a Refinitiv methodology change

The previous subsection shows that MSCI's committee adjustments co-move with the divergence between MSCI's methodology–implied score and the cross–rater consensus. While this pattern is highly suggestive of the investor–pay alignment mechanism described in Section 2, one natural concern is that our estimates might simply capture common exposure to funda-

mentals that happen to be measured similarly by MSCI committee and its peers. In other words, MSCI and other raters could react in parallel to the same underlying information, without MSCI explicitly conditioning its overrides on peer assessments.

In this subsection we sharpen the identification by exploiting a major revision of the Refinitiv ESG methodology that became effective in April 2020. As documented by Berg et al. (2021), Refinitiv replaced its previous aggregation scheme with a new proprietary "magnitude matrix" that assigns industry–specific weights to ESG indicators and, at the same time, changed the treatment of non–reported qualitative items from a neutral value of 0.5 to a score of zero. Crucially for our purposes, the new methodology was not only applied going forward but also retroactively to all historical fiscal years in the database. The scores that users can download today are therefore a rewritten under the post–2020 methodology. Berg et al. (2021) show that this rewriting generates sizable changes in firms' historical ESG levels and rankings.

This retroactive harmonization implies that the ESG assessments actually available to MSCI in, say, 2014–2018 are no longer observable in current Refinitiv databases. In our setting, however, this is a useful feature rather than a drawback. We have access to both (i) a historical snapshot of the pre–2020 Thomson Reuters / Refinitiv ESG scores that were available to data users at the time, and (ii) the currently downloadable Refinitiv ESG Combined Scores, which apply the post–2020 methodology to the entire history. For the pre–change period we can therefore contrast how MSCI reacts to divergence from the *original* scores with how it would have reacted to divergence from the *rewritten* series that did not exist when MSCI made its rating decisions.

To isolate the role of this single peer, we focus on the divergence between MSCI and Thomson Reuters / Refinitiv. As in Section 4, let $\widehat{\text{IAS}}_{it}$ denote the methodology–implied MSCI score for firm i in year t. We construct two measures of absolute disagreement,

$$\Delta_{it}^{TR} = \left|\widehat{\text{IAS}}_{it} - y_{it}^{TR}\right|, \qquad \Delta_{it}^{Ref} = \left|\widehat{\text{IAS}}_{it} - y_{it}^{Ref}\right|,$$

where y_{it}^{TR} is the legacy Thomson Reuters ESG rating from the pre–2020 snapshot and y_{it}^{Ref}

is the Refinitiv ESG Combined Score from the current (post–2020) database. Both series are linearly rescaled to the common $[0, 10]$ scale to ensure comparability with MSCI scores.

We obtain year–specific effects by interacting each divergence measure with year dummies in a single panel regression per provider $k \in \{\text{TR}, \text{Ref}\}$. We estimate

$$\text{Wedge}_{it} = \sum_{\tau} \beta_{\tau}^{k} \left(\Delta_{it}^{k} \times \mathbf{1}\{\text{Year} = \tau\} \right) + X_{it}' \delta^{k} + \gamma_i + \varepsilon_{it}^{k},$$

where Wedge_{it} is the realized wedge between MSCI's published score and its model–implied benchmark, X_{it} is the vector of firm controls introduced in Section 3.4, and γ_i are firm fixed effects. The coefficients β_{τ}^{k} are therefore the year–specific slopes linking MSCI's wedge to divergence from provider k in year τ.

Figure 8 plots these coefficients together with 95% confidence intervals. The green markers correspond to the legacy Thomson Reuters ratings ($k = \text{TR}$), while the blue markers correspond to the rewritten Refinitiv scores ($k = \text{Ref}$). Two patterns emerge clearly.

First, in the years before the 2020 methodology change, the coefficients on divergence from the legacy Thomson Reuters ratings, β_{τ}^{TR}, are positive and statistically different from zero, whereas the corresponding coefficients for the rewritten Refinitiv series, β_{τ}^{Ref}, are statistically indistinguishable from zero. During this period the MSCI wedge widens systematically when MSCI's own model deviates more from the *then available* Thomson Reuters ESG assessment, but not when it deviates from the *future* Refinitiv series that would only be constructed after the 2020 rewriting. This is exactly what one would expect if MSCI bases its overrides on contemporaneous Thomson Reuters / Refinitiv data.

Second, the years around the 2020 change behave as a natural transition. The methodology revision and retroactive backfilling were implemented in 2020, and starting with that year the coefficients on divergence from the Refinitiv scores, β_{τ}^{Ref}, turn positive and statistically significant.

This event–study style evidence strengthens the herding interpretation of Section 4. The

retroactively recomputed Refinitiv history, though highly correlated with the original series by construction, has no explanatory power before it exists in the information set. This pattern is hard to reconcile with a story in which the wedge simply reflects common exposure to slowly moving ESG fundamentals, since those fundamentals are embedded in both the legacy and the rewritten series. Instead, it points to an information channel: MSCI appears to condition its overrides on the signals provided by peers when those signals are actually available.

Lastly, the methodology change and subsequent backfilling create an implicit placebo test. If one were to ignore the existence of the legacy Thomson Reuters ratings and work only with today's Refinitiv database, one would (incorrectly) conclude that divergence from Refinitiv had little or no predictive power for MSCI's adjustments before 2020. Our dual–snapshot evidence shows that this is an artefact of the retroactive harmonization of the Refinitiv series, and that the relevant benchmark for MSCI's behavior in the pre–change period is the original Thomson Reuters rating. Overall, the Refinitiv episode reinforces the finding that MSCI's traffic–light adjustments are responsive to contemporaneous evaluations by major peer providers.

4.3 Financial Impact

We now examine the financial consequences of the traffic light effect by analyzing stock returns following downgrade signals. Specifically, we compare the cumulative abnormal returns associated with downgrades that are *implemented* in MSCI's published rating to those that are *suppressed*, i.e. those that are implied by MSCI methodology but not reflected in the published letter grade. To do so, we estimate horizon-specific regressions of cumulative abnormal returns on event indicators for implemented and suppressed downgrades. For each horizon $H \in \{0, 1, \ldots, 12\}$ months, we run:

$$\text{Return}_{H,i,t} = \beta_H^{\text{imp}} \cdot D_i^{\text{imp}} + \beta_H^{\text{sup}} \cdot D_i^{\text{sup}} + \text{Overlap}_H + X_{i,t-1}'\gamma_H + \mu_i + \theta_t + \varepsilon_{H,i,t}, \qquad (15)$$

where $\text{Return}_{H,i,t}$ is either the cumulative abnormal return (CAR) or abnormal buy-and-hold return (ABHR) from month t through $t+H$, D_i^{imp} equals one if firm i receives an implemented downgrade, and D_i^{sup} equals one if the model-implied rating suggests a downgrade but the published rating remains unchanged. Events are dated to MSCI's ESG index quarterly review dates (February 15, May 15, August 15, November 15).

The vector $X_{i,t-1}$ includes lagged firm characteristics: size, leverage (debt-to-assets), book-to-market, and profitability (income before extraordinary items scaled by total assets), all obtained from Compustat and winsorized at the 1st and 99th percentiles within each quarter. Firm fixed effects μ_i absorb time-invariant heterogeneity, and month fixed effects θ_t control for aggregate shocks. Standard errors are two-way clustered by firm and month. We further include six months of leads and lags of the implemented downgrade and upgrade indicators, following the approach of Jordà (2005) in local projection methods.

The coefficient β_H^{imp} captures the H-month cumulative abnormal return following an implemented downgrade, while β_H^{sup} captures the corresponding return following a suppressed downgrade. The key parameter of interest is the difference $\beta_H^{\text{imp}} - \beta_H^{\text{sup}}$, which measures whether implemented downgrades are associated with systematically different return dynamics than suppressed ones, conditional on observables.

Table 5 reports the estimated coefficients for implemented downgrades (β_H^{imp}), suppressed downgrades (β_H^{sup}), and their difference, using both CAR and ABHR as dependent variables. Figure 9 plots the cumulative return profiles and their difference over the 12-month horizon.

Stocks receiving implemented downgrades experience economically meaningful and persistent underperformance over the year following the rating change. In the CAR specifications, abnormal returns fall immediately (about -1.3% at month 1), deepen to roughly -2.1% by month 2, and remain negative thereafte. By month 12 the estimate is around -2.2% and is marginally significant at the 10% level. The ABHR results are similar, with a 12-month estimate of about -2.8% ($t = 1.97$). By contrast, suppressed downgrades exhibit no statistically detectable return response at any horizon: coefficients are small (often slightly positive) and never significant. Consistent with this pattern, the gap $\widehat{\beta}_H^{\text{imp}} - \widehat{\beta}_H^{\text{sup}}$ is most pronounced in

the first quarter, on the order of −3 to −4 percentage points in months 1–3, but becomes less precisely estimated at longer horizons. At month 12 the point estimate still implies an economically sizable gap of about −4%, but the confidence interval is wide and statistical significance is not retained. This is also due to power issues arising from the imbalance in event frequencies: implemented downgrades are roughly four times more common than suppressed ones, yielding tighter standard errors for the implemented coefficient than for the suppressed coefficient.

This null result is consistent with the hypothesis that suppression prevents the transmission of negative information to the market, either because investors do not observe the model-implied signal or because the published rating is what drives index flows and institutional mandates. Also, the first three months provide clear evidence that implementation matters. The fact that negative abnormal returns emerge immediately and significantly only when MSCI publicly implements the downgrade suggests that the decision to suppress effectively insulates firms from the market consequences that would otherwise follow.

