WORKING PAPER
Draft Date: August 22, 2026

Easy Markers, Tough Graders, and True Patriots: Decomposing generosity, nationalistic bias, and gamesmanship in the scoring of elite figure skating

Marc J. Champigny
PhenoLogic Co.
570 Montee des Seigneurs
Granby, QC, J2J 0A1
Canada
www.phenologic.net

Email for correspondence: champignym@phenologic.net

JEL Classification Codes:
Z20 (Sports Economics: General)
M54 (Labor Management: Performance Evaluation)
D73 (Bureaucracy; Administrative Processes; Corruption)
D82 (Asymmetric and Private Information; Mechanism Design)
D91 (Role and Effects of Psychological, Emotional, Social, and Cognitive Factors on Decision Making)

Keywords:
Sports analytics, human behavior, figure skating, nationalistic bias, subjective evaluation, Technical Panel, econometrics, judging.

Abstract
The subjective scoring system in competitive figure skating has historically been vulnerable to nationalistic bias. Despite the implementation of the International Judging System (IJS) to curb such anomalies, the scoring framework remains structurally vulnerable to behavioral adaptations that compromise its competitive integrity. This study presents a comprehensive, multi-methodological analysis of nationalistic bias and other behaviors, leveraging for the first time a fully de-anonymized dataset of 14,382 International Skating Union-sanctioned performances spanning the 2022–2026 Olympic cycle. To ensure a strictly objective assessment, I deployed a double-normalized baseline model that disentangles targeted nationalistic bias from general evaluation noise by explicitly controlling for both the skater's underlying performance consensus and the official's historical baseline leniency. Correcting for multiple testing via the False Discovery Rate (FDR), I find that 34 of 40 national federations and 157 of 290 individual judges exhibit statistically significant in-group favoritism, and 159 of 312 skaters/teams were verified recipients of nationalistic premiums. Furthermore, non-parametric tests reveal advanced gamesmanship, including synchronous reciprocal point-trading and intra-country point-spread adjustment, that successfully evades standard mean-based detection. Expanding beyond the primary judging panel, I introduce a novel Expected Value (EV) model to conduct the first quantitative assessment of the sport's Technical Panel. Results demonstrate that nationalistic bias systematically affects the deterministic technical foundation of the sport via unpenalized technical infractions and difficulty-level boosting. To quantify the tangible competitive consequences of these dual evaluation mechanics, I introduce counterfactual simulations that separate an evaluator's *Attempted Premium* from their *Realized Impact*. The comprehensive dual-panel simulation shows that neutralizing mathematically verified bias alters final placements in 40.7% of affected competitions and distorts medal outcomes in 11.8% of events. To mitigate these systemic vulnerabilities, this study outlines a series of data-driven administrative reforms, such as blind flight scoring, randomized sub-panels, and automated technical calling, highlighting how quantitative diagnostics can actively guide the design of more resilient, objective officiating systems.

Keywords: Sports analytics, human behavior, figure skating, nationalistic bias, subjective evaluation, Technical Panel, econometrics, judging.

1. Introduction

Subjectively judged sports offer a unique window into human decision-making, institutional design, and strategic behavior. While objective sports determine outcomes through absolute metrics like a stopwatch or a measuring tape, sports like figure skating, gymnastics, and diving rely entirely on human judges to translate complex physical movement into numerical scores. Because human evaluators are inherently susceptible to external pressures, these sports provide a laboratory for researchers seeking to understand how and why incentives, such as nationalistic bias, operate.

Empirical evidence demonstrates that this vulnerability is not an isolated quirk of any single sport, but a systemic feature of subjective evaluation across the Olympic spectrum (reviewed in Braeunig, 2025). Researchers have consistently documented that when nationality is salient and scoring is flexible, evaluators systematically inflate the scores of their compatriots. In artistic gymnastics, for example, early studies by Ansorge and Scheer (1988) identified significant national bias at the 1984 Olympic Games, a finding supported by modern econometric analyses showing that gymnastics judges reliably award higher execution scores to domestic athletes (Heiniger and Mercier 2019; 2021).

This phenomenon is further illustrated in equestrian dressage, another highly subjective Olympic discipline. Lacking the objective metrics of a stopwatch or tape measure, dressage relies on judges' holistic evaluation of movement, rhythm, and harmony. Empirical assessments of international dressage panels confirm that judges systematically award higher scores to compatriots (Sandberg 2018, Wolframm, 2023**)**. Like figure skating, dressage represents an environment where athletic execution is judged as a continuous, aesthetic product, making the baseline of evaluation highly susceptible to nationalistic distortion.

The magnitude of this bias is influenced by the degree of subjectivity permitted by the scoring rules. For example, researchers investigating Olympic diving have consistently identified significant patriotic bias despite the sport's relatively brief, isolated maneuvers **(**Emerson et al. 2009, Emerson and Meredith 2011**)**. Conversely, in disciplines that blend objective measurements with highly subjective scoring, such as ski jumping, bias is particularly pronounced in the subjective "style" points. By using objective metrics like jump distance and speed as proxies for true performance quality, models reveal that ski jumping judges actively manipulate subjective scores to favor compatriots **(**Krumer et al. 2022, Zitzewitz, 2006**)**.

The broader literature on sports economics demonstrates that the magnitude of in-group bias is strictly governed by "verifiability", the degree to which an evaluator's decision can be objectively proven wrong. A classic example is found in Major League Baseball. Parsons et al. (2011) demonstrated that while umpires typically exhibited in-group bias when calling balls and strikes, this bias vanished entirely in stadiums equipped with electronic pitch-tracking systems. Because the pitch location was objectively verifiable, the umpires could not exhibit bias without compromising their own professional credibility.

The verifiability dynamic also modulates bias in aesthetic sports. As highlighted by Braeunig (2025), nationalistic premiums consistently emerge where evaluators are granted wide interpretive latitude without deterministic boundaries. Conversely, when scrutiny increases, bias is mathematically constrained. For example, Zitzewitz (2006) observed that during the highest-stakes figure skating events, patriotic bias actually decreased, suggesting that judges actively suppress their favoritism to avoid the intense public and institutional scrutiny that accompanies major Olympic broadcasts. Similarly, Lee (2008) identified 'outlier aversion' among judges, who actively rein in their scores to avoid standing out from the panel. Across all these domains, a unifying economic principle emerges. Bias scales directly with the degree of subjectivity permitted by the rules and shrinks when a judge's actions are highly verifiable.

1.1 The Structure of Figure Skating Competition and its Officiating Mechanics

Competitive figure skating is contested across four disciplines: Men's Singles, Women's Singles, Pairs, and Ice Dance. A competition within any discipline is divided into two parts: a shorter, more constrained segment (the Short Program or Rhythm Dance) and a longer, more flexible segment (Free Skate or Free Dance).

Figure skating differs fundamentally from other judged sports like snowboarding, diving or gymnastics, where athletes typically execute a sequence of brief, independent skills. A figure skating routine is a synthesized artistic product where the whole is perceived to be greater than the sum of its parts. Competitors select from a menu of permitted technical elements (e.g., jumps, spins, lifts) and weave them into a continuous performance set to music, complete with thematic choreography. Crucially, these performances are not random, independent outings. Skaters generally retain the same "program" for an entire competitive season. The music, element sequence, and choreography remain largely consistent from event to event, providing researchers with a unique longitudinal dataset of repeated, highly similar trials evaluated by varying panels of officials.

Under the modern International Judging System (IJS), figure skating performances are evaluated by two distinct groups of officials:

- **The Technical Panel:** A three-person panel responsible for identifying what the skater/team did. They a) call the specific elements performed, b) determine the Level of Difficulty (from Base to Level 4), and c) assign technical infractions, such as an under-rotated jump or an extended lift. These decisions establish the underlying base value of each element performed in the routine and sometimes incur mandatory reductions in quality (GOE) scores.
- **The Judging Panel:** A panel of between five and nine international judges responsible for evaluating the quality of the elements and the overall performance. For every element identified by the technical panel, the judges award a Grade of Execution (GOE) as an integer ranging from -5 to +5. Separately, they award three holistic Program Component Scores (PCS) named Composition, Presentation, and Skating Skills, which evaluate the overarching artistic and skating quality of the performance.

To mitigate the impact of outlier judges the governing body, the International Skating Union (ISU), uses a trimmed mean procedure to obtain the final performance score. For each element GOE and each Program Component, the highest and lowest scores awarded by the panel are discarded, and the remaining marks are averaged. The final Total Segment Score (TSS) is the sum of the GOE-adjusted technical element scores (the Total Element Score) plus the factored component averages (the Total Component Score), minus any mandatory deductions arising from occurrences such as falls or timing violations.

1.2 The Principal-Agent Problem and the Evolution of Bias

While this system was engineered to be objective, it suffers from a classic principal-agent problem rooted in the economics of subjective performance evaluation (Prendergast, 1999). In contract theory, when performance metrics are highly subjective and cannot be objectively verified, agents face significant incentives to align their evaluations with the preferences of their immediate sponsors rather than the mandates of the overarching principal (Prendergast, 2002). In this context, the International Skating Union (the principal) mandates fair, unbiased evaluation. However, the technical specialists and judges (the agents) are nominated, trained, and promoted by their respective national federations. Because these federations are incentivized to see their own athletes win medals and secure state funding, officials face institutional pressure to maximize the placements of their compatriots, generating a fundamental conflict of interest.

Economists and statisticians have long recognized this vulnerability. The broader literature on sports economics demonstrates that when tournament structures feature high financial rewards and asymmetric incentives, agents frequently engage in strategic, collusive behaviors to manipulate outcomes (Duggan & Levitt 2002). Under figure skating's historic 6.0 ordinal ranking system, Campbell and Galbraith (1996) found non-parametric evidence of systematic in-group favoritism. Zitzewitz (2006) took this further, using ordered probit models from the 2002 Winter Olympics to demonstrate that judges were both favoring their own skaters and actively colluding by trading votes in geopolitically aligned blocs.

When figure skating transitioned to the current IJS point system, researchers adapted their methods, moving to hierarchical and mixed-effects models. Emerson and Arnold (2011) demonstrated that while the new system changed the variance of the scores, nationalistic bias persisted. However, research during this era was severely limited by the ISU's policy of anonymous judging (2004–2016). Because individual judges' names and nationalities were withheld from the public record, Zitzewitz (2014) had to rely on aggregate bloc analysis rather than tracking specific judges. When judge anonymity was abolished in 2016, an enhanced level of empirical scrutiny became possible, highlighted by a 2018 data journalism investigation by *BuzzFeed News* (Templon, 2018), which used the newly transparent protocols to publicly flag individual judges who gave their compatriots inflated scores.

1.3 Opportunities in Comprehensive Data and Current Contributions

The complete abolition of anonymous judging, combined with modern computational methods to parse thousands of ISU protocols, presents unprecedented opportunities to expand the scope of behavioral analyses of evaluation panels in figure skating. Most importantly, fully deanonymized data allows researchers to track individual officials longitudinally and establish their baseline generosity, or their inherent tendency to be lenient or strict toward the general international field. The data reveal that baseline generosity is an important axis of variation among judges and technical specialists. An official who is universally lenient will naturally award high scores to their compatriots, which basic mean-deviation models might misinterpret as targeted nationalistic bias. Previous deanonymized investigations, such as Templon's (2018) analysis for *BuzzFeed News*, recognized this dynamic and incorporated international baseline comparisons to flag judges who favored their own skaters.

By utilizing a large, custom-scraped dataset of 14,382 fully deanonymized senior-level performances from the 2022–2026 Olympic cycle, it is possible to build upon this foundational concept. I expand the analysis of bias to encompass both the primary judging panel and the previously unexamined technical panel, allowing for a comprehensive deconstruction of how bias operates, its strategic mechanics, and its tangible impact on athletes' scores and placements.

Rather than treating scoring anomalies as isolated errors, his study models officiating patterns as systematic behavioral responses to institutional incentives and constraints. Since the abolition of judge anonymity, the strategic application of bias has shifted toward subjective channels in which high information asymmetry shields evaluation from objective verification. Guided by the economic theory of verifiability, this paper examines how nationalistic differentials manifest in highly interpretive metrics while remaining small in verifiable technical elements. To understand the structural dynamics of these patterns, I analyze complex scoring distributions, demonstrating how co-occurring reciprocal adjustments and selective point-spread variations can affect relative rankings while keeping overall panel averages statistically neutral.

The analysis then examines, for the first time, a historically overlooked dimension of officiating: the deterministic gatekeeping of the technical panel. Because technical specialists control the underlying Base Value of a performance, their decisions are highly consequential, yet their potential bias is difficult to untangle from a skater's actual physical mistakes. I address this identification problem by constructing an Expected Value (EV) framework that establishes baseline probabilities of element execution. This allows for the isolation of marginal leniency, exposing systemic patterns of infraction shielding and difficulty-level boosting that bypass the scoring system's built-in outlier filters.

Finally, I evaluate the broader mechanism design of the IJS scoring system by analyzing how bias is transmitted or suppressed through institutional rules. I introduce a key conceptual distinction between an evaluator's *Attempted Premium* and its *Realized Impact* on the leaderboard. By simulating a zero-bias counterfactual equilibrium, I demonstrate that while the trimmed-mean algorithm successfully filters out extreme outliers, fractional subjective bias and unmitigated technical favoritism accumulate to alter final standings in 40.7% of affected competitions and disrupt the podium in 11.8% of events. An applied case study of the 2026 Winter Olympic Games illustrates these dynamics, demonstrating that the highly compressed scoring environment of Ice Dance remained structurally vulnerable, leading to altered rankings in the final Olympic standings.

1.3 Ethical Considerations and Interpretive Framework

Before presenting the empirical findings of this study, it is necessary to establish a strict epistemological boundary regarding what statistical models can and cannot determine. Because this research involves the deanonymization of official scorecards and the naming of specific individuals and national federations, the results presented here should be interpreted through a precise, objective framework. The fundamental objective of this study is to explain and analyze structural vulnerabilities within the mechanism design of the International Judging System, not to prosecute the moral character of individual actors.

The Epistemological Limits of Statistical Evaluation

First, while the baseline-controlled models in this study identify systemic, mathematically significant deviations from objective scoring standards, these statistical anomalies must not be conflated with definitive proof of conscious malicious intent. Econometric analyses of subjective evaluation are inherently limited to observable outputs, as in this case, the points awarded on a scorecard. Data can expose highly robust, mathematically undeniable patterns of in-group favoritism and strategic point-spread variance, but it cannot peer into the conscious minds of the evaluators.

Extensive literature in cognitive psychology and sports sociology (e.g., Emerson et al. 2009, Lyngstad et al, 2020, Plessner and Haar, 2006) suggests that in-group bias frequently operates on a subconscious level. In highly pressurized, subjective environments, evaluators often rely on involuntary cognitive heuristics, such as the fluency heuristic, evaluating familiar, domestically trained athletes through a more forgiving lens than their international rivals (Ste-Marie 1996, Ste-Marie & Lee, 1991). Consequently, the statistical "premiums" documented in this study should be understood primarily as systemic measurement errors rather than definitive evidence of deliberate, pre-meditated attempts to manipulate standings.

Furthermore, any critique of individual judging behavior must be contextualized within the institutional architecture of the sport. Judges and technical specialists in elite figure skating currently operate in a profound informational vacuum. The overarching governing body does not provide officials with proactive, longitudinal statistical tracking of their own baseline leniency or infraction-calling habits. Without access to advanced programmatic data scraping and baseline-controlled feedback, it is highly probable that a majority of the officials identified in this study are entirely unaware of their own statistical deviations. Identifying specific judges and technical specialists by name in this research is a requirement for methodological transparency and scientific replicability; it is not an exercise in public shaming.

The Locus of Agency and Athlete Integrity

Second, it is crucial to strictly demarcate the athlete's execution from the evaluator's measurement error. The athletes identified in this analysis are the subjects of the evaluation process and possess absolutely no agency over the mechanics of their scoring. Skaters do not control international panel lotteries, the geographic distribution of seated officials, or the mathematical idiosyncrasies of the trimmed-mean algorithm.

