

The Galileo problem in expert panels: screening calibration under decision loss and blame avoidance

Krzysztof Kontek

kkontek2000@yahoo.com

Merope

Research Article

Keywords: constitutional design, evaluator screening, expert panels, decision loss, blame avoidance

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11040691/v1>

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

[Read Full License](#)

Additional Declarations: Competing interest reported. Krzysztof Kontek is the sole named inventor on a pending patent application concerning specific implementations of evaluator-level screening and exclusion methods for panel evaluation systems. The author has also contributed to the development of the Excluding Outlier Jurors (EOJ) procedure analysed in the cited literature. The present paper develops general theoretical and institutional analysis of evaluator-level screening and does not describe or rely on patent-specific extensions.

The Galileo problem in expert panels: screening calibration under decision loss and blame avoidance

Krzysztof Kontek

Merope, Warsaw, Poland

Corresponding author: Krzysztof Kontek

E-mail: kkontek2000@yahoo.com

ORCID: 0000-0002-6753-7393

Abstract

Institutions that rely on expert panels routinely limit individual influence by applying score-level corrections such as trimming and clipping; evaluator-level screening offers a newer alternative that operates on entire reporting profiles. Borowiecki and Law (2026) show that observational ambiguity prevents outlier rules from simultaneously guaranteeing manipulation control and protection of sincere minority judgment. This paper develops an expected-loss framework for choosing the severity of evaluator-level screening. The analysis distinguishes two objectives. The first is decision quality: screening should minimise expected decision loss. The second is the organiser's broader institutional objective, which may include reputational, procedural, and blame-avoidance costs generated by consequential retained anomalies and controversial exclusions of valuable dissent. The organiser may therefore rationally accept some additional decision loss in order to reduce expected institutional exposure, creating a Public Choice wedge between decision-quality and organiser-preferred rules. Across the central environments studied here, stronger screening can therefore be justified by improved decision quality, institutional protection, or both. To make calibration operational, the paper introduces a simple probability-of-exclusion (PE_x) heuristic that translates the preferred screening regime into an ex ante ordinary-profile exclusion probability and a transparent pre-announced rule.

JEL classification: D02, D71, D72, D81, D83, Z11

Keywords: constitutional design; evaluator screening; expert panels; decision loss; blame avoidance.

1. Introduction

A music competition could appoint a single eminent pianist to select the winner, a grant agency a single distinguished scientist, and a hiring committee a single senior expert. Instead, these institutions combine several judgments and limit the authority of any one evaluator. Their rules embody a constitutional choice about how individual information and influence should enter a collective decision.

The analysis applies more broadly to score-based expert panels in which multiple evaluators assess candidates, proposals, projects, or alternatives and an institution aggregates their numerical judgments under a pre-announced rule. Relevant settings include grant review, academic hiring and promotion, public appointments, procurement panels, prize committees, and other expert-selection environments.

This problem connects several established literatures. Social-choice theory studies strategic reporting and the limits of strategy-proof collective aggregation (Gibbard 1973; Satterthwaite 1975; Balinski and Laraki 2010). The literature on committees and information aggregation examines how dispersed private information is combined and how strategic incentives affect collective accuracy (Austen-Smith and Banks 1996; Feddersen and Pesendorfer 1997, 1998). Constitutional political economy emphasises the prior choice of rules and the value of constraining case-specific discretion before identities, interests, and outcomes become salient (Buchanan and Tullock 1962; Brennan and Buchanan 1985; Buchanan 1987; Vanberg 1994; Hamlin 2023). A related Public Choice literature examines blame avoidance and institutional self-protection when decision-makers anticipate political or reputational costs from visible failures (Weaver 1986; Hood 2011).

Classical music competitions provide the main empirical and institutional illustration because their scoring rules, individual evaluations, and controversies are unusually observable. Earlier studies document systematic biases and organisational effects in competition outcomes, including performance-order effects (Flôres and Ginsburgh 1996; Glejser and Heyndels 2001), while Ginsburgh and van Ours (2003) show that competition rankings can have substantial subsequent career consequences. Candidate-level procedures that discard or correct unusual scores have also long been used in competition practice and have been analysed explicitly in the literature on competition scoring systems (Sosnowska 2017; Courtin, Gohard and Laruelle 2025; Kontek and Kenner 2025). Evaluator-level screening provides a newer alternative. Instead of modifying individual scores candidate by candidate, it conditions the influence of an evaluator's whole reporting profile on an announced admissibility rule. Kontek and Kenner (2025) develop this architecture through Excluding Outlier Jurors (EOJ). The underlying design problem is not competition-specific.

The attraction of such screening is accompanied by an informational problem. An evaluator who differs sharply from the panel may be strategically distorting the result, may be poorly calibrated or mistaken, or may genuinely perceive something that the majority has missed. The last possibility creates the Galileo problem: a rule intended to protect collective decisions from anomalous influence can suppress valuable minority judgment. Borowiecki and Law (2026) formalise the central difficulty. Under observational ambiguity, the same submitted score configuration may arise from strategic manipulation or sincere minority judgment. A score-based outlier rule therefore cannot guarantee both manipulation control and pluralism protection. Their result turns attention from perfect classification to constitutional choice: if both errors are possible, an institution must decide how much screening to impose and where to place the boundary between retained and lost influence.

This paper develops an expected-loss framework for that choice. Weak screening preserves more submitted information but leaves more harmful anomalous influence in the decision. Strong screening reduces that exposure while increasing the risk of suppressing useful information. The first objective is decision quality. Exclusion itself is not the relevant loss: many exclusions do not alter the collective outcome, and an outcome change can either improve or worsen the decision. The relevant criterion is therefore the expected consequence of retention and exclusion for the institutional decision.

The paper then distinguishes decision quality from the organiser’s broader institutional objective. The organiser chooses the rule, administers it, and ultimately bears responsibility for its visible consequences. A consequential retained anomaly can expose the institution to blame for allowing an outlying evaluator to change the collective result. A controversial exclusion can create the opposite exposure, especially when the excluded evaluator is prominent or is later perceived to have been substantively correct. The organiser therefore faces a two-sided blame-avoidance problem. Consistent with the broader blame-avoidance literature, institutional self-protection can justify a rule that sacrifices some expected decision quality in order to reduce exposure to a more salient failure. The organiser-preferred screening level can therefore differ materially from the decision-quality benchmark, creating a Public Choice wedge.

This wedge also gives ex ante commitment a second role. A pre-announced, case-neutral rule constrains evaluators by making loss of influence credible, and it constrains the organiser by limiting the scope for changing treatment after evaluator identity and outcome consequences become known. Screening can additionally affect reporting before any exclusion occurs: a credible but imperfectly predictable boundary can induce greater caution even when realised exclusions are rare.

To make these calibration choices operational, the paper introduces probability of exclusion (PEX) as a simple heuristic. A PEX level records the probability that an ordinary evaluator profile crosses the announced boundary under a pre-specified ordinary-reporting benchmark. A conventional tail probability provides a statistical reference point, while the same scale can later express either a decision-quality or organiser-preferred screening regime.

The paper makes four contributions. First, it extends the observational-equivalence logic behind the Galileo problem to deterministic score-based intervention more generally. Second, it develops an expected-loss approach to evaluator-level screening. Third, it separates the decision-quality optimum from an organiser-preferred rule shaped by blame avoidance and institutional self-protection, deriving the resulting Public Choice wedge and its implications for constitutional commitment. Fourth, it introduces PEX as an ex ante implementation heuristic that expresses a selected screening regime in terms of ordinary-profile exclusion risk rather than a configuration-specific numerical cutoff.

Section 2 develops the institutional architecture and informational constraint. Section 3 develops the calibration problem within a common illustrative benchmark, moving from statistical and decision-value reference points to consequentiality and expected decision loss. Section 4 derives organiser-optimal screening and the Public Choice wedge. Section 5 turns to constitutional implementation and develops PEX as the operational translation of a selected screening regime. Section 6 considers behavioural feedback and scope conditions, and Section 7 concludes.

2. Institutional architectures and the outlier-rule trade-off

2.1. Existing ways of limiting extreme influence

Institutions that rely on expert panels employ several mechanisms to limit excessive individual influence. The most common procedures operate at the level of individual score coordinates. Under symmetric trimming, one or more observations from each tail are discarded separately for each candidate. Under mean-centred clipping, scores outside an announced candidate-specific band are moved toward its boundary. Let S_{aj} denote the raw score assigned by evaluator j to candidate a , where $a = 1, \dots, A$ and $j = 1, \dots, N$. If

$$\mu_a = \frac{1}{N} \sum_{j=1}^N S_{aj}$$

denotes the raw mean score for candidate a , an additive clipping rule with width k transforms

$$S_{aj}^{\text{clip}} = \mu_a + \text{clip}(S_{aj} - \mu_a, -k, k). \quad (1)$$

Both trimming and clipping operate coordinate by coordinate, so an evaluator may lose influence on some candidates while retaining it on others. As a result, the effective score matrix can be assembled from different subsets of evaluators across candidates and need not preserve any submitted profile as a whole. Nominal participation is therefore not equivalent to informational preservation.

Evaluator-level screening has a different information footprint and changes the unit of intervention. A reporting profile remains intact while admitted and loses influence as a whole only when it crosses the announced screening boundary. All remaining submitted scores are left unchanged. The intervention is conditional at the profile level rather than imposed separately on each candidate-specific score distribution. Table 1 summarises these differences in the unit and visibility of intervention.

Table 1. Unit and visibility of intervention

Architecture	Unit of intervention	Typical action	Personal visibility
Trimming	score coordinate	deletion	low
Clipping	score coordinate	transformation	low
Profile screening	evaluator profile	loss of profile influence	high

The difference in personal exposure creates a visibility paradox. Conventional score-level correction restricts extreme judgments without identifying the evaluator as the object of intervention. Profile screening attaches the cut to an identifiable evaluator, making the same epistemic ambiguity more salient. That visibility can increase scrutiny, but it also makes intervention attributable and auditable at the evaluator level.

The established use of trimming and clipping shows that expert institutions already accept some information loss to limit extreme influence even though affected assessments may contain valuable minority information. The Galileo concern is therefore not unique to evaluator-level screening; Section 2.3 formalises the common informational constraint.

2.2. Excluding Outlier Jurors

Among evaluator-level screening procedures, we use Excluding Outlier Jurors (EOJ) of Kontek and Kenner (2025) as the concrete architecture for the analysis that follows. EOJ shifts the intervention from individual score coordinates to the evaluator’s submitted profile: an evaluator whose overall scoring pattern is sufficiently distant from the panel benchmark loses influence in the final aggregation. Borowiecki and Law (2026) likewise discuss EOJ as a motivating example before formalising the broader manipulation-control versus pluralism-protection trade-off for score-based outlier rules.

EOJ is particularly useful here because exclusion creates a well-defined evaluator-level counterfactual. The same realised panel can be evaluated with and without the focal reporting profile, holding the reports of the remaining evaluators fixed. This makes the decision consequence of intervention attributable to a specific marginal treatment without introducing method-specific score transformations. Profile screening can also aggregate moderate deviations across several score coordinates, allowing profile-wide abnormality to become visible even when no single score is exceptionally extreme.