The cumulative return gap of 3 to 4 percentage points in the first quarter following the event is economically large. For a firm with $10 billion in market capitalization, this corresponds to roughly $300–$400 million in avoided market-value loss. Over the full 12-month horizon, the point estimate of the difference reaches approximately −4.3% for ABHR, though this estimate is imprecise. These magnitudes are comparable to those documented in the broader literature on credit rating changes (Dichev and Piotroski, 2001; Jorion and Zhang, 2007) and suggest that ESG rating agencies, like traditional credit rating agencies, exert meaningful influence over asset prices (when rating changes are actually published).

Finally, the evidence for upgrades is more nuanced. We do not find a persistent response following implemented or suppressed upgrades and the estimates even show some short-run negative abnormal returns in the first months after the event. This suggests that the market does not treat upgrades as the mirror image of downgrades. One possible interpretation is that upward ESG rating changes may sometimes be perceived as less informative, more anticipated, or partially offset by contemporaneous concerns that are not fully captured by

the rating itself. More broadly, the contrast with downgrades is consistent with the idea that negative ESG news is more salient for investors and more likely to trigger binding portfolio constraints, whereas upgrades do not generate an equally mechanical inclusion effect.

4.4 Robustness analysis

We conclude with a series of robustness checks reported in Table 4. For all the additional specifications, the estimated relationship between the wedge and divergence remains qualitatively unchanged. First, we replace the mean wedge by its median (column (1)). The divergence coefficient stays positive and statistically significant, suggesting that the disagreement-adjustments association is not driven by a few very extreme wedges.

We then vary the fixed-effects structure and remove controls. Employing Industry $\times$ Year fixed effects rather than the Country $\times$ Year ones absorbs common sectoral shocks, such as changes in materiality maps or industry-wide news. As we observe in column (2), the coefficient on divergence is virtually unchanged and still statistically significant. Furthermore, omitting all firm-level controls and simply regressing the wedge on divergence and fixed effects, as in column (3), produces a very similar estimate, despite the much larger sample. Both exercises suggest that neither time-varying industry conditions nor the choice of firm controls or the smaller sample size account for the main pattern.

Next, we consider sample composition and scaling. Limiting the sample to firms for which at least four other ESG raters also provide coverage (column (4)) focuses on the more well-known, more intensively followed portion of the universe. The estimated coefficient is smaller in magnitude but still positive and statistically significant. Standardizing divergence to a z-score (column (5)) shows that a one standard deviation increase in disagreement is associated with a significant change in the wedge, illustrating the economic importance of the effect.

Finally, we conduct a placebo test by randomly shuffling the divergence measure across firms, while preserving its marginal distribution (column (6)). In this specification, the coefficient on shuffled divergence is close to zero and statistically insignificant. This falsification exercise

confirms that the baseline relationship is not mechanically induced by the distribution of wedges or generic time-series co-movement. Taken together, the robustness tests reported in Table 4 indicate that the positive link between peer disagreement and committee adjustments is stable across alternative ways of measuring wedges and divergence, across different fixed-effects structures or sample restrictions and under a placebo reshuffling of the key explanatory variable. The pattern we document therefore appears as a stable feature of MSCI's rating process, rather than a fragile artifact of a particular specification.

5 Conclusions

This paper examines how the investor-pays compensation model affects the quality of ESG ratings. Because ESG ratings directly govern index construction and portfolio mandates, rating changes impose rebalancing costs on the very subscribers who fund the rater, creating a conflict of interest that favors stability over timeliness. We document a traffic light effect in MSCI ESG ratings: companies systematically bunch below or above letter-rating thresholds. We trace this pattern to committee-level adjustments that create wedges between methodology-implied and published scores. These wedges grow with divergence from peer consensus and shrink with peer dispersion, consistent with a model in which the rater trades methodological consistency for peer alignment. Further, the wedges also grow with the distance from the previous score threshold, suggesting that committee intervention moves the scores in the direction that prevents score changes.

Our findings carry several implications. First, the optimality of investor-pays compensation structures is more context-dependent than the credit ratings literature suggests: when ratings are embedded in portfolio rules, subscriber demand for stability can induce the same type of bias that the model is designed to avoid. Second, published ESG ratings often reflect a compromise between the rater's own methodology and the assessments of its peers, suppressing private information that our return analysis shows to be value-relevant. Third, the concentration of adjustments near letter-rating boundaries implies that ESG index com-

position is partly shaped by committee discretion rather than by the codified methodology alone, and that researchers should be cautious about treating ESG ratings as quasi-random near categorical thresholds in RDD designs.

Several limitations deserve mention. Our analysis requires harmonizing heterogeneous peer rating scales and restricting attention to firms covered by multiple raters; measurement error from rescaling and selection into multi-rater coverage could affect the estimated magnitude of peer alignment, even though the Refinitiv methodology-change design alleviates concerns about common fundamentals. The return evidence, while consistent with suppressed downgrades containing value-relevant signals, is sensitive to event timing and limited statistical power for suppressed events. Finally, because we do not observe internal committee deliberations or the contractual details of the investor-pay model, our conclusions are best interpreted as evidence of incentive-consistent behavior rather than a complete causal accounting of ESG rating discretion.

References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Amel-Zadeh, A. and Serafeim, G. (2018). Why and how investors use esg information: Evidence from a global survey. *Financial Analysts Journal*, 74(3):87–103.

Azarmsa, E. and Shapiro, J. (2025). The market for esg ratings. *Journal of Finance. Forthcoming*.

Beaver, W. H., Shakespeare, C., and Soliman, M. T. (2006). Differential properties in the ratings of certified versus non-certified bond-rating agencies. *Journal of Accounting and Economics*, 42(3):303–334.

Berg, F., Fabisik, K., and Sautner, Z. (2021). Is history repeating itself? the (un)predictable past of ESG ratings. Finance Working Paper 708/2020, European Corporate Governance Institute (ECGI).

Berg, F., Kölbel, J. F., and Rigobon, R. (2022). Aggregate confusion: The divergence of esg ratings. *Review of Finance*, 26(6):1315–1344.

Bonsall, S. B., Gillette, J. R., Pundrich, G., and So, E. (2024). Conflicts of interest in subscriber-paid credit ratings. *Journal of Accounting and Economics*, 77(1):101614.

Brandon, R. G., Krueger, P., and Schmidt, P. S. (2021). Esg rating disagreement and stock returns. *Financial Analysts Journal*, 77(4):104–127.

Bruno, V., Cornaggia, J., and Cornaggia, K. J. (2016). Does regulatory certification affect the information content of credit ratings? *Management Science*, 62(6):1578–1597.

Bushman, R. M. and Smith, A. J. (2001). Financial accounting information and corporate governance. *Journal of Accounting and Economics*, 32(1–3):237–333.

Chen, C., Dube, S., and Froymovich, S. (2025). Esg rating competition and rating quality. *Journal of Accounting Research*, 63(5):1995–2037.

Christensen, D. M., Serafeim, G., and Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? the case of esg ratings. *The Accounting Review*, 97(1):147–175.

Cornaggia, J. and Cornaggia, K. (2026). ESG ratings management. *Working Paper*.

Cornaggia, J. and Cornaggia, K. J. (2013). Estimating the costs of issuer-paid credit ratings. *The Review of Financial Studies*, 26(9):2229–2269.

Derrien, F., Krüger, P., Landier, A., and Yao, T. (2025). Esg news, future cash flows, and firm value. *The Journal of Finance*, 80(6):3499–3554.

Dichev, I. D. and Piotroski, J. D. (2001). The long-run stock returns following bond ratings changes. *The Journal of Finance*, 56(1):173–203.

Griffin, J. M., Nickerson, J., and Tang, D. Y. (2013). Rating shopping or catering?an examination of the response to competitive pressure for cdo credit ratings. *The Review of Financial Studies*, 26(9):2270–2310.

Griffin, J. M. and Tang, D. Y. (2012). Did subjectivity play a role in cdo credit ratings? *The Journal of Finance*, 67(4):1293–1328.

Healy, P. M. and Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, 31(1–3):405–440.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 10(1):74–91.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182.

Jorion, P. and Zhang, G. (2007). Good and bad credit contagion: Evidence from credit default swaps. *Journal of Financial Economics*, 84(3):860–883.

Kashyap, A. K. and Kovrijnykh, N. (2013). Who should pay for credit ratings and how? NBER Working Paper 18923, National Bureau of Economic Research.

Kreps, D. M. (1990). *A Course in Microeconomic Theory*. Princeton University Press, Princeton, NJ.

Krüger, P. (2015). Corporate goodness and shareholder wealth. *Journal of Financial Economics*, 115(2):304–329.

Lewellen, J. W. (2015). The cross section of expected stock returns. *Critical Finance Review*.

Lovo, S. and Olivier, J. (2025). Who should pay for ESG ratings? *Review of Finance*. Forthcoming.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

MSCI (2019a). Esg ratings methodology. Technical report, MSCI ESG Research LLC.

MSCI (2019b). Understanding msci esg indexes: Methodologies, facts and figures. Technical report, MSCI ESG Research LLC.

MSCI (2022). Esg ratings methodology. Technical report, MSCI ESG Research LLC.

MSCI (2023). Esg ratings process. Technical report, MSCI ESG Research LLC.

MSCI (2024). Esg industry materiality map. Accessed: 2024-08-01.

Opler, T., Pinkowitz, L., Stulz, R., and Williamson, R. (1999). The determinants and implications of corporate cash holdings. *Journal of financial economics*, 52(1):3–46.

