Public Choice
https://doi.org/10.1007/s11127-026-01444-z

Fairness trade-offs in expert evaluation: information, judgment, and procedural design

Karol J. Borowiecki[1] · **Marc T. Law**[2]

Received: 15 June 2026 / Accepted: 16 July 2026

Abstract
Evaluation systems often seek to reduce bias, prevent manipulation, preserve relevant information, maintain comparability, and treat candidates equally, but these fairness goals can conflict. This paper develops a framework for understanding such trade-offs in expert evaluation—including peer review, grant panels, academic hiring, and artistic competitions—where quality is multidimensional and legitimate disagreement among evaluators is possible. We treat the choice among evaluation procedures as a problem of constitutional design: rules must be selected before the specific candidates, evaluators, and circumstances that will test them are known. Classical music competitions are our primary analytical setting because their rules, scores, and controversies make these trade-offs unusually visible, but the framework applies to other settings in which expert judgments are aggregated under incomplete information. The framework highlights two core tensions: information control and outlier rules. First, when information is both quality-relevant and identity-revealing, procedures that suppress bias-relevant cues may also remove information with genuine evaluative content. Second, because sincere minority judgment and strategic manipulation can generate similar observed score patterns, score-based exclusion rules may reduce manipulation while limiting legitimate dissent. Two further trade-offs concerning transparency and standardization are developed as illustrations. The analysis shows that no procedure can eliminate all vulnerabilities at once. Beyond this design implication, the framework also helps explain why evaluation rules differ across stages and settings, and why well-intentioned reforms may create new vulnerabilities.

Keywords Constitutional design · Information aggregation · Fairness · Committee decision-making · Competitions · Bias · Manipulation

JEL Classification D02 · D63 · D71 · D72 · D82 · Z11

✉ Karol J. Borowiecki
kjb@sam.sdu.dk

Marc T. Law
marc.law@uvm.edu

[1] Department of Economics, University of Southern Denmark, Odense, DK, Denmark

[2] Department of Economics, University of Vermont, Burlington, VT, USA

Published online: 04 August 2026

"The central values by which most men have lived ... are not always harmonious with each other."

—Isaiah Berlin, "The Pursuit of the Ideal".

1 Introduction

Evaluation systems are often expected to satisfy multiple fairness objectives simultaneously. They should identify high-quality candidates, reduce bias, prevent manipulation, preserve independent judgment, maintain comparability, and treat candidates equally. These goals are all defensible but not always mutually compatible.

This problem arises in many expert-evaluation settings. Academic hiring committees, grant panels, peer review systems, prize committees, and artistic competitions all rank candidates, projects, or performances whose quality is multidimensional and only partially observable. In such settings, rules designed to make evaluation fairer may impose costs. Blinding limits exposure to bias-relevant information, but can also remove information relevant to assessment. Outlier rules discourage strategic scoring, yet may also exclude dissenting expert judgment. Transparency makes decisions easier to scrutinize, while changing how evaluators express their views. Standardized tasks improve comparability only by narrowing the dimensions of quality that can be observed.

These procedural choices are therefore not merely technical adjustments. They are rules for collective decision-making under uncertainty: they determine what information enters the process, how disagreement is treated, and which forms of strategic behavior remain possible. In this sense, choosing evaluation procedures is a problem of constitutional design: rules must be selected before the specific candidates, evaluators, and circumstances that will test them are known, and should be assessed by their general properties rather than by whether they produce the desired outcome in any particular case (Buchanan and Tullock 1962; Brennan and Buchanan 1985; Buchanan 1987; Vanberg 1994). This places the analysis within the constitutional political economy tradition, which treats rules themselves—their operating characteristics, enforcement, and supply—as central objects of analysis (Hamlin 2023).

Much of the existing literature evaluates procedural reforms according to whether they reduce particular vulnerabilities in evaluation systems. Blind auditions and double-blind review are often advocated to reduce gender, prestige, or identity-related bias (Goldin and Rouse 2000; Tomkins et al. 2017; Brodie et al. 2021). Open-review procedures are frequently defended on grounds of transparency, accountability, and openness (e.g., Ross-Hellauer 2017), while the literature on judged sports examines how scoring transparency and judge anonymity affect favoritism and strategic judging in expert evaluation (e.g., Zitzewitz, 2014). Statistical procedures for identifying anomalous evaluators or scores are increasingly proposed as tools for limiting strategic manipulation and conflicts of interest in expert-evaluation settings (Kontek and Kenner 2025), and structured evaluation procedures are frequently advocated as mechanisms for improving comparability and consistency across candidates (Levashina et al. 2014). These studies identify weaknesses in existing evaluation systems and show how targeted reforms may address them.

This paper develops a framework for understanding why procedural reforms may improve one dimension of fairness while weakening another, shifting attention from the vulnerability a reform is designed to reduce to the trade-offs it may introduce. Classical music competitions provide the paper's primary analytical setting because they offer clear, well-documented examples of these trade-offs. Major events such as the Chopin, Queen Elisabeth, Van Cliburn, Busoni, and Leeds competitions combine features that make fairness both important and difficult to secure. Prizes and rankings can affect careers and reputations (Ginsburgh and Van Ours 2003) and performances, jury decisions, and competition rules are increasingly scrutinized by candidates, audiences, media, and online communities (McCormick 2015). Yet artistic quality is multidimensional, and expert disagreement is not merely a nuisance to be eliminated.[1]

Similar problems arise in other expert-evaluation settings. Academic hiring committees consider job market papers, presentations, research agendas, publication potential, teaching ability, recommendation letters, and departmental fit. Grant panels and peer reviewers likewise weigh methodological rigor, feasibility, novelty, and broader significance. In each case, evaluators aggregate multidimensional judgments under incomplete information and legitimate disagreement.

In music competitions, disagreement among jurors may reflect genuine differences in artistic judgment, but it may also reflect bias, error, or strategic behavior. The 1980 International Chopin Competition in Warsaw offers a vivid illustration. Ivo Pogorelich's unconventional interpretations sharply divided the jury and led to his elimination before the final round. Martha Argerich resigned from the jury in protest, and the episode became internationally known as the "Pogorelich Affair" (McCormick 2018). The affair shows not simply that juries can be wrong, but that when artistic standards are genuinely contested, the same score pattern may look like manipulation or error to one observer and sincere minority judgment to another.

Recent concerns over gender disparities in piano competitions illustrate a different trade-off. As of 2024, men had won 82 percent of the 40 most recent piano competitions affiliated with the World Federation of International Music Competitions (WFIMC), and more than a third had all-male finals (Alberge 2024; Sinclair 2024). These concerns have contributed to policy discussions within the WFIMC concerning competition design (Asmat et al. 2023a). In response, the Leeds Competition has introduced blind pre-selection procedures and restrictions on biographical information available to jurors. These reforms reduce exposure to bias-relevant cues but may also limit the information available to jurors. Whether they improve fairness overall or merely trade one vulnerability for another is precisely the question this paper addresses.

The framework formalizes two core tensions illustrated by these examples. The first concerns information control: if information relevant to evaluation also reveals candidate identity, procedures designed to reduce identity-related bias may remove artistically relevant information. This trade-off arises even when jurors share a common standard of quality. The second concerns outlier rules. If sincere minority artistic judgment and strategic manipulation can generate similar observed score patterns, score-based exclusion rules may

[1] This difficulty may also help explain why competition outcomes do not always align with later assessments of artistic quality. Winning a prize reflects performance as assessed through particular rules, information constraints, and evaluative standards, which may diverge from the dimensions of quality associated with long-run artistic achievement (Ginsburgh 2003; Ginsburgh and Weyers 2014).

reduce manipulation while limiting legitimate dissent. This trade-off arises when artistic disagreement reflects differences in judgment that the evaluation process is meant to accommodate. The paper shows why both conditions are plausible and traces their implications for evaluation design.