Suppose focal evaluator i assigns score S_{ai} to candidate $a \in \{1, \dots, A\}$. Let

$$M_a = \text{Med}_j S_{aj} \quad (2)$$

be the median score of candidate a . EOJ defines the Manhattan profile distance

$$JD_i = \sum_{a=1}^A |S_{ai} - M_a|. \quad (3)$$

The cross-evaluator median distance is

$$MJD = \text{Med}_j JD_j, \quad (4)$$

and its median absolute deviation is

$$MAD_{JD} = \text{Med}_j |JD_j - MJD|. \quad (5)$$

The median/MAD construction also limits masking by a minority of extreme profiles; Appendix C gives a simple illustration and relates this choice to the robust-statistics literature (Hampel et al. 1986; Rousseeuw and Croux 1993).

For a tolerance multiplier $K \geq 0$, the realised upper boundary is

$$T_K = MJD + K MAD_{JD}. \quad (6)$$

Evaluator i loses influence when

$$JD_i > T_K. \quad (7)$$

Larger K is more permissive.

For $MAD_{JD} > 0$, define the dimensionless evaluator-level deviation index

$$z_i = \frac{JD_i - MJD}{MAD_{JD}}. \quad (8)$$

Then literal EOJ is simply

$$z_i > K. \quad (9)$$

Equation (8) is the measurement layer and (9) is the enforcement layer. The score-space boundary T_K is endogenous to the realised panel, while K is fixed ex ante in standardised z -space. The organiser can therefore commit to K before evaluations are submitted even though the realised numerical boundary in raw JD -space is not known in advance. The degenerate case $MAD_{JD} = 0$ is handled directly through (6)–(7).

EOJ is applied once to the submitted full-panel score matrix, and the welfare analysis below compares the decision with the focal evaluator retained or excluded.

2.3. The Borowiecki–Law impossibility result and its extension

Borowiecki and Law (2026) formulate a score-based outlier rule as a binary status $E(s_j, S_{-j}) \in \{0, 1\}$ for an evaluator’s entire score vector, with $E = 0$ covering exclusion or downweighting. Their manipulation-control property requires every manipulated score vector to receive the restrictive status $E = 0$, while pluralism protection requires every sincere score vector to receive $E = 1$. Under their Assumption 2, the same observed pair (s_j, S_{-j}) can arise from strategic manipulation or sincere minority judgment. The Borowiecki–Law impossibility result (their Proposition 2) therefore shows that no such outlier rule can guarantee both manipulation control and pluralism protection under observational ambiguity.

We extend the same observational-ambiguity logic beyond evaluator-level exclusion or downweighting to deterministic score-based interventions more generally.

Let ω_M and ω_G denote latent environments in which the same observable score matrix is generated by harmful manipulation and valuable sincere dissent, respectively:

$$S(\omega_M) = S(\omega_G) = \bar{S}. \quad (10)$$

Let $\mathcal{R}$ be any deterministic intervention based only on the submitted score matrix, mapping submitted reports into the effective inputs used by the final aggregation rule. This class includes trimming, clipping, score replacement, evaluator-level downweighting and evaluator exclusion. Say that $\mathcal{R}$ restricts report j at S if at least one component of j ’s submitted contribution is deleted, transformed, replaced, or assigned less than its original weight. Extending the benchmark properties of Borowiecki and Law (2026) to this broader class, $\mathcal{R}$ satisfies manipulation control if every manipulated report is restricted, and pluralism protection if no sincere report is restricted.

Proposition 1. Intervention invariance under observational equivalence

Let $\mathcal{R}$ be deterministic. If two latent environments generate the same submitted score matrix,

$$S(\omega_M) = S(\omega_G),$$

then

$$\mathcal{R}(S(\omega_M)) = \mathcal{R}(S(\omega_G)). \quad (11)$$

If, in addition, Borowiecki and Law’s Assumption 2 holds so that harmful manipulation and sincere minority judgment can generate such observationally equivalent environments, no deterministic $\mathcal{R}$ in this broader class can satisfy both manipulation control and pluralism protection: any restriction applied to the observationally identical report in ω_M is also applied in ω_G , while preserving the report fully in ω_G also preserves it fully in ω_M .

Proof. Equation (11) follows directly from determinism: the intervention receives the same observable input in both latent environments and therefore returns the same effective aggregation input. Borowiecki and Law’s Assumption 2 supplies the observationally equivalent manipulation and sincere-dissent states needed for the impossibility conclusion. ▀

Proposition 1 extends the Borowiecki–Law impossibility result from whole-profile treatment to the broader class of deterministic partial and coordinate-level interventions. Trimming, clipping, score replacement and evaluator-level screening therefore inherit the same observational-ambiguity constraint.

The unit and extent of intervention differ across architectures, but no deterministic procedure can recover latent information absent from the observed scores. Trimming can discard, and clipping can transform, precisely the assessment that is valuable in the Galileo state. Moving from score-level correction to profile screening changes the unit and visibility of intervention, not the underlying observational ambiguity.

2.4. The remaining calibration problem

After establishing their impossibility result, Borowiecki and Law (2026) turn to its practical implication. They note that exclusion rules can be useful when observed patterns are strongly suggestive of manipulation. The remaining problem is calibration: a stricter rule will exclude some sincere dissenters, whereas a more permissive rule will admit some strategic distortion. They therefore identify the remaining design question as where to set the threshold, noting that the choice depends on the relative costs of suppressing legitimate dissent and allowing strategic distortion. They leave open how these costs should be translated into an operational calibration criterion.

3. Where should the screening boundary be?

The calibration problem is developed within a common illustrative benchmark. We first establish two reference points: a statistical benchmark based on ordinary-profile exclusion risk and a decision-value benchmark based on the marginal contribution of a sincere evaluator. We then introduce the latent states used in the welfare analysis and follow screening from statistical exclusion through pivotality and consequential exclusion to expected decision loss.

3.1. PEx and the PE5 statistical benchmark

Let H denote the ordinary sincere-reporting benchmark and let

$$\alpha_H(K) = P(z_i > K \mid H)$$

be the marginal probability that an ordinary evaluator profile crosses cutoff K . Define the PEx level associated with K by

$$x(K) = 100 \alpha_H(K) = 100 P(z_i > K \mid H). \quad (12)$$

Conversely, $K_{\text{PEx}}(N, A)$ denotes the cutoff calibrated to PEx level x for panel size N and scorecard width A . Its dependence on the chosen ordinary-reporting reference model H is suppressed in the notation. PEx is an ex ante operating characteristic of the announced rule, not a target fraction of evaluators to be excluded from a realised panel. It is calibrated from a pre-specified all- H environment before evaluations are submitted; realised scores determine the EOJ centre, dispersion, and raw-score boundary but do not update x .

The definition in (12) does not require a particular distribution for H . Numerical calibration of K_{PEx} , however, requires an ordinary-reporting reference model. For Table 2 and the benchmark below, H is specified by independent Gaussian reporting errors across evaluators and score coordinates with a common variance. The resulting cutoffs do not depend on the value of that common variance, but may differ under other distributional assumptions.

A natural statistical reference point is the conventional five-per-cent tail probability, denoted PE5. Table 2 reports the corresponding cutoffs for selected panel sizes when $A = 6$.

Table 2. Approximate literal-EOJ cutoffs for PE5, $A = 6$

N	3	5	7	9	11	13	17
K_{PE5}	5.85	3.95	3.48	3.26	3.14	3.07	2.97

The table shows why the PEx scale is more portable than a raw K : the same ex ante ordinary-profile exclusion probability maps into different EOJ cutoffs as panel size changes. In the main benchmark, with $N = 17$ and $A = 6$,

$$K_{\text{PE5}} \approx 2.97. \quad (13)$$

The value $K = 3$, used in Kontek and Kenner (2025), therefore lies almost exactly at the PE5 statistical benchmark in this environment.

3.2. Decision loss and the marginal value of an evaluator

PE5 provides a transparent statistical reference point; whether it is also desirable from the perspective of decision quality remains an open question. To address that question, let $W_j(\omega)$ denote the outcome generated by institutional rule J in environment ω , and let $\ell_Q(W_j, \omega) \geq 0$ denote realised decision loss. Expected decision loss is

$$L_Q(J) = E[\ell_Q(W_j, \omega)]. \quad (14)$$

The loss function may represent a wrong winner, ranking loss, shortlist error, or another decision criterion. The numerical benchmark below uses binary winner loss.

Sincerity alone does not guarantee positive marginal decision value. Suppose n ordinary evaluators assess two alternatives whose expected score difference is $\Delta > 0$. Each ordinary evaluator has independent per-score error variance σ_H^2 . Add one sincere, unbiased, independent focal evaluator with variance σ_i^2 , with all evaluators carrying equal formal weight.

Lemma 1. Marginal decision value of a sincere evaluator

Adding the focal evaluator decreases pairwise decision loss if and only if

$$\sigma_i^2 < \left(2 + \frac{1}{n}\right) \sigma_H^2. \quad (15)$$

At equality the evaluator has zero marginal decision value; above the boundary, retention increases expected pairwise decision loss.

Proof. See Appendix A.

The corresponding decision-neutral noise ratio is

$$r_D = \frac{\sigma_i}{\sigma_H} = \sqrt{2 + \frac{1}{n}}. \quad (16)$$

With $N = 17$, hence $n = 16$, $r_D \approx 1.436$. Thus a sincere focal evaluator reaches the decision-neutral boundary when the standard deviation of the evaluator's reporting error is approximately 43.6% higher than that of an ordinary evaluator. Above this level, retaining the focal evaluator increases expected pairwise decision loss.

We can therefore ask where an evaluator at the decision-neutral noise ratio r_D is located on the EOJ scale. The following proposition gives this relationship for the population benchmark.

Proposition 2. Marginal decision value and the population EOJ scale

For a sincere evaluator with independent Gaussian reporting errors and an ordinary population EOJ reference, increasing the evaluator's relative error scale shifts the EOJ statistic outward. At the decision-neutral ratio r_D , its median location is

$$K_D^{\text{pop}} = c_A(r_D - 1), \quad (17)$$

where $c_A > 0$ depends on scorecard width.

Proof. See Appendix B.

For $A = 6$, numerical evaluation gives $K_D^{\text{pop}} \approx 2.05$. The exact location of the corresponding evaluator on the literal finite-panel EOJ scale depends on how the joint finite-panel centre and dispersion are summarised; Appendix B discusses this point. We therefore refer conservatively to a decision-neutral region around $K \approx 2$, clearly below the PE5 reference near $K = 3$.

The two benchmarks answer different questions. The PE5 benchmark controls the statistical rarity of excluding an ordinary profile, whereas the decision-neutral benchmark marks the approximate point at which retaining an additional sincere evaluator ceases to improve expected decision quality. In the illustrative configuration,

$$K_D \approx 2 < K_{\text{PE5}} \approx 3.$$

Neither reference point is an optimum. Welfare calibration also requires specifying the possible decision-relevant states of the focal evaluator.

3.3. Latent states and state-specific exclusion

Before the focal evaluator's state is known, let

$$T \in \{H, F, M, G\}, \quad \pi_T = P(T), \quad \sum_T \pi_T = 1. \quad (18)$$

Thus, π_T is the ex ante probability that the focal evaluator is in state T in a given evaluation instance. Equivalently, the π_T form mixture weights over the latent states represented by the model. They aggregate state-contingent consequences,

$$E[\ell] = \sum_T \pi_T E[\ell | T],$$

and are neither aggregation weights nor permanent evaluator types. They describe the ex ante incidence of the four decision-relevant states in the modelled institutional environment and are treated below as scenario inputs rather than as quantities inferred from the realised EOJ statistics.

State H represents ordinary competent judgment. The evaluator is sincere and has the benchmark precision of the ordinary panel.