S&P (2023). Global industry classification standard. Technical report, S&P Dow Jones Indices.

Thistlewaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Observational Studies*, 3:119 – 128.

Tsallis, C. and Stariolo, D. A. (1996). Generalized simulated annealing. *Physica A: Statistical Mechanics and its Applications*, 233(1):395–406.

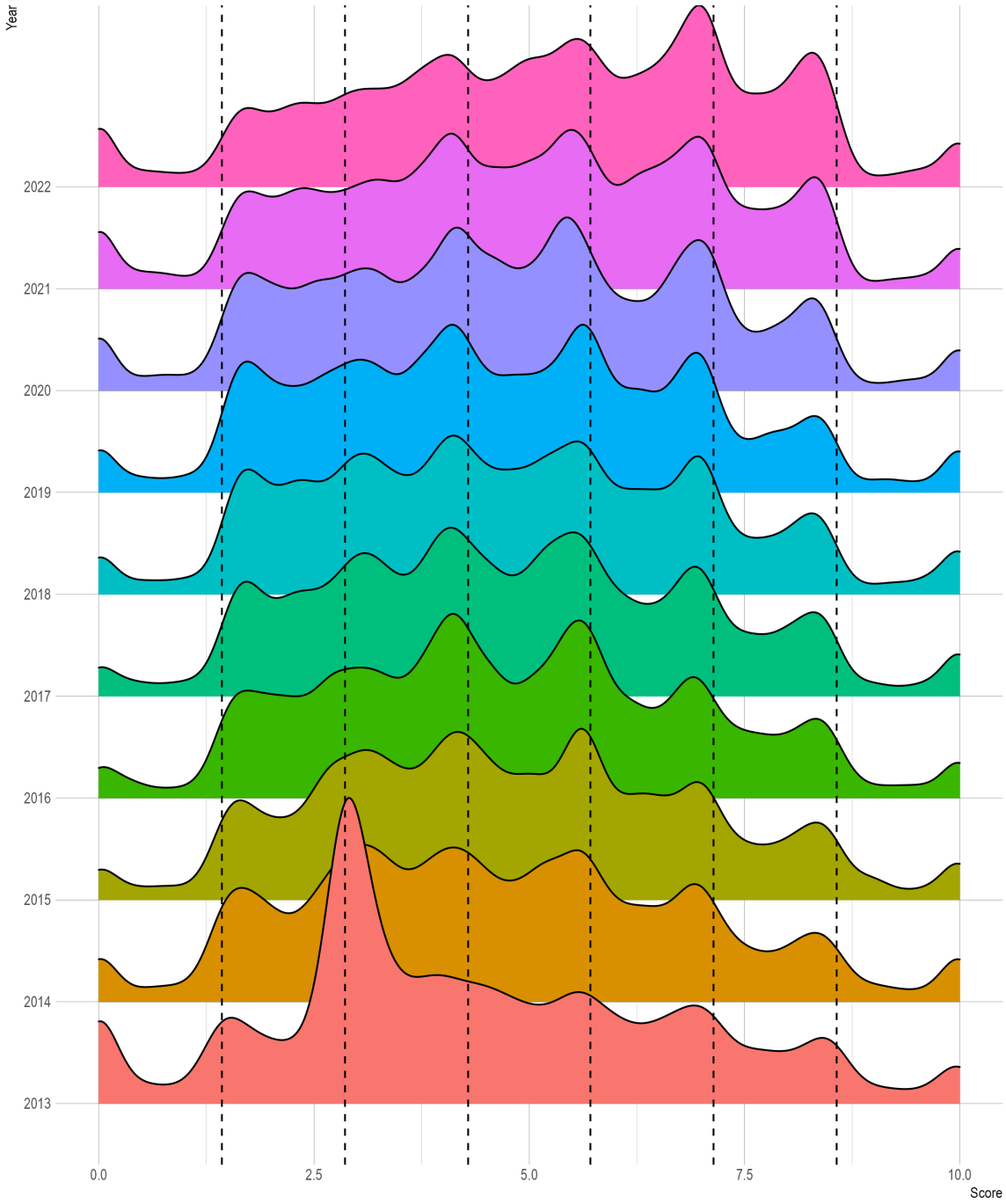

Figure 1: **MSCI's Industry Adjusted Score distribution**. Densities exhibit pronounced peaks near the thresholds between letter ratings. Similar to cars queuing before a traffic-light, firms cluster just below a threshold, as if awaiting a "green light" to move into the group of the next letter rating.

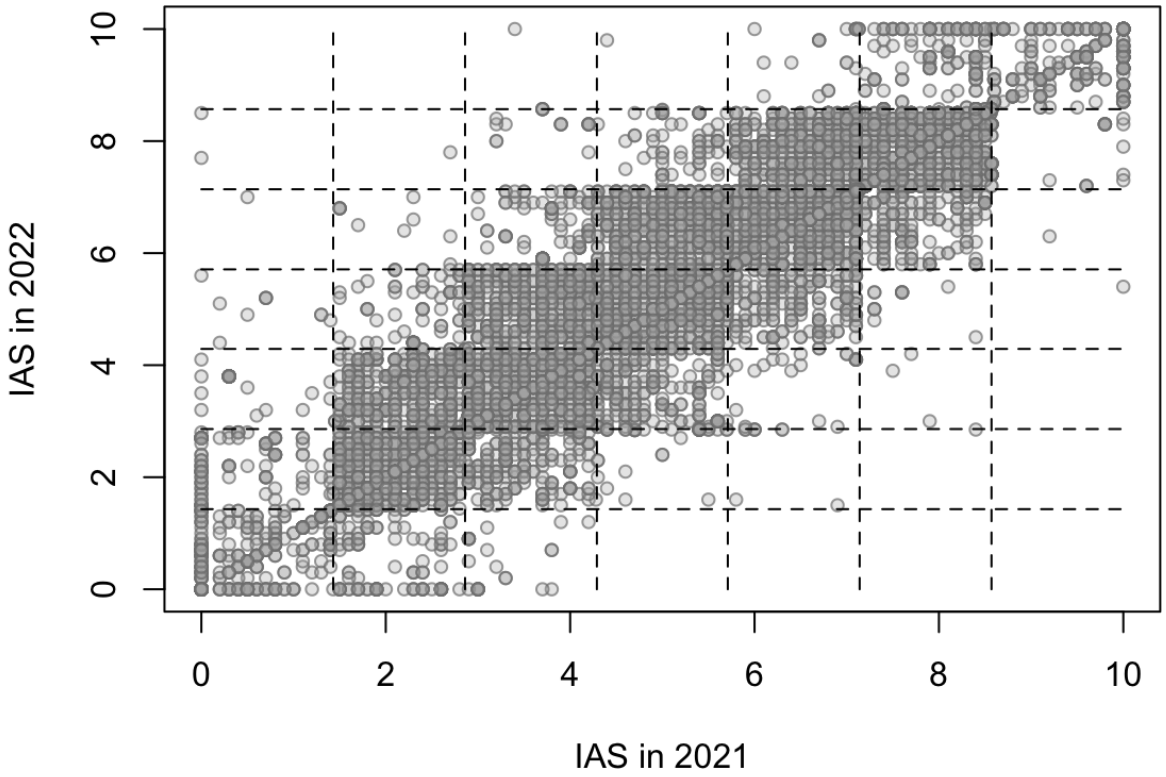

Figure 2: **Change of Industry Adjusted Scores across two consecutive years.** Scores appear to be bounded by letter rating thresholds, suggesting the presence of a mechanism that constrains transitions between letter ratings.