When this study identifies a specific skater or team as a "verified beneficiary" of nationalistic bias or demonstrates via counterfactual simulation that an Olympic placement was altered by in-group premiums, this designation is a strict reflection of systemic variations in *judging behavior*. It is not an indictment of the athlete's performance, effort, or integrity. In a perfectly efficient scoring system, an athlete's expected score should remain constant regardless of the national composition of the panel.

By separating the structural critique of the evaluation mechanics from the athletic merits of the skaters, this research seeks to empower the global figure skating community. Acknowledging that the current system is vulnerable to both subconscious heuristics and strategic gamesmanship is the necessary first step toward implementing data-driven reforms required to protect the athletes and guarantee a truly level playing field.

2. Data Composition

To ensure a comprehensive and representative evaluation of judging and technical panel habits, this study utilizes a multi-season dataset spanning the entire 2022–2026 Olympic cycle. A multi-season approach is critical to ensure that observed scoring anomalies are generalized traits of the evaluating officials, rather than isolated reactions to specific, idiosyncratic programs from a single season.

The raw data comprises the complete scoring record contained in protocols from international, senior-level, competitions overseen by the ISU. To capture a broad spectrum of athletic ability, scoring scales, and evaluation panel compositions, the scope includes four competitive tiers:

1. Championships: e.g., Olympic Games, World Championships, Four Continents.
2. Grand Prix Series: e.g., Skate Canada, Internationaux de France, Grand Prix Final.
3. Challenger Series: e.g., Nebelhorn Trophy, Autumn Classic, Lombardia Trophy.
4. "B-Series" Internationals: e.g., Volvo Cup, Cranberry Cup, Tayside Trophy.

2.1 Data Extraction Architecture

Data extraction utilized an AI-driven, schema-constrained parsing architecture designed to overcome the structural inconsistencies inherent in official ISU PDF protocols. The raw protocol documents were processed using the Google GenAI SDK (Gemini-3.1-Flash-Lite) bound to strict data extraction schemas via the Pydantic library. To maximize contextual accuracy and prevent token-limit degradation, documents were processed in two-page chunks.

The model was programmed with domain-specific extraction rules, including the capture of specialized element annotations (e.g., edge calls, under-rotations), the identification of second-half bonus multipliers, and the mapping of aborted or invalid element marks to numerical zero. Extracted hierarchical data (segmented by performance, element, and component) was subsequently flattened into a granular, row-level dataset representing individual marks given by specific judges. To prevent identification collisions, unique identifiers were generated for every performance and element using MD5 hashing of concatenated metadata.

To ensure the highest fidelity of officials' data, the athlete scoring records extracted from the PDFs were merged with judging panel and technical committee data extracted directly from structured

HTML event pages using the BeautifulSoup Python library. This bypassed the ambiguity of PDF-based optical character recognition (OCR) for official names and national affiliations. Entity names (athletes and officials) were then rigorously standardized using a nation-bounded fuzzy string-matching algorithm (thefuzz) to detect and correct typographical variations and transposed names (e.g., First-Last vs. Last-First ordering) prior to final manual verification.

2.2 Automated Integrity Pipeline and Validation

Due to the unstandardized nature of event protocols, the raw extracted dataset was subjected to a strict, multi-stage automated integrity pipeline. Rather than omitting entire competitions due to isolated parsing errors, the pipeline utilized a targeted audit log to isolate and exclude only specific corrupted performances. Performances were excluded from the analytical sample if they failed any of the following deterministic checks:

1. Structural Consistency**:** Performances were required to exhibit an unbroken numerical sequence of executed technical elements, exactly three program components (characteristic of the updated ISU judging system), and continuous event ranking sequences.

2. Panel Dimensionality**:** Extracted score arrays were required to precisely match standard ISU panel dimensions. Anomalies such as "ghost columns" or asymmetric panel sizes across different elements within the same performance resulted in exclusion.

3. Top-Down Mathematical Reconciliation**:** The sum of the Total Element Score (TES) and Total Component Score (TCS), minus recorded deductions, was required to perfectly match the reported Total Segment Score (TSS).

4. Bottom-Up Mathematical Validation**:** To safeguard against undetected LLM hallucinations or extraction artifacts, summary metrics were empirically reconstructed from the raw individual judge inputs. Component scores were recalculated using the ISU's official trimmed-mean algorithm multiplied by the appropriate component factor. Element base values and panel Grades of Execution (GOE) were similarly aggregated. Performances in which the mathematically reconstituted summary scores deviated from the official reported scores by a tolerance exceeding 0.05 points were excluded.

Finally, strict row-level deduplication was executed to eliminate "fan-out" errors, ensuring that every judge is represented by exactly one score per technical element or component.

Table 1.

a)

Metric	Count
Unique Performances	14,382
Unique Skaters/Teams	1,044
Unique Judges	502
Unique Technical Specialists	165

b)

Discipline	Segment	Performances	Score (Mean ± SD)
Ice Dance	Free	1,573	99.08 ± 15.06
Ice Dance	Rhythm	1,669	64.07 ± 10.61
Men	Free	1,993	131.45 ± 27.43
Men	Short	2,093	68.84 ± 13.92
Pairs	Free	795	109.39 ± 17.39
Pairs	Short	846	58.63 ± 9.10
Women	Free	2,623	95.69 ± 22.70
Women	Short	2,790	50.57 ± 11.13

Table 1: Composition and descriptive statistics of the final analytical dataset (2022–2026 Olympic Cycle). (a) Aggregate unique counts of evaluated performances, athletes/teams, seated judges, and primary technical specialists. **(b)** Segment-level breakdown of the final sample across the four competitive disciplines, reporting the total number of unique performances and the corresponding mean Total Segment Scores (TSS).

The finalized, fully verified analytical dataset encompasses N = 14,382 unique performances across all four disciplines (Men, Women, Pairs, Ice Dance) and their respective segments (Short/Rhythm and Free) (Table 1). The data captures the evaluations of 502 unique judges and 165 unique technical specialists evaluating 1044 distinct athletes or teams (Supplementary Table 1). Examining the intersection of officials reveals significant cross-discipline overlap. While many judges are specialized, a substantial cohort actively officiates across multiple disciplines, necessitating the discipline- and segment-specific baselines utilized in subsequent statistical modeling (Figure 1).

Figure 1.

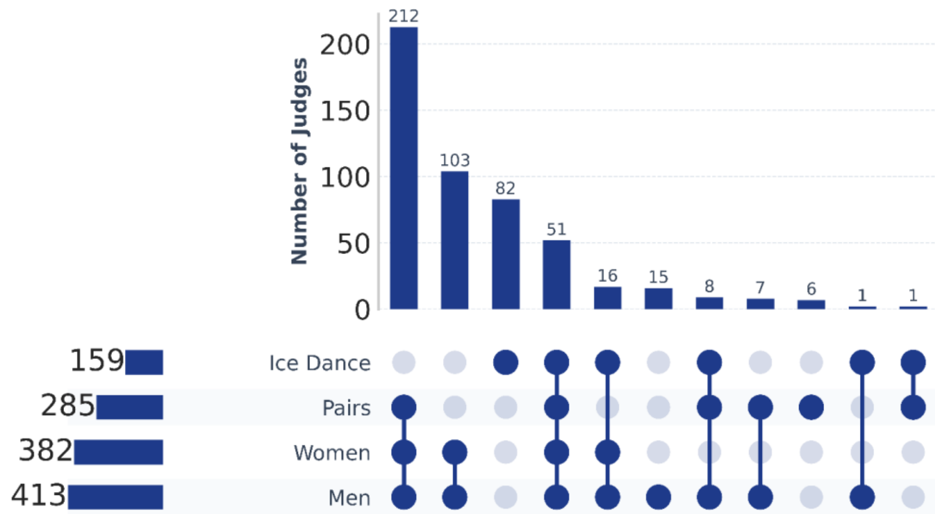


Figure 1. Cross-discipline certification profiles and overlap of seated international judges.
UpSet plot visualizes the shared credential profiles of the judging pool across the four competitive disciplines (Ice Dance, Pairs, Women's Singles, and Men's Singles). The horizontal bar chart at bottom-left displays the total number of unique judges certified to evaluate each individual discipline (set sizes). The main vertical bar chart displays the size of each exclusive intersection subset (the number of judges sharing that exact combination of credentials). The connected dark-blue circles beneath the vertical bars denote the specific combination of disciplines represented in each intersection.

3. Global Evidence of In-Group Favoritism

Before isolating the specific evaluation habits of individual federations or officials, I first establish the macroscopic presence and magnitude of nationalistic bias across the sport as a whole. At the aggregate level, the analytical objective is to determine whether the global population of ISU judges awards systematically higher Total Segment Scores (TSS) to compatriots compared to the international consensus.

3.1 Macro-Level Methodology

To quantify this global effect, the raw point differential for every evaluation in the dataset was calculated. For each performance, I isolated the score awarded by the home-country judge (if present) and compared it against the mean score awarded by the out-of-group (international) panel members evaluating that same performance.

This generates two aggregate distributions of scoring deviations: one representing in-group same-nation evaluations and one representing out-of-group evaluations. To ensure the robustness of the findings against potential non-normality in scoring extremes, I applied both parametric (Welch’s unequal variances t-test) and non-parametric (Mann-Whitney U) tests. To control for multiple comparisons across the eight specific discipline segments, all p-values are adjusted using the

Benjamini-Hochberg False Discovery Rate (FDR). Finally, I calculated Cohen's *d* to quantify the standardized effect size of the in-group premium.

3.2 Segment-Level Results

The aggregate analysis reveals a pervasive nationalistic premium across every discipline and segment in elite figure skating. As illustrated by the shifted frequency density distributions in the dataset, home-country judges systematically award higher scores to compatriots than their international peers evaluating the same performances (Figure 2).

Figure 2.

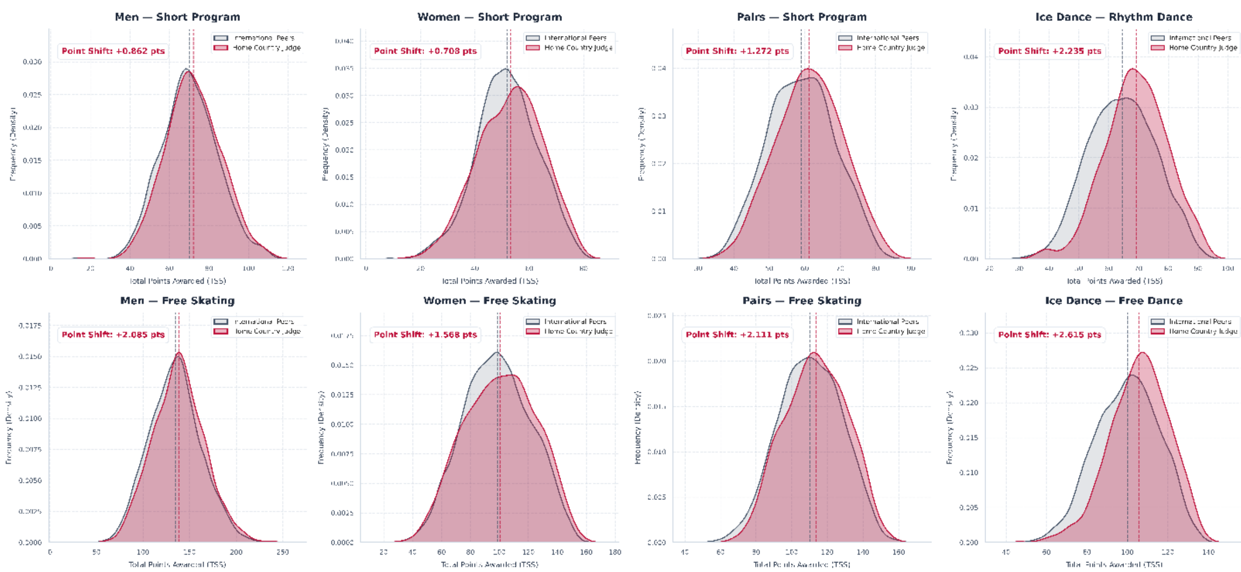


Figure 2: Total segment score distributions by competitive segment comparing home-country judges and international peers. Kernel density estimation curves displaying the distribution of Total Segment Scores (TSS) awarded by home-country judges (shaded in red) against the evaluations of the remaining international panel members (shaded in grey) across all eight discipline segments. The black vertical dashed line represents the mean score awarded by international peers, while the red vertical dashed line represents the mean score awarded by the home-country judge. The highlighted Point Shift value in each sub-panel quantifies the raw point differential between these two means.

This in-group premium is statistically significant across all eight segments (FDR-adjusted $p < 10^{-31}$ in all tests). The magnitude of the bias contributes an average home-judge advantage ranging from 0.708 to 2.615 points, depending on the discipline and segment (Table 2). The magnitude of the bias correlates with the subjective nature of the discipline. Ice Dance, which lacks discrete, easily quantifiable jumping elements and relies heavily on holistic subjective evaluation, exhibits the most severe in-group favoritism. The Ice Dance Free Dance yields the highest absolute point inflation (2.615 points), while the Ice Dance Rhythm Dance yields the highest standardized effect size, Cohen's $d = 0.774$ (Table 2). Conversely, the Men's and Women's Short Programs, segments constrained by strict required jumping passes and shorter durations, exhibit the lowest absolute premiums (0.862 and 0.708 points, respectively), alongside the lowest effect sizes, Cohen's d = 0.347 and 0.356 (Table 2).

Table 2.

Segment	Avg Home Bonus (Pts)	Sample Size (Home)	Sample Size (Intl)	FDR T-Test p-value	FDR M-W p-value	Effect Size (Cohen's d)	Significant?
Ice Dance Free Dance	+2.615	940	10,102	8.8561e-101	2.1428e-103	0.754	YES
Ice Dance Rhythm Dance	+2.235	978	10,922	2.7138e-112	6.8235e-109	0.774	YES
Pairs Free Skating	+2.111	578	5,513	4.8145e-40	6.1720e-36	0.583	YES
Men Free Skating	+2.085	1,174	12,900	6.5037e-55	1.6678e-51	0.476	YES
Women Free Skating	+1.568	1,433	15,920	2.9624e-64	2.5135e-58	0.461	YES
Pairs Short Program	+1.272	600	5,877	2.6430e-40	6.0552e-39	0.575	YES
Men Short Program	+0.862	1,218	13,771	8.1968e-33	1.7132e-31	0.347	YES
Women Short Program	+0.708	1,514	17,208	1.3394e-41	9.0224e-41	0.356	YES

Table 2: Statistical significance of nationalistic scoring premiums across competitive segments. Segment-level statistical analysis comparing the Total Segment Scores (TSS) awarded by home-country judges against the consensus of their international peers. For each segment, the table reports the average absolute point premium (average home bonus), sample sizes for both groups, and the adjusted p-values derived from both parametric (Welch's unequal variances t-test) and non-parametric (Mann-Whitney U) tests. Both sets of p-values are adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) procedure to control for multiple testing across the eight segments. Standardized effect sizes are quantified using Cohen's d.