State F represents sincere but sufficiently harmful imprecision. The evaluator remains unbiased, but reporting precision lies in the region identified by Lemma 1 in which retention increases expected decision loss. Limited familiarity with the task, misunderstanding of criteria, instability, error, or other sources of low decision-relevant precision can generate this state.

State M represents harmful strategic distortion. The evaluator shifts decision-relevant scores towards a preferred outcome.

State G represents Galileo-type informed dissent. The evaluator differs substantially from the panel because the evaluator has identified something that the rest of the panel has missed. An observable profile generated under G can therefore resemble, or even coincide with, one generated under M , while carrying the opposite decision value.

At the level of expected decision value, H and G are positive-value states, whereas F and M are negative-value states. These are decision-relative analytical states, not labels assigned by EOJ. The same evaluator may occupy different states across tasks or decision environments, and the mechanism neither infers intent nor adjudicates substantive truth. A profile can be procedurally inadmissible under a pre-announced rule without any claim that the evaluator acted in bad faith.

Conditional on state T , define the state-specific probability that the focal evaluator crosses the EOJ boundary as

$$\alpha_T(K) = P(z_i > K \mid T). \quad (19)$$

Lower K corresponds to more aggressive screening.

3.4. Statistical extremity and consequentiality

The analysis proceeds through three distinct layers:

statistical exclusion $\rightarrow$ outcome relevance $\rightarrow$ decision loss.

Crossing the screening boundary does not by itself imply that the collective outcome changes. EOJ evaluates the evaluator's full A -dimensional reporting profile, whereas decision relevance depends on how that profile affects the particular collective choice under consideration.

To make this distinction transparent, consider a binary decision between candidates 1 and 2 in a panel of N evaluators. Let i denote the focal evaluator. For evaluator j , define

$$d_j = S_{1j} - S_{2j}, \quad (20)$$

the evaluator's score difference between the two candidates. The aggregate contrast with the focal evaluator retained is

$$D^{+i} = \sum_{j=1}^N d_j, \quad (21)$$

whereas the corresponding contrast after excluding evaluator i is

$$D^{-i} = \sum_{j \neq i}^N d_j. \quad (22)$$

Positive values favour candidate 1 and negative values favour candidate 2. The focal evaluator is pivotal when retention and exclusion produce different winners:

$$W^{+i} \neq W^{-i} \Leftrightarrow \text{sgn}(D^{+i}) \neq \text{sgn}(D^{-i}). \quad (23)$$

The illustrative benchmark now specialises this general setup to

$$N = 17, \quad A = 6, \quad \Delta = 0.30, \quad \sigma_H = 0.60. \quad (24)$$

The full six-candidate location vector and simulation design are reported in Appendix D.

For an anomalous focal profile, the two contested scores are shifted by

$$(-q, +q), \quad q \geq 0. \quad (25)$$

Here q measures the magnitude of the anomalous displacement applied to each contested score; the resulting pairwise score difference is shifted by $2q$ towards candidate 2.

We distinguish three increasingly decision-relevant quantities. First, the state-specific statistical exclusion probability is

$$\alpha_T(q; K) = P(z_i > K \mid T, q). \quad (26)$$

This is the probability that the focal evaluator crosses the EOJ boundary in state T at anomaly magnitude q , irrespective of whether the exclusion affects the collective decision.

Second, potential pivotality is

$$P_{\text{piv}}(q) = P(W^{+i} \neq W^{-i} \mid q). \quad (27)$$

This is the probability that retaining rather than excluding the focal evaluator changes the selected winner. It measures the evaluator's potential decision leverage independently of whether EOJ actually excludes the evaluator.

Third, consequential exclusion is

$$\chi_T(q; K) = P(z_i > K, W^{+i} \neq W^{-i} \mid T, q). \quad (28)$$

A consequential exclusion occurs only when the focal evaluator both crosses the screening boundary and changes the collective decision. Hence

$$\chi_T(q; K) \leq \min\{\alpha_T(q; K), P_{\text{piv}}(q)\}. \quad (29)$$

A profile can therefore be statistically extreme without being decision-relevant, and an evaluator can be potentially pivotal without being excluded. Only their intersection produces a consequential exclusion.

Informational distinctiveness and decision value are separate objects. A report can contain highly distinctive information while having little expected effect on the institutional decision when it is rarely pivotal; conversely, a less extreme report can have substantial decision value near a close margin. The same statistical exclusion probability can therefore have very different consequences across decision environments. Pivotality is the transmission mechanism between statistical intervention and decision loss.

Under the displacement in (25), the expected full-panel contrast is $N\Delta - 2q$. It therefore changes sign at the balance point

$$q_B = \frac{N\Delta}{2}. \quad (30)$$

Thus q_B is the anomaly magnitude at which the expected full-panel decision is exactly balanced. For the benchmark $N = 17$ and $\Delta = 0.30$, $q_B = 2.55$. Under the Gaussian benchmark, potential pivotality can also be derived analytically; Appendix E gives the closed-form expression used to construct Figure 1 and discusses the behaviour of the pivotality curve at both endpoints, $q = 0$ and $q \rightarrow \infty$.

Figure 1 shows why statistical extremity and outcome relevance must be separated. At $q = 2$ and $K = 3$, marked by the dotted vertical line, the anomalous profile is excluded in about 78% of realisations, yet the evaluator is potentially pivotal in only about 29.8%, consistent with $q = 2$ remaining below the balance point $q_B = 2.55$. Consequential exclusion occurs in about 24% of realisations. The anomaly is therefore already highly detectable before it becomes large enough to place the expected full-panel decision at the balance point.

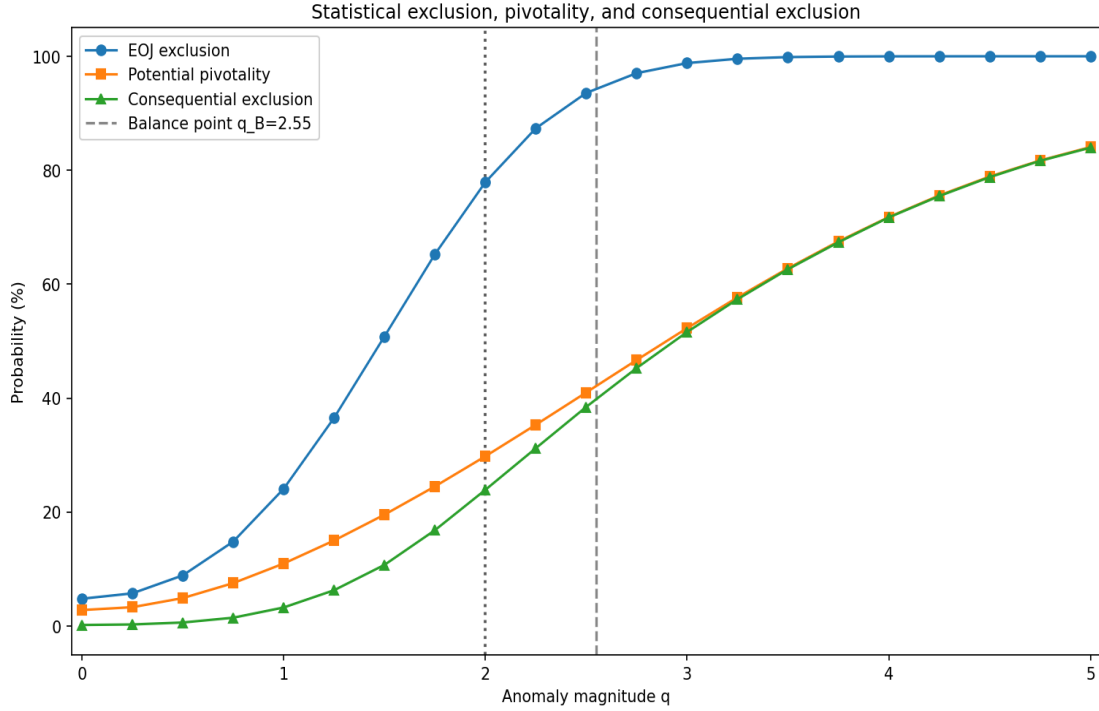


Figure 1. Statistical exclusion, potential pivotality, and consequential EOJ exclusion. The benchmark uses $N = 17$, $A = 6$, $\Delta = 0.30$, $\sigma_H = 0.60$, and $K = 3$. The focal anomaly changes the two contested scores by $(-q, +q)$; EOJ uses all six scores. The pivotality curve is calculated from the closed-form expression derived in Appendix E. The dotted vertical line marks the analysed benchmark $q = 2$, while the dashed vertical line marks the balance point $q_B = 2.55$, at which the expected full-panel contrast is zero.

3.5. From consequential exclusion to decision loss

To complete the transition from outcome relevance to decision loss, a consequential exclusion must be classified by whether it improves or worsens the decision. For each latent state T , let W_T^* denote the benchmark winner implied by the underlying decision environment, against which binary decision loss is evaluated. In the numerical benchmark, candidate 1 is the benchmark winner in states H , F , and M , whereas state G reverses the benchmark winner.

A harmful consequential exclusion occurs when retention would select W_T^* but exclusion changes the outcome to the other candidate. Define its probability as

$$h_T(K) = P(z_i > K, W^{+i} = W_T^*, W^{-i} \neq W_T^* | T). \quad (31)$$

Conversely, a beneficial consequential exclusion occurs when retention would select the other candidate and exclusion restores W_T^* . Its probability is

$$b_T(K) = P(z_i > K, W^{+i} \neq W_T^*, W^{-i} = W_T^* | T). \quad (32)$$

Hence, in the binary benchmark,

$$h_T(K) + b_T(K) = \chi_T(K) \leq \alpha_T(K). \quad (33)$$

Every consequential exclusion therefore falls into exactly one of two categories: it either improves or worsens the binary decision. Consequential exclusions are themselves only a subset of all exclusions, because many excluded profiles do not change the selected winner. The welfare effect of screening depends on the balance between beneficial and harmful consequential exclusions.

Proposition 3. Binary-loss decomposition

Under binary decision loss, the reduction in expected decision loss generated by screening at cutoff K , relative to retaining the focal evaluator, is

$$G_Q(K) = L_Q(\infty) - L_Q(K) = \sum_T \pi_T [b_T(K) - h_T(K)]. \quad (34)$$

Here $L_Q(\infty)$ denotes expected decision loss when the focal evaluator is always retained, corresponding to the no-screening limit $K \rightarrow \infty$.

Proof. Conditional on state T , a beneficial consequential exclusion reduces binary decision loss by one, whereas a harmful consequential exclusion increases it by one. All other realisations leave binary decision loss unchanged. The expected gain in state T is therefore $b_T(K) - h_T(K)$. Weighting by the state probabilities π_T and summing over T gives (34). $\square$

Proposition 3 establishes the decision-quality criterion used below. Screening improves expected decision quality whenever

$$G_Q(K) > 0.$$

The distinction between the three layers is now complete. The exclusion probability $\alpha_T(K)$ measures how often the rule intervenes; $\chi_T(K)$ measures how often that intervention changes the decision; and $b_T(K) - h_T(K)$ determines whether those decision changes improve expected decision quality.

3.6. Decision-quality calibration and the Galileo trade-off

To illustrate the calibration problem, we assign the following ex ante state probabilities to the benchmark environment:

$$(\pi_H, \pi_F, \pi_M, \pi_G) = (0.85, 0.03, 0.07, 0.05). \quad (35)$$

These probabilities are scenario inputs to the expected-loss calculation, not estimates recovered from the submitted scores. The particular ordering $\pi_M > \pi_G$ is illustrative. Its institutional motivation is that expert competitions have long maintained systematic safeguards such as trimming and clipping against harmful extreme influence, indicating that such influence is treated as a sufficiently salient risk to justify intervention. This practice does not identify the underlying state probabilities separately and therefore cannot establish the numerical values of π_M and π_G .