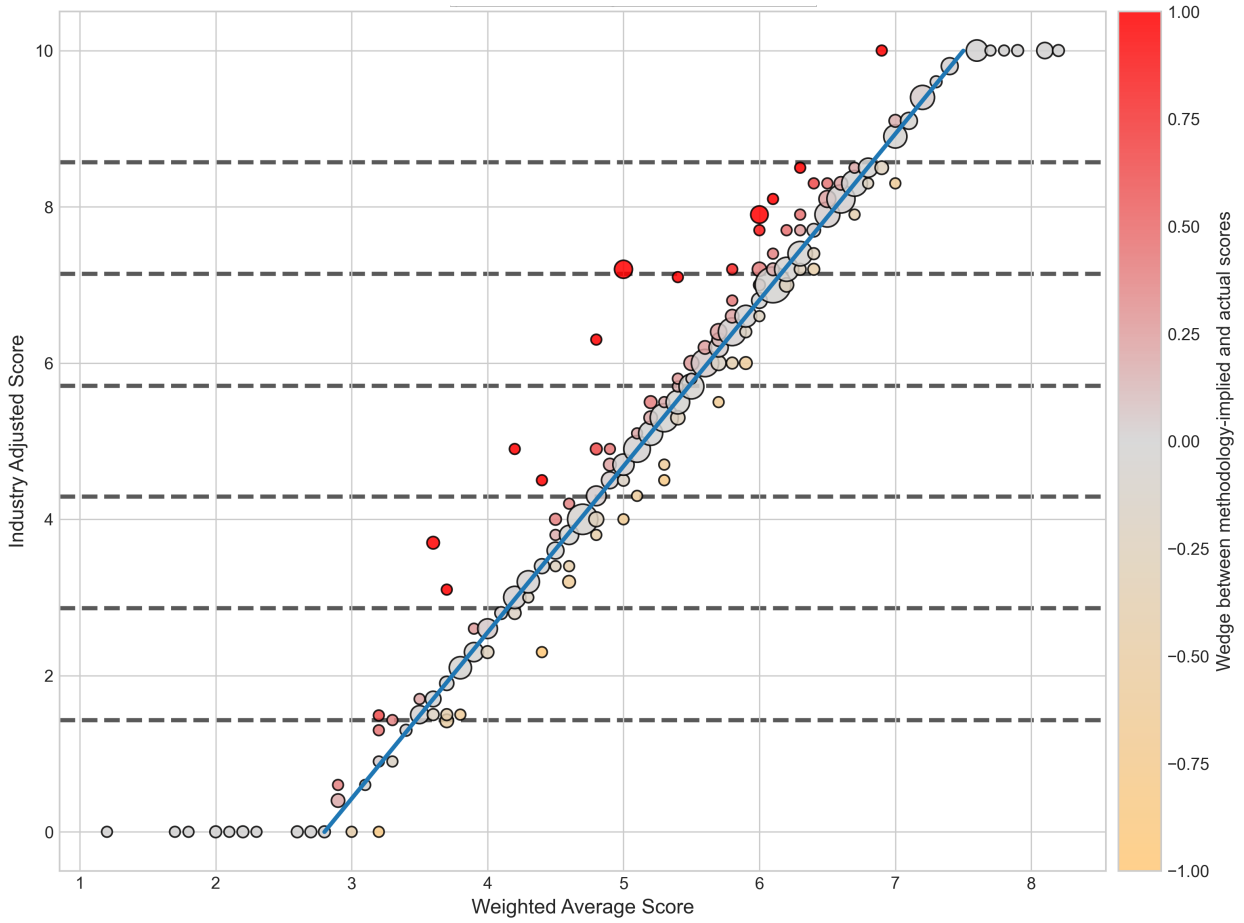

Figure 3: **Scatterplot of Industry Adjusted Scores against Weighted Average Score.** The industry benchmarking process does not appear to be linear as deviations from the methodology-implied formula occasionally arise. In the figure, gray dots represent firms whose scores show no deviation from the methodology-implied rating. Dot sizes are proportional to the number of firms per (WAS, IAS) pair.

Figure 4: **Frequency of rating changes by Industry Adjusted Score.** For each IAS, the bins indicate how many companies, among those with that IAS, were subject to a score change. Extensive adjustments refer to changes that result in a letter rating modification relative to the methodology-implied evaluation. Intensive adjustments are score changes that occur within the same letter rating range.

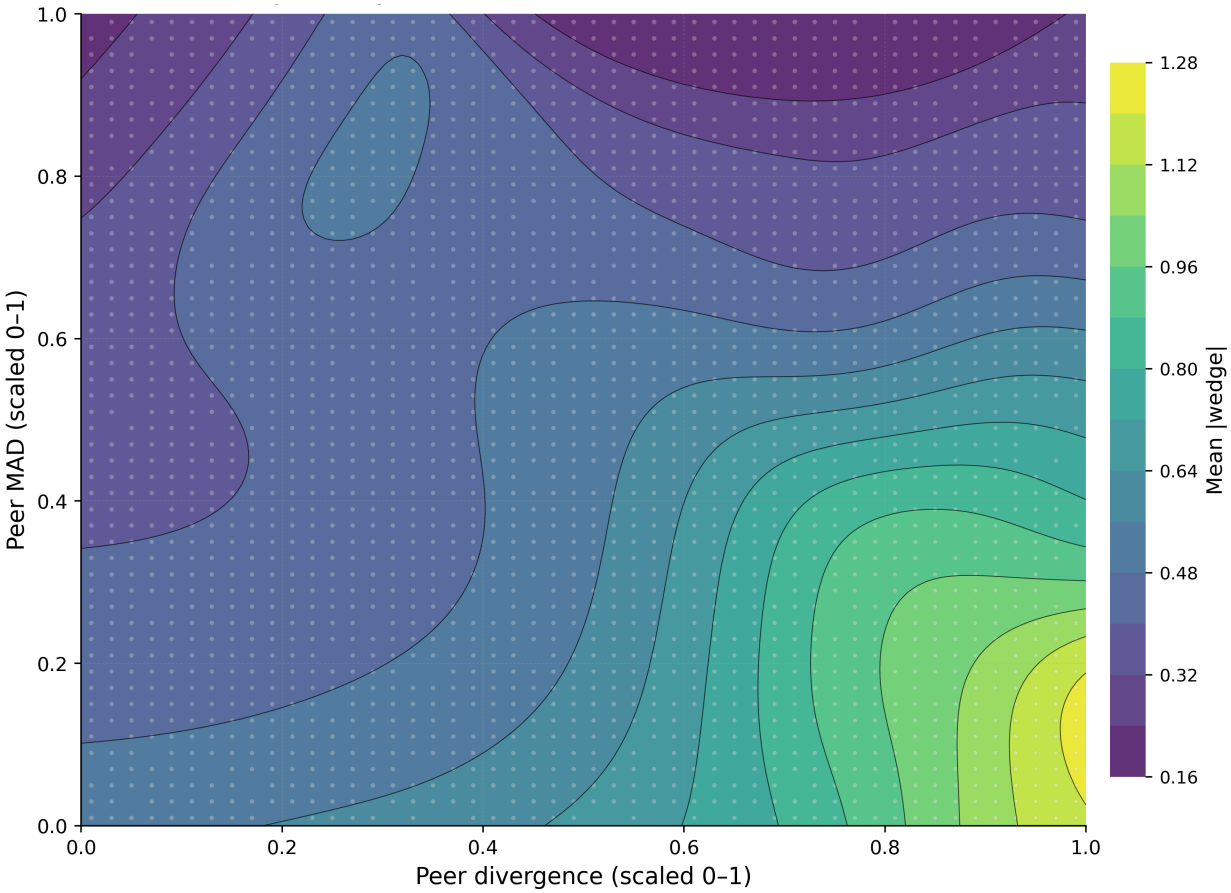

Figure 5: **Contour plot of the average wedge as a function of peer rating divergence and disagreement.** On average, firms whose methodology-implied score deviates more from the average score assigned by other rating agencies with low dispersion, receive a larger score adjustment.

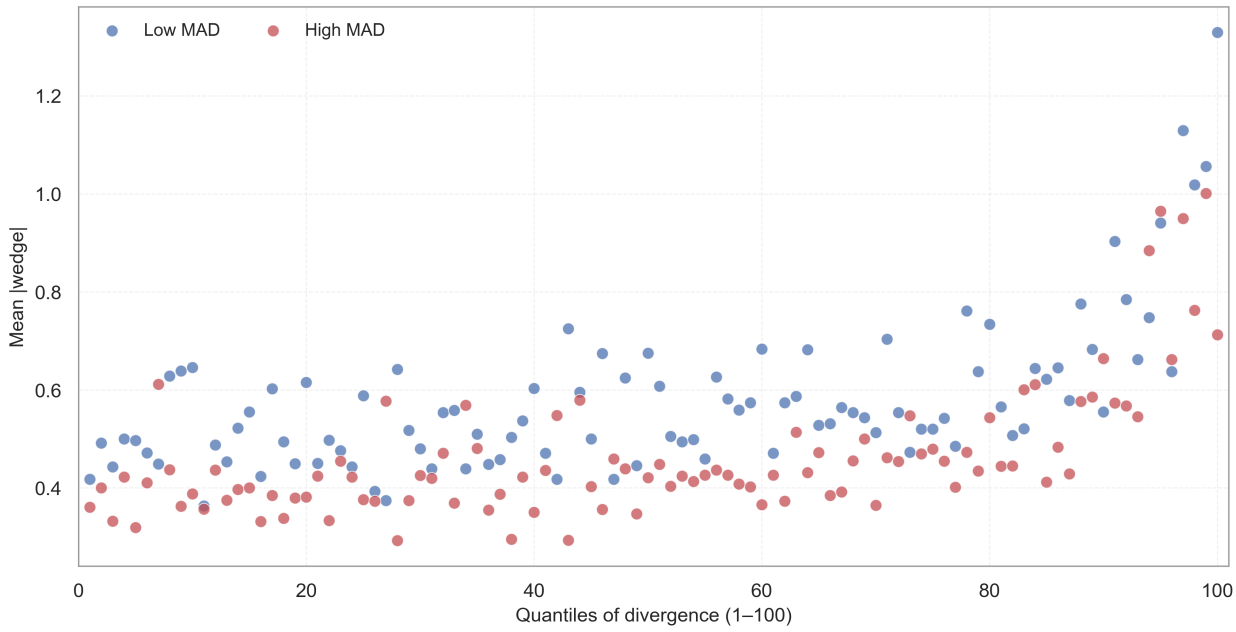

Figure 6: **Scatterplot of the average wedge across divergence quantiles.** Firms whose methodology-implied score deviates further from the average peer score receive a larger adjustment. Moreover, for the same level of divergence, firms whose peers exhibit greater consensus on a rating receive a larger adjustment.