The same logic extends to two further design trade-offs.[2] First, transparency may improve accountability but weaken independent artistic judgment by exposing jurors to reputational pressure. Second, standardized evaluation requirements may improve comparability across candidates but narrow the range of qualities they can demonstrate. These trade-offs depend on additional behavioral and contextual assumptions, and are developed as illustrations rather than formal results.

This does not imply that meaningful fairness is unattainable: rules matter because they reduce particular vulnerabilities. The central point is that, under the conditions identified here, no procedure can eliminate them all at once. Like other problems of constitutional design, the task is not to identify a procedure that is optimal under all conditions, but to choose among imperfect rules: which risks to reduce, which forms of information or discretion to preserve, and which values to prioritize given the purpose of the evaluation. The broader implication is familiar from constitutional political economy: well-intentioned reforms may work against their purpose, not because designers act in bad faith, but because the information and incentive structures of the decision environment constrain what any procedure can achieve.[3]

The framework is diagnostic rather than comprehensive: it does not provide a complete theory of music competitions, or expert evaluation. Instead, under explicit and defensible assumptions, it identifies the structural features that make different dimensions of fairness difficult to secure simultaneously. The models are simple by design and yield conditional rather than general results: they show when familiar reforms generate trade-offs, what features of the evaluation environment produce them, and why those features are plausible in the settings under study. This matters because the problems prompting new rules may be more visible than the trade-offs those rules introduce.

The remainder of the paper proceeds as follows. Section 2 situates the argument in the literature on social choice, information aggregation, and expert evaluation, focusing on bias, anonymization, and outlier scoring in artistic competitions. Section 3 examines the trade-off between information control and bias reduction; Section 4, the tension between outlier rules and artistic pluralism; and Section 5, the application of these trade-offs to transparency and repertoire standardization. Section 6 draws implications for competition design and expert evaluation more broadly, including peer review, grant panels, and academic hiring. It then shows how these trade-offs help explain variation in evaluation rules across stages and settings, before discussing sortition and the political economy of procedural reform. Section 7 concludes.

[2] The framework may also apply to other procedural dimensions of evaluation not explicitly analyzed here, including jury composition, discussion versus anonymous voting, speaking order during deliberation, performance order, and other procedural features that shape how information is revealed, aggregated, and interpreted during evaluation.

[3] Examples include balanced-budget rules and constitutional limits on taxation. The former may constrain the political incentive to finance current spending through future taxes, but may also limit fiscal flexibility in wars or other emergencies (Buchanan and Wagner 1977). The latter may reduce the scope for fiscal exploitation, but they can also encourage substitution toward less transparent instruments, including regulation, mandates, or off-budget forms of redistribution (Brennan and Buchanan 1980; Posner 1971).

2 Related literature and conceptual background

2.1 Expert evaluation, aggregation, and procedural design

Expert-evaluation systems must aggregate heterogeneous judgments when quality is multidimensional and information incomplete. These conditions raise broader questions of design: how to aggregate judgments, limit strategic behavior, and structure procedures when fairness objectives conflict. The aggregation problem is compounded by the fact that the information relevant to evaluation is itself distributed across evaluators: no single designer or procedure can capture the dispersed and often tacit knowledge that individual experts bring to their judgments (Hayek 1945).

A natural starting point is the social choice literature on preference aggregation. Arrow's impossibility theorem shows that no social welfare rule can satisfy a set of seemingly reasonable axioms when aggregating individual rankings over three or more alternatives (Arrow 1951). The Gibbard-Satterthwaite theorem further shows that, under broad conditions, non-dictatorial voting rules are vulnerable to strategic manipulation (Gibbard 1973; Satterthwaite 1975). Together, these results establish general limits on collective decision rules and imply that evaluation procedures necessarily involve trade-offs among desirable properties.

A second strand examines when voting rules successfully aggregate dispersed private information into correct collective decisions. The Condorcet jury theorem shows that, under suitable independence conditions, majority voting aggregates private signals and large juries converge on the correct outcome (Ladha 1992, 1993). Subsequent work shows that strategic behavior can undermine these properties: sincere voting is generally not a Nash equilibrium (Austen-Smith and Banks 1996), and rules intended to prevent one kind of evaluative error may worsen another through strategic responses (Feddersen and Pesendorfer 1997, 1998).

Our approach differs from these classical results in two respects. First, unlike the Condorcet framework, there is no uniquely correct ranking to recover; quality is multidimensional and evaluators may legitimately disagree. Second, the trade-offs identified here arise before final aggregation, from constraints on what information is observable and how score patterns can be interpreted. These are problems no choice of voting rule can resolve.

2.2 Music competitions: voting rules, information, and bias

A growing literature shows that music-competition outcomes depend not only on performance quality but also on rules, information, and juror behavior. Studies of the Queen Elisabeth Competition find that performance order and gender affect scores (Flôres and Ginsburgh 1996; Glejser and Heyndels 2001), while work on the Chopin Competition emphasizes prior information, performance order, and halo effects (Chmurzyńska 2015). Together, these studies show that evaluation rules, the information available to jurors, and competition structure can all influence outcomes.

A related strand applies these concerns directly to music competitions. Sosnowska (2017) compares voting methods used in classical music competitions, while Kontek and Sosnowska (2020) show how strategic voting can affect outcomes under the Borda count in the Wieniawski Violin Competition. Kontek and Kenner (2025) propose procedures for identifying outlier scores and jurors. Their Excluding Outlier Jurors (EOJ) method uses

median-based statistics to identify jurors whose scoring patterns deviate markedly from the jury's central tendency and excludes their votes from the final aggregation. These contributions are especially relevant because they treat fairness as a problem of aggregation and manipulation. They also raise the question illustrated by EOJ itself: when does an unusual score pattern reflect manipulation, and when does it reflect sincere disagreement about artistic quality? Section 4 shows that any score-based exclusion rule—including EOJ—faces a trade-off between eliminating manipulation and preserving sincere minority judgment.

A related set of papers examines who evaluates performances and how evaluators differ. Asmat et al. (2023b) compare expert and lay evaluations in international classical music competitions and show that the two groups do not always rank competitors alike. Artistic disagreement, then, need not reflect only noise or error; it may also arise from differences in expertise, information, or evaluative criteria. In a related study of international piano competitions, Asmat et al. (2024) find that increasing the share of female jurors does not necessarily improve outcomes for female competitors and may sometimes have the opposite effect. The broader lesson is that changing jury composition may affect evaluation, but not always as a simple representation argument would predict.

2.3 Information control and expert evaluation

A separate literature studies information control and bias. Goldin and Rouse (2000) show that screens in orchestral auditions increased women's chances of advancement and hiring. Although music competitions differ from orchestral hiring, blind or audio-only pre-selection applies a similar principle. Recent reforms at the Leeds International Piano Competition, including blind pre-selection and reduced biographical information, illustrate how organizers have adapted this approach in response to gender disparities in piano competitions (Sinclair 2024).

The difficulty of separating quality-relevant from identity-revealing information extends beyond the arts. In peer review, Blank (1991) shows that double-blind reviewing affected acceptance rates in ways that varied with the prestige of the author's university affiliation. This is consistent with the view that identity-revealing cues can shape evaluation, whether because they introduce bias, affect perceived quality, or proxy for information evaluators believe to be relevant. Related experimental evidence shows that observable identity can introduce bias directly: male peer graders assign lower scores to female-authored work even when an independent blind assessment of the same work is available for comparison (Saygin and Knight 2026). Similar tensions arise in grant evaluation and academic hiring, where track records, affiliations, advisor identities, and networks may inform assessments of quality while exposing evaluators to bias-relevant cues.