The benchmark further sets $\sigma_F^2 = 3\sigma_H^2$ and uses mirrored M/G profiles at

$$q = 2. \quad (36)$$

Together, these assumptions define the illustrative benchmark environment. Within the EOJ threshold family, define the decision-quality optimum by

$$K_Q^* \in \arg \min_{K \geq 0} L_Q(K). \quad (37)$$

Thus K_Q^* is the cutoff within the EOJ threshold family that minimises expected decision loss.

Table 3 compares three operating points. Because states M and G are observationally aliased in this benchmark, $\alpha_M = \alpha_G$ and $\chi_M = \chi_G$; Appendix D.1 gives the formal decomposition. The table reports $100G_Q(K)$ in percentage points (pp). The middle row is the representative centre of the benchmark low-loss region; the other rows correspond approximately to the PE5 and decision-neutral references.

Table 3. Selected EOJ operating points in the illustrative benchmark

K	Interpretation	α_H	χ_H	α_F	χ_F	$\alpha_M = \alpha_G$	$\chi_M = \chi_G$	$G_Q(K)$
3.0	PE5 region	4.85%	0.21%	50.04%	2.95%	77.84%	24.00%	0.437 pp
2.2	representative low-loss point	9.52%	0.40%	62.31%	3.52%	88.73%	27.03%	0.458 pp
2.0	decision-neutral region	11.22%	0.47%	65.41%	3.66%	90.84%	27.60%	0.452 pp

The comparison reveals why a moderately stricter cutoff can improve decision quality. Moving the screening boundary inward captures a substantially larger share of profiles whose retention is potentially harmful, including sincere but decision-harmful imprecision and strategic distortion. Most additional ordinary profiles that cross the boundary, however, are not pivotal, so the increase in ordinary-profile flagging translates into only a small increase in harmful decision changes. Screening can therefore become materially more effective against harmful influence before the cost of additional ordinary exclusions becomes comparable. Once the boundary moves further inward, the incremental gain from further F- and M-state screening begins to flatten while exclusions of ordinary and Galileo-type profiles continue to accumulate. This produces the broad low-loss region around the representative point reported in the table.

The substantive result is therefore the flatness of the loss surface around its best-performing region, not a uniquely identified decimal cutoff. Independent high-replication runs place the grid maximum at nearby values around $K = 2.1$ – 2.3 , while neighbouring cutoffs produce economically indistinguishable loss. We therefore use

$$K_{\text{rep}} \approx 2.2 \quad (38)$$

as a representative centre of the low-loss region rather than interpreting it as a precisely estimated universal optimum. The difference in decision loss between $K = 2$ and nearby points around the centre is below the resolution at which the benchmark simulation reliably distinguishes them. Figure 2 displays this broad low-loss region.

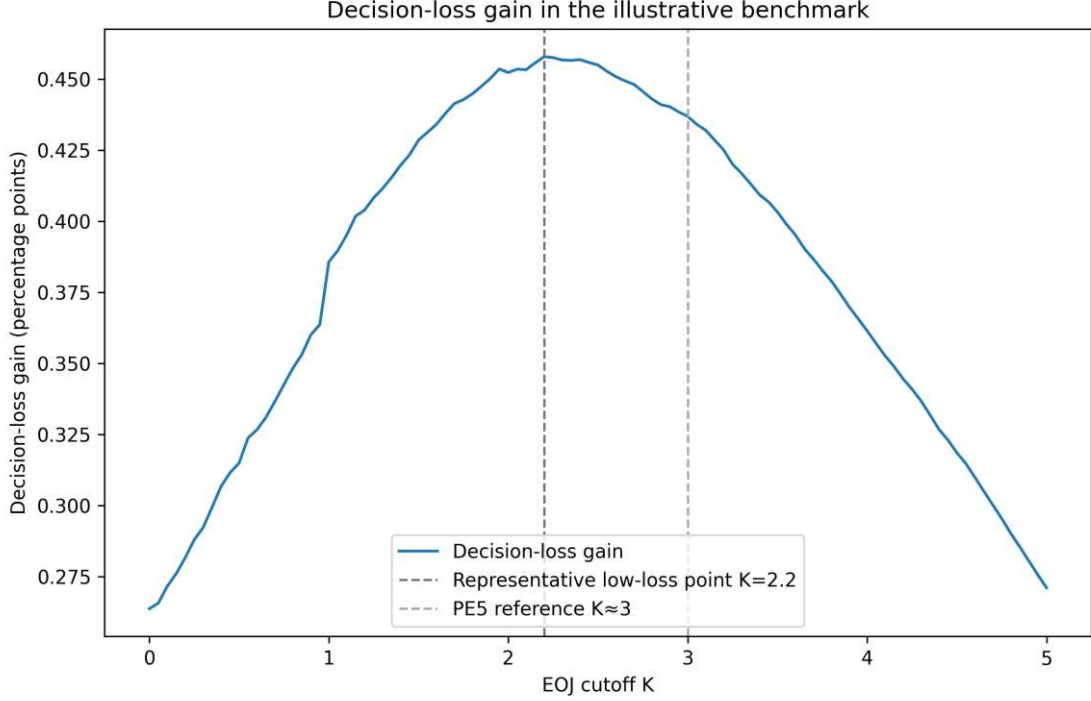


Figure 2. Decision-loss gain in the illustrative benchmark. Positive values indicate lower expected binary decision loss relative to retaining the focal evaluator. The solid curve and Table 3 are generated from the same common-random-number benchmark run. The dashed vertical lines mark $K = 2.2$, used as a representative centre of the low-loss region, and the PE5 region near $K = 3$. Appendix D gives the numerical design and simulation details.

The benchmark anchors can therefore be summarised as

$$K_D \approx 2 \lesssim K_{\text{rep}} \approx 2.2 < K_{\text{PE5}} \approx 3. \quad (39)$$

The ordering in (39) highlights the distinction between statistical rarity and decision value. The PE5 reference is defined by ordinary-profile exclusion frequency, whereas K_D and the low-loss region are determined by the consequences of retaining or excluding the focal evaluator. In the symmetric benchmark, the observationally identical M and G states necessarily generate the same exclusion and consequential-exclusion probabilities. This is the Galileo constraint in calibration form: screening severity can change how often the common observable profile is retained or excluded, but it cannot resolve the latent M/G ambiguity itself.

3.7. Marginal characterisation of the optimum

The numerical analysis above identifies a broad low-loss region rather than a sharply determined cutoff. It is nevertheless useful to characterise the decision-loss optimum theoretically.

For a focal evaluator with realised EOJ statistic z , let ℓ_R denote decision loss under retention and ℓ_E decision loss under exclusion. Define the conditional marginal advantage of exclusion as

$$D_Q(z) = E[\ell_R - \ell_E \mid z_i = z]. \quad (40)$$

Thus $D_Q(z) > 0$ means that exclusion is preferable on decision-quality grounds for a profile observed at z , whereas $D_Q(z) < 0$ means that retention is preferable.

Proposition 4. Marginal-loss characterisation

At a regular interior decision-loss optimum,

$$D_Q(K_Q^*) = 0. \quad (41)$$

Proof. See Appendix G.

The condition says that the expected marginal loss from retaining a boundary profile equals that from excluding it. It is a necessary first-order condition at a regular interior optimum, not a sufficiency or uniqueness result; a flat loss surface can generate multiple solutions, and atoms can replace the differential condition by a discrete comparison. Direct implementation is demanding because consequence-based calibration can depend on panel size, scorecard width, decision margins, evaluator precision, anomaly magnitude and breadth, dependence, the ex ante state probabilities, and the relative losses associated with consequential retention and exclusion. Several of these quantities concern latent states and decision consequences that submitted scores alone generally do not identify.

Proposition 4 should therefore be understood primarily as a theoretical characterisation of the decision-quality optimum. It identifies the economic object that an ideal calibration would balance at the screening boundary, even when that object cannot be estimated directly from the realised score matrix. Because exact consequence-based optimisation is demanding and environment-specific, Section 5.1 develops PEx as a simple operational heuristic for translating a defensible screening regime into a transparent pre-announced rule.

4. The Public Choice problem: organiser loss and blame avoidance

Section 3 identifies the screening rule that minimises expected decision loss. A different question arises when the institution that selects and enforces the rule also bears the political, reputational, and procedural consequences of visible failures. This creates the Public Choice problem studied in this section.

4.1. Institutional exposure

The institution that chooses, announces, administers, and certifies the rule is also exposed to the public consequences of visible failures. Its objective can therefore include blame avoidance and institutional self-protection in addition to decision quality (Weaver 1986; Hood 2011).

In the present setting, blame exposure is two-sided. Retaining a harmful anomalous evaluator can become institutionally costly when that evaluator changes the result and the organiser is criticised for allowing an unusual profile to overturn the collective judgment of the remaining panel. Excluding a valuable dissenter can generate a different controversy, especially when the excluded evaluator is prominent or when subsequent scrutiny supports the dissenter's judgment. In highly visible expert competitions, either channel can take the more acute form of scandal avoidance.

Let C_M denote the additional institutional cost associated with a consequential retained M -type anomaly, and let C_G denote the additional institutional cost associated with consequentially excluding a valuable G -type evaluator. Let $C_Q > 0$ denote the cost assigned to one unit of decision loss; in the binary benchmark, one unit corresponds to an incorrect competition outcome. To avoid double counting, C_M and C_G measure only reputational, procedural, and related institutional costs over and above the decision loss already valued through C_Q . They can incorporate the expected visibility, salience, and institutional severity of the corresponding failure.

The two failures can have different institutional consequences. Excluding a Galileo-type evaluator leaves the decision with the remaining panel, allowing the organiser to defend the result as the collective verdict generated by the announced rule. The additional institutional cost may then arise primarily from the treatment of the excluded expert. By contrast, retaining an anomalous M -type evaluator who changes the result can expose the organiser to criticism for allowing a single reporting profile to overturn the choice supported by the remaining panel. In public expert competitions, this asymmetry can make $C_M > C_G$ plausible, although the model imposes no ordering.

The ex ante state probabilities π_M and π_G describe incidence, while C_M and C_G describe conditional institutional severity. The relevant expected institutional exposures are therefore $\pi_M C_M$ and $\pi_G C_G$. Screening severity reallocates exposure between them: stricter screening increases exposure to consequential Galileo exclusion, while more permissive screening increases exposure to consequential retained manipulation.

4.2. Institutional loss

Under the observationally aliased M/G benchmark, $\chi_M(K) = \chi_G(K) = \chi_{MG}(K)$. Lower K raises this common consequential-exclusion probability: it reduces consequential retained manipulation under M while increasing consequential exclusion under G . Let $\mathcal{L}_O(K)$ denote the organiser's expected total loss, combining decision loss valued at C_Q with the two institutional-exposure components defined above. Appendix H gives the full decomposition; after dropping the component that is constant in K , minimising $\mathcal{L}_O(K)$ is equivalent to the objective below.

Proposition 5. Organiser-optimal screening

The organiser's preferred cutoff satisfies

$$K_O^* \in \arg \min_{K \geq 0} \{C_Q \mathcal{L}_O(K) - (\pi_M C_M - \pi_G C_G) \chi_{MG}(K)\}. \quad (42)$$

Proof. See Appendix H.