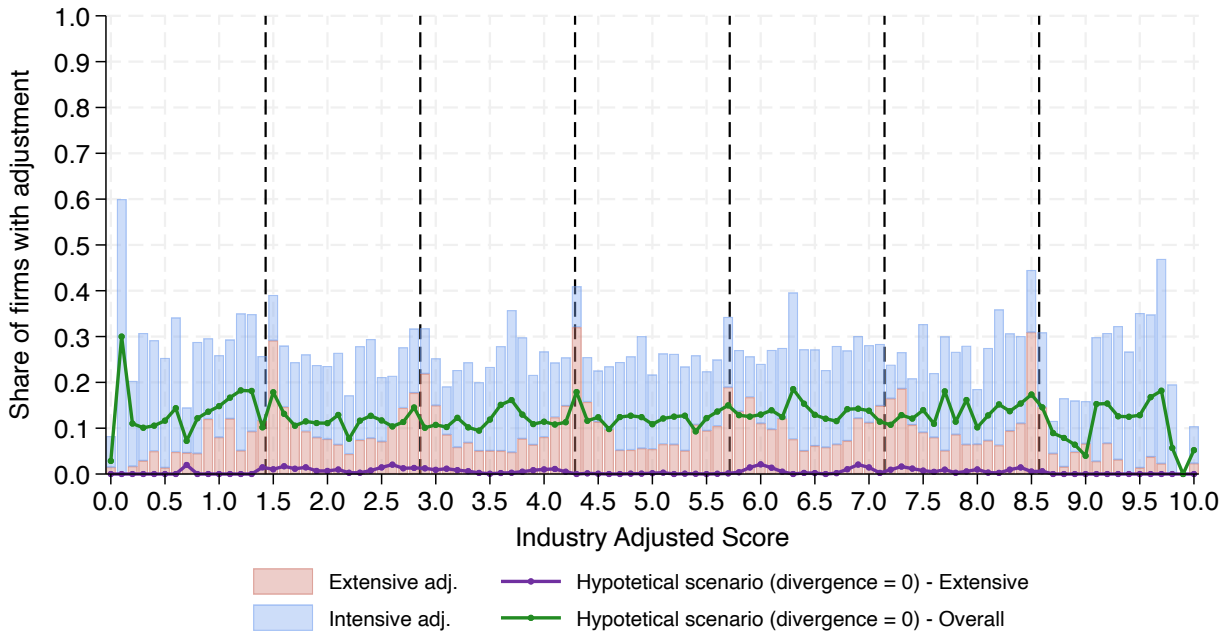

Figure 7: **Frequency of adjustments in a hypothetical scenario with zero divergence.** Setting the wedge to zero and computing the methodology-implied score from the fitted model leads to nearly all letter rating changes to disappear.

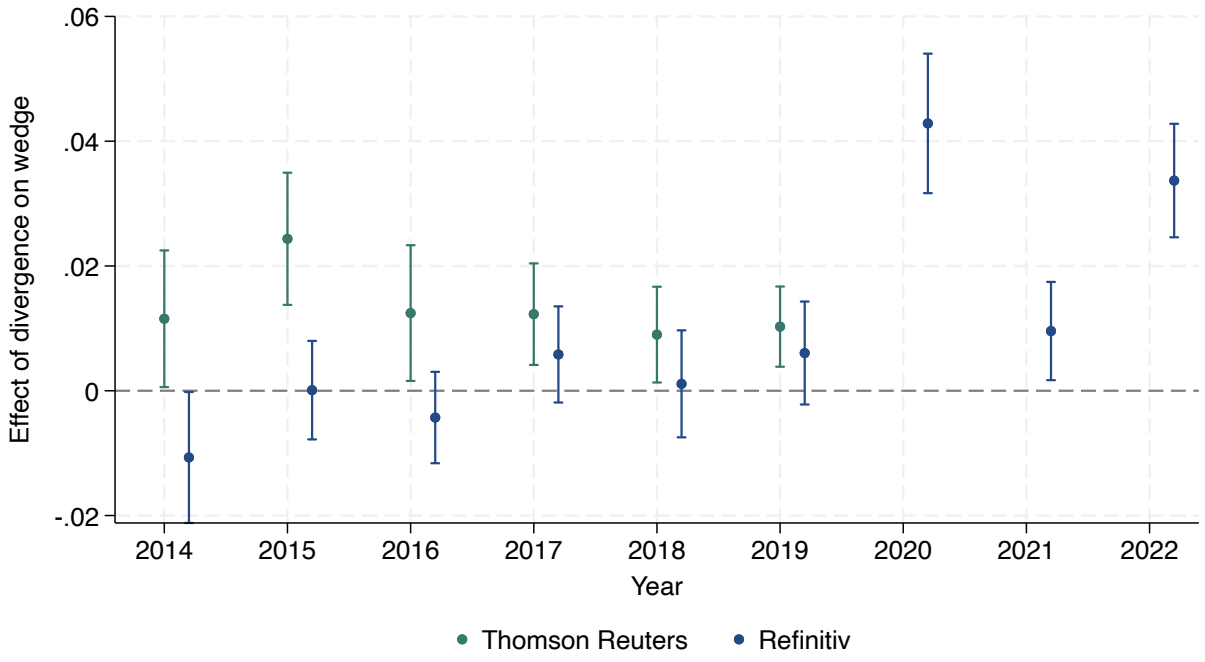

Figure 8: **Year-specific effect of divergence on the wedge (Refinitiv vs. Thomson Reuters).** Before 2020, the effect of Refinitiv's divergence from MSCI methodology-implied score is statistically indistinguishable from zero when using retroactive data, as MSCI did not have access to those ratings (which did not exist until 2020). Instead, they relied on the scores available until 2019 (Thomson Reuters ESG).

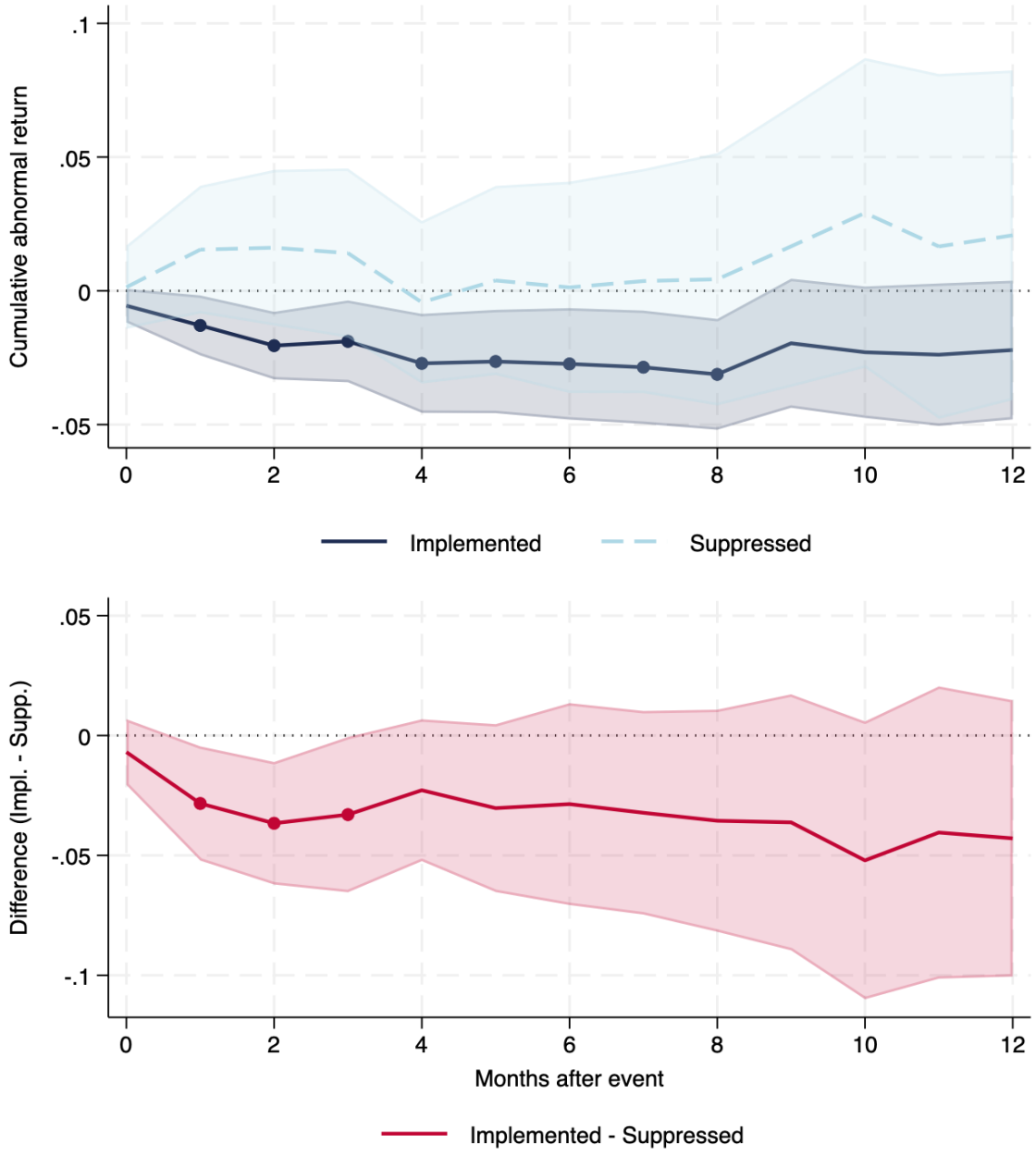

Figure 9: **Returns following downgrade signals: implemented vs. suppressed.** The top chart plots cumulative abnormal buy-and-hold returns (ABHR) for implemented downgrades (solid line) and suppressed downgrades (dashed line) over a 12-month horizon following the rating event. The bottom chart plots the difference between implemented and suppressed returns. Shaded regions represent 95% confidence intervals.