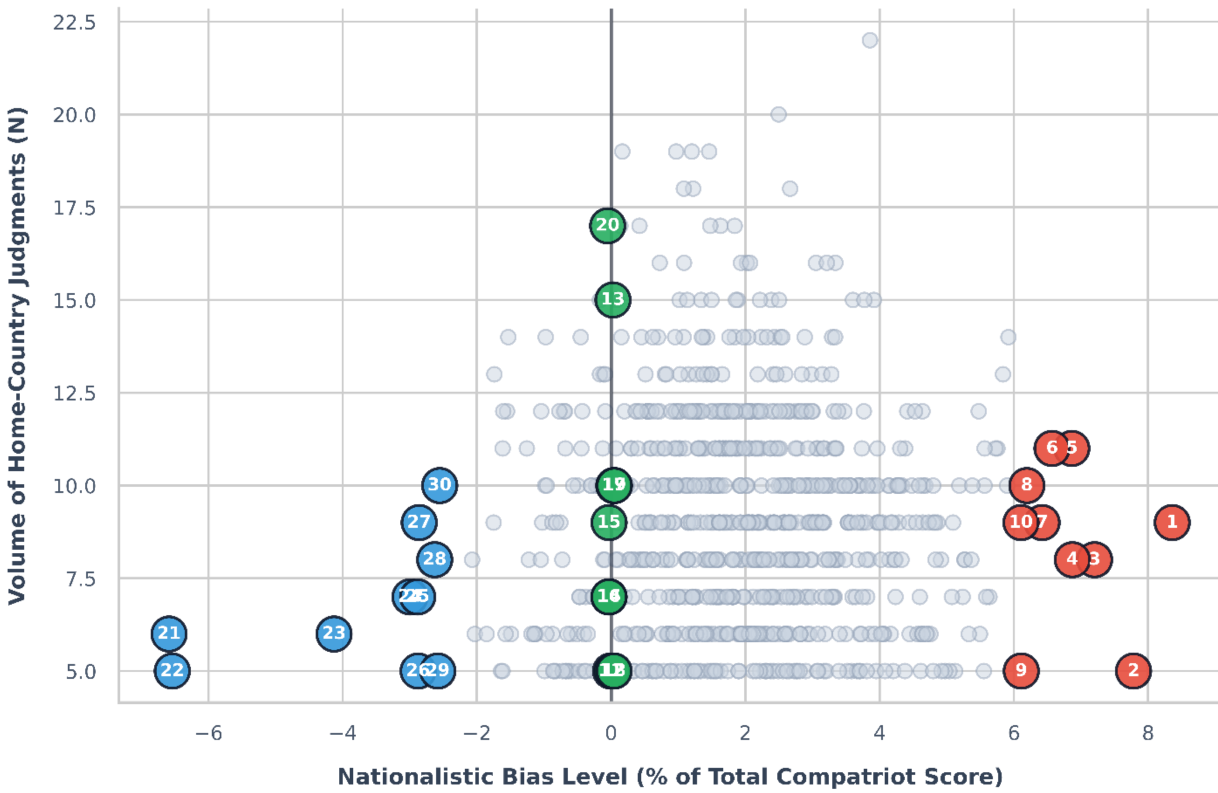
3.3 Competition-Level Volatility and Evaluation Equilibrium

While the segment-level aggregates confirm a systemic in-group nationalistic premium across the sport, this bias is not uniformly distributed across the competitive calendar. Aggregating the scoring deviations by specific event and segment reveals substantial volatility in both the magnitude and direction of nationalistic bias. Evaluating the nationalistic impact of each competition segment, expressed as the percentage shift in Total Segment Score (TSS) awarded to home-country athletes in the competition, demonstrates a wide dispersion across the dataset (Figure 3a, Supplementary Table 2).

The right tail of the distribution reveals competitions characterized by extreme in-group favoritism, where domestic judges awarded large premiums over the international panel consensus. For example, the upper bounds of the dataset include the 2023/2024 Golden Spin of Zagreb (Rhythm Dance) and the 2024/2025 Golden Spin of Zagreb (Rhythm Dance), which exhibited score inflations of 8.36% and 7.78%, respectively. Conversely, a distinct cluster of events sits closely at the zero-variance axis, demonstrating evaluation equilibrium. The data shows that near-perfect objectivity is achievable within the current judging framework. For instance, the 2025/2026 Challenge Cup (Free Dance) and the 2023/2024 Jelgava Cup (Women's Short) recorded net domestic bias impacts of -0.01% and +0.02%, respectively. Interestingly, the distribution is not strictly right-skewed, demonstrated by a significant left tail showing negative bias. In these specific environments, home-country judges applied a punitive standard to their compatriots. The most extreme recorded instance occurred at the 2022/2023 Philadelphia Summer International (Women's Short), where the domestic evaluation was 6.59% below the panel consensus (Figure 3a).

Figure 3.

a.



Most Punitive / Deflation			Most Objective / Equilibrium			Most Generous / Inflation		
21.	PHIINT (2022-23) women Short	[-6.59%]	11.	CHACUP (2025-26) dance Free	[-0.01%]	1.	GLDSPN (2023-24) dance Short	[+8.36%]
22.	CUPNIC (2025-26) men Short	[-6.54%]	12.	JELCUP (2023-24) women Short	[+0.02%]	2.	GLDSPN (2024-25) dance Short	[+7.78%]
23.	DIASPN (2023-24) women Free	[-4.13%]	13.	DIASPN (2024-25) women Free	[+0.03%]	3.	NEBTRO (2022-23) dance Short	[+7.20%]
24.	BAVOPN (2025-26) men Free	[-3.00%]	14.	CUPNIC (2024-25) women Short	[-0.03%]	4.	FINTRO (2023-24) dance Short	[+6.87%]
25.	LOUTRO (2024-25) women Short	[-2.88%]	15.	JNPCHA (2025-26) pairs Short	[-0.03%]	5.	LOMTRO (2024-25) dance Short	[+6.86%]
26.	SKTHEL (2023-24) women Free	[-2.88%]	16.	LOUTRO (2024-25) women Free	[-0.03%]	6.	LOMTRO (2025-26) dance Short	[+6.57%]
27.	SOFTRO (2022-23) women Free	[-2.86%]	17.	LOMTRO (2023-24) pairs Free	[+0.04%]	7.	GLDSPN (2022-23) dance Short	[+6.42%]
28.	TRITRO (2022-23) women Free	[-2.63%]	18.	DRATRO (2022-23) women Free	[+0.04%]	8.	BUDTRO (2022-23) dance Short	[+6.19%]
29.	CUPNIC (2025-26) men Free	[-2.58%]	19.	OLSMEM (2022-23) women Short	[+0.05%]	9.	CUPNIC (2023-24) women Free	[+6.11%]
30.	ROAD26 (2024-25) men Free	[-2.55%]	20.	4CC (2024-25) women Short	[-0.05%]	10.	CUPNIC (2024-25) dance Short	[+6.11%]

b.

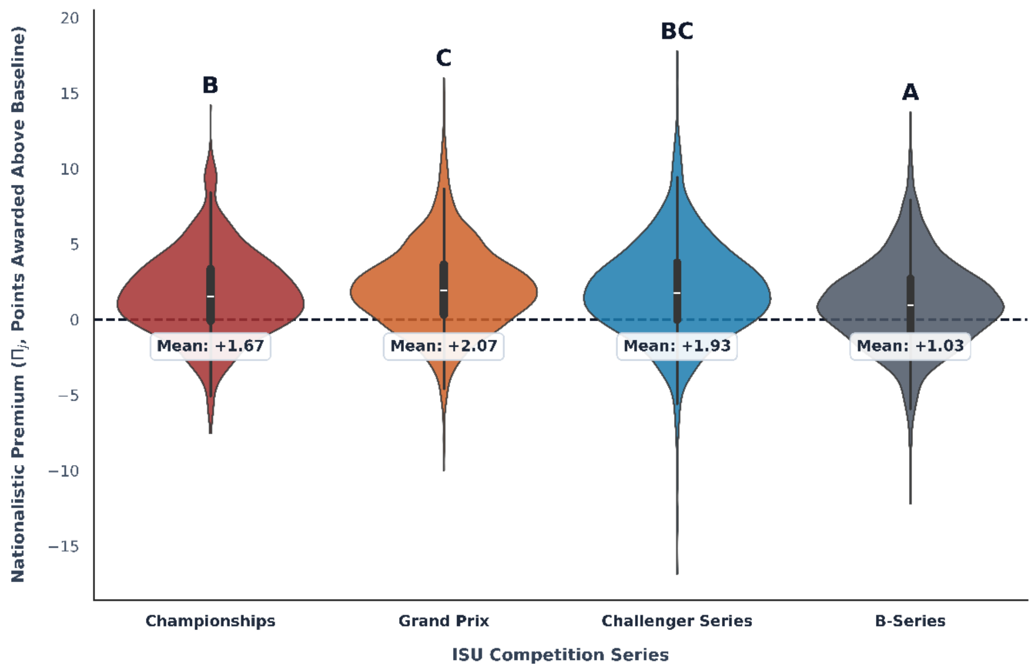


Figure 3: Competition-level distribution of nationalistic premiums. a) Scatter plot illustrating variation in nationalistic premiums across individual competition segments. The horizontal axis measures the nationalistic bias level, calculated as the percentage shift in Total Segment Score (TSS) awarded to home-country athletes relative to the international panel consensus. The vertical axis represents the total volume of home-country judgments recorded within each competition segment. The solid grey vertical line marks the axis of perfect evaluation equilibrium. Colored and numbered nodes highlight the extreme tails and center of the distribution. Specific competitive events and their corresponding percentage deviations are indexed below the plot. **b)** Violin plots showing distributions and average nationalistic premiums within competitive tiers. Letters above violins represent significant differences according to Tukey's highly significant differences (HSD) test.

Although there are significant differences in the magnitude of the average nationalistic premiums awarded between competitive tiers, the pattern observed in our comprehensive, multi-season dataset does not strongly reinforce previous investigations, such as Zitzewitz, 2006. The highest-profile, televised events comprising the Championship series show a mean nationalistic premium per skater of 1.67 points, which is significantly lower than the Grand Prix series at 2.07 points. However, the smallest mean nationalistic premium per skater is found for B-Series events at 1.03 points. Overall, the magnitude of nationalistic bias was highest in competitions of intermediate tier. A plausible interpretation of these results is that biased judging at competitions in the Grand Prix and Challenger tiers might help to secure athlete participation in the premier events such as the World Championships (Figure 3b).

This extreme competition-level volatility presents a critical methodological challenge. If judging behavior fluctuates aggressively based on the specific event, and if some officials actively penalize their compatriots in certain contexts, a simple global mean-deviation model is insufficient. To isolate whether an official is making strategic adjustments or simply adapting to a universally inflated or deflated event environment, we must transition from macroscopic observation to an individualized, baseline-controlled statistical decomposition.

4. Evaluating the Judging Panel: General Leniency vs. Targeted Bias

The foundational challenge in quantifying evaluation bias is disentangling a judge's systematic measurement error (general leniency or strictness) from targeted nationalistic favoritism a well as other strategic behaviors. A judge who exhibits high baseline leniency will naturally award higher scores to all athletes, including compatriots, generating false positives under simple mean-deviation models. Conversely, a highly monitored officiating environment may induce sophisticated evaluators to suppress the scores of their compatriots' closest competitors rather than directly inflating their compatriots' scores, thereby keeping their own mean deviations statistically neutral.

To isolate these distinct behavioral phenotypes, I transition from macroscopic observations to an individualized, baseline-controlled statistical decomposition. This framework evaluates judges along three orthogonal axes of variation:

1. **Baseline Leniency Offset (L_j^S):** The evaluator's natural propensity for systematic generosity or strictness relative to the panel consensus when judging international skaters.
2. **Adjusted In-Group Premium ($\pi_{j,p}$):** The deviation awarded exclusively to compatriots, isolating favoritism from baseline leniency.
3. **Rival Suppression (Σ_j):** The potential suppression of scores awarded to direct international rivals who threaten the leaderboard standing of the judge's compatriots.

4.1 Statistical Bias Decomposition

The primary unit of analysis for the judging panel is the Total Segment Score (TSS) awarded by individual judges. To isolate judging behavior independent of skater ability, we first calculate the **Performance-Level Deviation ($\Delta_{j,p}$)**:

$$\Delta_{j,p} = S_{j,p} - \bar{P}_p$$

where $S_{j,p}$ is the total score awarded by judge j to performance p, and $\bar{P}_p$ is the consensus mean score awarded by the entire judging panel for that specific performance. This metric mathematically neutralizes the objective quality of the performance, isolating the judge's individual variance.

Axis 1: Baseline Leniency Offset (L_j^S)

To avoid confounding domestic bias with general grading habits, we establish the judge’s baseline leniency strictly utilizing out-of-group (international) evaluations. Let I_j^S represent the set of all international (non-compatriot) performances evaluated by judge j within a specific discipline segment s (e.g., Men's Free Skate). The **Baseline Leniency Offset (L_j^s)** is defined as:

$$L_j^s = \frac{1}{|\ I_j^s\ |} \sum_{p \in I_j^s} \Delta_{j,p}$$

This value establishes the official's baseline grading bias (generosity if positive, strictness if negative) while neutralizing discipline-specific scoring scales.

Axis 2: Adjusted In-Group Premium ($\pi_{j,p}$)

With the baseline established, we isolate the targeted nationalistic effect. For any performance p where the athlete shares a nationality with judge j, the **Adjusted In-Group Premium ($\pi_{j,p}$)** is the residual scoring variance not explained by the panel consensus or the judge's general grading baseline:

$$\pi_{j,p} = \Delta_{j,p} - L_j^s$$

To evaluate systemic nationalistic bias at the aggregate official or federation level, we calculate the **Mean Nationalistic Premium (Π_j)**, representing the average adjusted premium across all compatriot evaluations:

$$\Pi_j = \frac{1}{|\ C_j\ |} \sum_{p \in C_j} \pi_{j,p}$$

where C_j represents the set of all compatriot performances evaluated by judge j. Positive values of Π_j represent targeted in-group favoritism, whereas negative values suggest a punitive domestic standard.

Axis 3: Strategic Rival Suppression (Σ_j)

While the Mean Nationalistic Premium captures blunt score inflation, actors operating under public and institutional scrutiny may utilize more discrete strategies. By strategically undervaluing direct competitors, an official can maximize the ordinal point-spread in favor of their compatriot without triggering outlier filters on the compatriot's own scorecard.

To model this behavior, we define a compatriot's direct competitor using segment standings. Let $Rank(p)$ denote the final ordinal rank of performance p within its specific competitive segment, calculated using the uncorrupted panel averages. For each segment s where judge j evaluates at least one compatriot $c \in C_{j,s}$, we partition the set of international evaluations $I_{j,s}$ into two mutually exclusive subsets:

Rivals ($R_{j,s}$): International skaters whose final segment rank falls within $\pm\delta$

places of any evaluated compatriot:

$$R_{j,s} = \{p \in I_{j,s} \mid \exists c \in C_{j,s} \text{ s.t. } \mid Rank(p) - Rank(c) \mid \leq \delta\}$$

where $\delta$is the ordinal proximity threshold (empirically set to $\delta = 2$).

Non-Rivals ($NR_{j,s}$): The remaining international skaters who do not pose an immediate threat to the compatriot's standing:

$$NR_{j,s} = I_{j,s} \setminus R_{j,s}$$

Using these subsets, we compute the judge's baseline-controlled grading behaviors toward rivals ($\bar{\pi}_{j,R}$) and non-rivals ($\bar{\pi}_{j,NR}$):

$$\bar{\pi}_{j,R} = \frac{1}{\mid R_j \mid} \sum_{p \in R_j} \pi_{j,p}$$

$$\bar{\pi}_{j,NR} = \frac{1}{\mid NR_j \mid} \sum_{p \in NR_j} \pi_{j,p}$$

The **Mean Rival Suppression (Σ_j)** metric is defined as the difference between the judge's adjusted scoring of non-threat international skaters and their scoring of direct rivals:

$$\Sigma_j = \bar{\pi}_{j,NR} - \bar{\pi}_{j,R}$$

A positive value ($\Sigma_j > 0$) demonstrates that the evaluator scored direct competitive threats harsher than they did other international skaters, relative to their own established baseline.

4.2 Statistical Testing and Hypothesis Framework

To ensure statistical robustness and prevent false positives, we impose strict evaluation thresholds:

- **Individual Judges:** Restricted to officials with at least 10 compatriot evaluations ($\mid C_j \mid \geq 10$) and at least 15 international evaluations ($\mid I_j \mid \geq 15$).
- **National Federations:** Restricted to federations with at least 20 compatriot evaluations ($\mid C_{\text{fed}} \mid \geq 20$).

We test two distinct null hypotheses for each officiating entity to verify the presence of active bias:

Hypothesis 1: Testing for Nationalistic Premium (Π_j):

The first hypothesis tests whether the judge evaluates compatriots under the same scalar parameters as the rest of the international field:

$$H_0: \mu_{\text{compatriot}} = \mu_{\text{international}}$$

$$H_1: \mu_{\text{compatriot}} > \mu_{\text{international}}$$

Because variance naturally differs between domestic and international evaluation sets, we deploy Welch's unequal variances t-test comparing the distribution of the judge's adjusted home-country scores ($\pi_{j,p}$ for $p \in C_j$) against their international baseline deviations ($\pi_{j,p} = 0$ for $p \in I_j$).