The first term values decision quality. The second captures the organiser-specific balance between the cost of retaining a consequential M -type anomaly and the cost of consequentially excluding a G -type evaluator. The organiser-preferred cutoff therefore coincides with the decision-quality optimum only when this additional institutional exposure is neutral.

4.3. The Public Choice wedge

For any candidate cutoff K , let

$$\Delta_Q(K) = L_Q(K) - L_Q(K_Q^*) \geq 0 \quad (43)$$

denote the additional decision loss from using K instead of the decision-quality optimum. Similarly,

$$\Delta_X(K) = \chi_{MG}(K) - \chi_{MG}(K_Q^*) \quad (44)$$

measures the change in consequential M/G exclusion relative to the decision-quality optimum. Unlike $\Delta_Q(K)$, which is non-negative by construction, $\Delta_X(K)$ can be positive or negative.

Corollary 1. Departure from the decision-loss optimum

The organiser prefers K to K_Q^* if and only if

$$(\pi_M C_M - \pi_G C_G) \Delta_X(K) > C_Q \Delta_Q(K). \quad (45)$$

Proof. See Appendix H.

The right-hand side is the decision-quality price of departing from K_Q^* ; the left-hand side is the organiser's net institutional gain from the resulting change in consequential screening. When organiser-specific exposure costs are small relative to C_Q , decision quality dominates the objective and K_Q^* remains at or close to K_Q^* . As these costs become more important, the direction of the departure depends on the balance between expected exposure to retained harmful anomalies, $\pi_M C_M$, and expected exposure to Galileo-type exclusions, $\pi_G C_G$. Table 4 summarises the resulting cases.

Table 4. Organiser incentives and the screening boundary

Institutional environment	Screening implication
$\pi_M C_M > \pi_G C_G$	pressure toward stricter screening; K_Q^* may fall below K_Q^*
$\pi_M C_M = \pi_G C_G$	organiser-specific exposure is neutral; decision quality determines the optimum
$\pi_M C_M < \pi_G C_G$	pressure toward more permissive screening; K_Q^* may rise above K_Q^*

At the neutral boundary,

$$\pi_M C_M = \pi_G C_G, \quad (46)$$

so organiser-specific exposure cancels and decision quality alone determines the threshold. For the illustrative benchmark probabilities $\pi_M = 0.07$ and $\pi_G = 0.05$, neutrality requires

$$\frac{C_M}{C_G} = \frac{\pi_G}{\pi_M} = \frac{5}{7} \approx 0.714. \quad (47)$$

Equivalently, $C_G/C_M = 1.4$. Because the benchmark has $\pi_M/\pi_G = 1.4$, equal per-event institutional costs imply $\pi_M C_M > \pi_G C_G$; retained harmful anomalies therefore receive greater expected institutional weight than Galileo-type exclusions.

4.4. Institutional interpretation

The Public Choice wedge can be substantive. Weaver (1986) emphasises that blame-avoiding decision-makers may choose differently from actors focused primarily on the quality of the underlying policy, while Hood (2011) treats blame management and organisational self-protection as recurring features of institutional behaviour. The organiser may accordingly choose K_Q^* such that

$$L_Q(K_Q^*) > L_Q(K_Q^*),$$

when the reduction in expected institutional exposure is sufficiently valuable.

This departure can point in either direction. Concern about a consequential retained M -type anomaly pushes the organiser toward stricter screening, while concern about a controversial Galileo-type exclusion pushes it toward a more permissive rule. Sufficiently strong exposure can move the organiser beyond the broad low-loss region identified in Section 3 and justify a measurable deterioration in expected decision quality.

The two risks may also differ in controllability. The asymmetry between C_M and C_G can already arise from the different institutional consequences described above; differences in controllability can amplify it. A consequential retained anomaly that becomes visible through the outcome can be difficult to contain, whereas the public salience of an exclusion can sometimes be limited by reporting that enforcement occurred without directly identifying the affected evaluator, although retrospective reconstruction may remain possible. Where

retained harmful anomalies are already more costly to the organiser, this difference can widen the gap between C_M and C_G . Settings in which exclusion of valuable dissent carries particularly high reputational, procedural, or informational costs can produce the opposite ordering. Section 5.5 develops the disclosure margin, and Section 6.2 considers the heightened stakes in small panels.

The central Public Choice implication is that the rule-maker's preferred constitution can reflect institutional self-protection as well as decision quality. A rule that is not decision-optimal may still minimise the organiser's broader loss. This divergence strengthens the case for constitutional commitment because the actor selecting the rule is also the actor later exposed to the consequences of its enforcement.

5. Constitutional implementation

5.1. PEx as an operational calibration heuristic

The preceding analysis identifies preferred screening regimes in terms of the EOJ cutoff: Section 3 studies the decision-quality benchmark, while Section 4 allows the organiser's broader institutional objective to shift that benchmark. A raw cutoff such as $K = 2.2$ or $K = 3$, however, is not a portable institutional recommendation. Its operating meaning changes with panel size, scorecard width, and the reference distribution of ordinary reports. PEx provides a convenient way to bridge this gap between theoretical calibration and implementable rule design.

Using the definition in (12), any selected cutoff K^* can be expressed by its implied ordinary-profile exclusion probability $x^* = x(K^*)$. Here K^* may represent the decision-quality choice or the organiser-preferred choice. PEx does not select between those objectives and does not create a new optimum. It translates a cutoff selected on substantive loss grounds into an operating characteristic with an immediate institutional interpretation: under the announced ordinary-reporting benchmark, an ordinary evaluator profile crosses the boundary with ex ante probability x^* per cent.

This translation is useful in two immediate ways. First, it separates the substantive choice of screening severity from the numerical implementation of EOJ: the institution can choose a defensible operating point and then calibrate the corresponding $K_{\text{PEx}}(N, A)$ for the panel configuration in use. Second, it makes screening severity comparable across settings for institutional designers because the same PEx level retains the same ex ante interpretation even when the numerical cutoff changes with N or A .

Third, PEx gives screening severity an intuitive operating meaning. A statement such as $K = 2.3$ says little by itself about how restrictive the rule is, whereas a PEx level states the ex ante probability that an ordinary evaluator profile will lose influence under the reference environment: higher PEx means more aggressive screening and greater ordinary-profile exclusion risk. This interpretation is useful for internal design and communication whether or not PEx is later disclosed publicly. Fourth, the PEx idea is not tied to the particular EOJ statistic. Any profile-screening rule with a pre-announced severity parameter can in principle be indexed by the ordinary-profile probability of losing influence under a specified reference environment. PEx can therefore provide a common operating language across different implementations of evaluator-level screening, while the mapping from PEx to a numerical cutoff remains rule-specific and must be calibrated separately. Table 2 and Appendix D illustrate this mapping for EOJ.

In the illustrative benchmark, the broad low-loss region maps to an ordinary-profile exclusion probability around ten per cent, and the PE9 cutoff $K_{\text{PE9}} \approx 2.27$ lies within that region. PE9 is therefore a useful provisional implementation benchmark, not a universal optimum. Its role is to provide a simple, communicable operating point supported by the benchmark analysis while preserving the distinction between a general PEx scale and a configuration-specific EOJ cutoff.

The resulting heuristic is deliberately simple: identify a defensible screening regime from decision and institutional consequences, translate it into a PEx level under the ordinary-reporting benchmark, and calibrate the corresponding numerical cutoff ex ante for the announced panel configuration. This two-stage structure—substantive choice followed by statistical translation—is the main operational contribution of PEx.

5.2. Ex ante calibration and realised enforcement

The distinction between choosing a rule ex ante and acting under it ex post is central to constitutional political economy (Buchanan and Tullock 1962; Brennan and Buchanan 1985; Buchanan 1987; Vanberg 1994; Hamlin 2023). EOJ separates these stages cleanly.

Under the PEx heuristic, the institution specifies the ordinary-reporting benchmark H , panel configuration, and PEx level x . It then calibrates $K_{\text{PEx}}(N, A)$ before evaluations are submitted. The realised score matrix subsequently determines M_a, JD_j, MJD, MAD_{JD} , and therefore

$$T_K = MJD + K MAD_{JD}.$$

The operational sequence is

choose $(H, N, A, x) \rightarrow$ calibrate and announce $K \rightarrow$ observe reports $\rightarrow$ compute $T_K \rightarrow$ apply EOJ mechanically $\rightarrow$ aggregate. (48)

The organiser chooses x , not the realised number of exclusions. Endogeneity of the raw-score boundary is compatible with commitment because the mapping from reports to that boundary is fixed in advance. The realised panel determines the score-space location of the boundary; it does not determine the PEx level or retroactively recalibrate K . This distinction separates a pre-announced screening rule from an ex post discretionary decision about whether a particular evaluator should count.

The hard-threshold form is itself an important institutional feature of the rule. Once the cutoff has been announced, enforcement is deterministic, transparent, auditable ex post, and case-neutral. Randomised exclusion would undermine these properties because a more extreme profile could remain influential while a less extreme one was excluded purely by chance, creating an additional source of procedural controversy.

5.3. What calibration should report

A complete calibration should report the ordinary-profile exclusion coordinate supplied by PEx together with anomalous-state exclusion, consequentiality $\chi_T(K)$, harmful and beneficial decision changes $h_T(K)$ and $b_T(K)$, and the resulting decision loss. Historical or pilot score matrices can inform score dispersion, panel-size effects, decision margins, and the distribution of the EOJ statistic. Latent state frequencies require sensitivity analysis or external evidence.

Appendix D reports the simulation design and expanded PEx lookup table. Reporting the PEx coordinate together with consequentiality and decision-loss measures makes clear both how often the rule intervenes and how often those interventions matter for the collective decision.

5.4. Ex ante commitment and case-neutral enforcement

A rule chosen ex ante can become difficult to enforce once a particular case is known. Before reports are submitted, the future outlier is anonymous; after a profile crosses the announced boundary, the evaluator's identity, prestige, relationships, and influence on the outcome become salient. These considerations can create pressure for exceptions precisely when the commitment value of the rule becomes most important.

The informational problem does not disappear after the flag. The score profile cannot reveal whether the evaluator is strategic, imprecise, mistaken, or exceptionally well informed. Asking the excluded evaluator to justify the report does not resolve the ambiguity: both a strategic evaluator and a genuinely dissenting expert can say that the scores reflect what they heard, observed, or professionally preferred, and such judgments are not directly verifiable by the organiser. In an artistic setting, the evaluator can simply maintain that this was their musical assessment and that differences in taste are not subject to administrative adjudication. Referring the case to a second body does not solve the problem either, because reviewers face the same latent-state uncertainty. Allowing substantive exceptions on the basis of identity, reputation, or outcome consequences would therefore replace a pre-announced rule with discretionary treatment after the stakes are known.

Ex post review should be limited to verifiable matters covered by rules specified in advance, such as clerical errors, data integrity, eligibility conditions, or declared conflicts of interest. If selected experts are to receive greater authority, that weighting should likewise be determined before reports are submitted.

Ex ante commitment thus constrains both evaluators and the organiser. It makes loss of influence credible while limiting the organiser's ability to suspend, reinterpret, or selectively waive the rule once identities and consequences become salient. Case-neutral enforcement is therefore part of the institutional value of the screening constitution.

5.5. Transparency and disclosure

Transparency has two distinct objects: the rule and its enforcement. An institution can announce the screening procedure and operative cutoff K . PEx may be used internally to calibrate that cutoff, but it need not itself be disclosed; its principal role is to provide institutional designers with a common operating scale.