Variable	Mean	SD	Min	Max
Main MSCI ESG Score variables				
IAS	4.73	2.28	0	10
WAS	4.64	1.02	0	9.1
E Score	4.88	2.23	0	10
S Score	4.48	1.63	0	10
G Score	5.05	1.66	0	9.7
Financial variables				
Log(Assets)	8.28	1.84	0	15.26
CAPEX/Assets	0.04	0.04	0	0.20
Cash/Assets	0.12	0.13	0.0008	0.68
Debt/Assets	0.26	0.20	0	0.86
EBIT/Assets	0.06	0.10	-0.43	0.35
PPET/Assets	0.25	0.23	0	0.86
Sales growth	0.11	0.34	-0.63	2.10
Log(MktCap)	7.88	1.52	1.42	11.87
Book-to-market	0.54	0.48	0.01	9.89
Profitability (IB/Assets)	0.01	0.04	-1.40	0.26

Table 1: **Descriptive statistics**. Summary statistics (mean, standard deviation, minimum and maximum) for the key variables in the MSCI data sample. Financial variables, defined in Section 3.4 are winsorized at the 1% and 99% levels.

	Stability channel			Peers channel			Both channels		
	Abs	+	−	Abs	+	−	Abs	+	−
Retain gap	0.077***						0.075***		
	(9.26)						(9.20)		
|Retain gap|		0.065***	0.056***					0.048***	0.045***
		(11.07)	(7.63)					(7.78)	(6.26)
|Divergence|				0.028***			0.025***		
				(4.65)			(4.45)		
Divergence					-0.041***	-0.022***		-0.029***	-0.011***
					(-10.37)	(-6.41)		(-7.73)	(-4.20)
Peer MAD				-0.001	-0.003*	0.002	-0.002	-0.003***	0.001
				(-0.74)	(-2.36)	(1.84)	(-0.86)	(-2.84)	(1.37)
|Divergence| × MAD				-0.002**			-0.002**		
				(-3.00)			(-2.85)		
Divergence × MAD					0.002***	-0.000		0.002***	-0.000
					(3.88)	(-1.09)		(3.83)	(-1.11)
Observations	16699	14985	14466	16699	14985	14466	16699	14985	14466
R^2	0.445	0.464	0.426	0.435	0.464	0.419	0.449	0.472	0.429
Firm FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year × Country FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓

t-statistics in parentheses; SEs clustered at firm level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: **Main results.** The table reports panel regressions of committee adjustments to MSCI's ESG scores. For each mechanism, absolute wedge (Abs), positive wedge (+), and negative wedge (−) regressions are reported. The wedge is defined as the difference between the published Industry Adjusted Score and the methodology-implied score, $Wedge_{it} = IAS_{it} - \widehat{IAS}_{it}$. *Retain gap* is the minimum adjustment needed to move the methodology-implied score back into last year's actual MSCI letter-rating bucket; in the signed specifications, *retain gap signed* is positive when preserving last year's bucket requires an upward correction and negative when it requires a downward correction. *Divergence* is the difference between the methodology-implied score and the peer consensus, $\Delta_{it} = \widehat{IAS}_{it} - y^C_{it}$, where y^C_{it} is the simple average ESG score across other major providers. *Peer MAD* is the mean absolute dispersion of peer scores around that consensus, so lower values indicate tighter peer agreement. All specifications include firm controls *capx*, *ch*, *debt*, *ebit*, and *ppet*.

	(1) (Buffer)	(2) (Bare cross)	(3) (Buffer)	(4) (Bare cross)
Peer wins	0.128***	-0.108***	0.149***	-0.137***
	(6.53)	(-7.53)	(3.86)	(-4.87)
Observations	1893	1893	530	530
R^2	0.183	0.216	0.234	0.211
Firm Controls			✓	✓
Year × Country FE	✓	✓	✓	✓

Standard errors are clustered at the firm level.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3: **Overrides when peer alignment conflicts with stability.** The table reports regressions estimated on the subset of firm-years in which the stability channel and the peer-alignment channel point in opposite directions and an actual rating change occurs. Columns (1) and (2) present baseline specifications with year × country fixed effects only. Columns (3) and (4) add firm-level controls. The dependent variable in columns (1) and (3) is the post-cross buffer, defined as the distance between the realized score and the nearest crossed threshold measured inside the new rating bucket. The dependent variable in columns (2) and (4) is an indicator for a bare cross, equal to one when the realized score lies within 0.20 points of the crossed threshold.

	(1) (Median)	(2) (FE ind)	(3) (No controls)	(4) (≥4 peers)	(5) (z-scale)	(6) (Placebo)
Divergence (median)	0.008***					
	(4.31)					
Divergence		0.020***	0.024***	0.015***		
		(6.63)	(10.83)	(3.85)		
Divergence (z-score)					0.040***	
					(4.20)	
Divergence (shuffled)						0.003
						(1.46)
Observations	21589	21912	58100	15790	21589	15703
R^2	0.542	0.603	0.424	0.318	0.542	0.606
Firm Controls	✓	✓		✓	✓	✓
Firm FE	✓	✓	✓	✓	✓	✓
Year × Country FE	✓		✓	✓	✓	✓
Year × Ind FE		✓				

t-statistics in parentheses; SEs clustered at firm level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: **Robustness analysis.** The table reports panel regressions of the wedge on ESG score divergence under alternative specifications. Columns (1)–(6) successively use the median wedge, include Industry × Year fixed effects, exclude firm-level controls, restrict the sample to firms with at least four peers, rescale divergence to a standardized (z-score) measure, and implement a placebo test based on shuffled divergence.

	CAR			ABHR		
	(1)	(2)	(3)	(4)	(5)	(6)
H	Implemented	Suppressed	Difference	Implemented	Suppressed	Difference
1	-0.0129*	0.0154	-0.0283*	-0.0133*	0.0161	-0.0294*
	(-2.26)	(1.27)	(-2.36)	(-2.33)	(1.26)	(-2.39)
2	-0.0205**	0.0161	-0.0366**	-0.0197**	0.0171	-0.0369*
	(-3.18)	(1.09)	(-2.84)	(-2.96)	(1.02)	(-2.52)
3	-0.0189*	0.0141	-0.0330*	-0.0176*	0.0130	-0.0306
	(-2.41)	(0.87)	(-2.02)	(-2.20)	(0.68)	(-1.55)
4	-0.0271**	-0.0043	-0.0228	-0.0265**	-0.0069	-0.0196
	(-2.87)	(-0.25)	(-1.53)	(-2.78)	(-0.40)	(-1.13)
5	-0.0291**	-0.0070	-0.0221	-0.0280**	-0.0031	-0.0250
	(-2.98)	(-0.39)	(-1.53)	(-2.83)	(-0.17)	(-1.31)
6	-0.0288**	0.0011	-0.0299	-0.0284**	0.0034	-0.0319
	(-2.78)	(0.05)	(-1.81)	(-2.66)	(0.16)	(-1.49)
7	-0.0284*	0.0034	-0.0318	-0.0284*	0.0047	-0.0332
	(-2.51)	(0.15)	(-1.83)	(-2.37)	(0.20)	(-1.46)
8	-0.0306*	0.0041	-0.0346	-0.0295*	0.0039	-0.0334
	(-2.30)	(0.17)	(-1.84)	(-2.12)	(0.16)	(-1.49)
9	-0.0193	0.0202	-0.0395	-0.0190	0.0208	-0.0398
	(-1.29)	(0.74)	(-1.83)	(-1.22)	(0.72)	(-1.48)
10	-0.0229	0.0291	-0.0521	-0.0295*	0.0210	-0.0505
	(-1.84)	(1.00)	(-1.79)	(-2.30)	(0.55)	(-1.34)
11	-0.0239	0.0166	-0.0405	-0.0287*	0.0070	-0.0357
	(-1.75)	(0.51)	(-1.32)	(-2.10)	(0.19)	(-0.98)
12	-0.0221	0.0208	-0.0429	-0.0277*	0.0131	-0.0408
	(-1.68)	(0.66)	(-1.48)	(-1.97)	(0.36)	(-1.16)
Observations	356176	356176	356176	356187	356187	356187
Controls	✓	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓

t-statistics in parentheses; two-way clustered SEs by firm and month. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5: **Returns following downgrade signals, pooled local projections with firm controls.** Each row reports coefficients from separate horizon-H regressions of CAR and ABHR on implemented vs. suppressed downgrade indicators, controlling for lagged firm characteristics (size, leverage, book-to-market, profitability) and absorbing firm and month fixed effects. The "Difference" columns report $\beta_{imp\_dn} - \beta_{sup\_dn}$ at each horizon.

Appendix

We report below the proofs to the propositions in Section 2.

Lemma 1 (Concavity and interior candidates)**.** *On the letter-preserving region, the smooth objective is* $\Pi_{PA,\mathrm{sm}}(\varphi)$*, strictly concave with unique global maximum* $\varphi^{PA} := A/(A+h) \in [0,1)$*. On the letter-changing region, the smooth objective* $\Pi^{-}_{\mathrm{sm}}(\varphi) := \Pi_{PA,\mathrm{sm}}(\varphi) - p\sigma/d_A(\varphi)$ *is strictly concave with unique interior maximum* φ^* *satisfying:* $\varphi^* < \varphi^{PA}$ *when* $\Delta_0 > 0$ *(reinforcing channels);* $\varphi^* > \varphi^{PA}$ *when* $\Delta_0 < 0$ *(opposing channels). For* $\sigma = 0$*:* $\varphi^* = \varphi^{PA}$ *and* $\Delta\Pi_{PA} := A^2/[2(A+h)] > 0$*.*

Proof. Letter-preserving region. Differentiating $\Pi_{PA,\mathrm{sm}}$ twice: $\Pi''_{PA,\mathrm{sm}}(\varphi) = -2p\rho\omega(m_C)\Delta_0^2 - h = -(A+h) < 0$. Setting $\Pi'_{PA,\mathrm{sm}}(\varphi) = 0$: $A(1-\varphi) = h\varphi$, giving $\varphi^{PA} = A/(A+h) \in [0,1)$ since $h > 0$. The quadratic representation centred at φ^{PA}, $\Pi_{PA,\mathrm{sm}}(\varphi) = \Pi_{PA,\mathrm{sm}}(\varphi^{PA}) - \frac{A+h}{2}(\varphi - \varphi^{PA})^2$, evaluated at $\varphi = 0$ gives $\Delta\Pi_{PA} = \frac{A+h}{2}(\varphi^{PA})^2 = A^2/[2(A+h)]$.