Hypothesis 2: Testing for Rival Suppression (Σ_j)

The second hypothesis tests whether the judge evaluates direct competitive rivals under the same parameters as non-threat international skaters:

$$H_0: \mu_{\text{rival}} = \mu_{\text{non-rival}}$$

$$H_1: \mu_{\text{rival}} < \mu_{\text{non-rival}}$$

We deploy a second, independent Welch's unequal variances t-test comparing the distribution of adjusted deviations awarded to rivals ($\pi_{j,p}$ for $p \in R_j$) against those awarded to non-rivals ($\pi_{j,p}$ for $p \in NR_j$). A minimum threshold of $|R_j| \geq 5$ and $|NR_j| \geq 5$ is required to execute the test.

Multiple Testing Correction

Because these tests are conducted simultaneously across a high volume of judges and federations, the probability of encountering Type I errors (false discoveries) increases. To control the false discovery rate (FDR), the Benjamini-Hochberg procedure is applied independently to all generated p-values for both tests. Officiating entities are flagged as exhibiting statistically significant bias on either axis only if their respective adjusted p-value satisfies:

$$p_{\text{FDR}} < 0.05$$

Using these dual significance tests, judges and federations are objectively categorized into distinct bias profiles: "Premium & Suppression" (exhibiting both direct inflation and rival suppression), "Premium Only", "Suppression Only", or "None".

4.3 Official-Level Results: Federations and Individuals

Applying the 3-axis baseline decomposition model to the dataset reveals that in-group favoritism is a highly pervasive, structural feature of the officiating pool, rather than an isolated anomaly driven by a small subset of rogue actors. Furthermore, the inclusion of the Rival Suppression metric (Σ_j) demonstrates that a measurable subset of evaluators engages in strategic non-compatriot evaluation that maximize domestic outcomes.

Federation-Level Phenotypes

At the national level, 34 of the 40 evaluated ISU member federations exhibit statistically significant evaluation bias after FDR correction (Table 3). Plotting each federation's Baseline Leniency Offset (L_j^s)against their Mean Nationalistic Premium (Π_j) in Panel A (Figure 4) reveals distinct evaluator phenotypes. The most severe federational inflations are observed for Georgia (2.846 points) and Hungary (2.737 points) The majority of biased federations cluster in the upper-left "Punitive Biased" quadrant. These federations operate with a strict baseline against the general international field ($L_j^s < 0$) while simultaneously awarding significant point premiums to their own athletes ($\Pi_j > 0$). Conversely, federations such as Ukraine and Kazakhstan occupy the "Generous Biased" quadrant, exhibiting systematic leniency to the entire international field but affording an additional, mathematically distinct premium to compatriots. Only six federations (Monaco, Serbia, Hong Kong, Denmark, Croatia, and Australia) failed to cross the threshold of statistical significance, suggesting that objective domestic evaluation is a minority behavior within the sport (Table 3).

Panel B (Figure 4) introduces the third axis of variation, mapping Rival Suppression (Σ_j) against the Nationalistic Premium. This reveals that the mechanics of bias are not uniform. Twenty-nine federations exhibit a "Premium Only" profile (shaded green); their judges systematically inflate compatriots but evaluate direct international rivals neutrally relative to their established baselines. However, three federations, Georgia (GEO), Hungary (HUN), and China (CHN), exhibit a "Dual-Axis" profile (shaded orange). These federations not only award premiums to their own skaters (GEO: +2.846 pts; HUN: +2.737 pts; CHN: +1.889 pts) but simultaneously suppress the scores of direct international threats. For example, Georgian judges systematically dock rivals an average of -0.733 points relative to their scoring of non-threat international skaters. This dual-action approach effectively maximizes the ordinal point-spread between their compatriots and the field. Interestingly, Azerbaijan (AZE) occupies a unique "Premium & Elevation" profile (shaded purple), awarding a significant domestic premium (+0.828 pts) but actively elevating the scores of their rivals (+0.698 pts), effectively compressing the point-spread.

Table 3.

Country	Home Asg.	Intl Asg.	Baseline Leniency Offset (L_j^s)	Mean Nat. Premium (Π_j)	Premium FDR p-value	Mean Rival Suppression (Σ_j)	Suppression FDR p-value	Evaluation Profile
GEO	109	1,586	-0.188 pts	+2.846 pts	**1.90e-21**	+0.733 pts	**1.05e-05**	**Premium & Suppression**
HUN	183	3,163	+0.019 pts	+2.737 pts	**3.47e-25**	+0.451 pts	**3.63e-03**	**Premium & Suppression**
LTU	98	1,308	-0.092 pts	+2.624 pts	**1.09e-15**	+0.263 pts	4.11e-01	**Premium Only**
ISR	33	1,013	+0.061 pts	+2.512 pts	**1.03e-06**	-0.526 pts	1.81e-01	**Premium Only**
UKR	160	2,592	+0.545 pts	+2.280 pts	**5.38e-16**	+0.346 pts	6.38e-02	**Premium Only**
KOR	232	2,343	+0.013 pts	+2.262 pts	**5.99e-24**	+0.174 pts	4.28e-01	**Premium Only**
ITA	672	5,564	-0.424 pts	+2.260 pts	**1.25e-62**	+0.033 pts	8.45e-01	**Premium Only**
KAZ	97	1,174	+0.418 pts	+2.178 pts	**6.05e-09**	+0.182 pts	6.84e-01	**Premium Only**
LAT	82	910	-1.179 pts	+2.175 pts	**1.24e-07**	-0.185 pts	6.84e-01	**Premium Only**
ESP	67	1,153	-0.626 pts	+2.147 pts	**2.72e-06**	+0.195 pts	6.84e-01	**Premium Only**
CAN	628	4,812	-0.546 pts	+2.101 pts	**1.32e-56**	+0.001 pts	9.94e-01	**Premium Only**
GER	337	4,509	-0.184 pts	+2.024 pts	**1.14e-26**	+0.128 pts	5.10e-01	**Premium Only**
AUT	209	3,587	-0.448 pts	+1.945 pts	**6.27e-19**	-0.052 pts	8.45e-01	**Premium Only**
CHN	210	1,653	+0.054 pts	+1.890 pts	**1.09e-15**	+0.475 pts	**8.59e-03**	**Premium & Suppression**
USA	1,300	6,530	-0.262 pts	+1.719 pts	**1.62e-64**	+0.071 pts	6.84e-01	**Premium Only**
NED	173	2,771	-1.067 pts	+1.689 pts	**1.19e-13**	+0.236 pts	3.47e-01	**Premium Only**
FRA	550	4,547	-0.031 pts	+1.673 pts	**2.04e-25**	+0.072 pts	6.84e-01	**Premium Only**
GBR	220	2,566	+0.072 pts	+1.630 pts	**3.35e-12**	-0.027 pts	9.14e-01	**Premium Only**
ROU	32	704	+0.267 pts	+1.619 pts	**3.21e-03**	+0.322 pts	4.79e-01	**Premium Only**
CZE	172	3,035	-0.144 pts	+1.535 pts	**8.47e-15**	+0.098 pts	6.84e-01	**Premium Only**
BEL	37	962	+0.008 pts	+1.476 pts	**3.39e-04**	+0.272 pts	6.21e-01	**Premium Only**
SVK	139	2,914	-0.069 pts	+1.287 pts	**9.05e-07**	-0.102 pts	7.05e-01	**Premium Only**
TUR	126	1,468	+0.792 pts	+1.189 pts	**1.48e-06**	+0.102 pts	7.56e-01	**Premium Only**
FIN	231	2,410	-0.534 pts	+1.160 pts	**1.61e-08**	+0.034 pts	8.94e-01	**Premium Only**
MEX	30	915	-0.194 pts	+1.089 pts	**3.16e-02**	+0.297 pts	6.84e-01	**Premium Only**
MON	23	771	-0.571 pts	+1.028 pts	7.08e-02	+0.012 pts	9.94e-01	N/A
NOR	69	1,029	-0.281 pts	+0.989 pts	**1.36e-03**	+0.074 pts	8.48e-01	**Premium Only**
SRB	28	894	+0.056 pts	+0.947 pts	7.08e-02	+0.359 pts	4.28e-01	N/A
SWE	129	1,510	-0.515 pts	+0.908 pts	**1.47e-04**	+0.305 pts	3.21e-01	**Premium Only**
SUI	148	2,415	-0.188 pts	+0.903 pts	**1.64e-04**	+0.023 pts	9.14e-01	**Premium Only**
HKG	26	291	-0.188 pts	+0.856 pts	1.25e-01	-0.185 pts	7.56e-01	N/A
POL	480	3,755	+0.082 pts	+0.836 pts	**2.30e-09**	-0.150 pts	4.28e-01	**Premium Only**
AZE	47	1,297	-0.675 pts	+0.828 pts	**4.77e-02**	-0.698 pts	**4.92e-02**	**Premium & Suppression**
JPN	624	3,341	-0.155 pts	+0.803 pts	**8.95e-10**	+0.153 pts	4.28e-01	**Premium Only**
SLO	87	1,121	+0.107 pts	+0.683 pts	**2.05e-02**	+0.031 pts	9.14e-01	**Premium Only**
BUL	143	1,350	+0.585 pts	+0.658 pts	**7.34e-03**	-0.062 pts	8.45e-01	**Premium Only**
EST	225	2,401	+0.399 pts	+0.415 pts	**4.82e-02**	+0.122 pts	6.84e-01	**Premium Only**
DEN	22	522	+0.217 pts	+0.322 pts	5.65e-01	+0.179 pts	6.84e-01	N/A
CRO	37	1,089	-0.350 pts	+0.311 pts	5.14e-01	-0.169 pts	6.84e-01	N/A
AUS	102	2,165	+0.267 pts	+0.298 pts	3.73e-01	+0.485 pts	5.43e-02	N/A

Table 3: Federation-level baseline leniency, mean nationalistic premiums and rival suppression metrics. Federation-level statistical decomposition of baseline leniency offsets (L_j^s), mean nationalistic deviations (Π_j and Rival Suppression (Σ_j) across 40 national federations meeting the minimum sample threshold of N ≥ 20 compatriot evaluations. Adjusted p-values are generated using Welch's unequal variances t-test with Benjamini-Hochberg False Discovery Rate (FDR) corrections applied to control for multiple testing. A significance threshold of adjusted FDR $p < 0.05$ was utilized to determine statistical significance.

Figure 4.

Figure 4: Dual-panel bivariate mapping of federation-level evaluation phenotypes across three orthogonal axes of bias.

Dual-panel bubble plots illustrating the distribution of in-group scoring deviations and relative point-spread variance among federations during the 2022–2026 Olympic cycle. In both panels, the vertical axis measures the Mean Nationalistic Premium (Π_j, in points), representing the average baseline-adjusted point deviation awarded to compatriots relative to the international consensus. Positive values on this axis indicate relative in-group point premiums, while negative values indicate relative in-group strictness. (a) Panel A maps the Baseline Leniency Offset (L_j^S, horizontal axis) against the Mean Nationalistic Premium, illustrating the relationship between a judge's compatriot evaluation and their inherent strictness or leniency toward the general international field. (b) Panel B maps the Mean Rival Suppression metric (Σ_j, horizontal axis) against the Mean Nationalistic Premium. Across both panels, individual bubble diameters are scaled proportionally to the federation's total volume of domestic evaluations. Three-letter country codes identify specific national federations. Node coloration denotes the federation's statistically significant bias profile after controlling for multiple comparisons via Benjamini-Hochberg FDR adjustments: orange denotes "Premium & Suppression" (Dual-Axis bias), green denotes "Premium Only", purple denotes "Premium & Elevation", pink denotes "Suppression Only", and grey denotes federations that did not meet the threshold for statistical significance on either axis.

Decomposing the data to the individual judge level (N = 290) further elucidates this divergence and exposes the full theoretical spectrum of judging behavior. The dual-normalization protocol identifies 157 individual judges who award mathematically significant in-group premiums (Supplementary Table 3.).

Figure 5.

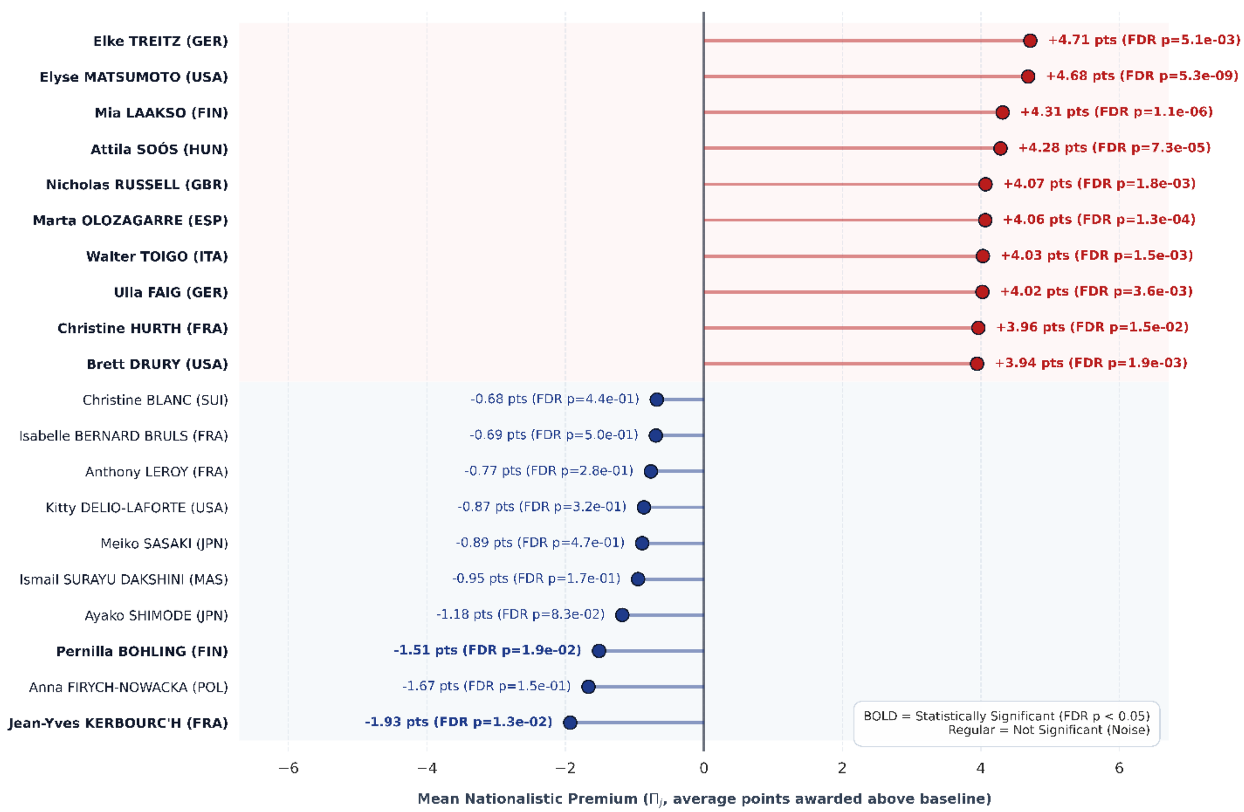


Figure 5: Extreme bounds of the individual judge nationalistic premium spectrum.
Lollipop chart displaying the ten highest (shaded in red, positive) and ten lowest (shaded in blue, negative) individual judge nationalistic deviations (Π_j recorded during the 2022–2026 Olympic cycle. The horizontal axis measures the mean adjusted points awarded to compatriots above the panel consensus. Bold typography denotes officials whose deviations are statistically significant after controlling for multiple comparisons via the Benjamini-Hochberg False Discovery Rate (FDR).

Examining the absolute extremes of this spectrum, the "Top 10" officials exhibiting the largest positive deviations versus the "Bottom 10" officials exhibiting the most extreme negative deviations, reveals a stark asymmetry in both magnitude and statistical significance. The upper bounds of individualized bias exceed federation averages. The highest systemic premiums in the 2022–2026 cycle were recorded by Elke Treitz (GER, +4.71 points, FDR $p = 5.1 \times 10^{-3}$), Elyse Matsumoto (USA, +4.68 points, FDR $p = 5.3 \times 10^{-9}$), and Mia Laakso (FIN, +4.31 points, FDR $p = 1.1 \times 10^{-6}$) (Figure 5). Conversely, the lower bound of the spectrum captures officials who ostensibly disadvantage their own skaters, awarding negative in-group premiums. Statistically significant punitive deviations were recorded by judges Pernilla Bohling (FIN, -1.51 pts) and Jean-Yves Kerbourc'h (FRA, -1.93 pts). While the entire "Top 100" positively biased cohort easily clears the FDR significance threshold, judges exhibiting statistically significant negative bias are exceedingly rare. This asymmetry confirms that awarding positive nationalistic premiums is a systemic, mathematically undeniable behavior, whereas punitive domestic judging is infrequent.