After a round, the organiser can disclose that screening has actually been enforced without naming the affected evaluators, for example by reporting how many profiles crossed the pre-announced boundary and lost influence. Such aggregate disclosure makes enforcement observable and credible without publicly attributing motive or competence to particular individuals.

This distinction matters because EOJ identifies a boundary crossing rather than its cause. A flagged profile may reflect strategic distortion, sincere imprecision, informed dissent, or an ordinary tail realisation. Public identification can nevertheless be interpreted as an accusation of manipulation or incompetence and can create reputational and interpersonal costs unrelated to the statistical purpose of the rule.

This disclosure design also affects the organiser's blame exposure discussed in Section 4. Reporting that the rule was enforced without directly identifying the affected evaluators can reduce the immediate reputational cost of a Galileo-type exclusion. The protection is incomplete when detailed score data and the announced rule permit

retrospective reconstruction of the excluded profile, but direct public identification remains an additional and avoidable source of controversy. Aggregate disclosure can preserve enforcement credibility while limiting the organiser’s role in making a particular exclusion publicly salient.

A practical principle is therefore: disclose enforcement, not identities. This applies both to the public and to the evaluators themselves. In multi-stage settings, evaluators may be informed that the rule was activated, and possibly how many profiles lost influence, without learning whose profiles they were. Aggregate disclosure makes enforcement credible while limiting strategic learning about the realised boundary: learning that one’s own or another evaluator’s profile remained admissible could reveal additional room for more extreme reporting in later stages. The disclosure policy itself should therefore be fixed *ex ante*.

6. Behavioural feedback, small panels, and scope conditions

6.1. Structural opacity, non-monotonic incentives, and behavioural feedback

The analysis above concerns *ex post* operating characteristics. Evaluator-level admissibility can also affect reports before they are submitted. The rule and cutoff may be public, while the realised boundary remains uncertain because candidate medians, profile distances, MJD , and MAD_{JD} depend on contemporaneous peer reports. Individual evaluators do not observe those reports when submitting their own scores. We refer to this uncertainty as structural opacity. Experience from earlier rounds does not remove it, because reporting dispersion, the candidate pool, and evaluator behaviour can change across stages.

This uncertainty interacts with a non-monotonic incentive. Suppose an evaluator favours candidate A by raising A ’s scores and lowering those of competitors. Greater distortion initially moves the aggregate in the intended direction. If the profile crosses the screening boundary, however, the entire report loses influence: A loses the favourable scores and its competitors lose the unfavourable ones. Further manipulation can therefore reverse its own intended effect by eliminating the evaluator’s influence altogether. Structural opacity makes it difficult to identify a safe point immediately inside the realised boundary.

A second channel arises from the evaluator’s relationship with the organiser. Even when flagged identities are not disclosed publicly or to other evaluators, the organiser knows which profiles lost influence. Repeated or conspicuous flags may reduce confidence in an evaluator and affect future jury appointments. Individual enforcement status may also become known after the competition, creating an additional reputational risk. The strength of this channel depends on institutional practice, whereas the loss-of-influence channel follows directly from the mechanism.

Anonymous enforcement disclosure can reinforce both incentives. If the organiser reports that one or more profiles were excluded without identifying them, every evaluator who reported strategically must consider the possibility that their own profile was affected. The signal confirms that the rule is operative while revealing neither the current boundary nor an evaluator’s personal safety margin. It can therefore induce broader caution in subsequent stages.

A round with no exclusions provides a weaker signal in the opposite direction. Evaluators may infer that the risk of exclusion is lower and become more willing to report aggressively, but zero exclusions do not reveal whether all profiles were comfortably admissible or some were close to the boundary. Positive anonymous enforcement is therefore more informative about effective discipline than a zero-exclusion round is about its absence.

Anonymous disclosure consequently serves two functions. It limits reputational attribution while preserving uncertainty about individual enforcement status. In multi-stage settings, this allows the organiser to demonstrate that screening is active without enabling evaluators to learn how far they can safely push their reports. Realised reporting behaviour is therefore partly endogenous to the rule: credible enforcement can reduce strategic extremity before exclusion occurs, so fewer later flags may reflect deterrence rather than weak enforcement.

6.2. Small panels and conservative screening

Small panels magnify both sides of the screening trade-off. Excluding one evaluator removes a large share of the panel’s submitted information, but retaining one harmful evaluator also gives that profile correspondingly greater leverage over the collective outcome. In a three-member panel, one evaluator represents one third of the formal decision authority. Such an evaluator can make a substantial informational contribution, but a sufficiently unreliable or distorted report can also generate substantial decision loss.

Two distinct small-panel effects should be separated. The first is statistical and is already built into PEx calibration. For the same ordinary-profile exclusion probability, the corresponding EOJ cutoff rises sharply as panel size falls because the panel-relative centre and scale are less stable. With $A = 6$, Table 2 gives approximately

$$K_{PE5} \approx 5.85 \quad \text{for } N = 3,$$

compared with

$$K_{\text{PE5}} \approx 2.97 \quad \text{for } N = 17.$$

Thus the same five-per-cent ordinary-profile exclusion probability requires a much more permissive numerical cutoff in a three-member panel. A high K in a small panel should therefore not be interpreted as abandonment of screening; it is partly the mechanical consequence of implementing the same ex ante PEx level in a less stable panel-relative environment.

The second effect concerns the institutional choice of the PEx level itself. Strong prior confidence in a small and carefully selected panel, together with the large informational and procedural cost of excluding one of only a few experts, can justify a more conservative operating point. At the same time, the greater leverage of each retained evaluator makes harmful imprecision or strategic distortion especially consequential. Lemma 1 shows that even a sincere and unbiased evaluator can have negative marginal decision value when sufficiently imprecise. Small panel size therefore increases the stakes on both sides of the boundary rather than providing a general argument against screening.

The behavioural channel identified in Section 6.1 is especially relevant in small panels because each evaluator has greater leverage over the collective outcome. A higher cutoff need not eliminate the disciplining role of screening. Even when K is set conservatively and realised exclusions are rare, a credible loss-of-influence boundary can continue to constrain strategic extremism.

The relevant small-panel choice is therefore primarily the severity of screening, not its presence. Small N can justify a substantially higher cutoff and, when exclusion costs are sufficiently important, can produce $K_O^* > K_Q^*$. The mechanism can nevertheless retain both its safeguard function against exceptional profiles and the ex ante behavioural discipline created by the possibility of losing influence.

6.3. Scope conditions

The welfare analysis is marginal. EOJ flags are computed simultaneously from the full submitted score matrix; exclusion does not trigger recursive re-estimation of the medians, MJD , or MAD_{JD} . The comparison then retains or excludes one focal evaluator while peer reports are fixed. The states H, F, M, G describe alternative interpretations of that evaluator, not the composition of the entire panel. Multiple simultaneously consequential anomalous evaluators require a joint system-level model.

The main numerical benchmark uses a binary winner because it yields transparent pivotality and loss decompositions. Exclusions can also affect margins, qualification cutoffs, and full rankings; Appendix F relates the pairwise benchmark to broader decision criteria.

Exact M/G aliasing is a stress test in which submitted scores contain no information for distinguishing manipulation from informed dissent. Under exact aliasing, the submitted scores themselves also cannot identify π_M and π_G . Operational calibration therefore requires external evidence, historical or experimental information, or robustness analysis over plausible state-frequency ranges. Additional signals could create partial separation, but the quantitative location of the preferred cutoff remains environment-specific.

Proposition 1 is architecture-general, whereas the quantitative results are specific to evaluator-level EOJ screening. Trimming, clipping, and other interventions require their own counterfactual loss models. The institutional logic extends beyond competitions to other score-based expert panels whenever evaluators submit comparable numerical judgments, individual influence is aggregated mechanically, and the institution can commit ex ante to a rule governing anomalous influence.

7. Conclusion

As Borowiecki and Law (2026) show for outlier rules, observational ambiguity prevents score-based rules from perfectly distinguishing harmful distortion from valuable minority judgment. That classification limit does not remove the institutional need to choose a rule. Evaluator-level screening therefore creates a calibration problem: the institution must compare the decision consequences of retaining harmful anomalous influence with those of excluding ordinary or unusually valuable reports.

The decision-loss analysis shows why exclusion frequency alone is an inadequate criterion. Many ordinary exclusions are non-consequential, while excluding a harmful anomalous evaluator can prevent a substantially larger decision error. In the illustrative benchmark, the broad low-loss region occurs at lower K than the conventional five-per-cent statistical anchor, while consequential exclusion of ordinary evaluators remains very small.

The organiser may choose differently from a decision-quality planner. Consequential retained anomalies and controversial exclusions generate distinct procedural, reputational, and blame-avoidance exposures, creating a Public Choice wedge between the two objectives. In the environments analysed here, the main pressures generally favour stronger screening: decision quality can improve when harmful anomalous influence is removed, institutional protection can favour stricter intervention when retained anomalies are costly, and

credible loss of influence can discipline strategic reporting before exclusion occurs. Across the central environments studied here, stronger screening can therefore be justified by improved decision quality, institutional protection, or both. More permissive screening remains possible when the expected institutional cost of excluding valuable dissent is sufficiently large. Small panels require separate calibration because a given PEx level maps into a more permissive numerical cutoff, but this does not by itself imply weaker screening in operating terms.

The same constitutional problem extends to other score-based expert settings in which institutions must decide ex ante how much influence an anomalous evaluation should retain. The practical prescription is to calibrate screening before the case is known, announce its operative form, enforce it case-neutrally, and separate transparency about enforcement from public attribution of motive or competence. Evaluator-level screening cannot eliminate the Galileo ambiguity; it makes the resulting trade-off explicit, auditable, and governable.

Statements and Declarations

Funding

No funding was received for this study.

Competing interests

The author is the sole named inventor on a pending patent application concerning specific implementations of evaluator-level screening and exclusion methods for panel evaluation systems. The author has also contributed to the development of the Excluding Outlier Jurors (EOJ) procedure analysed in the cited literature. The present paper develops general theoretical and institutional analysis of evaluator-level screening and does not describe or rely on patent-specific extensions.

Data availability

No external datasets were used in this study. The numerical results are based on analytically specified models and Monte Carlo simulations. Replication code and supporting simulation outputs will be made available upon acceptance of the manuscript.