A growing body of evidence shows that fairness reforms can have unexpected effects. Studies of scientific and academic hiring committees show that changing committee composition does not mechanically eliminate bias and, in some settings, can produce less favorable outcomes for female candidates (e.g., Bagues and Esteve-Volart 2010; Bagues et al. 2017; Deschamps, 2024). Experimental evidence from music composition likewise cautions against simple predictions about identity cues and procedural reform: Ting et al. (2026) find no pro-male bias in composition evaluation, and structured evaluation does not eliminate the observed effects of gendered names. These findings reinforce the broader point that fair-

ness reforms should be judged not only by whether they reduce a targeted vulnerability, but also by the information, discretion, or disagreement they suppress.

A related literature on information design asks which information structures are available to a designer and what outcomes they can support (e.g., Kamenica and Gentzkow 2011; Bergemann and Morris 2019). Our question is different. Rather than ask what a designer can achieve by choosing among feasible information structures, we ask what constraints the evaluation environment imposes on that choice. Section 3 formalizes one such constraint. When quality-relevant and identity-revealing information are bundled, suppressing identity cues may also suppress information relevant to evaluation. The resulting trade-off is therefore a feature of the evaluation setting, not simply a design variable that organizers can optimize away.

2.4 From piecemeal reforms to fairness trade-offs

This literature shows that rules matter. Competition outcomes can depend on performance order, the information available to jurors, jury composition, voting rules, and procedures for identifying anomalous scores. Existing work has identified specific vulnerabilities in competition design. Building on this work, we ask a different question: when can fairness goals be achieved together, and when do they conflict? Our contribution is also distinct from the classical impossibility results of Arrow and Gibbard-Satterthwaite, which concern the aggregation of given preferences. The trade-offs identified here arise earlier in the evaluation process, from constraints on observable information and the interpretation of score patterns—problems that precede aggregation and that no choice of voting rule can resolve.

Table 1 illustrates the framework's broader applicability. The analogies are not exact: music competitions often rely on formal scores and public rankings, while peer review, grant panels, and academic hiring rely more heavily on reports, deliberation, and informal judgment. Across these settings, however, rules intended to improve one dimension of fairness may weaken another.

3 Information control and bias

We begin with the first tension identified above: information that helps evaluators assess quality may also reveal characteristics that should not affect evaluation. The problem is not simply that evaluators may be biased; it is that bias-relevant cues and quality-relevant information may be bundled together. This difficulty can arise even under a common evaluative benchmark. When the two kinds of information cannot be separated, removing identity cues may also remove information relevant to assessment, while preserving fuller information may expose jurors to bias-relevant cues. The problem is familiar from debates over blind auditions, anonymous applications, and the suppression of biographical information.

The following model formalizes this trade-off in a deliberately simple way. The result should be read as a conditional one: if quality-relevant and bias-relevant information cannot be fully separated, then a procedure cannot simultaneously use all relevant information and eliminate all exposure to bias-relevant cues.

Consider a competition with candidates $i = 1, \ldots, N$ and jurors $j = 1, \ldots, J$. Candidate i's quality q_i is a function of two components,

Table 1 Fairness trade-offs across expert-evaluation settings

Setting	Information control	Outlier evaluations	Transparency	Standardization
Music competitions	Blinding may reduce bias exposure but suppress visible dimensions of performance	Unusual scores may reflect manipulation or sincere minority judgment	Published scores may improve accountability but encourage conformity	Prescribed repertoire improves comparability but narrows what can be observed
Peer review	Double-blind review suppresses identity, affiliation, and reputation cues	Unusually positive or negative reports may reflect expertise, idiosyncrasy, or bad faith	Open or signed review may improve accountability but discourage candor	Methodological requirements improve comparability but may disadvantage unconventional work
Grant panels	Anonymization may reduce prestige and network cues but obscure capacity to execute	Divergent assessments may reflect minority insight, idiosyncrasy, or conflicts of interest	Public disclosure may reduce arbitrariness but discourage support for risky proposals	Deliverable requirements improve comparability but may favor incremental work
Academic hiring	Blinding is difficult because quality, networks, and pedigree are intertwined	Strong support or opposition may reflect judgment, personal ties, or strategic advocacy	Transparent procedures may improve accountability but discourage candid reservations	Narrow criteria improve comparability but may miss broader strengths

Notes: The paper's formal framework is developed using music competitions as the primary setting. The other settings are illustrative applications. The outlier-evaluation dimension maps most directly to environments with formal scoring, while in peer review, grant panels, and academic hiring similar tensions often arise through reports, deliberation, or informal committee judgment

$$q_i = (x_i, z_i). \tag{1}$$

The component $x_i \in \mathbb{R}$ denotes qualities that can be evaluated under blind conditions. In a music competition, these may include intonation, timing, phrasing, and technical control. The component $z_i \in \mathbb{R}$ denotes qualities that require fuller information. These may

include stage presence, physical expressiveness, visual dimensions of performance, or other elements observable only when the performer is seen or when contextual information is available.[4]

Candidate i also has an identity or affiliation signal, $g_i \in \{0, 1\}$. This signal may represent gender, nationality, race, conservatory affiliation, teacher lineage, prior reputation, or another characteristic that may be salient to jurors.

The central assumption is that some quality-relevant information is bundled with identity-relevant information.

Assumption 1 *(Bundled information)* The component z_i is quality-relevant. A procedure that allows jurors to observe z_i also reveals, or permits inference about, g_i.

This assumption captures a common feature of artistic evaluation. Live performance may reveal stage presence and physical expressiveness, but it also reveals gender, race, age, or other visible characteristics. Biographical information may reveal training and professional experience, but it may also reveal nationality, conservatory pedigree, or teacher networks. The binary formulation is stark, but it isolates the core problem: some information may be valuable for evaluation and problematic for fairness at the same time.

Let the full quality of candidate i be

$$Q_i = x_i + z_i. \tag{2}$$

This normalization is not essential. The important point is that z_i is relevant to evaluation, so that a procedure that suppresses z_i does not use all information relevant to assessing quality.

Suppose juror j's score for candidate i is

$$s_{ij} = x_i + \lambda z_i + \beta_j g_i + \varepsilon_{ij}, \tag{3}$$

where $\lambda = 1$ if z_i is observed and $\lambda = 0$ otherwise. The parameter β_j captures juror j's bias associated with the identity or affiliation signal g_i, and ε_{ij} is an idiosyncratic evaluation error.[5]

Scores are aggregated by their mean:

$$S_i = \frac{1}{J} \sum_{j=1}^{J} s_{ij}. \tag{4}$$

Let $\bar{\beta} = (1/J) \sum_{j=1}^{J} \beta_j$ and $\bar{\varepsilon}_i = (1/J) \sum_{j=1}^{J} \varepsilon_{ij}$.

[4] A similar information problem arises in other expert-evaluation settings. In peer review, x_i corresponds to the quality of the argument and evidence as assessed from the manuscript alone, while z_i corresponds to author reputation and track record. In grant evaluation, x_i corresponds to the scientific content of the proposal itself, while z_i corresponds to the research team's prior record, institutional affiliation, and demonstrated capacity to execute the project. In academic hiring, x_i corresponds to the quality of the job market paper on its own merits, while z_i corresponds to the candidate's advisor identity, training, and departmental pedigree.

[5] The formal argument abstracts from realized idiosyncratic evaluation error. Equivalently, one may set $\varepsilon_{ij} = 0$, or interpret the result as applying to expected scores when ε_{ij} has mean zero. The trade-off does not depend on stochastic error; it arises from the bundling of z_i and g_i.