Mapping these premiums against the judges' Baseline Leniency Offsets in a bivariate scatterplot (Figure 6A) demonstrates the underlying mechanics of this favoritism. The majority of the highest-magnitude officials cluster in the "Punitive Biased" quadrant. These judges operate with a strict baseline against the general international field but award premiums to their compatriots, effectively maximizing the relative point advantage for their own skaters. Judges exhibiting significant positive nationalistic premiums evaluate performances at the highest possible levels of competition, exemplified by 31 of these judges who participated in judging panels at the 2026 Olympic Games (Figure 6).

The inclusion of the Rival Suppression metric (Σ_j) (Figure 6B) reveals more complex, multi-directional evaluation patterns. While the majority of officials exhibiting in-group favoritism demonstrate a "Premium Only" profile (shaded green), a distinct cohort exhibits statistically significant rival suppression. Nine officials exhibit a "Dual-Axis" profile (shaded orange), characterized by the simultaneous inflation of compatriot scores and the relative deflation of scores awarded to direct international rivals. This specific evaluation phenotype is observed for officials such as Salome Chigogidze (GEO), Richard Kosina (CZE), and Jeroen Prins (NED) (Figure 6, Supplementary Table 3).

Furthermore, Panel B isolates officials such as Valerie Grelet-Lambert (CYP), Todd Bromley (USA), and Ayako Shimode (JPN) who exhibit a "Suppression Only" profile (shaded pink). These judges do not award statistically significant premiums to their own skaters. However, they systematically award significantly lower scores to their compatriots' direct rivals relative to their own baseline evaluations of non-threat skaters (e.g., Grelet-Lambert evaluates rivals an average of -3.055 points lower, and Bromley by -2.784 points) (Figure 6B, Supplementary Table 3). Under standard mean-deviation analyses focused solely on domestic score inflation, these evaluation patterns would remain undetected. By decomposing evaluation into three orthogonal axes, this framework successfully captures the full spectrum of scoring variance, independently measuring a judge's general leniency, their baseline-adjusted evaluation of compatriots, and their relative scoring of direct competitors.

Figure 6.

** Point size is proportional to the number of home-country evaluations (N_{Home})*

*** Thick blue border denotes judges assigned to a 2026 Olympic panel**

Figure 6: Dual-panel bivariate mapping of individual judge evaluation phenotypes across three orthogonal axes of scoring variance.
Dual-panel bubble plots illustrating the distribution of in-group scoring deviations and relative point-spread variance among individual judges during the 2022–2026 Olympic cycle. In both panels, the vertical axis measures the Mean Nationalistic Premium (Π_j, in points), representing the average baseline-adjusted point deviation awarded to compatriots relative to the international consensus. Positive values on this axis indicate relative in-group point premiums, while negative values indicate relative in-group strictness. (a) Panel A maps the Baseline Leniency Offset (L_j^s, horizontal axis) against the Mean Nationalistic Premium, illustrating the relationship between a judge's compatriot evaluation and their inherent strictness or leniency toward the general international field. (b) Panel B maps the Mean Rival Suppression metric (Σ_j, horizontal axis) against the Mean Nationalistic Premium. Across both panels, individual bubble diameters are scaled proportionally to the judge's total volume of domestic evaluations. Node coloration denotes the official's statistically significant evaluation profile after controlling for multiple comparisons via Benjamini-Hochberg FDR adjustments): orange denotes "Premium & Suppression", green denotes "Premium Only", purple denotes "Premium & Elevation", pink denotes "Suppression Only", and grey denotes judges whose evaluations did not meet the threshold for statistical significance on either axis. Numbered nodes correspond to specific Judge IDs, which are fully indexed in Supplementary Table 3 located in the Appendix. Nodes highlighted with a thick blue border denote officials assigned to judging panels at the 2026 Milan-Cortina Winter Olympic Games.

5. Category-Level Bias Decomposition: The Distribution of Premiums

Having established the presence and magnitude of the Total Segment Score premium, I next investigated its internal distribution. The International Judging System is highly granular, requiring judges to evaluate many different types of distinct technical elements (e.g., jumps, spins, lifts, step sequences) as well as the holistic Program Components (Composition, Presentation, Skating Skills).

Economic models of evaluation suggest that in-group preferences will disproportionately manifest in subjective categories characterized by high information asymmetry, as evaluators possess far greater interpretive latitude there than in verifiable, objective technical elements. To test this hypothesis, I decomposed the nationalistic premium across individual scoring categories.

5.1 Category-Specific Methodology

Because judging habits might vary between technical and subjective metrics (e.g., a judge may be inherently strict on jumps but generous on more subjective components), the baseline decomposition must be calculated independently for every scoring category c within each discipline d.

First, we calculate the **Category-Specific Deviation ($\Delta_{j,p}^c$)**:

$$\Delta_{j,p}^c = S_{j,p}^c - \bar{P}_p^c$$

Where $S_{j,p}^{c}$ is the score awarded by judge j to performance p in category c, and $\bar{P}_{p}^{c}$ is the panel average for that specific category.

Next, we establish the **Category-Specific Leniency Offset ($L_j^{c,d}$)** strictly utilizing international evaluations ($I_j^{c,d}$) to capture the judge's baseline strictness for that specific element type within that discipline:

$$L_j^{c,d} = \frac{1}{|\, I_j^{c,d} \,|} \sum_{p \in I_j^{c,d}} \Delta_{j,p}^{c}$$

We then isolate the **Category-Specific Premium ($\pi_{j,p}^{c}$)**:

$$\pi_{j,p}^{c} = \Delta_{j,p}^{c} - L_j^{c,d}$$

Because scoring scales vary drastically between categories (e.g., A singles Free program comprises 7 jumping passes that are worth significantly more absolute points than the three spins), comparing raw point premiums is insufficient. To allow for cross-category comparison, we calculate the **Relative Bias Intensity (Φ_c)**, representing the mean category premium normalized as a percentage of the global average score awarded for that category:

$$\Phi_c = \frac{Mean(\pi_{j,p}^{c})}{Mean(\bar{P}_{p}^{c})} \times 100$$

Statistical significance is determined via Welch's unequal variances t-test comparing the distribution of domestic category premiums against international category premiums, with Benjamini-Hochberg FDR adjustments applied across all evaluated categories within each discipline (threshold: adjusted $p < 0.05$).

5.2 Results: Subjectivity and Verifiability

The category-level decomposition strongly confirms the verifiability hypothesis in that the magnitude of nationalistic inflation scales directly with the subjectivity of the element being evaluated. While all categories exhibit statistically significant in-group premiums, the distribution of bias is heavily skewed toward interpretive elements and the more artistic Component scores (Figure 7).

Across the Men's, Women's, and Pairs disciplines, the highest Relative Bias Intensities (Φ_c) are universally found in the Choreographic Sequence (Men: +3.0%, Women: +2.7%, Pairs: +4.2%). This element lacks strict geometric requirements and relies heavily on holistic visual impression, providing a mathematically safe vector for point inflation. Similarly, the Composition Component score consistently absorbs the second-highest levels of bias in these disciplines. Conversely, discrete technical elements governed by strict, verifiable biomechanical criteria exhibit the lowest levels of bias. Jumps are the least inflated elements in both Men's (+0.9%) and Women's (+0.9%) singles. In Pairs, Jump elements and Pairs Spins anchor the bottom of the bias distribution (+1.3% and +1.2%, respectively).

Figure 7.

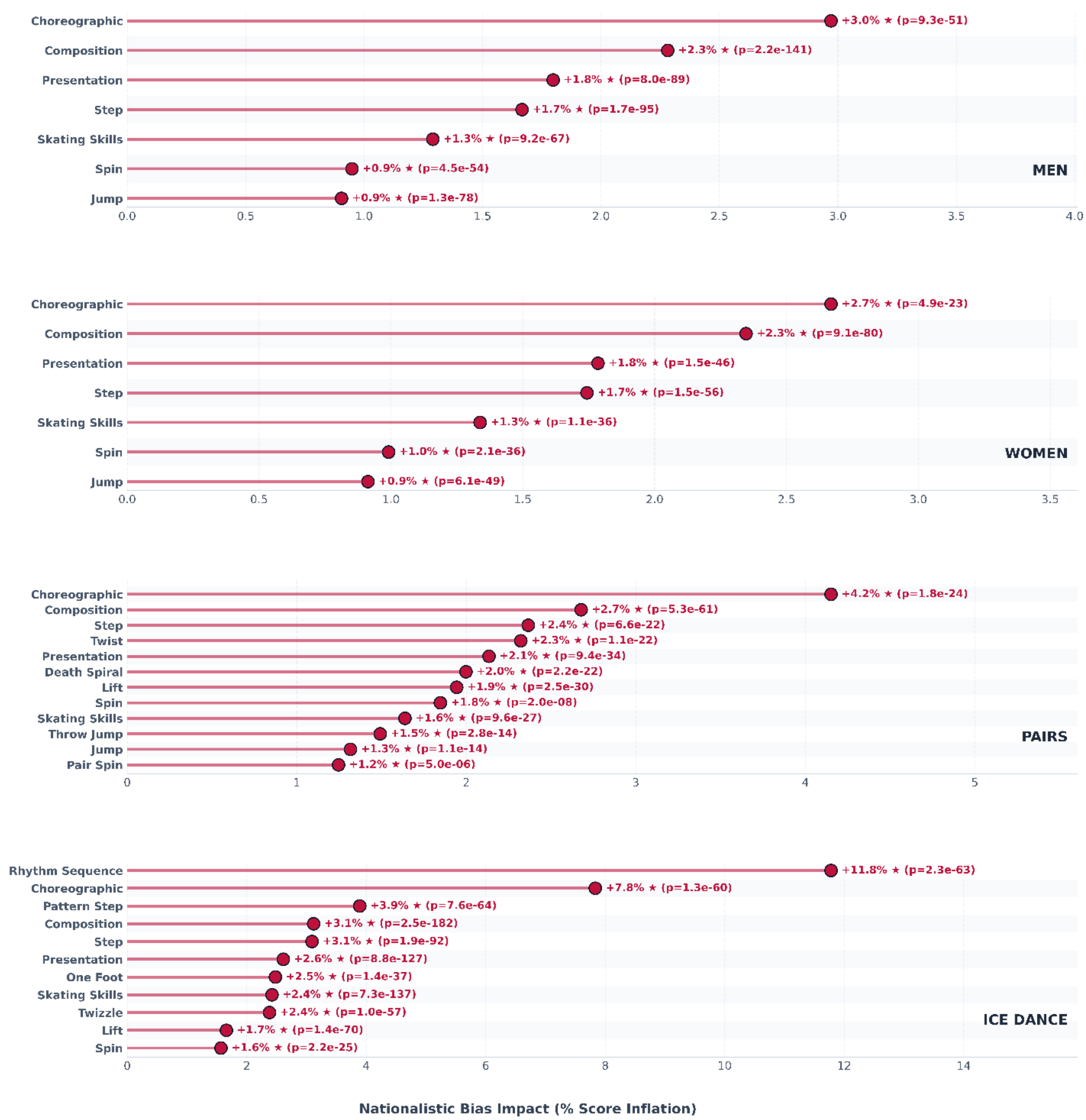


Figure 7: Category-specific decomposition of nationalistic bias intensity across disciplines.
Lollipop charts displaying the Relative Bias Intensity (Φ_c for individual technical elements and holistic Program Components across the four figure skating disciplines. The horizontal axis measures the mean category-specific premium normalized as a percentage of the global average score awarded for that specific category. Dark-blue bars marked with an asterisk (*) denote scoring categories that are statistically significant after controlling for multiple testing via Benjamini-Hochberg FDR adjustments (FDR $p < 0.05$). Light-grey bars denote categories that did not clear the significance threshold (e.g., Pair Spin).

The most extreme manifestation of category-specific bias occurs in Ice Dance, which is devoid of jump elements. The "Rhythm" sequence element in the Rhythm Dance segment exhibits an unprecedented +11.8% score inflation (FDR $p = 2.3 \times 10^{-63}$), alongside large inflations in the Choreographic (+7.8%) and Pattern (+3.9%) elements (Figure 7). This hyper-inflation underscores a fundamental vulnerability of the IJS. In the absence of objective, verifiable technical constraints, subjective scores can be adjusted to maximize domestic outcomes.

6. Identifying Significant Beneficiaries: The Skater Assessments

While demonstrating that judges systematically exhibit scoring deviations is critical for understanding evaluation behavior, the fundamental economic consequence of this bias lies in its impact on the athletes. In elite figure skating, placements determine not only championship medals but also national federation funding, international assignments, and endorsement viability.

To quantify the real-world impact of nationalistic bias, I shift the unit of analysis from the *evaluator* to the *athlete*. However, simply identifying a skater who receives high scores from a compatriot judge is insufficient to demonstrate bias. To rigorously eliminate false positives, I introduce a double-verification protocol requiring simultaneous statistical significance across two independent methodologies.

6.1 The Double-Verification Protocol

For an athlete (x) to be classified as a verified beneficiary of nationalistic bias, their domestic evaluations must pass two distinct intersectional thresholds.

Method 1: Panel Consensus Deviation **(t_1):** This method tests whether compatriot judges score the athlete systematically higher than the rest of the international panel evaluating the same performances. Welch's t-test was used to compare the distribution of home-country deviations (Δ_{home}) against international deviations (Δ_{intl}). This filter eliminates naturally strict home judges who merely graded the compatriot at the panel average.

Method 2: Baseline Departure **(t_2):** A skater might be detected as a significant beneficiary using Method 1 simply because their home-country judge is universally lenient to all competitors. To control for this, Method 2 tests whether the home judge broke their *own* established grading habits specifically for this athlete. A 1-sample t-test was used to evaluate whether the Adjusted In-Group Premiums ($\pi_{j,p}$, as calculated in Section 4.1) awarded to skater x are significantly greater than zero ($\mu_0 = 0$).

A skater/team is flagged only if both tests achieve statistical significance after applying the Benjamini-Hochberg False Discovery Rate (FDR) correction ($Adj.p_1 < 0.05 \cap Adj.p_2 < 0.05$). For verified athletes, I calculated the Expected Nationalistic Premium per Outing **($E[\Pi_x]$)**, representing the average expected point bonus the skater mathematically expects to receive from compatriot judges, entirely independent of their actual performance quality.

6.2 Results: A Non-Level Playing Field for Athletes

Applying the double-verification protocol to the dataset identifies 159 individual athletes or teams who were mathematically significant beneficiaries of systematic nationalistic inflation during the 2022–2026 Olympic cycle. These athletes compete at the highest possible level, as 38 individual athletes or teams participated in the 2026 Olympic Games (Figure 8, Supplementary Table 4).

Figure 8.