Letter-changing region. The smooth objective is $\Pi^{-}_{\mathrm{sm}}(\varphi) = \Pi_{PA,\mathrm{sm}}(\varphi) - p\sigma/d_A(\varphi)$ with $d_A(\varphi) = g - \varphi\Delta_0 > 0$ on this region. Differentiating twice: $(\Pi^{-}_{\mathrm{sm}})''(\varphi) = -(A+h) - 2p\sigma\Delta_0^2/d_A^3 < -(A+h) < 0$, establishing strict concavity. To determine the direction of φ^* relative to φ^{PA}, evaluate $(\Pi^{-}_{\mathrm{sm}})'$ at φ^{PA}: since $\Pi'_{PA,\mathrm{sm}}(\varphi^{PA}) = 0$, we have $(\Pi^{-}_{\mathrm{sm}})'(\varphi^{PA}) = -p\sigma\Delta_0/d_A^2$. For $\Delta_0 > 0$: this is negative, so Π^{-}_{sm} is decreasing at φ^{PA}, giving $\varphi^* < \varphi^{PA}$. For $\Delta_0 < 0$: this is positive, giving $\varphi^* > \varphi^{PA}$. Boundedness follows since $h > 0$ ensures $\Pi^{-}_{\mathrm{sm}} \to -\infty$ as $|\varphi| \to \infty$. □

Proposition 1 (Optimal committee adjustment)**.** *When* $g = 0$ *(stability channel inactive), the global maximizer of* (4) *is*

$$\varphi^\star = \begin{cases} \varphi^{PA} = \dfrac{A}{A+h} & \text{if } \dfrac{A^2}{2(A+h)} \geq F, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

When $g > 0$*,* $\Delta_0 > 0$ *(reinforcing channels) and* $\varphi^{PA} \geq \varphi^{RG}$*, the stability penalty is avoided at* φ^{PA}*: the net gain from overriding rises to* $\Delta\Pi_{PA} + p\sigma/g \geq \Delta\Pi_{PA}$*, weakly lowering the*

effective fixed-cost hurdle.

Proof. *Case $g = 0$.* The stability penalty is absent: $\Pi(\varphi) = \Pi_{PA,\mathrm{sm}}(\varphi) - F \cdot \mathbf{1}_{\varphi \neq 0}$, with unique smooth maximizer φ^{PA}. Comparing candidates: $\Pi(\varphi^{PA}) - \Pi(0) = \Delta\Pi_{PA} - F$. Override optimal iff $\Delta\Pi_{PA} \geq F$.

Case $g > 0$, $\Delta_0 > 0$, $\varphi^{PA} \geq \varphi^{RG}$. At φ^{PA}: the letter is retained (no stability penalty), giving $\Pi(\varphi^{PA}) = \Pi_{PA,\mathrm{sm}}(\varphi^{PA}) - F$. At $\varphi = 0$: the letter changes and the stability penalty $p\sigma/d_A(0) = p\sigma/g$ applies, giving $\Pi(0) = \Pi_{PA,\mathrm{sm}}(0) - p\sigma/g$. Hence $\Pi(\varphi^{PA}) - \Pi(0) = \Delta\Pi_{PA} + p\sigma/g - F$, and override is optimal iff $\Delta\Pi_{PA} + p\sigma/g \geq F$. □

Proposition 2 (Comparative statics)**.** *If $\varphi^\star = \varphi^{PA}$ (i.e. $\sigma = 0$), then*

$$\frac{\partial \varphi^{PA}}{\partial \Delta_0^2} \;=\; \frac{2p\rho\,\omega(m_C)\,h}{(A+h)^2} \;>\; 0, \qquad \frac{\partial \varphi^{PA}}{\partial m_C} \;=\; \frac{2p\rho\,\Delta_0^2\,h}{(A+h)^2}\,\omega'(m_C) \;<\; 0.$$

The realized wedge $|\Delta_A| = \varphi^{PA}|\Delta_0|$ satisfies $\partial|\Delta_A|/\partial|\Delta_0| > 0$ and $\partial|\Delta_A|/\partial m_C < 0$. The event $\{\varphi^\star \neq 0\}$ is (weakly) increasing in $|\Delta_0|$ and (weakly) decreasing in m_C.

Proof. With $A = 2p\rho\,\omega(m_C)\Delta_0^2$: $\partial A/\partial \Delta_0^2 = 2p\rho\omega(m_C) > 0$ and $\partial A/\partial m_C = 2p\rho\Delta_0^2\omega'(m_C) < 0$. Since $\partial\varphi^{PA}/\partial A = h/(A+h)^2 > 0$, the chain rule gives the stated derivatives of φ^{PA}. For $|\Delta_A| = \varphi^{PA}|\Delta_0|$, using $\partial\varphi^{PA}/\partial|\Delta_0| = 2|\Delta_0| \cdot h \cdot 2p\rho\omega(m_C)/(A+h)^2 > 0$ and the product rule:

$$\frac{\partial|\Delta_A|}{\partial|\Delta_0|} = \varphi^{PA} + |\Delta_0| \cdot \frac{\partial\varphi^{PA}}{\partial|\Delta_0|} > 0, \qquad \frac{\partial|\Delta_A|}{\partial m_C} = |\Delta_0| \cdot \frac{\partial\varphi^{PA}}{\partial m_C} < 0.$$

Monotonicity of $\{\varphi^\star \neq 0\}$ follows from Proposition 3. □

Proposition 3 (Peer-alignment activation threshold)**.** *A peer-alignment override ($g = 0$, $\varphi^\star = \varphi^{PA} > 0$) occurs if and only if*

$$\omega(m_C) \;\geq\; \bar{\omega}(\Delta_0) \;:=\; \frac{F + \sqrt{F^2 + 2Fh}}{2p\rho\,\Delta_0^2}. \tag{17}$$

For $\omega(m_C) = 1/(1+m_C)$, this is equivalent to $m_C \leq \bar{m}_C(\Delta_0) := 1/\bar{\omega}(\Delta_0) - 1$. $\bar{\omega}(\Delta_0)$ is strictly decreasing in $|\Delta_0|$ and strictly increasing in F and h.

Proof. For $g = 0$ and $\sigma = 0$: the override condition $\Delta\Pi_{PA} \geq F$ becomes $A^2/[2(A+h)] \geq F$, i.e. $A^2 - 2FA - 2Fh \geq 0$, yielding $A \geq F + \sqrt{F^2 + 2Fh}$ (positive root). Dividing by $2p\rho\Delta_0^2 > 0$ gives (17). For $\omega(m_C) = 1/(1+m_C)$: $\omega \geq \bar{\omega}$ iff $m_C \leq 1/\bar{\omega} - 1 = \bar{m}_C(\Delta_0)$. Monotonicities of $\bar{\omega}$ follow by inspection of the closed-form expression. □

Proposition 4 (Stability bias)**.** *Suppose $g > 0$. Define the* retention gain

$$\Gamma := \frac{p\sigma}{g} - \frac{A+h}{2}(\varphi^{RG} - \varphi^{PA})^2, \tag{18}$$

and let $\tilde{\Delta}(\varphi^{RG}) := \Pi_{PA,\mathrm{sm}}(\varphi^{RG}) - \Pi_{PA,\mathrm{sm}}(0)$. Then:

1. Reinforcing channels ($\Delta_0 > 0$, $\varphi^{PA} < \varphi^{RG}$):

$$\varphi^\star = \begin{cases} \varphi^{RG} & \text{if } \Gamma \geq 0 \text{ and } \tilde{\Delta}(\varphi^{RG}) + p\sigma/g \geq F, \\ \varphi^* & \text{if } \Gamma < 0 \text{ and } \Pi^-_{\mathrm{sm}}(\varphi^*) - \Pi^-_{\mathrm{sm}}(0) \geq F, \\ 0 & \text{otherwise.} \end{cases}$$

   When $\varphi^\star = \varphi^{RG}$: $|\Delta_A| = g$. Since $p\sigma/g$ is decreasing in g, the retention incentive is strongest for marginal threshold crossings.

2. Opposing channels ($\Delta_0 < 0$): $\varphi^* > \varphi^{PA}$ *on the letter-changing region. The stability override $\varphi^{RG} = g/\Delta_0 < 0$ eliminates the penalty but sets $|\Delta_A| = g$ at alignment cost.*

Proof. Reinforcing case ($\Delta_0 > 0$, $\varphi^{PA} < \varphi^{RG}$). On $[0, \varphi^{RG})$: letter changes, smooth objective Π^-_{sm} with unique interior maximum $\varphi^* \in (0, \varphi^{PA})$ (Lemma 1). On $[\varphi^{RG}, \infty)$: letter preserved, objective $\Pi_{PA,\mathrm{sm}}$; since $\varphi^{PA} < \varphi^{RG}$, this is decreasing on $[\varphi^{RG}, \infty)$, so the constrained maximum on this region is φ^{RG}. The global optimum is found by comparing $\{0, \varphi^*, \varphi^{RG}\}$.