Under full-information evaluation, jurors observe x_i, z_i, and, by Assumption 1, g_i. Aggregate scores are therefore

$$S_i^F = x_i + z_i + \bar{\beta} g_i + \bar{\varepsilon}_i. \tag{5}$$

Under blind evaluation, jurors observe x_i but not z_i or g_i. Aggregate scores are

$$S_i^B = x_i + \bar{\varepsilon}_i. \tag{6}$$

The competition selects a candidate with the highest aggregate score.

3.1 A formal trade-off

We now define four desirable properties of an evaluation procedure.

Definition 1 *(Full relevant information)* An evaluation procedure satisfies full relevant information if scores condition on all quality-relevant components of performance, including both x_i and z_i.

This definition requires that jurors observe and use z_i—the quality-relevant component that requires fuller information—in forming their scores.

Definition 2 *(Identity-blindness)* An evaluation procedure satisfies identity-blindness if jurors are not exposed to information that reveals, or permits inference about, g_i.

This definition captures blinding as an information-control procedure: jurors cannot use g_i because the procedure prevents them from observing information that reveals it.

Definition 3 *(Correct full-information selection)* An evaluation procedure satisfies correct full-information selection if, for every candidate profile, the selected candidate i^* satisfies $i^* \in \arg\max_i Q_i$.

This definition sets the benchmark for evaluative accuracy: a procedure meets it if it always selects the candidate with the highest full evaluative quality, using all quality-relevant information including z_i.

Definition 4 *(Bias immunity)* An evaluation procedure satisfies bias immunity if the selected candidate is invariant to the values of β_j.

This definition captures the fairness requirement that the outcome should not depend on how biased individual jurors are toward candidates with a particular identity signal g_i. A procedure satisfies it if changing the β_j values—making jurors more or less biased—leaves the selected candidate unchanged.

Proposition 1 *(Information-control trade-off) Suppose Assumption 1 holds. Then an evaluation procedure cannot simultaneously satisfy full relevant information, identity-blindness, correct full-information selection, and bias immunity.*

Proof Suppose first that the procedure satisfies full relevant information. Then jurors observe z_i. By Assumption 1, observing z_i reveals, or permits inference about, g_i. If $\beta_j \neq 0$ for some jurors, then aggregate scores under full information are

$$S_i^F = x_i + z_i + \bar{\beta} g_i + \bar{\varepsilon}_i. \tag{7}$$

Thus, whenever $\bar{\beta} \neq 0$, aggregate scores depend on g_i. For some candidate profiles, this dependence changes the ranking of candidates. The selected candidate may therefore vary with the values of β_j, violating bias immunity.

Now suppose instead that the procedure satisfies identity-blindness by suppressing information that reveals g_i. By Assumption 1, this requires suppressing z_i. Scores are then based on

$$S_i^B = x_i + \bar{\varepsilon}_i. \tag{8}$$

But since full evaluative quality is $Q_i = x_i + z_i$, there may exist candidates i and k such that

$$\begin{aligned} x_i &> x_k, \\ x_i + z_i &< x_k + z_k. \end{aligned} \tag{9}$$

Ignoring idiosyncratic error for simplicity, blind evaluation selects candidate i, while candidate k has higher full evaluative quality. Hence the procedure violates correct full-information selection.

Thus a procedure that uses full relevant information cannot guarantee bias immunity, while a procedure that guarantees identity-blindness cannot guarantee correct full-information selection. Under Assumption 1, the four properties cannot be satisfied simultaneously.

The additive specification and mean aggregation are used only for transparency. The same trade-off arises whenever z_i has positive weight in evaluation and observing z_i reveals, or permits inference about, g_i. □

The result is not an argument against blind procedures. Rather, it clarifies the trade-off that such procedures involve. When quality-relevant information is bundled with identity-revealing information, suppressing that information can reduce exposure to bias, but only by limiting the information used in evaluation. Conversely, preserving fuller quality-relevant information may expose jurors to identity cues. Assumption 1 is in this sense minimal: if quality-relevant and identity-revealing information could be fully separated—if observing z_i conveyed no information about g_i—no trade-off would arise. Bundling is the necessary condition for the conflict between full quality-relevant information and bias immunity. The relevant design question is therefore not whether blinding is fair or unfair in general, but where in the evaluation process the informational cost of blinding is likely to be smallest relative to its bias-reducing benefit.

4 Outlier rules, manipulation, and expert pluralism

The first trade-off concerned information control. The second concerns a different problem: how to interpret disagreement among jurors. In music competitions, an unusual score pattern may indicate manipulation, favoritism, or bias. But it may also reflect sincere disagreement about artistic quality. This distinction matters because rules designed to identify outlier jurors often use distance from the jury's central tendency as evidence of problematic scoring. In artistic settings, however, independent judgment may itself appear as distance from the consensus.

The following model formalizes this interpretive problem. The point is not to provide a complete model of strategic voting. Rather, it is to show why score-based outlier rules face a trade-off whenever the same observed score pattern can plausibly arise from either manipulation or sincere minority judgment.

Let candidate i's artistic characteristics be represented by a vector $q_i \in \mathbb{R}^K$. Juror j assigns nonnegative weights to these K characteristics, with the weights summing to one. Formally, $w_j \in \Delta_K$, where $\Delta_K = \{w \in \mathbb{R}^K_+ : \sum_{k=1}^K w_k = 1\}$. Juror j's sincere valuation of candidate i is

$$v_{ij} = w_j \cdot q_i. \tag{10}$$

Jurors may therefore disagree sincerely because they place different weights on technical control, stylistic fidelity, originality, or other dimensions of performance.[6]

Observed scores may differ from sincere valuations because of manipulation. Let

$$s_{ij} = v_{ij} + m_{ij} + \varepsilon_{ij}, \tag{11}$$

where m_{ij} is a manipulation term. A sincere juror has $m_{ij} = 0$ for all candidates. A strategic juror has $m_{ij} \neq 0$ for at least one of the candidates. The competition organizer observes the score vector

$$s_j = (s_{1j}, \ldots, s_{Nj}), \tag{12}$$

and the scores submitted by other jurors, but does not observe w_j, v_{ij}, or m_{ij} separately.

The important feature of this environment is not that every score vector can be rationalized in every possible way. It is that some score patterns that look anomalous relative to the rest of the jury may have more than one plausible interpretation. A juror who assigns unusually high scores to one candidate and unusually low scores to others may be attempting to manipulate the outcome. But a juror may also sincerely assign such scores if she places unusual weight on dimensions of performance that other jurors discount.

[6] In peer review, q_i corresponds to the quality of a submission across multiple dimensions—originality, rigor, and relevance—while w_j corresponds to the referee's methodological or disciplinary priorities. In grant evaluation, q_i corresponds to dimensions such as scientific merit, societal impact, and implementation quality, while w_j reflects the panel member's views on what constitutes promising research.

4.1 Observational ambiguity

We capture this problem with the following assumption.

Assumption 2 *(Observational ambiguity)* There exists a score vector s_j and a profile of scores by the other jurors S_{-j} such that s_j can arise either from sincere minority judgment or from strategic manipulation. That is, the same observed pair (s_j, S_{-j}) is consistent with both $m_{ij} = 0$ for all i and $m_{ij} \neq 0$ for at least one i.

This assumption is the central substantive restriction in this section. It does not say that all outlier scores are ambiguous. Some score patterns may be so extreme, or so closely tied to conflicts of interest, that manipulation is the most plausible interpretation. The assumption only requires that there exist some cases in which score data alone do not distinguish bad-faith deviation from sincere dissent. This is precisely the difficulty in artistic competitions: disagreement is not merely noise. It may be part of what expert judgment is meant to contribute.