Method 2: Baseline Departure (π_{home}, Points)
Method 1: Panel Consensus Deviation ($\Delta_{home} - \Delta_{intl}$, Points)
Significance Classes
Double-Verified (Δ & π Sig.)
Method 1 Only (Consensus Dev.)
Method 2 Only (Baseline Departure)
Not Significant (Noise)
N (Home Assignments)
10
20
30

KEY: Double-Verified Skaters (Significant in Both Methods)

★ = Assigned to an Olympic Panel | Blue Text = Competed at 2026 Winter Olympics | Disciplines: D = Ice Dance, P = Pairs, M = Men's, W = Women's

R	Skater	Cty	π_h
1	Paulina RAMANAUSKAITEDeividas KIZALA (D)	LTU	+4.59
2	★ Hyungyeom KIM (M)	KOR	+4.58
3	Russalina SHAKROVA (W)	KAZ	+4.47
4	★ Allison REEDSaulius AMBRULEVICIUS (D)	LTU	+4.39
5	Lucy HANCOCKIlias FOURATI (D)	HUN	+4.33
6	Phebe BEKKERJames HERNANDEZ (D)	GBR	+4.26
7	★ Charlene GUIGNARDMarco FABBRI (D)	ITA	+4.15
8	★ Daniel GRASSL (M)	ITA	+4.14
9	Marjorie LAJOIEZachary LAGHA (D)	CAN	+4.00
10	Hannah LIMYe QUAN (D)	KOR	+3.94
11	Emilea ZINGASVadym KOLESNIK (D)	USA	+3.89
12	Letizia ROSCHERLuis SCHUSTER (P)	GER	+3.86
13	Maria PAVLOVAAlexei SVIATCHENKO (P)	HUN	+3.83
14	Christina CARREIRAAnthony PONOMARENKO (D)	USA	+3.80
15	Lily HENSENNathan LICKERS (D)	CAN	+3.76
16	Natacha LAGOUGEArnaud CAFFA (D)	FRA	+3.75
17	Emily BRATTIlan SOMERVILLE (D)	USA	+3.60
18	Jiaxuan ZHANGYihang HUANG (P)	CHN	+3.44
19	★ Lorine SCHILD (W)	FRA	+3.41
20	Lou TERREAUXNoe PERRON (D)	FRA	+3.40
21	Aleksandr VLASENKO (M)	HUN	+3.40
22	★ Piper GILLESPaul POIRIER (D)	CAN	+3.36
23	Olivia SMARTTim DIECK (D)	ESP	+3.36
24	Caroline GREENMichael PARSONS (D)	USA	+3.30
25	Wesley CHIU (M)	CAN	+3.29
26	Natalie TASCHLEROVAFilip TASCHLER (D)	CZE	+3.16
27	★ Boyang JIN (M)	CHN	+3.16
28	Lilah FEARLewis GIBSON (D)	GBR	+3.12
29	Junfei RENJianing XING (D)	CHN	+3.10
30	Diana DAVISGleb SMOLKIN (D)	GEO	+3.10
31	Oona BROWNGage BROWN (D)	USA	+3.09
32	★ Andrew TORGASHEV (M)	USA	+3.08
33	Utana YOSHIDAMasaya MORITA (D)	JPN	+3.07
34	Alicia FABBRIPaul AYER (D)	CAN	+3.06
35	★ Julia SAUTER (W)	ROU	+3.04
36	Minerva Fabienne HASENikita VOLODIN (P)	GER	+2.99
37	Sara CONTINIccolo MACII (P)	ITA	+2.96
38	Marina PIREDDA (W)	ITA	+2.96
39	Madison CHOCKEvan BATES (D)	USA	+2.94
40	Yudong CHEN (M)	CHN	+2.91
41	Evgeniia LOPAREVAGeoffrey BRISSAUD (D)	FRA	+2.88
42	Starr ANDREWS (W)	USA	+2.88
43	★ Camille KOVALEVPavel KOVALEV (P)	FRA	+2.86
44	Gabriele FRANGIPANI (M)	ITA	+2.85
45	★ Lara Naki GUTMANN (W)	ITA	+2.81
46	Laurence FOURNIER BEAUDRYNikolaj SOERENSEN (D)	CAN	+2.81
47	Charise MATTHAEIMax LIEBERS (D)	GER	+2.80
48	Daria ZSIRNOV (W)	HUN	+2.78
49	★ Nika EGADZE (M)	GEO	+2.78
50	Niki WORIES (W)	NED	+2.78
51	Casper JOHANSSON (M)	SWE	+2.78
52	Nikita STAROSTIN (M)	GER	+2.77
53	Giulia Isabella PAOLINOAndrea TUBA (D)	ITA	+2.75
54	Sofiya FARAFONOVA (W)	KAZ	+2.68
55	★ Matteo RIZZO (M)	ITA	+2.66
56	Lela DOZZIPietro PAPETTI (D)	ITA	+2.63
57	Ivan SHMURATKO (M)	UKR	+2.63
58	Mariia IGNATEVADanijil Leonyidovics SZ..	HUN	+2.63
59	★ Deniss VASILJEVS (M)	LAT	+2.62
60	Elyce LIN-GRACEY (W)	USA	+2.60
61	Daniel MARTYNOV (M)	USA	+2.59
62	Lorraine McNAMARAAnton SPIRIDONOV (D)	USA	+2.58
63	★ Kevin AYMOZ (M)	FRA	+2.55
64	Sofiia HOLICHENKOArtem DARENSKYI (P)	UKR	+2.55
65	Loicia DEMOUGEOTTheo le MERCIER (D)	FRA	+2.52
66	Laurence FOURNIER BEAUDRYGuillaume CIZERON (D)	FRA	+2.50
67	Zixi XIAOLinghao HE (D)	CHN	+2.50
68	Ahsun YUN (W)	KOR	+2.49
69	Maria Sofia PUCHEROVANikita LYSAK (D)	SVK	+2.47
70	Gabriella IZZOLuc MAIERHOFER (P)	AUT	+2.47
71	★ Andreas NORDEBACK (M)	SWE	+2.45
72	★ Olga MIKUTINA (W)	AUT	+2.42
73	Seoyeong WI (W)	KOR	+2.41
74	Daria DANILOVAMichel TSIBA (P)	NED	+2.40
75	Eva PATELogan BYE (D)	USA	+2.38
76	Celina FRADJIJean-Hans FOURNEAUX (D)	FRA	+2.35
77	Flora Marie SCHALLER (W)	AUT	+2.34
78	Ginevra Lavinia NEGRELLO (W)	ITA	+2.34
79	Lucrezia BECCARIMatteo GUARISE (P)	ITA	+2.33
80	Shiyue WANGXinyu LIU (D)	CHN	+2.32
81	Maurizio ZANDRON (M)	AUT	+2.32
82	★ Mikhail SHAIDOROV (M)	KAZ	+2.31
83	Mariia PINCHUKMykyta POGORIELOV (D)	UKR	+2.30
84	Tomoki HIWATASHI (M)	USA	+2.23
85	★ Deanna STELLATO-DUDEKMaxime DESCHAMPS ..	CAN	+2.23
86	Liam KAPEIKIS (M)	USA	+2.22
87	Emese CSISZERMark SHAPIRO (D)	HUN	+2.21
88	Katerina MRAZKOVADaniel MRAZEK (D)	CZE	+2.20
89	Marie DUPAYAGEThomas NABAIS (D)	FRA	+2.18
90	Nikolaj MEMOLA (M)	ITA	+2.15
91	Koshiro SHIMADA (M)	JPN	+2.12
92	Valtter VIRTANEN (M)	FIN	+2.12
93	★ Anastasiia METELKINALuka BERULAVA (P)	GEO	+2.12
94	★ Niina PETROKINA (W)	EST	+2.11
95	Katie McBEATHDaniil PARKMAN (P)	USA	+2.10
96	★ Lia PEREIRATrennt MICHAUD (P)	CAN	+2.10
97	★ Anastasiia GUBANOVA (W)	GEO	+2.09
98	★ Maxim NAUMOV (M)	USA	+2.09
99	Stefanie PESENDORFER (W)	AUT	+2.07
100	Emily CHANSpencer Akira HOWE (P)	USA	+2.07
101	Hana CVIJANOVIC (W)	CRO	+2.06
102	★ Ilia MALININ (M)	USA	+2.05
103	Katinka Anna ZSEMBERY (W)	HUN	+2.03
104	Yuka ORIHARAJuho PIRINEN (D)	FIN	+2.03
105	Jakub LOFEK (M)	POL	+2.02
106	Sophia SCHALLERLivio MAYR (P)	AUT	+1.99
107	Juulia TURKKILAMatthias VERSLUIS (D)	FIN	+1.97
108	Rebecca GHILARDIFilippo AMBROSINI (P)	ITA	+1.95
109	Annika HOCKERobert KUNKEL (P)	GER	+1.93
110	Oceane PIEGADDenys STREKALIN (P)	FRA	+1.92
111	Jennifer JANSE van RENSBURGBenjamin STEF..	GER	+1.92
112	Zofia GRZEGORZEWSKAOleg MURATOV (D)	POL	+1.91
113	Anna SIMOVAKirill AKSENOV (D)	SVK	+1.91
114	Filip SCERBA (M)	CZE	+1.90
115	Catherine-Charlotte LAPIERRE (W)	POL	+1.89
116	★ Loena HENDRICKX (W)	BEL	+1.86
117	★ Junhwan CHA (M)	KOR	+1.83
118	Fatma Yade KARLIKLI (W)	TUR	+1.82
119	★ Madeline SCHIZAS (W)	CAN	+1.81
120	Kaiya RUITER (W)	CAN	+1.79
121	★ Amber GLENN (W)	USA	+1.77
122	★ Adam SIAO HIM FA (M)	FRA	+1.75
123	Zoe LARSONAndrii KAPRAN (D)	UKR	+1.75
124	★ Haein LEE (W)	KOR	+1.70
125	Aleksa RAKIC (M)	CAN	+1.68
126	Rion SUMIYOSHI (W)	JPN	+1.66
127	Wakaba HIGUCHI (W)	JPN	+1.65
128	Irma CALDARARiccardo MAGLIO (P)	ITA	+1.59
129	Mai MIHARA (W)	JPN	+1.59
130	Lea SERNA (W)	FRA	+1.57
131	★ Anastasia GOLUBEVAHektor GIOTOPOULOS MOOR..	AUS	+1.56
132	★ Donovan CARRILLO (M)	MEX	+1.55
133	★ Kimmy REPOND (W)	SUI	+1.54
134	Kristina ISAEV (W)	GER	+1.54
135	★ Kaori SAKAMOTO (W)	JPN	+1.52
136	Laura SZCZESNA (W)	POL	+1.50
137	Mihhail SELEVKO (M)	EST	+1.49
138	Valentina PLAZASMaximiliano FERNANDEZ (P)	USA	+1.45
139	★ Adam HAGARA (M)	SVK	+1.44
140	Chaeyeon KIM (W)	KOR	+1.42
141	Ellie KAMDanny O'SHEA (P)	USA	+1.42
142	Ceyda SAGLAM (W)	TUR	+1.39
143	Julija LOVRENCIC (W)	SLO	+1.38
144	Jacob SANCHEZ (M)	USA	+1.38
145	Maia MAZZARA (W)	FRA	+1.38
146	Kelly Ann LAURINLoucas ETHIER (P)	CAN	+1.35
147	Yagmur Derin KEVINC (W)	TUR	+1.34
148	Victoria MANNICarlo ROETHLISBERGER (D)	ITA	+1.34
149	★ Livia KAISER (W)	SUI	+1.33
150	Holly HARRISJason CHAN (D)	AUS	+1.32
151	David SEDEJ (M)	SLO	+1.21
152	★ Isabeau LEVITO (W)	USA	+1.20
153	Riku MIURARyuichi KIHARA (P)	JPN	+1.15
154	Sara-Maude DUPUIS (W)	CAN	+1.14
155	Camden PULKINEN (M)	USA	+1.12
156	Filip KAYMAKCHIEV (M)	BUL	+1.03
157	Eliska BREZINOVA (W)	CZE	+1.01
158	Yelim KIM (W)	KOR	+0.94
159	Jolanda VOS (W)	NED	-1.40

Figure 8: Bivariate mapping of individual skater and team assessments to identify verified beneficiaries of nationalistic bias. Bivariate bubble plot mapping individual skaters and teams across the two independent statistical thresholds of the double-verification protocol during the 2022–2026 Olympic cycle. The horizontal axis measures the panel consensus deviation (Method 1), comparing domestic deviations against international consensus deviations. The vertical axis measures the domestic baseline departure (Method 2), evaluating whether compatriot judges broke their own established grading habits specifically for that athlete (adjusted in-group premium). Individual bubble diameters are scaled proportionally to the total

volume of domestic evaluations received. Bubbles are color-coded by significance class: Double-Verified (shaded in dark red, nodes 1–40): Skaters who meet the significance threshold of adjusted FDR $p < 0.05$ under *both* independent methods. Method 1 Only (shaded in light blue): Skaters significant only under the panel consensus deviation test. Method 2 Only (shaded in orange): Skaters significant only under the baseline departure test. Not Significant (shaded in light grey): Skaters not significant in either test. Numbered nodes correspond to the forty double-verified skaters indexed in the key below. Athletes highlighted in blue font with an asterisk (*) represent the twenty-one unique skaters or teams who competed at the 2026 Winter Olympic Games.

The skater/team receiving the largest Expected Nationalistic Premium per Outing is the Ice Dance team of Paulina Ramanauskaite and Deividas Kizala (LTU, + 4.68 points). Previous analyses demonstrated that the magnitude of nationalistic bias is highest in the Ice Dance discipline and, unsurprisingly, there is a significant enrichment of Ice Dance teams on this list, as 7 of the top 10 participate in this discipline (Figure 8). An additional 38 skaters/teams achieved marginal significance as they were significant beneficiaries of bias according to one test but not both (Supplementary Table 4).

The empirical presence of these targeted premiums highlights a divergence from the ideal of a uniform evaluation environment. In a perfectly efficient scoring system, an athlete's expected score should remain constant regardless of the national composition of the panel. The data indicates this assumption does not hold; rather, systemic evaluation patterns generate measurable, unequal baseline conditions for specific competitors depending on the origin of the seated officials.

Furthermore, these results reveal that in-group premiums are not distributed homogeneously, even among athletes representing the same federation. Comparative analysis within the Women's Singles discipline illustrates this intra-country variance. Kaori Sakamoto, a multiple Olympic medalist and four-time World Champion, emerges as a double-verified recipient of positive in-group evaluation, receiving a significant point premium-per-outing of +1.52 points from Japanese officials above her international baseline (Figure 8). Conversely, her compatriot Rinka Watanabe exhibits a highly divergent evaluation profile, demonstrating marginal significance for punitive domestic grading, whereby she is systematically evaluated more strictly (-0.61 points) by domestic officials than by the international consensus. Jolanda Vos, representing the Netherlands, was the sole athlete in the dataset receiving a statistically significant punitive nationalistic premium of -1.40 points-per-outing. Conversely, her compatriot Niki Wories receives a generous average premium of + 2.78 points (Figure 8, Table 4, Supplementary Table 4).

Table 4.

Skater Name	Cty	Discipline	N Home	M1 Panel Consensus Dev.	M1 FDR p	M2 Baseline Departure	M2 FDR p	Double Ver?
Paulina RAMANAUSKAITE Deividas KIZALA	LTU	dance	13	+4.533 pts	9.16e-05	+4.591 pts	1.69e-04	YES
Hyungyeom KIM	KOR	men	15	+5.397 pts	2.41e-03	+4.575 pts	2.04e-03	YES
Russalina SHAKROVA	KAZ	women	10	+4.907 pts	4.48e-05	+4.474 pts	1.46e-03	YES
Allison REED Saulius AMBRULEVICIUS	LTU	dance	30	+4.390 pts	1.62e-15	+4.394 pts	1.43e-12	YES
Lucy HANCOCK Ilias FOURATI	HUN	dance	12	+4.975 pts	4.98e-04	+4.332 pts	1.66e-03	YES
Phebe BEKKER James HERNANDEZ	GBR	dance	20	+4.622 pts	2.93e-04	+4.260 pts	1.76e-04	YES
Nadiia BASHYNSKA Peter BEAUMONT	CAN	dance	10	+2.780 pts	5.97e-02	+4.172 pts	6.34e-03	NO
— SPECTRUM GAP: MIDDLE SKATERS OMITTED FOR PRESENTATION —								
Milania VAANANEN Filippo CLERICI	FIN	pairs	14	-1.607 pts	6.92e-03	-0.655 pts	3.09e-01	NO
Linzy FITZPATRICK Keyton BEARINGER	USA	pairs	12	-1.192 pts	2.51e-01	-0.952 pts	1.70e-01	NO
Olesja LEONOVA	EST	women	14	-0.186 pts	6.94e-01	-1.106 pts	1.16e-02	NO
Jolanda VOS	NED	women	14	-1.338 pts	4.53e-02	-1.396 pts	3.49e-02	YES
Xavier VAUCLIN	FRA	men	14	-1.237 pts	4.56e-01	-1.410 pts	3.89e-01	NO
Grace HANNS Danny NEUDECKER	USA	pairs	11	-2.351 pts	8.10e-02	-1.554 pts	2.16e-01	NO
Maria CHERNYSHOVA	AUS	women	10	-3.493 pts	2.43e-02	-1.952 pts	8.02e-02	NO

Table 4: Individual skater and team double-verification results showing extreme bounds of the scoring spectrum. Detailed breakdown of individual athlete and team evaluations across the two independent statistical thresholds of the double-verification protocol. The upper portion displays the athletes receiving the largest positive scoring deviations (favoritism), while the lower portion displays those experiencing the largest negative scoring deviations (punitive evaluation. For each athlete or team, the table reports their country of representation (Cty), discipline, total home judgments received (N), panel consensus deviation (Method 1), baseline departure (Method 2), and their corresponding FDR-adjusted p-values. Green-shaded cells highlight statistically significant results (FDR$p < 0.05$), while grey-shaded cells represent non-significant results.