Comparing φ^{RG} and φ^*. Since $\Delta_0 > 0$ and $\varphi^* \in [0, \varphi^{RG})$, we have $d_A(\varphi^*) = g - \varphi^*\Delta_0 \leq g$,

so $p\sigma/d_A(\varphi^*) \geq p\sigma/g$. Together with $\Pi_{PA,\text{sm}}(\varphi^*) \leq \Pi_{PA,\text{sm}}(\varphi^{PA})$:

$$\Pi^-_{\text{sm}}(\varphi^*) = \Pi_{PA,\text{sm}}(\varphi^*) - \frac{p\sigma}{d_A(\varphi^*)} \;\leq\; \Pi_{PA,\text{sm}}(\varphi^{PA}) - \frac{p\sigma}{g}.$$

Therefore, using the quadratic representation of Lemma 1:

$$\Pi(\varphi^{RG}) - \Pi(\varphi^*) \;\geq\; \Pi_{PA,\text{sm}}(\varphi^{RG}) - \Pi_{PA,\text{sm}}(\varphi^{PA}) + \frac{p\sigma}{g} = -\frac{A+h}{2}(\varphi^{RG} - \varphi^{PA})^2 + \frac{p\sigma}{g} = \Gamma.$$

Hence $\Gamma \geq 0$ is sufficient for $\varphi^{RG} \succ \varphi^*$.

Comparing φ^{RG} and 0. At φ^{RG}: letter retained, no stability penalty. At 0: letter changes, penalty $p\sigma/g$.

$$\Pi(\varphi^{RG}) - \Pi(0) = [\Pi_{PA,\text{sm}}(\varphi^{RG}) - F] - [\Pi_{PA,\text{sm}}(0) - p\sigma/g] = \tilde{\Delta}(\varphi^{RG}) + p\sigma/g - F.$$

Override to φ^{RG} profitable iff $\tilde{\Delta}(\varphi^{RG}) + p\sigma/g \geq F$.

Comparing φ^* and 0 (when $\Gamma < 0$). Using $\Pi^-_{\text{sm}}(0) = \Pi_{PA,\text{sm}}(0) - p\sigma/g$:

$$\Pi(\varphi^*) - \Pi(0) = \Pi^-_{\text{sm}}(\varphi^*) - \Pi^-_{\text{sm}}(0) - F.$$

Override to φ^* profitable iff $\Pi^-_{\text{sm}}(\varphi^*) - \Pi^-_{\text{sm}}(0) \geq F$. When $\varphi^\star = \varphi^{RG}$: $|\Delta_A| = \varphi^{RG}|\Delta_0| = (g/\Delta_0) \cdot \Delta_0 = g$.

Opposing case ($\Delta_0 < 0$). Since $\Delta_0 < 0$: $d_A(\varphi) = g + \varphi|\Delta_0|$ increases in φ for $\varphi > 0$, so the stability term $-p\sigma/d_A$ is increasing in φ, reinforcing the peer-alignment gradient at every $\varphi > \varphi^{RG}$. By Lemma 1: $\varphi^* > \varphi^{PA}$. On the letter-preserving region ($\varphi \leq \varphi^{RG} = g/\Delta_0 < 0$): the objective is $\Pi_{PA,\text{sm}}(\varphi) - F$, increasing toward $\varphi^{PA} > 0$, so the constrained maximum is φ^{RG} with $|\Delta_A| = |g/\Delta_0| \cdot |\Delta_0| = g$. The committee compares $\Pi^-_{\text{sm}}(\varphi^*) - F$ against $\Pi_{PA,\text{sm}}(\varphi^{RG}) - F$ and $\Pi^-_{\text{sm}}(0)$; a peer-alignment-dominated committee selects φ^*, while a stability-dominated committee may select φ^{RG}. □

Proposition 5 (Excess peer alignment condition)**.** *An investor-pay model exhibits excess*

peer alignment relative to the planner benchmark when

$$\varphi^{PA} \;\geq\; \varphi^{\mathrm{pl}} \quad\Longleftrightarrow\quad 2p\,\rho\,\omega(m_C)\,\Delta_0^2 \;\geq\; \frac{v_M}{v_C}\,h.$$

The stability channel ($g > 0$) makes this condition more likely to hold: when $\Delta_0 > 0$, by raising the net gain from any positive override through $p\sigma/g$; when $\Delta_0 < 0$, by amplifying the peer-alignment override above φ^{PA} within the letter-changing region.

Proof. $\varphi^{PA} \geq \varphi^{\mathrm{pl}}$ iff $A/(A+h) \geq v_M/(v_M+v_C)$ iff $Av_C \geq v_M h$, i.e. $2p\rho\,\omega(m_C)\Delta_0^2 \geq (v_M/v_C)h$. When $\Delta_0 > 0$ and $g > 0$: from Proposition 1, the stability benefit $p\sigma/g$ lowers the effective fixed-cost hurdle, making the override and hence excess peer alignment weakly more likely. When $\Delta_0 < 0$ and $g > 0$: by Lemma 1, the stability term pushes $\varphi^* > \varphi^{PA}$, further exceeding φ^{pl} whenever φ^{PA} does. □

Additional Figures

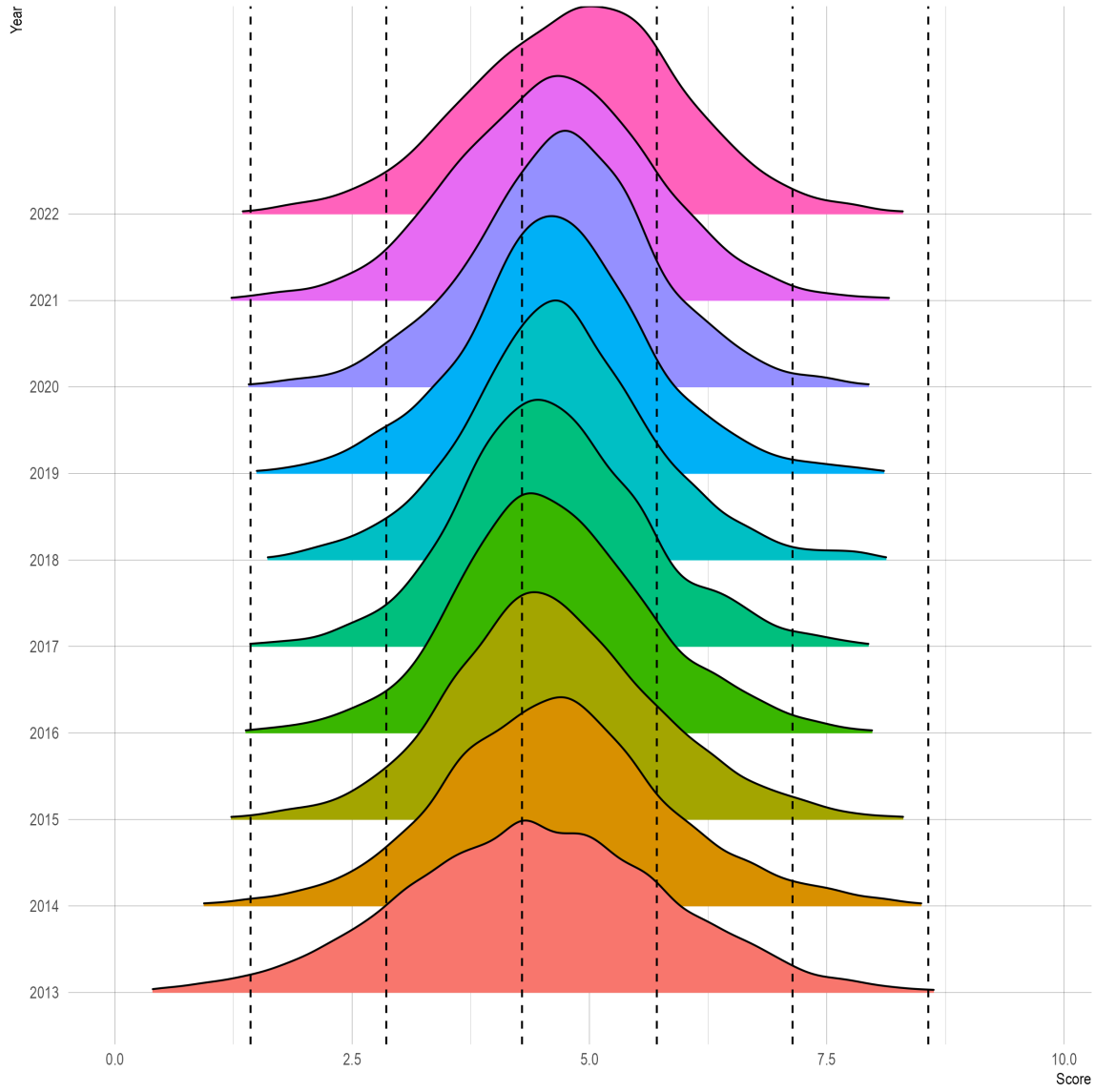

Figure 10: **Distribution of Weighted Average Scores (WAS) by year.** The figure shows yearly MSCI's Weighted Average Score (WAS) densities (the pre-benchmarking continuous score before the industry-adjustment step). Vertical dashed lines indicate the IAS cutoffs associated with MSCI's letter ratings. In contrast to the IAS densities in Figure 1, the WAS distributions exhibit no pronounced bunching around these thresholds. Because the letter-rating cutoffs are applied only after benchmarking, this pattern suggests that the threshold bunching observed in IAS is not a mechanical feature of the underlying raw score distribution, but instead arises later in the score-construction process.