4.2 Outlier exclusion

An outlier rule is a procedure that identifies and excludes or downweights evaluators whose scores deviate markedly from the rest of the group. Let an outlier rule be a function

$$\mathcal{E}(s_j, S_{-j}) \in \{0, 1\}, \tag{13}$$

where S_{-j} denotes the scores submitted by all jurors other than j. If

$$\mathcal{E}(s_j, S_{-j}) = 1, \tag{14}$$

juror j's scores are included. If

$$\mathcal{E}(s_j, S_{-j}) = 0, \tag{15}$$

juror j's scores are excluded or downweighted. The rule is score-based in the sense that it uses only the submitted scores and their relationship to the scores of the other jurors.[7]

We define two benchmark properties of such a rule, each stated in deliberately strong form to clarify the trade-off.

Definition 5 *(Manipulation control)* A score-based outlier rule satisfies manipulation control if every manipulated score vector is excluded.

[7] Analogous procedures arise in peer review, where editors who must decide how much weight to place on a divergent report face the same interpretive problem, even in the absence of a formal exclusion rule. In grant panels that use numerical scores, a program officer may discount assessments that deviate markedly from the panel's central tendency. Similar tensions arise in academic hiring when committee members interpret unusually strong support or opposition for a candidate as evidence either of special insight or of personal bias, strategic advocacy, or conflict of interest.

Definition 6 *(Pluralism protection)* A score-based outlier rule satisfies pluralism protection if every sincere score vector is included, even when it departs from the jury's central tendency.

Proposition 2 *(Outlier-rule trade-off) Suppose Assumption 2 holds. Then no score-based outlier rule can simultaneously satisfy manipulation control and pluralism protection.*

Proof By Assumption 2, there exists an observed pair (s_j, S_{-j}) that can arise either from sincere minority judgment or from strategic manipulation. A score-based outlier rule observes only (s_j, S_{-j}), so it must assign the same decision in both cases.

If $\mathcal{E}(s_j, S_{-j}) = 0$, then the rule excludes the strategic juror in the manipulation case, but it also excludes the juror in the sincere-judgment case. It therefore violates pluralism protection.

If $\mathcal{E}(s_j, S_{-j}) = 1$, then the rule includes the juror in the sincere-judgment case, but it also includes the strategic juror in the manipulation case. It therefore violates manipulation control.

Thus, when sincere minority judgment and manipulation are observationally ambiguous in score data, no score-based outlier rule can satisfy both properties. □

When score patterns are strongly suggestive of manipulation—especially when combined with contextual evidence about conflicts of interest, teacher-student relationships, or reciprocal voting—exclusion rules can be a useful corrective. The argument is therefore not that such rules should be avoided. Rather, the point is that scores alone cannot always reveal whether an unusual scoring pattern reflects manipulation or sincere minority judgment. A rule strict enough to catch all manipulation will also exclude some sincere dissenters, while a rule permissive enough to protect all sincere dissenters will also admit some manipulation. The design question is where to set the threshold, and that choice depends on the relative costs of the two errors: suppressing legitimate dissent on one side, and allowing strategic distortion on the other.

5 Further trade-offs in evaluation design

The preceding sections developed the paper's two core trade-offs. This section discusses two additional design choices that arise in many expert-evaluation settings: transparency and repertoire (i.e., standardization) requirements. These trade-offs depend on additional behavioral and contextual assumptions, and are developed as illustrations rather than formal results. They nevertheless show how the broader logic of the paper applies to other choices in evaluation design: reforms that improve one dimension of fairness may weaken another.

5.1 Transparency and juror independence

Transparency plays an important role in many expert-evaluation settings, including music competitions, peer review, grant evaluation, and committee decision-making more broadly. Public disclosure of rules, rankings, scores, reports, or individual votes can improve accountability, reduce suspicion, and allow participants and observers to understand how

decisions were reached. Yet transparency may also change juror behavior. If individual scores are made public, jurors may anticipate criticism from candidates, teachers, audiences, colleagues, or online commentators. Public disclosure can therefore create reputational pressure to avoid scores that appear idiosyncratic, controversial, or difficult to justify. Prat (2005) shows that this kind of pressure—where observability of an expert's action creates incentives to conform rather than follow private judgment—can arise across a range of settings where expert agents face reputational concerns.[8]

A simple way to see the trade-off is to suppose that jurors care both about expressing their sincere evaluation and about avoiding visible deviation from the consensus they perceive to be forming around a candidate. Let $t \in [0, 1]$ denote the degree of transparency or public exposure. A higher value of t means that juror scores are more visible to candidates, audiences, colleagues, or the broader public. Juror j's sincere valuation of candidate i is v_{ij}, and c_{ij} denotes juror j's perceived consensus score for candidate i: the score juror j believes would approximate the panel's consensus assessment. This perceived consensus need not equal the score other jurors would actually assign, nor need such perceptions be accurate on average. Jurors may form beliefs about it from other jurors' reputations, prior scoring patterns, comments during deliberation, shared professional norms, or visible reactions to the candidate's performance. Suppose the juror chooses a submitted score s_{ij} to minimize

$$(s_{ij} - v_{ij})^2 + t(s_{ij} - c_{ij})^2. \tag{16}$$

The first term penalizes deviations from sincere artistic judgment. The second term captures reputational pressure to remain close to the individual juror's perceived consensus score; the importance of this term increases with transparency. The first-order condition implies

$$s_{ij}(t) = \frac{v_{ij} + tc_{ij}}{1 + t}. \tag{17}$$

The submitted score is therefore a weighted average of the juror's sincere valuation and the juror's perceived consensus score. As transparency increases, more weight is placed on the perceived consensus.

Averaging across jurors gives

$$\bar{s}_i(t) = \frac{\bar{v}_i + t\bar{c}_i}{1 + t}, \tag{18}$$

where bars denote averages across jurors. This expression clarifies two effects. If $\bar{c}_i = \bar{v}_i$ —that is, if the average perceived consensus equals the average sincere valuation—transparency does not change candidate i's mean submitted score, and therefore does not affect rankings based only on mean scores through this channel. It may still reduce dispersion by pulling individual jurors' reports toward their perceived consensus scores. If $\bar{c}_i \neq \bar{v}_i$,

[8] The conformity-pressure concern that arises from transparency in competitions has close analogs in other settings. Publishing individual voting records on monetary policy committees, disclosing referee identities in peer review, publishing referee reports and author responses in open-review systems, and making committee votes public in academic hiring may all create pressure to conform rather than express independent judgment.

transparency changes candidate i's mean submitted score by pulling it toward the average perceived consensus. Rankings based on mean submitted scores may then differ from rankings based on mean sincere valuations when $\bar{c}_i - \bar{v}_i$ differs across candidates.

Transparency therefore changes not only who observes a juror's score, but also the incentives under which that score is produced. Published scores may discipline arbitrary or biased behavior, but they may also make jurors reluctant to submit evaluations that visibly depart from what they believe others will regard as defensible. The design problem is to choose the form and timing of transparency so that scrutiny does not turn independent judgment into reputational risk management.

5.2 Standardization and evaluative breadth

A second design choice concerns repertoire and performance conditions. Requiring candidates to perform the same work, or the same type of work, can improve comparability. Jurors can compare performances more directly when candidates face similar tasks. Yet repertoire requirements can also narrow the dimensions of artistic ability that the competition observes. If candidates have heterogeneous strengths across repertoire domains, a narrow repertoire may identify the best candidate within the prescribed domain while failing to reveal distinctive strengths in others.[9]

Suppose, for example, that there are two repertoire domains, A and B. Candidate i's artistic ability in these domains is (q_{iA}, q_{iB}). A competition chooses a repertoire weight $r \in [0,1]$, where $r = 1$ means that evaluation is based entirely on domain A, and $r = 0$ means that evaluation is based entirely on domain B. Candidate i's evaluated score is

$$V_i(r) = rq_{iA} + (1-r)q_{iB}. \tag{19}$$

A more standardized repertoire makes candidates more directly comparable along the selected dimension. But if the competition's broader aim is to identify breadth or versatility, then a narrow repertoire may fail to reveal some candidates' strongest qualities.