7. Strategic Evaluation Models: Reciprocal Exchange and Ordinal Variance

The baseline decomposition models deployed in Section 4 are effective at identifying systemic, mean-based evaluation inflation. However, focusing exclusively on mean deviations assumes that domestic evaluation anomalies are blunt, unilateral occurrences. In reality, officials operating within highly monitored environments like the International Judging System (IJS) function as sophisticated, rational actors. Extensive econometric literature demonstrates that when evaluators are constrained by institutional oversight, evaluation patterns often adapt into strategic behaviors that net to a mean bias of zero (e.g., Courty and Marschke 2004; Coupé and Gergaud 2013; Zitzewitz 2014**)**. Within the context of subjective measurement, these strategic anomalies typically manifest in two distinct forms: reciprocal score exchange (historically referred to as voting blocs) and intra-country ordinal variance (point-spread adjustment). Because both statistical anomalies can be executed without significantly inflating an official's overall mean nationalistic bias, they remain largely invisible to standard regulatory checks. To identify these scoring deviations, this study introduces two specialized, non-parametric methodologies.

7.1 Synchronous Reciprocal Exchange

To evaluate the presence of mutual point exchange between federations, the analysis examines the distribution of cross-national evaluation premiums. In game theory, bilateral favor-trading thrives when agents can establish long-term, repeated interactions with highly predictable partners, a dynamic empirically demonstrated in the reciprocal collusion of professional sports (Duggan and Levitt 2002). This behavior is closely mirrored in subjective, non-sporting peer evaluations, such as the Eurovision Song Contest, where network analyses and regression models have consistently exposed persistent geopolitical 'bloc voting' cliques that operate independently of performance quality (Fenn et al. 2006; Clerides and Stengos 2012).

However, accurately modeling reciprocal exchange in the current figure skating era requires a significant departure from historical statistical frameworks, necessitating strict controls for the spatial and temporal constraints of the current officiating environment. Historically, the concentration of elite talent among a smaller cohort of national federations resulted in highly predictable, recurring judging panel compositions. This predictability allowed previous researchers, such as Zitzewitz (2006), to successfully model global bloc voting, demonstrating that geopolitical alliances engaged in widespread, asynchronous vote-trading.

As of 2026, the pool of participating federations and athletes has expanded significantly, and the specific panel evaluating any given segment is selected via a randomized, blind lottery immediately prior to the event. From a game-theoretic perspective, this randomized temporal constraint renders traditional, season-long bloc voting models highly suspect. An official is unlikely to systematically inflate a foreign athlete as a "loan" for future reciprocation when the lottery system offers no guarantee that the allied federation will be seated on a future panel to return the evaluation premium. Consequently, rigorous analysis of modern reciprocal exchange must reject continuous market assumptions and focus strictly on *synchronous* exchange, isolating stochastic instances where officials from allied federations are simultaneously drawn to the same panel.

To isolate genuine synchronous reciprocity from unilateral favoritism, this study analyzes point exchanges strictly bounded by concurrent panel assignments. We define the Reciprocal Exchange Metric ($R_{A,B}$) between any two federations, A and B, as the minimum shared premium:

$$R_{A,B} = \min\left(\Pi_{A\to B}, \Pi_{B\to A}\right)$$

Here, $\Pi_{A\to B}$ represents the mean Adjusted Premium awarded by judges from Country A to skaters from Country B, calculated *only* during segments where both federations are seated on the panel. Taking the minimum of the two directional premiums establishes the mathematical floor of the exchange, effectively filtering out unilateral affinity. Standard parametric statistics cannot account for the constraints of panel lotteries, so we map these interactions onto an Opportunity Matrix (M) and employ a 100,000-iteration Monte Carlo permutation test. In each iteration, skater nationalities are randomly shuffled strictly within the confines of their specific competitive segments, generating a null distribution of randomized reciprocal premiums (R^{sim}) that preserves the real-world panel lotteries, event structures, and the judges' baseline strictness. The empirical p-value is calculated as the proportion of simulations where the simulated reciprocal premium equals or exceeds the observed real-world premium. Benjamini-Hochberg FDR adjustments are then applied across all pairwise federation combinations to isolate partnerships that are mathematically irreconcilable with random lottery chance.

Table 5.

a.

Federation A	Federation B	$\Pi_{A \to B}$	$\Pi_{B \to A}$	$R_{A,B}$	Z-Score	FDR p-value	Significant?
AUT	**FIN**	+1.020	+1.127	1.020	**10.21**	8.3999e-04	**YES**
AZE	**GEO**	+3.460	+3.549	3.460	**10.01**	8.3999e-04	**YES**
FRA	**MON**	+0.917	+0.873	0.873	**9.44**	4.6200e-03	**YES**
GEO	**LAT**	+0.879	+1.525	0.879	**9.09**	2.6600e-02	**YES**
GEO	**ISR**	+1.633	+1.723	1.633	**8.39**	2.2400e-03	**YES**
FRA	**UKR**	+0.986	+0.843	0.843	**7.31**	5.0399e-03	**YES**
ROU	**UKR**	+0.867	+1.013	0.867	**6.13**	4.2700e-02	**YES**
BUL	**GEO**	+1.313	+2.890	1.313	**5.59**	3.9709e-02	**YES**
ESP	**USA**	+0.802	+1.185	0.802	**5.42**	2.7090e-02	**YES**
GEO	**HUN**	+1.394	+3.367	1.394	4.93	5.0012e-02	NO

b.

Partner	Baseline Anomaly (Before Removals)	Judges Driving the Anomaly (Removed to Break Significance)	Adjusted Reciprocity (After Removals)	Anomaly Classification
AZE	**R = 3.460 Z-Score = 10.00 p = 1.00e-05**	**S. CHIGOGIDZE (+4.43 pts, N=5) Y. LEVSHUNOVA (+4.08 pts, N=5) E. KHMYZENKO (+2.89 pts, N=5)**	R = 0.000 Z-Score = 0.00 p = 1.00e+00	**Multi-Judge Influence (N=3)**
LAT	**R = 0.879 Z-Score = 9.09 p = 8.00e-04**	**S. CHIGOGIDZE (+1.78 pts, N=13)**	R = 0.000 Z-Score = 0.00 p = 1.00e+00	**Single-Judge Influence**
ISR	**R = 1.633 Z-Score = 8.31 p = 5.00e-05**	**E. KHMYZENKO (+5.01 pts, N=2) S. CHIGOGIDZE (+2.08 pts, N=6) Y. LEVSHUNOVA (+1.00 pts, N=15)**	R = 0.000 Z-Score = 0.00 p = 1.00e+00	**Multi-Judge Influence (N=3)**
BUL	**R = 1.313 Z-Score = 5.62 p = 2.62e-03**	**S. CHIGOGIDZE (+4.48 pts, N=6)**	R = 0.000 Z-Score = 0.00 p = 1.00e+00	**Single-Judge Influence**

Table 5: Pairwise federation-level analyses of synchronous reciprocal scoring exchange.
a) Pairwise analysis of synchronous reciprocal scoring exchange between national federations during concurrent panel assignments. The table reports the directional adjusted premiums ($\Pi_{A\to B}$) and ($\Pi_{B\to A}$) as well as the resulting Reciprocal Exchange Metric ($R_{A,B}$), defined as the minimum shared premium between Country A and Country B. Standardized Z-scores and FDR-adjusted p-values are derived from a 100,000-iteration Monte Carlo permutation test that randomly shuffles skater nationalities within competitive segments while preserving the empirical event structures and random panel lotteries. **b)** Statistical significance testing of synchronous scoring anomalies between federations following step-wise removal of Georgian judges.

The permutation test identifies 9 instances of highly significant reciprocal point exchange that cannot be explained by standard evaluation noise (Table 5a). Of these 9 instances, 4 involve pairings including the Georgian figure skating federation. The most statistically extreme synchronous exchange identified in the 2022–2026 cycle occurs between the federations of Azerbaijan (AZE) and Georgia (GEO). The empirical data demonstrates a highly synchronized point exchange: during the exact segments where officials from both nations are seated on the same panels, AZE judges award a mean premium of +3.46 points to GEO athletes, while GEO judges reciprocate with a +3.55-point premium to AZE athletes (Z =10.01, FDR $p < 8.40 \times 10^{-4}$).

Having identified federation pairs exhibiting statistically significant reciprocal scoring at the aggregate level, I performed a judge-level decomposition to determine whether a given anomaly reflects a broad, panel-wide pattern or is driven by a small number of individual actors. This analysis was run against Georgia (GEO), the federation with the largest number of flagged reciprocal exchanges in the strict-world dataset.

The drill-down proceeds in two stages. First, an isolation analysis subsets all external scorecards in which a Georgian judge evaluated a skater from a partner federation, and groups them by individual judge. For each judge, I record the mean premium along with the number of scorecards contributing to that mean. This produces a per-judge ledger of exactly who contributed to the aggregate reciprocal signal, and by how much.

Second, a stepwise elimination procedure tests the structural dependence of the reciprocity result on specific individuals. Judges are ranked by their isolation-analysis premium scores (highest first) and removed one at a time from the Georgian side of the calculation only; the partner federation's judges, and their scoring of Georgian skaters, are held fixed throughout. At each step, the reciprocal premium statistic is recomputed on the restricted judge pool and retested via a 100,000-iteration permutation test that shuffles skater nationality within competition (preserving panel and event structure), using the same pseudo-count-corrected p-value estimator. Elimination continues until the observed statistic falls below significance ($p \geq 0.05$), at which point the partnership is classified by how many judges were required for removal: single-judge influence ($n = 1$) versus multi-judge influence ($n > 1$). Because only the Georgian-side judge pool is pruned, this design specifically tests whether the anomaly is attributable to identifiable individuals on one side of the exchange, rather than testing whether the reciprocal relationship itself dissolves.

The stepwise procedure fully neutralized all four anomalies. In every case, reciprocity collapsed to R = 0.000 (p = 1.0) once the implicated judges were removed, indicating the signal in each partnership is concentrated among a small, identifiable subset of the Georgian panel rather than diffused across it. Two partnerships required removing only a single judge: Salome Chigogidze accounted for the entire LAT anomaly (+1.78 pts, N = 13) and the entire BUL anomaly (+4.48 pts, N = 6). The remaining two anomalies required removing three judges each: the AZE anomaly (baseline R = 3.460, the largest in the full audit) was driven jointly by Chigogidze (+4.43 pts, N = 5), Y. Levshunova (+4.08 pts, N = 5), and E. Khmyzenko (+2.89 pts, N = 5); the ISR anomaly (baseline R = 1.633) by Khmyzenko (+5.01 pts, N = 2), Chigogidze (+2.08 pts, N = 6), and Levshunova (+1.00 pts, N = 15) (Table 5b).

Notably, Salome Chigogidze appears as a contributing judge in all four of Georgia's significant exchanges, the only individual common to every case, and is independently sufficient to explain two of them outright. Levshunova and Khmyzenko each appear in two of the four exchanges (AZE and ISR). This concentration suggests the aggregate Georgian reciprocity signal is not evenly distributed across the judging panel but is substantially attributable to a small, overlapping set of individuals.

I performed an event-level decomposition of the AZE/GEO anomaly to examine the specific judge pairing characterized by the largest reciprocal score anomalies. Specifically, the synchronous evaluation patterns of Yuri Kliushnikov (AZE) and Salome Chigogidze (GEO) help drive the statistical significance of this federation-level result (Table 6).

Table 6.

Event	Discipline	Segment	Skater Name	Judge Name	Raw Deviation (pts)	Baseline Leniency (L_j^b)	Nationalistic Premium (Π_j)
Denis Ten Memorial Challenge (2023-24)	Dance	Free	Adrienne CARHART Oleksandr KOLOSOVSKYI (AZE)	Salome CHIGOGIDZE (GEO)	+3.23	-0.64	+3.88
Denis Ten Memorial Challenge (2023-24)	Men	Free	Nika EGADZE (GEO)	Yuri KLIUSHNIKOV (AZE)	+5.41	-0.29	+5.70
Denis Ten Memorial Challenge (2023-24)	Men	Short	Nika EGADZE (GEO)	Yuri KLIUSHNIKOV (AZE)	+1.88	-0.29	+2.17
Denis Ten Memorial Challenge (2023-24)	Women	Free	Alina URUSHADZE (GEO)	Yuri KLIUSHNIKOV (AZE)	-2.11	-1.15	-0.96
Denis Ten Memorial Challenge (2023-24)	Women	Short	Alina URUSHADZE (GEO)	Yuri KLIUSHNIKOV (AZE)	+1.03	-1.15	+2.19
European Figure Skating Championships 2025	Dance	Free	Diana DAVIS Gleb SMOLKIN (GEO)	Yuri KLIUSHNIKOV (AZE)	+5.37	-1.27	+6.64
European Figure Skating Championships 2025	Men	Free	Vladimir LITVINTSEV (AZE)	Salome CHIGOGIDZE (GEO)	+6.64	-1.01	+7.65
European Figure Skating Championships 2025	Men	Short	Vladimir LITVINTSEV (AZE)	Salome CHIGOGIDZE (GEO)	+4.86	-1.01	+5.87
Skate Milano (2025-26)	Women	Short	Anastasiia GUBANOVA (GEO)	Yuri KLIUSHNIKOV (AZE)	+0.83	-1.15	+1.98
Skate Milano (2025-26)	Women	Short	Nargiz SUELEYMANOVA (AZE)	Salome CHIGOGIDZE (GEO)	+4.51	+0.14	+4.37

Table 6: Event-level decomposition of the synchronous reciprocal scoring exchange between Azerbaijan and Georgia. Event-level micro-decomposition of the statistically significant reciprocal scoring exchange observed between the federations of Azerbaijan (AZE) and Georgia (GEO). The table isolates the concurrent panel evaluations of two officials: Yuri Kliushnikov (AZE) and Salome Chigogidze (GEO) across three international events: the Denis Ten Memorial (2023/2024), the European Figure Skating Championships (2024/2025) and the Skate to Milano Olympic qualifying event (2025/2026). For each evaluated skater or team, the table reports the raw score deviation from the panel consensus, the judge's segment-specific out-of-group leniency baseline, and the resulting adjusted reciprocal premium (calculated as Raw Deviation minus Panel Baseline).

The data illustrate both the breadth and the bidirectional symmetry of this reciprocal scoring exchange between the Azerbaijani and Georgian officials. This anomaly is not confined to a single skater, discipline, or event, but manifests consistently across multiple competitive seasons. For instance, during the 2023/2024 Denis Ten Memorial Challenge, this synchronization is evident across disciplines. The scorecard of Yuri Kliushnikov (AZE) demonstrates an isolated Nationalistic Premium of +5.70 points awarded to Georgian Men's Singles skater Nika Egadze in the Free Skate. Concurrently, Salome Chigogidze (GEO) evaluated the Azerbaijani Ice Dance team of Adrienne Carhart and Oleksandr Kolosovskyi with a reciprocal premium of +3.88 points in their Free Dance.

This mirrored behavior is even observable within the same competitive segment. During the Women's Short Program at Skate to Milano (2025/2026), Kliushnikov (AZE) awarded Georgian skater Anastasiia Gubanova a +1.98-point premium, while Chigogidze (GEO) simultaneously awarded Azerbaijani skater Nargiz Sueleymanova a +4.37-point premium.