References

- Austen-Smith, David, and Jeffrey S. Banks. 1996. "Information aggregation, rationality, and the Condorcet jury theorem." *American Political Science Review* 90 (1): 34–45.
- Balinski, Michel, and Rida Laraki. 2010. *Majority Judgment: Measuring, Ranking, and Electing*. Cambridge, MA: MIT Press.
- Borowiecki, Karol J., and Marc T. Law. 2026. "Fairness trade-offs in expert evaluation: information, judgment, and procedural design." *Public Choice*. <https://doi.org/10.1007/s11127-026-01444-z>.
- Brennan, Geoffrey, and James M. Buchanan. 1985. *The Reason of Rules: Constitutional Political Economy*. Cambridge: Cambridge University Press.
- Buchanan, James M. 1987. "The constitution of economic policy." *American Economic Review* 77 (3): 243–250.
- Buchanan, James M., and Gordon Tullock. 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press.
- Courtin, Sébastien, Marin Gohard, and Annick Laruelle. 2025. "Queen Elisabeth Competition: A social choice approach." *The B.E. Journal of Theoretical Economics* 25 (2): 289–318.
- Feddersen, Timothy J., and Wolfgang Pesendorfer. 1997. "Information aggregation and voting behavior in elections." *Econometrica* 65 (5): 1029–1058.
- Feddersen, Timothy J., and Wolfgang Pesendorfer. 1998. "Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting." *American Political Science Review* 92 (1): 23–35.
- Flôres, Renato G., and Victor A. Ginsburgh. 1996. "The Queen Elisabeth musical competition: How fair is the final ranking?" *The Statistician* 45 (1): 97–104.
- Gibbard, Allan. 1973. "Manipulation of voting schemes: A general result." *Econometrica* 41 (4): 587–601.
- Ginsburgh, Victor A., and Jan C. van Ours. 2003. "Expert opinion and compensation: Evidence from a musical competition." *American Economic Review* 93 (1): 289–296.
- Gleijser, Herbert, and Bruno Heyndels. 2001. "Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth Music Contest." *Journal of Cultural Economics* 25 (2): 109–129.
- Hamlin, Alan. 2023. "The rule of rules." *Public Choice* 195: 231–250.
- Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

- Hood, Christopher. 2011. *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government*. Princeton, NJ: Princeton University Press.
- Kontek, Krzysztof, and Kevin Kenner. 2025. "Identifying outlier scores and outlier jurors to reduce manipulation in classical music competitions." *Journal of Cultural Economics* 49 (1): 49–98.
- Kontek, Krzysztof, and Honorata Sosnowska. 2020. "Specific tastes or cliques of jurors? How to reduce the level of manipulation in group decisions?" *Group Decision and Negotiation* 29: 1057–1084.
- Rousseeuw, Peter J., and Christophe Croux. 1993. "Alternatives to the median absolute deviation." *Journal of the American Statistical Association* 88 (424): 1273–1283.
- Satterthwaite, Mark Allen. 1975. "Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions." *Journal of Economic Theory* 10 (2): 187–217.
- Sosnowska, Honorata. 2017. "Comparison of voting methods used in some classical music competitions." *Transactions on Computational Collective Intelligence XXVII*: 108–117.
- Vanberg, Viktor J. 1994. *Rules and Choice in Economics: Essays in Constitutional Political Economy*. London: Routledge.
- Weaver, R. Kent. 1986. "The politics of blame avoidance." *Journal of Public Policy* 6 (4): 371–398.

Appendix A. Proof of Lemma 1

For each ordinary evaluator, let the two candidate scores be

$$S_{1j} = \theta_1 + \varepsilon_{1j}, \quad S_{2j} = \theta_2 + \varepsilon_{2j},$$

where

$$\theta_1 - \theta_2 = \Delta > 0,$$

and the score errors are independent with

$$\varepsilon_{1j}, \varepsilon_{2j} \sim N(0, \sigma_H^2).$$

The pairwise contrast

$$d_j = S_{1j} - S_{2j}$$

has mean Δ and variance $2\sigma_H^2$. The sum of n ordinary contrasts therefore has mean $n\Delta$ and variance $2n\sigma_H^2$, giving

$$L_n = \Phi\left(-\frac{\sqrt{n}\Delta}{\sqrt{2}\sigma_H}\right).$$

Let the additional sincere evaluator have independent per-score error variance σ_i^2 . Its pairwise contrast has mean Δ and variance $2\sigma_i^2$. With that evaluator retained,

$$L_{n+1} = \Phi\left(-\frac{(n+1)\Delta}{\sqrt{2n\sigma_H^2 + 2\sigma_i^2}}\right).$$

Because Φ is increasing, $L_{n+1} < L_n$ if and only if

$$\frac{n+1}{\sqrt{n\sigma_H^2 + \sigma_i^2}} > \frac{\sqrt{n}}{\sigma_H}.$$

Squaring and rearranging gives

$$(n+1)^2\sigma_H^2 > n(n\sigma_H^2 + \sigma_i^2),$$

hence

$$\sigma_i^2 < \left(2 + \frac{1}{n}\right)\sigma_H^2.$$

Equality gives equal decision loss; reversing the inequality makes retention of the additional evaluator harmful. This proves Lemma 1.

Appendix B. Proof of Proposition 2

Consider a sincere focal evaluator with

$$S_{ai} = \theta_a + \sigma_i Z_{ai}, \quad Z_{ai} \stackrel{iid}{\sim} N(0,1).$$

For the population benchmark, the candidate reference is $M_a = \theta_a$. Hence the EOJ profile distance is

$$JD_i = \sum_{a=1}^A |S_{ai} - \theta_a| = \sigma_i U_A,$$

where

$$U_A = \sum_{a=1}^A |Z_a|, \quad Z_a \stackrel{iid}{\sim} N(0,1).$$

Let

$$u_A = \text{Med}(U_A), \quad v_A = \text{MAD}(U_A),$$

and define

$$c_A = \frac{u_A}{v_A}. \quad (\text{B1})$$

For the ordinary population with error standard deviation σ_H , the population counterparts of the EOJ distance centre and scale are

$$MJD^{\text{pop}} = \sigma_H u_A, \quad MAD_{JD}^{\text{pop}} = \sigma_H v_A.$$

The corresponding population EOJ statistic for a sincere focal evaluator is therefore

$$z_i^{\text{pop}} = \frac{\sigma_i U_A - \sigma_H u_A}{\sigma_H v_A} = \frac{(\sigma_i / \sigma_H) U_A - u_A}{v_A}. \quad (\text{B2})$$

For each realised $U_A > 0$, expression (B2) is increasing in σ_i / σ_H . The distribution of z_i^{pop} therefore shifts outward as the evaluator's relative error scale increases. Because (B2) is an increasing affine transformation of U_A , its median is

$$\text{Med}(z_i^{\text{pop}}) = c_A \left(\frac{\sigma_i}{\sigma_H} - 1 \right). \quad (\text{B3})$$

At the decision-neutral noise ratio r_D from Lemma 1, (B3) gives

$$K_D^{\text{pop}} = c_A (r_D - 1),$$

which proves Proposition 2.

For $A = 6$, numerical evaluation gives

$$u_6 \approx 4.685, \quad v_6 \approx 0.996, \quad c_6 \approx 4.70.$$

Together with $r_D \approx 1.436$, this gives $K_D^{\text{pop}} \approx 2.05$.

The scorecard-width comparative static follows from the half-normal central-limit approximation. Since $E|Z| = \sqrt{2/\pi}$ and $\text{Var}(|Z|) = 1 - 2/\pi$, for moderate and large A , U_A is approximately normal with location proportional to A and dispersion proportional to $\sqrt{A}$. Using the normal MAD constant 0.67449 gives

$$c_A \approx 1.96\sqrt{A}. \quad (\text{B4})$$

Literal EOJ uses a finite self-inclusive panel rather than the population reference, so the focal JD_i , MJD , and MAD_{JD} are jointly determined from the same realised score matrix. There is therefore no unique finite-panel plug-in analogue of (17). In an all- H benchmark with $N = 17$ and $A = 6$, direct simulation of a decision-neutral focal evaluator gives a median literal-EOJ statistic of about 1.73 and a mean of about 2.10; alternative natural plug-in conventions place the corresponding location within roughly 1.7–2.3. The robust conclusion is qualitative: the decision-neutral region remains well below the PE5 cutoff near $K = 3$.

Appendix C. Robustness of the precision-to-distance mapping

A simple limiting configuration illustrates the masking issue noted in Section 2.2. Suppose $m < N/2$ evaluator distances equal $x > 0$ and the remaining $N - m$ distances equal zero. Under mean/population-standard-deviation standardisation, the mean is

$$\bar{x} = \frac{m}{N} x,$$

and the population standard deviation is

$$s_x = x \frac{\sqrt{m(N-m)}}{N}.$$

Hence the standardised distance of each anomalous profile is

$$z_{\text{out}}^{\text{SD}} = \frac{x - \bar{x}}{s_x} = \sqrt{\frac{N-m}{m}}. \quad (\text{C1})$$

The standardised value is therefore bounded independently of how large x becomes. For $N = 17$ and $m = 2$, (C1) equals approximately 2.74, so two arbitrarily distant profiles can remain below a threshold of $K = 3$ under this non-robust standardisation.

The zero-centre configuration above is used only to display the mean/standard-deviation masking calculation. For the median/MAD comparison, consider instead any non-degenerate central cloud of $N - m$ distances with positive MAD, together with a minority $m < N/2$ of anomalous distances that increase with x . The median centre and MAD remain anchored to the central cloud, while the anomalous numerator grows with x ; the corresponding robust standardised distance therefore diverges. The illustration is a special case of the standard masking logic motivating robust location and scale estimators; see Hampel et al. (1986) and Rousseeuw and Croux (1993).

Appendix D. Numerical benchmark

The benchmark uses $N = 17$ evaluators and $A = 6$ candidates. Ordinary score errors are independent Gaussian draws with per-score standard deviation $\sigma_H = 0.60$. Candidate locations may be written as

$$\theta = \left(\frac{\Delta}{2}, -\frac{\Delta}{2}, -1, -2, -3, -4 \right), \quad \Delta = 0.30.$$

Only the first two components determine the stylised binary winner problem; all six coordinates enter EOJ profile distances. State H uses the ordinary noise distribution. State F remains unbiased with $\sigma_F^2 = 3\sigma_H^2$. States M and G use observationally identical focal-score distributions with the two contested scores displaced by $(-q, +q)$, $q = 2$, while the benchmark winner is reversed in state G . The illustrative state probabilities are

$$(\pi_H, \pi_F, \pi_M, \pi_G) = (0.85, 0.03, 0.07, 0.05).$$

D.1. Aliased M/G benchmark

States M and G are constructed to generate observationally identical focal score profiles while carrying opposite decision value. Consequently,

$$z_M \stackrel{d}{=} z_G, \tag{D1}$$

and every common EOJ cutoff satisfies

$$\alpha_M(K) = \alpha_G(K). \tag{D2}$$

Where the common density is positive, Bayes' rule gives

$$\frac{P(M \mid z_i = z)}{P(G \mid z_i = z)} = \frac{\pi_M}{\pi_G}.$$

Exact aliasing therefore leaves the posterior M/G odds equal to the prior odds: greater extremity can strengthen the evidence that a profile is anomalous without identifying whether the latent state is harmful manipulation or valuable dissent. Under the symmetric benchmark, consequential-exclusion probabilities also coincide,

$$\chi_M(K) = \chi_G(K) \equiv \chi_{MG}(K). \tag{D3}$$

In the directional benchmark, a consequential exclusion improves the binary decision in state M and worsens it in state G whenever the focal contrast has the intended sign. Gaussian noise has unbounded support, so this statement is not literally exact: at $q = 2$, the opposite-sign event has probability about 6.5×10^{-6} . Ignoring this negligible sign-reversal event,

$$b_M(K) \simeq h_G(K) \simeq \chi_{MG}(K), \quad h_M(K) \simeq b_G(K) \simeq 0. \tag{D4}$$

Accordingly, to benchmark numerical accuracy, the combined contribution of the aliased M/G block to the decision-quality gain is

$$G_Q^{MG}(K) \simeq \chi_{MG}(K)(\pi_M - \pi_G). \tag{D5}$$

This decomposition is specific to the symmetric benchmark. It shows that screening severity cannot resolve the latent ambiguity; it only changes how often the common observable profile is retained or excluded. In this benchmark component, stronger screening improves expected decision quality when $\pi_M > \pi_G$, and the direction reverses when $\pi_G > \pi_M$.