To see the point, consider two candidates. Candidate 1 has $q_{1A} = 10$ and $q_{1B} = 0$, while candidate 2 has $q_{2A} = 8$ and $q_{2B} = 20$. If the competition evaluates only repertoire domain A, so that $r = 1$, then candidate 1 is selected because $V_1(1) > V_2(1)$. If the competition instead gives equal weight to both domains, so that $r = 0.5$, then candidate 2 is selected because $V_2(0.5) > V_1(0.5)$. Standardization on a common repertoire domain therefore improves comparability within that domain, but it may change rankings when the selected domain captures only part of the quality the competition aims to reward.

Repertoire requirements therefore do more than standardize the conditions of performance; they define what kinds of excellence the competition is able to observe. In a specialized competition, this is appropriate: the Chopin Competition is designed to evaluate pianists within the expressive, technical, and stylistic world of Chopin, not to compare candidates across the broader piano repertoire. But in competitions that claim to reward more

[9] Repertoire requirements are a form of standardization, and the underlying tension arises in other expert-evaluation settings as well. Top-publication requirements in tenure cases impose a common metric on research records but may penalize scholars working in subfields or using methodological approaches that top journals publish less frequently. In peer review, privileging particular methodological approaches sharpens the evaluative criterion while narrowing what counts as valid work.

general artistic excellence, the choice of repertoire necessarily privileges some dimensions of performance over others. Expanding the repertoire menu may allow candidates to reveal more varied artistic strengths, while a narrower menu makes direct comparison easier.

6 Implications

6.1 Rule choice across stages and settings

The framework identifies two recurring tensions in expert evaluation. Some information cannot be stripped of identity cues without losing evaluative content, and some score patterns cannot be classified as manipulative without risking the exclusion of sincere minority judgment. In both cases, a procedure designed to secure one dimension of fairness may weaken another.

These trade-offs have both normative and positive implications. Normatively, they clarify the choices evaluators face when selecting procedures. Positively, they help explain why evaluation systems use different rules across stages and settings. Rules matter because they reduce particular vulnerabilities, but they do so by changing the evaluative environment itself. Blind procedures limit exposure to bias-relevant information; outlier rules constrain strategic scoring or bad-faith evaluation; transparency makes decisions easier to scrutinize; and standardization improves comparability across candidates. None of these reforms, however, is costless. Each changes the information available to jurors, the discretion they exercise, the types of disagreement that remain admissible, or the dimensions of quality candidates can reveal. This setting-specific approach is consistent with work in constitutional political economy that treats procedural features of collective decision-making as objects of constitutional choice rather than fixed features of the institutional environment (Tridimas 2017).

The severity of each trade-off depends on features of the evaluation environment that vary across rounds, competitions, and settings. For the information-bundling trade-off, three features matter: the importance of z_i for evaluation, how strongly it reveals or permits inference about g_i, and whether identity-related bias, summarized by $\bar{\beta}$, is large enough to affect rankings. The trade-off is least severe when z_i contributes little to the competition's evaluative purpose, reveals little about candidate identity, or when such bias is unlikely to change the ordering of candidates. It becomes most severe in live finals, where stage presence, physical expressiveness, and other visible dimensions of performance are central to what is being evaluated, and where jurors observe the full range of identity-relevant cues. In those settings, procedures that suppress identity-revealing information carry a real informational cost, and that cost rises as z_i becomes more central to the competition's evaluative purpose.

Music competitions illustrate how this logic explains observed procedural variation: more restrictive procedures are often used in early rounds, followed by fuller information in later ones. This can be understood as a rational response to the trade-off in Proposition 1: competition designers accept the informational cost of restricting z_i when that cost is smallest and relax the restriction as the competition's evaluative purpose demands fuller information. For the outlier-rule trade-off, what matters is the degree of genuine evaluative pluralism relative to the scope for strategic behavior. Divergent scores are harder to interpret when jurors come from diverse national and stylistic traditions because sincere minority

positions are then more varied and harder to distinguish from manipulation on the basis of scores alone. Stronger exclusion rules may therefore be more appropriate when juries are homogeneous and contextual evidence of conflicts of interest is available, and less appropriate when heterogeneous artistic standards are a deliberate feature of jury design.

A similar logic applies in other expert-evaluation settings. For the information-bundling trade-off, the key question is how much evaluative weight the suppressed information carries. In peer review, double-blind procedures are most costly when author reputation and prior output are genuinely informative about submission quality, as may be the case when editors assess a research agenda rather than a single self-contained contribution. In grant evaluation, anonymization is most costly when panels assess research capacity rather than a narrowly specified deliverable, since capacity is difficult to evaluate without knowing who will do the work. In academic hiring, the bundling problem is especially severe because research quality, advisor identity, and academic pedigree are deeply intertwined and difficult to disentangle even in principle. For the outlier trade-off, the severity depends on the range of legitimate disagreement among evaluators. In peer review and grant panels, methodological and disciplinary heterogeneity plays an analogous role to national or stylistic heterogeneity among competition jurors. The wider the legitimate range of expert opinion, the harder it is to treat a divergent assessment as evidence of bad faith rather than genuine minority judgment.

6.2 Sortition as a boundary case

The framework also helps explain why sortition is attractive as a boundary case. Sortition—random selection by lot—replaces fine-grained evaluation with a draw among eligible candidates. In the settings considered here, the relevant form is usually a qualified lottery: candidates first clear a qualification threshold, and final selection is then made by random draw (Frey et al. 2022). Public choice scholars have long considered lotteries as alternatives to conventional selection procedures (Mueller et al. 1972). Conditional on the eligible pool, a random draw is unbiased across candidates, cannot be manipulated through strategic scoring, and eliminates the need to decide how different dimensions of quality should be weighted beyond the threshold stage. As Stone (2009) argues, random selection has a "sanitizing effect": at the stage where the lottery is used, it prevents bad reasons from influencing the choice by preventing any further reasons, good or bad, from mattering. Sortition therefore occupies an extreme position in the space of evaluation procedures. It avoids the two core trade-offs emphasized in this paper—between bias reduction and information preservation, and between manipulation prevention and legitimate dissent—only by relocating quality identification to the threshold, where versions of the same bundling and aggregation problems reappear,[10]

Whether that relocation is acceptable depends on what remains to be learned after the eligibility threshold is applied. If the threshold screens out clearly weaker candidates and

[10] Historical examples illustrate this logic. In ancient Athens, many public offices were selected by lot, but citizens first had to pass a *dokimasia* which screened for citizenship, civic obligations, and basic competence (Hansen 1991). The Venetian republic's *ballotta* system used lot in the selection of nominating committees, reducing opportunities for factional control (Finlay 1980). Modern jury selection similarly begins with a random draw from an eligible population, followed by screening for competence and impartiality. In each case, sortition reduces some forms of strategic selection only by shifting evaluative judgment to the eligibility or screening stage.

leaves a closely matched pool, the cost of random selection may be small. Fine-grained ranking may then turn on distinctions that are noisy, contestable, or especially vulnerable to bias, manipulation, favoritism, or false precision. If quality differences among eligible candidates remain large and consequential, replacing evaluation with a lottery is correspondingly costly. Sortition is therefore most attractive when the risks of fine-grained ranking outweigh the value of the distinctions it would preserve, and least attractive when expert judgment can still identify consequential differences beyond the threshold.

6.3 The political economy of procedural reform

The preceding discussion treats the choice of evaluation rules as a function of the evaluation environment. A further question is how such rules change over time. Procedures adopted in response to specific failures—for instance, a documented gender gap or a notorious jury scandal—are often calibrated to the cases that motivated them. A constitutional economics perspective suggests that the value of a rule lies in its general properties across the range of cases it will govern, not its performance in the particular cases that prompted its adoption (Hayek 1945, 1960; Buchanan 1987, 1990). A procedure optimized to correct a known bias may perform poorly against unforeseen ones, while a rule designed to catch a specific pattern of manipulation may suppress sincere dissent where no manipulation is occurring.

Evaluating rules by their general properties becomes more important as the stakes rise. When selection confers valuable rewards—prizes, grants, jobs, prestige, or career opportunities—candidates and interested parties have stronger incentives to influence both outcomes and procedures. Higher payoffs make strategic scoring, coalition formation, informal vote trading, lobbying over rules, and efforts to shape the eligible pool more attractive. Some of these efforts may improve the underlying quality being evaluated. Others are redistributive, as when actors devote resources to changing the rules, cultivating influence, or manipulating the margin on which selection occurs, a logic familiar from the rent-seeking literature (Tullock 1967; Krueger 1974). Evaluation rules are therefore not merely devices for aggregating judgment; in high-stakes settings, they may become objects of contestation in their own right. The greater the payoff, the stronger the incentive to exploit whatever vulnerability the procedure leaves open.[11]

These incentives operate in an environment where some interests are better represented than others. Salient failures may generate organized pressure for targeted reforms from those most directly affected, while the diffuse costs of those reforms—costs that may take the form of suppressed information, constrained dissent, or narrowed evaluation—often fall on less organized constituencies (Olson 1965). The Leeds Competition's introduction of blind pre-selection procedures illustrates this asymmetry. The gender disparity motivating reform was visible and publicly salient, while the informational costs identified in Proposition 1 are less directly observable and are unlikely to be represented by an equally organized constituency. Similar asymmetries may arise whenever evaluation systems come under public scrutiny—in peer review, grant panels, and academic hiring as much as in music competitions.

[11] In some settings, disputes over evaluation procedures are simultaneously disputes over which dimensions of quality the system should recognize. Debates over admissions criteria at selective universities are a case in point: at issue is both whether agreed standards are applied consistently across groups and how to weight academic preparation, extracurricular achievement, leadership, and other criteria against one another. This is a form of the standardization trade-off identified in Section 5.

Evaluation systems may therefore accumulate targeted rules over time, each locally defensible but collectively drifting away from the general properties a constitutional perspective would recommend. Once adopted, moreover, such rules tend to persist: beneficiaries have strong incentives to resist their removal, while those who bear the costs are too diffuse to mount comparable opposition (Tullock 1975). Rules adopted in response to different failures may pull in conflicting directions; a procedure introduced to reduce bias may sit uneasily alongside one designed to limit manipulation, while a transparency requirement may undercut a rule meant to protect independent judgment.

The result is an evaluation system whose procedural architecture reflects the sequence of salient failures that prompted each rule's adoption rather than any coherent set of design principles (Hayek 1973; Brennan and Buchanan 1985). The dynamic is structurally analogous, though narrower in scope, to the institutional sclerosis identified by Olson (1982), in which concessions to organized constituencies accumulate and create systemic rigidity. This suggests that the constitutional perspective offers a useful corrective to treating each salient failure as a warrant for targeted procedural adjustment, directing attention instead to the general properties a procedure should have across the full range of evaluation environments it will face.

A final implication is that the evaluation environment is itself endogenous. The framework has treated the information structure and distribution of juror types as features of the setting that evaluation rules must navigate. But rules also shape the environments they govern. As the economics of policy evaluation emphasizes, agents adapt their behavior to rules, sometimes in ways that undermine the properties those rules were intended to have (Lucas 1976; Goodhart 1984). If blind procedures become standard, candidates may invest more heavily in dimensions of performance that remain visible and less in those that are suppressed. If outlier rules become widely known, strategic jurors may adjust their scoring to avoid detection. The trade-offs identified here are therefore not static: they evolve as participants adapt to the procedures designed to govern them. This strengthens the case for evaluating rules by their general and robust properties rather than by their performance against the specific behaviors that prompted their adoption.

7 Conclusion

Expert evaluation systems are sometimes treated as though fairness problems are primarily failures of implementation: failures to apply good rules consistently, identify biased evaluators, or ensure that procedures are followed as designed. The analysis here points to a prior difficulty. Some fairness problems are not implementation failures but consequences of the constraints under which evaluation procedures operate—constraints on what can be observed and what observed data can reveal about the judgments that produced them. No procedure can fully escape these constraints.

Although the paper develops the framework through music competitions, the argument applies more broadly to evaluation systems in which expert judgment cannot be reduced to a single objective metric. Similar tensions arise in peer review, grant evaluation, academic hiring, other artistic competitions, and judged sports such as figure skating, gymnastics, and freestyle snowboarding, where evaluators aggregate multidimensional judgments under incomplete information and heterogeneous standards.

Like other problems of constitutional design, evaluation design is not a matter of finding a procedure that performs best in every case, but of choosing among imperfect rules that trade off competing values. The practical question is not whether a rule improves fairness in some general sense, but which fairness problem it addresses and what other values it may weaken. Fairness reforms should therefore be evaluated not only by whether they reduce bias, manipulation, opacity, or inconsistency, but also by the information, discretion, or disagreement they suppress.

The framework also helps explain why evaluation systems use different rules across stages and settings and why those rules change over time. Procedures differ not simply because some are fairer than others, but because the relevant trade-offs vary with the information available, the purpose of evaluation, the scope for legitimate disagreement, and the incentives for strategic behavior. Reforms adopted in response to salient failures may therefore create new vulnerabilities even when they address the problem that prompted them.

The broader lesson is that fairness in evaluation is not achieved by finding a rule that eliminates all vulnerabilities. It requires choosing among imperfect procedures, confronting their trade-offs honestly, and aligning them with the purpose of the evaluation. Because these trade-offs vary across settings and evolve as rules change, fairness remains a problem of procedural design rather than a property of any single procedure.

Acknowledgements We are grateful to Lee Benham, Hiromi Fukuda, Associate Editor Ennio Piano, and an anonymous referee for helpful comments on earlier drafts of this paper.

Author contributions M.T.L. conceived the study and developed the theoretical framework. K.J.B. and M.T.L. developed the implications, wrote the manuscript, and analyzed the results. Both authors revised, reviewed, and approved the final manuscript.

Funding Open access funding provided by University of Southern Denmark. The authors received no funding in support of this research.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Alberge, Dalya. 2024, February. 'We're going to blame the women, not our sexism': bias holding back top female pianists. *The Observer*. Published February 3, 2024. Accessed: 20 August 2025. https://www.theguardian.com/music/2024/feb/03/sexism-bias-top-female-pianists-classical-music.

Arrow, Kenneth J. 1951. *Social Choice and Individual Values*. New York: John Wiley & Sons.