For Proposition 2, numerical evaluation gives $u_6 \approx 4.685$, $v_6 \approx 0.996$, and $K_D^{\text{pop}} \approx 2.05$. As discussed in Appendix B, the literal finite-panel location depends on the summary convention; direct simulation gives a median near 1.73 and a mean near 2.10, with the qualitative conclusion that the decision-neutral region lies well below PE5.

PEX cutoffs are calibrated under literal full-panel EOJ from all- H panels, with each evaluator's distance entering the cross-evaluator median and MAD. Table D1 reports representative mappings for selected panel

sizes, scorecard widths, and PEx levels. Values are empirical quantile approximations rounded to two decimals. The reported PEx lookup values are benchmark-specific calibration values, not distribution-free critical values.

Table D1. Approximate $K_{\text{PEx}}(N, A)$ under all- H calibration

N	A	PE3	PE5	PE7	PE9
5	4	5.64	4.08	3.21	2.64
5	6	5.51	3.95	3.12	2.56
5	20	5.15	3.74	2.96	2.43
7	4	4.63	3.57	2.93	2.49
7	6	4.52	3.48	2.86	2.43
7	20	4.28	3.29	2.72	2.32
9	4	4.21	3.35	2.81	2.42
9	6	4.08	3.26	2.74	2.36
9	20	3.90	3.10	2.61	2.26
11	4	4.02	3.22	2.73	2.38
11	6	3.89	3.14	2.67	2.33
11	20	3.70	2.99	2.55	2.23
13	4	3.86	3.14	2.69	2.35
13	6	3.74	3.07	2.63	2.30
13	20	3.56	2.92	2.51	2.20
17	4	3.69	3.05	2.63	2.32
17	6	3.59	2.97	2.57	2.27
17	20	3.41	2.83	2.46	2.18

Table 3 and Figure 2 use the same common-random-number benchmark run with 1,000,000 realisations per focal-state block. In that run, $G_Q(2.2) \approx 0.458$ percentage points, $G_Q(2.0) \approx 0.452$, and $G_Q(3.0) \approx 0.437$. The paired decision-loss difference between $K = 2.2$ and $K = 3$ is approximately 0.021 percentage points with Monte Carlo standard error 0.004; the corresponding difference in consequential M/G exclusion is approximately 3.03 percentage points with standard error 0.02. Independent runs place the grid maximum at nearby values within the broad low-loss region, so the manuscript reports $K \approx 2.2$ as a representative centre and does not use local differences inside the broad low-loss region to estimate precise institutional-cost thresholds.

Appendix E. Derivation and endpoint behaviour of potential pivotality

Let

$$X = D^{-i}$$

be the aggregate score contrast of the $N - 1$ peers. Under the Gaussian benchmark,

$$X \sim N((N - 1)\Delta, 2(N - 1)\sigma_H^2). \quad (\text{E1})$$

The focal evaluator's contrast under the displacement $(-q, +q)$ is

$$d_i = \Delta - 2q + \eta_i, \quad \eta_i \sim N(0, 2\sigma_H^2), \quad (\text{E2})$$

independently of X . Hence

$$Y = D^{+i} = X + d_i$$

satisfies

$$Y \sim N(N\Delta - 2q, 2N\sigma_H^2). \quad (\text{E3})$$

The correlation between X and Y is

$$\rho = \sqrt{\frac{N-1}{N}}. \quad (\text{E4})$$

Define the standardised zero thresholds

$$a = -\frac{\sqrt{N-1}\Delta}{\sqrt{2}\sigma_H}, \quad b(q) = \frac{2q - N\Delta}{\sqrt{2N}\sigma_H}. \quad (\text{E5})$$

Potential pivotality is

$$P_{\text{piv}}(q) = P(X < 0, Y > 0) + P(X > 0, Y < 0).$$

Using the bivariate normal distribution function,

$$P(X < 0, Y > 0) = \Phi(a) - \Phi_2(a, b(q); \rho),$$

and

$$P(X > 0, Y < 0) = \Phi(b(q)) - \Phi_2(a, b(q); \rho).$$

Therefore

$$P_{\text{piv}}(q) = \Phi(a) + \Phi(b(q)) - 2\Phi_2(a, b(q); \rho). \quad (\text{E6})$$

Substituting $N = 17$, $\Delta = 0.30$, and $\sigma_H = 0.60$ gives the closed-form pivotality curve used in Figure 1. The expected full-panel contrast is zero when

$$N\Delta - 2q = 0,$$

so the balance point is

$$q_B = \frac{N\Delta}{2} = 2.55.$$

Endpoint behaviour

At $q = 0$, the focal evaluator has no systematic directional displacement, but ordinary score noise remains. The evaluator can therefore still be pivotal when the full-panel and remaining-panel contrasts fall on opposite sides of zero. In the benchmark, potential pivotality at this endpoint is about 2.8%. The EOJ statistic is also random, so under $K = 3$ some otherwise ordinary profiles cross the screening boundary at approximately the PE5 rate; consequential exclusion is rarer because both exclusion and pivotality must occur together.

At the opposite endpoint, even an arbitrarily large focal displacement is not necessarily pivotal. As $q \rightarrow \infty$, the retained focal evaluator favours candidate 2 so strongly that candidate 2 wins whenever the evaluator is retained. Candidate 2 also wins in those realisations where the remaining panel already favours it. Exclusion changes the winner only when the remaining panel favours candidate 1. Hence

$$\lim_{q \rightarrow \infty} P_{\text{piv}}(q) = P(D^{-i} > 0) = \Phi\left(\frac{\sqrt{N-1}\Delta}{\sqrt{2}\sigma_H}\right). \quad (\text{E7})$$

For the benchmark parameters,

$$\lim_{q \rightarrow \infty} P_{\text{piv}}(q) = \Phi(\sqrt{2}) \approx 0.921.$$

Thus the pivotality curve need not start at zero and does not converge to one: ordinary noise can make a non-displaced evaluator pivotal, while an arbitrarily extreme evaluator is irrelevant whenever the remaining panel already favours the same alternative.

Appendix F. From pairwise loss to broader decision outcomes

The binary benchmark is a building block rather than a claim that expert competitions care only about the identity of the winner. Let $E_{m,m'}$ denote reversal of the benchmark ordering between alternatives m and m' , and define

$$L_{m,m'} = P(E_{m,m'}).$$

If ranking loss is measured by normalised Kendall distance,

$$L_{\text{rank}} = \frac{2}{A(A-1)} \sum_{m < m'} L_{m,m'}. \quad (\text{F1})$$

If m^* is the benchmark winner,

$$P(\text{benchmark-inconsistent winner}) \leq \sum_{j \neq m^*} L_{m^*,j}. \quad (\text{F2})$$

If F_k^* is the benchmark top- k set,

$$P(\text{benchmark-inconsistent top-}k) \leq \sum_{m \in F_k^*} \sum_{j \notin F_k^*} L_{m,j}. \quad (\text{F3})$$

These relations connect the focal pairwise building block to ranking, winner, and shortlist decisions without requiring the main text to impose a unique objective artistic ranking over the full field. Even when no pairwise reversal occurs, removing a focal evaluator can change score levels and margins. Accordingly, “non-consequential” in the main binary benchmark means only that the selected winner between candidates 1 and 2 is unchanged.

Appendix G. Proof of Proposition 4

Let $f_T(z)$ denote the density of the EOJ statistic in latent state T at regular points, and let

$$f(z) = \sum_T \pi_T f_T(z) \quad (G1)$$

be the corresponding mixture density. Conditional on state T and realised statistic z , let $\ell_R(T, z)$ and $\ell_E(T, z)$ denote expected decision loss under retention and exclusion, and define

$$D_T(z) = \ell_R(T, z) - \ell_E(T, z). \quad (G2)$$

The posterior state probability at z is

$$\rho_T(z) = P(T | z) = \frac{\pi_T f_T(z)}{f(z)}. \quad (G3)$$

Hence the conditional marginal advantage of exclusion introduced in Section 3.7 can be written as

$$D_Q(z) = \sum_T \rho_T(z) D_T(z). \quad (G4)$$

Under a hard threshold K , profiles with $z > K$ are excluded and profiles with $z \leq K$ are retained. At a regular point of the distribution, increasing K by an infinitesimal amount changes treatment only for profiles at the boundary. Therefore

$$L_Q'(K) = \sum_T \pi_T f_T(K) D_T(K) = f(K) D_Q(K). \quad (G5)$$

At an interior optimum K_Q^* with $f(K_Q^*) > 0$, the first-order condition $L_Q'(K_Q^*) = 0$ implies

$$D_Q(K_Q^*) = 0,$$

which proves Proposition 4. The result applies at regular interior points. When the EOJ statistic has an atom, the derivative representation in (G5) does not apply at that point and the relevant comparison is the change in expected loss across the discontinuity. $\square$

Appendix H. Proof of Proposition 5 and Corollary 1

Under state M , a decision-changing focal evaluator is either excluded or retained. In the symmetric aliased benchmark, consequential exclusion occurs with probability $\chi_{MG}(K)$, while potential pivotality is independent of the screening threshold. Hence the probability of consequential retained manipulation is

$$P_{\text{piv},M} - \chi_{MG}(K).$$

Under state G , consequential exclusion occurs with the same probability $\chi_{MG}(K)$. The organiser's expected total loss can therefore be written as

$$\mathcal{L}_O(K) = C_Q L_Q(K) + \pi_M C_M [P_{\text{piv},M} - \chi_{MG}(K)] + \pi_G C_G \chi_{MG}(K). \quad (H1)$$

The term $\pi_M C_M P_{\text{piv},M}$ does not depend on K . Dropping this constant gives

$$K_O^* \in \arg \min_{K \geq 0} \{C_Q L_Q(K) - (\pi_M C_M - \pi_G C_G) \chi_{MG}(K)\}, \quad (H2)$$

which proves Proposition 5.

For Corollary 1, recall that K_Q^* is defined in (37) as a global minimiser of L_Q , so $\Delta_Q(K) \geq 0$ for every admissible K . Subtract the objective in (H2) evaluated at K_Q^* from the objective at K . Using $\Delta_Q(K)$ and $\Delta_\chi(K)$, the difference is

$$C_Q \Delta_Q(K) - (\pi_M C_M - \pi_G C_G) \Delta_\chi(K). \quad (H3)$$

Threshold K is preferred exactly when (H3) is negative, yielding Corollary 1.

The benchmark decision-loss surface is too flat around its centre to support stable numerical cost ratios for moves among neighbouring cutoffs such as $K = 2, 2.2$, and 2.3 . Those local ratios are therefore deliberately not reported. Numerical cost thresholds become informative for departures large enough to generate decision-loss differences clearly resolved by the simulation.

As a sensitivity illustration of the reverse, Galileo-dominant direction, consider the move from $K = 2.2$ to $K = 3$. The common-random-number run gives

$$\Delta_Q(3) \approx 0.00021 \text{ (= 0.021 pp; MC SE 0.004 pp)}, \quad \chi_{MG}(2.2) - \chi_{MG}(3) \approx 0.03031. \quad (H4)$$

If $C_G = 2C_M$, then $\pi_M C_M - \pi_G C_G = -0.03C_M$, and Corollary 1 implies

$$\frac{C_M}{C_Q} > \frac{0.00021}{0.03 \times 0.03031} \approx 0.23. \quad (H5)$$

In the limiting case in which manipulation exposure is negligible relative to Galileo exposure, the condition becomes

$$\frac{C_G}{C_Q} > \frac{0.00021}{0.05 \times 0.03031} \approx 0.14. \quad (H6)$$

These calculations illustrate a departure that is large enough to be resolved numerically; they are not intended as universal institutional cost ratios.