Asmat, Roberto, Karol J. Borowiecki, and Marc T. Law. 2023. Competing for equality: The gender gap in piano competitions since 1890. Research Summary for the WFIMC General Assembly, World Federation of International Music Competitions (WFIMC).
Asmat, Roberto, Karol J. Borowiecki, and Marc T. Law. 2023. Do experts and laypersons differ? Some evidence from international classical music competitions. *Journal of Economic Behavior & Organization* 214:270–290.
Asmat, Roberto, Karol J. Borowiecki, and Marc T. Law. 2024. Competing for equality: Gender bias among juries in international piano competitions, 1890–2023. SSRN Working Paper No. 4910483.
Austen-Smith, David, and Jeffrey S. Banks. 1996. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review* 90 (1): 34–45.
Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva. 2017. Does the gender composition of scientific committees matter? *American Economic Review* 107 (4): 1207–1238.
Bagues, Manuel F., and Berta Esteve-Volart. 2010. Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. *Review of Economic Studies* 77 (4): 1301–1328.
Bergemann, Dirk, and Stephen Morris. 2019. Information design: A unified perspective. *Journal of Economic Literature* 57 (1): 44–95.
Blank, Rebecca M. 1991. The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. American Economic Review 81 (5): 1041–106
Brennan, Geoffrey, and James M. Buchanan. 1980. *The Power to Tax: Analytical Foundations of a Fiscal Constitution*. Cambridge: Cambridge University Press.
Brennan, Geoffrey, and James M. Buchanan. 1985. *The Reason of Rules: Constitutional Political Economy*. Cambridge: Cambridge University Press.
Brodie, Stephanie, André Frainer, Maria Grazia Pennino, Shuheng Jiang, Laura Kaikkonen, Javier Lopez, Kelly Ortega-Cisneros, Cara A. Peters, Sara A. Selim, and Nicoleta Văidianu. 2021. Equity in science: advocating for a triple-blind review system. *Trends in Ecology & Evolution* 36 (11): 957–959. https://doi.org/10.1016/j.tree.2021.07.011.
Buchanan, James M. 1987. The constitution of economic policy. *American Economic Review* 77 (3): 243–250.
Buchanan, James M. 1990. The domain of constitutional economics. *Constitutional Political Economy* 1 (1): 1–18.
Buchanan, James M., and Gordon Tullock. 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press.
Buchanan, James M., and Richard E. Wagner. 1977. *Democracy in Deficit: The Political Legacy of Lord Keynes*. New York: Academic Press.
Chmurzyńska, Małgorzata. 2015. Influence of a priori information on music performance assessment. In *Proceedings of the Ninth Triennial Conference of the European Society for the Cognitive Sciences of Music*, 292–295. Manchester.
Deschamps, Pierre. 2024. Gender quotas in hiring committees: A boon or a bane for women? *Management Science* 70 (11): 7486–7505.
Feddersen, Timothy J., and Wolfgang Pesendorfer. 1997. Voting behavior and information aggregation in elections with private information. *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.
Finlay, Robert. 1980. *Politics in Renaissance Venice*. New Brunswick: Rutgers University Press.
Flôres, Renato GFlôres., and Victor A. Ginsburgh. 1996. The Queen Elisabeth musical competition: How fair is the final ranking? *Journal of the Royal Statistical Society: Series D (The Statistician)* 45 (1): 97–104.
Frey, Bruno S., Margit Osterloh, and Katja Rost. 2022. The rationality of qualified lotteries. *European Management Review*. https://doi.org/10.1111/emre.12550.
Gibbard, Allan. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41 (4): 587–601.
Ginsburgh, Victor, and Sheila Weyers. 2014. Nominees, winners, and losers. *Journal of Cultural Economics* 38 (4): 291–313. https://doi.org/10.1007/s10824-014-9210-7.
Ginsburgh, Victor A. 2003. Awards, success and aesthetic quality in the arts. *Journal of Economic Perspectives* 17 (2): 99–111. https://doi.org/10.1257/089533003765888458.
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.
Glejser, 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.
Goldin, Claudia D., and Cecilia Rouse. 2000. Orchestrating impartiality: the impact of "blind" auditions on female musicians. *American Economic Review* 90 (4): 715–741.
Goodhart, C. A. E. 1984. Problems of monetary management: The U.K. experience. In *Monetary Theory and Practice*, ed. C. A. E. Goodhart, 91–121. London: Macmillan.

Hamlin, Alan. 2023. The rule of rules. *Public Choice* 195:231–250. https://doi.org/10.1007/s11127-021-00872-3.
Hansen, Mogens Herman. 1991. *The Athenian Democracy in the Age of Demosthenes: Structure, Principles and Ideology*. Oxford: Blackwell.
Hayek, Friedrich A. 1945. The use of knowledge in society. *American Economic Review* 35 (4): 519–530.
Hayek, Friedrich A. 1960. *The Constitution of Liberty*. Chicago: University of Chicago Press.
Hayek, Friedrich A. 1973. *Law, Legislation and Liberty, Volume 1: Rules and Order*. Chicago: University of Chicago Press.
Kamenica, Emir, and Matthew Gentzkow. 2011. Bayesian persuasion. *American Economic Review* 101 (6): 2590–2615.
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 (5): 1057–1084.
Krueger, Anne O. 1974. The political economy of the rent-seeking society. *American Economic Review* 64 (3): 291–303.
Ladha, Krishna K. 1992. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science* 36 (3): 617–634.
Ladha, Krishna K. 1993. Condorcet's jury theorem in light of de Finetti's theorem: Majority-rule voting with correlated votes. *Social Choice and Welfare* 10 (1): 69–85.
Levashina, Julia, Christopher J. Hartwell, Frederick P. Morgeson, and Michael A. Campion. 2014. The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology* 67 (1): 241–293. https://doi.org/10.1111/peps.12052.
Lucas, Robert E. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* 1:19–46.
McCormick, Lisa. 2015. *Performing Civility: International Competitions in Classical Music*. Cambridge: Cambridge University Press.
McCormick, Lisa. 2018. Pogorelich at the Chopin: Towards a sociology of competition scandals. *The Chopin Review* 1 (1): 1–16.
Mueller, Dennis C., Robert D. Tollison, and Thomas D. Willett. 1972. Representative democracy via random selection. *Public Choice* 12 (1): 57–68.
Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.
Olson, Mancur. 1982. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. New Haven: Yale University Press.
Posner, Richard A. 1971. Taxation by regulation. *Bell Journal of Economics and Management Science* 2 (1): 22–50.
Prat, Andrea. 2005. The wrong kind of transparency. *American Economic Review* 95 (3): 862–877.
Ross-Hellauer, Tony. 2017. What is open peer review? *A systematic review. F1000Research* 6:588.
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.
Saygin, Perihan O., and Thomas Knight. 2026. Gender bias in peer performance evaluations. *Journal of Behavioral and Experimental Economics* 122: 102568. https://doi.org/10.1016/j.socec.2026.102568.
Sinclair, Fiona. 2024, March. Why piano competitions must address the gender gap. *The Guardian*. Published March 20, 2024. Accessed: 20 August 2025. https://www.theguardian.com/music/2024/mar/20/leeds-piano-fiona-sinclair-why-piano-competitions-must-address-the-gender-gap.
Sosnowska, Honorata. 2017. Comparison of voting methods used in some classical music competitions. In *Transactions on Computational Collective Intelligence XXVII, Lecture Notes in Computer Science*, vol. 10480, ed. Thanh Nguyen Ngoc, 108–117. Cham: Springer.
Stone, Peter. 2009. The logic of random selection. *Political Theory* 37 (3): 375–397.
Ting, Chaowen, Yating Chuang, and John Chung-En. Liu. 2026. Gender bias in competitive music composition evaluation: An experimental study. *Public Choice*. https://doi.org/10.1007/s11127-026-01405-6.
Tomkins, Andrew, Min Zhang, and William D. Heavlin. 2017. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences* 114 (48): 12708–12713.
Tridimas, George. 2017. Constitutional choice in ancient Athens: The evolution of the frequency of decision making. *Constitutional Political Economy* 28:209–230. https://doi.org/10.1007/s10602-017-9241-2.
Tullock, Gordon. 1967. The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal* 5 (3): 224–232.
Tullock, Gordon. 1975. The transitional gains trap. *Bell Journal of Economics* 6 (2): 671–678.

Vanberg, Viktor J. 1994. *Rules and Choice in Economics: Essays in Constitutional Political Economy*. London: Routledge.

Zitzewitz, Eric. 2014. Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging. *Journal of Sports Economics* 15 (1): 3–30.