The most significant observation of this synchronous point exchange occurred at the 2025 European Figure Skating Championships, an event of major international consequence. The data isolates extreme, mathematically anomalous premiums that far exceed standard evaluation variance. In the Men's Free Skate, Chigogidze (GEO) evaluated Vladimir Litvintsev (AZE) with a raw deviation of +6.64 points above the panel consensus. Because Chigogidze exhibited a strict baseline leniency of -1.01 points during that segment, subtracting this baseline reveals a large, isolated reciprocal premium of +7.65 points. Demonstrating exact reciprocal behavior at the same championship, Kliushnikov (AZE) evaluated the Georgian Ice Dance team of Diana Davis and Gleb Smolkin in the Free Dance with an isolated premium of +6.64 points.

The magnitude, cross-disciplinary frequency, and precise synchronicity of these concurrent reciprocal premiums mathematically preclude random evaluation noise. Instead, this decomposition provides empirical evidence that the current scoring system remains highly vulnerable to localized, synchronous evaluation alliances.

7.2 Intra-Country Point-Spread Variance

The second mechanism of strategic evaluation involves intra-country point-spread variance. In situations where a federation fields multiple athletes in the same segment, an official can secure a desired domestic hierarchy through a zero-sum ordinal adjustment. By simultaneously evaluating a primary compatriot above the panel consensus and a secondary compatriot below it, the official creates a score gap while their net nationalistic premium remains perfectly aligned with the international baseline. To detect this variance, the analysis isolates all instances where a judge evaluates multiple compatriots in a single segment (s). We calculate the Strategy Index ($I_{j,s}$) as the difference in premiums awarded to the highest- and lowest-evaluated compatriots:

$$I_{j,s} = \pi_{Primary} - \pi_{Secondary}$$

Because selecting the maximum and minimum values inherently creates an order statistic that is necessarily positive, we must establish a true out-of-group baseline. We calculate the expected order statistic mean ($\mu_{foreign}^{s}$) and standard deviation ($\sigma_{foreign}^{s}$) strictly using foreign skater pairings evaluated by the same judge in the same discipline, successfully neutralizing natural evaluation variance. The anomaly is then quantified using a standardized T-Statistic:

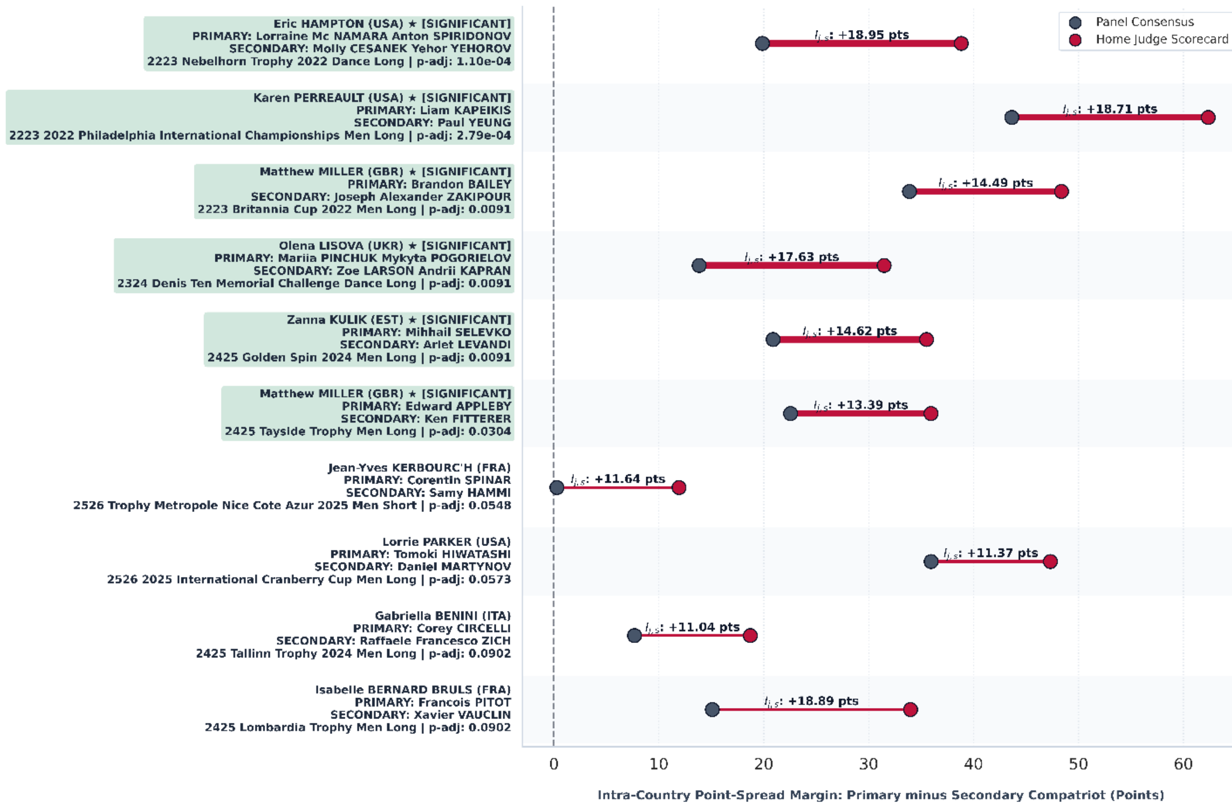
$$T_{j,s} = (I_{j,s} - \mu_{foreign}^{s})/\sigma_{foreign}^{s}$$

Scorecards are flagged only if the FDR-adjusted p-value across all domestic evaluation opportunities is < 0.05. This methodology detects two distinct classes of statistically significant ordinal variance that otherwise evade standard mean-bias detection (Supplementary Table 5).

The first class of ordinal variance is ordinal expansion ($T > 0$), where a judge's scorecard increases the score gap between compatriots relative to the panel consensus, effectively stabilizing or securing a specific domestic hierarchy. The analysis identified six statistically significant examples of ordinal expansion (Figure 9a). The most statistically extreme, FDR-significant example of this in the dataset is observed on the scorecard of Eric Hampton (USA) at the 2022/2023 Nebelhorn Trophy (FDR $p = 1.10 \times 10^{-4}$). While the international panel consensus established a baseline margin of 19.87 points between the teams of Lorraine McNamara/Anton Spirodonov (Primary) and Molly Cesankek/Yehor Yehorov (Secondary), Hampton's personal evaluation expanded this gap by 18.95 points (Figure 9a).

Figure 9.

a)


Eric HAMPTON (USA) ★ [SIGNIFICANT]
PRIMARY: Lorraine Mc NAMARA Anton SPIRIDONOV
SECONDARY: Molly CESANEK Yehor YEHOROV
2223 Nebelhorn Trophy 2022 Dance Long | p-adj: 1.10e-04
Karen PERREAULT (USA) ★ [SIGNIFICANT]
PRIMARY: Liam KAPEIKIS
SECONDARY: Paul YEUNG
2223 2022 Philadelphia International Championships Men Long | p-adj: 2.79e-04
Matthew MILLER (GBR) ★ [SIGNIFICANT]
PRIMARY: Brandon BAILEY
SECONDARY: Joseph Alexander ZAKIPOUR
2223 Britannia Cup 2022 Men Long | p-adj: 0.0091
Olena LISOVA (UKR) ★ [SIGNIFICANT]
PRIMARY: Mariia PINCHUK Mykyta POGORIELOV
SECONDARY: Zoe LARSON Andrii KAPRAN
2324 Denis Ten Memorial Challenge Dance Long | p-adj: 0.0091
Zanna KULIK (EST) ★ [SIGNIFICANT]
PRIMARY: Mihhail SELEVKO
SECONDARY: Arlet LEVANDI
2425 Golden Spin 2024 Men Long | p-adj: 0.0091
Matthew MILLER (GBR) ★ [SIGNIFICANT]
PRIMARY: Edward APPLEBY
SECONDARY: Ken FITTERER
2425 Tayside Trophy Men Long | p-adj: 0.0304
Jean-Yves KERBOURC'H (FRA)
PRIMARY: Corentin SPINAR
SECONDARY: Samy HAMMI
2526 Trophy Metropole Nice Cote Azur 2025 Men Short | p-adj: 0.0548
Lorrie PARKER (USA)
PRIMARY: Tomoki HIWATASHI
SECONDARY: Daniel MARTYNOV
2526 2025 International Cranberry Cup Men Long | p-adj: 0.0573
Gabriella BENINI (ITA)
PRIMARY: Corey CIRCELLI
SECONDARY: Raffaele Francesco ZICH
2425 Tallinn Trophy 2024 Men Long | p-adj: 0.0902
Isabelle BERNARD BRULS (FRA)
PRIMARY: Francois PITOT
SECONDARY: Xavier VAUCLIN
2425 Lombardia Trophy Men Long | p-adj: 0.0902
$I_{j,s}$: +18.95 pts
$I_{j,s}$: +18.71 pts
$I_{j,s}$: +14.49 pts
$I_{j,s}$: +17.63 pts
$I_{j,s}$: +14.62 pts
$I_{j,s}$: +13.39 pts
$I_{j,s}$: +11.64 pts
$I_{j,s}$: +11.37 pts
$I_{j,s}$: +11.04 pts
$I_{j,s}$: +18.89 pts
Panel Consensus
Home Judge Scorecard
0
10
20
30
40
50
60
Intra-Country Point-Spread Margin: Primary minus Secondary Compatriot (Points)

b)

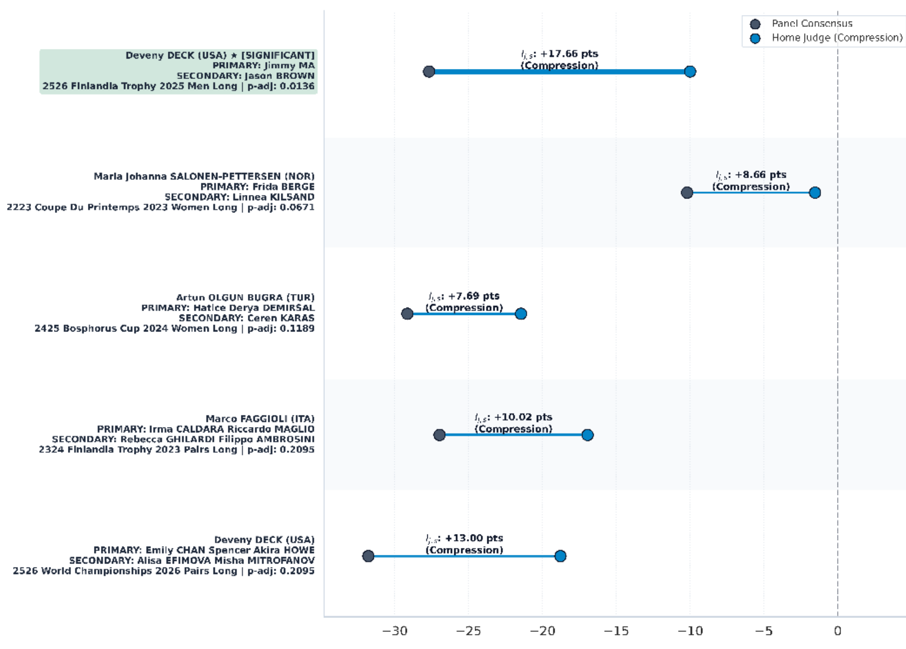

Deveny DECK (USA) ★ [SIGNIFICANT]
PRIMARY: Jimmy MA
SECONDARY: Jason BROWN
2526 Finlandia Trophy 2025 Men Long | p-adj: 0.0136
Maria Johanna SALONEN-PETTERSEN (NOR)
PRIMARY: Frida BERGE
SECONDARY: Linnea KILSAND
2223 Coupe Du Printemps 2023 Women Long | p-adj: 0.0671
Artun OLGUN BUGRA (TUR)
PRIMARY: Hatice Derya DEMIRSAL
SECONDARY: Ceren KARAS
2425 Bosphorus Cup 2024 Women Long | p-adj: 0.1189
Marco FAGGIOLI (ITA)
PRIMARY: Irma CALDARA Riccardo MAGLIO
SECONDARY: Rebecca GHILARDI Filippo AMBROSINI
2324 Finlandia Trophy 2023 Pairs Long | p-adj: 0.2095
Deveny DECK (USA)
PRIMARY: Emily CHAN Spencer Akira HOWE
SECONDARY: Alisa EFIMOVA Misha MITROFANOV
2526 World Championships 2026 Pairs Long | p-adj: 0.2095
$I_{j,s}$: +17.66 pts (Compression)
$I_{j,s}$: +8.66 pts (Compression)
$I_{j,s}$: +7.69 pts (Compression)
$I_{j,s}$: +10.02 pts (Compression)
$I_{j,s}$: +13.00 pts (Compression)
Panel Consensus
Home Judge (Compression)
−30
−25
−20
−15
−10
−5
0
5
Intra-Country Point-Spread Margin: Primary minus Secondary Compatriot (Points)

Figure 9: Intra-country point-spread variance analysis illustrating strategic spread expansion and spread compression. Dumbbell plots mapping the strategic point-spread alterations introduced by home-country judges when simultaneously evaluating multiple compatriots within the same segment. The horizontal axis measures the point-spread margin (in points) between the primary (higher-ranked) and secondary compatriot. Grey circles represent the baseline margin established by the out-of-group international panel consensus, while colored circles represent the margin recorded on the home judge's scorecard. The connecting horizontal bar represents the Strategy Index ($I_{j,s}$), which quantifies the absolute point-spread deviation relative to the consensus. **(a)** Spread expansion ($T > 0$, shaded in red): Instances where an official's scorecard increases the gap between compatriots. **(b)** Spread compression ($T < 0$, shaded in blue): Instances where a primary compatriot is outskated by a domestic peer according to the international consensus, resulting in a compressed margin on the domestic official's scorecard.

The second, and more mathematically complex, class of ordinal variance is ordinal compression ($T < 0$), which may occur when a primary compatriot is outskated by a domestic peer according to the international panel consensus. In these scenarios, significant positive premiums are not identified in the scores of both domestic skaters. Instead, a compressed margin is observed, effectively mitigating the primary compatriot's deficit. A statistically significant instance of this compression is seen on the scorecard of Deveny Deck (USA) at the 2025/2026 Grand Prix of Finland Men's Free Skate (FDR *p* = 0.0174). The international panel consensus placed Jason Brown ahead of Jimmy Ma by a decisive margin of 27.66 points. However, Deck's scorecard evaluation compressed this deficit by 17.66 points (Figure 9b).

8. The Role and Influence of the Technical Panel

This study next expands the scope of bias analysis beyond the primary judging panel to provide the first comprehensive, quantitative evaluation of the Technical Panel. Despite holding deterministic control over the underlying Base Value of a performance, the Technical Panel has historically escaped statistical analysis, a limitation now overcome due to the unique granularity of the deanonymized dataset. Administratively, the panel is composed of three officials: the Technical Controller, the Technical Specialist, and the Assistant Technical Specialist, who often represent different national federations. Operationally, the Technical Specialist is primarily responsible for making the initial, real-time calls during a performance, while the other two members intervene in the event of a disagreement or query from the judging panel. Because these internal deliberations and voting splits are not recorded on the official competition protocols, it is impossible to reconstruct the panel's internal consensus. Consequently, I make a necessary methodological choice given the available data. All subsequent technical bias analyses are mapped strictly to the national affiliation of the primary Technical Specialist.

From an econometric perspective, mapping bias to the primary Specialist serves as a conservative test. If the secondary panel members were consistently and objectively intervening to correct a primary Specialist's nationalistic favoritism, the resulting signal would attenuate toward zero or fall below detectable thresholds. The empirical presence of statistically significant, high-magnitude deviations mapped to the primary Specialist (see below) demonstrates that internal panel oversight fails to neutralize this initial localized bias and this conservatism probably operates in one direction. Where corrections are imperfect rather than absent, the detected values represent a lower bound on the primary Specialist's true nationalistic deviation.

Modeling the decisions of the Technical Panel requires separating the two distinct types of evaluations they perform, namely technical infraction calls and Level of Difficulty assignments. Analyzing the frequency and score-effects of technical infraction calls (such as under-rotated jumps, incorrect take-off edges, and extended lifts) is relatively straightforward due to a logical assumption of athletic intent. In elite competition, an athlete's objective is always to execute clean elements such that no skater intends to incur a technical penalty. Therefore, any technical infraction called by the panel represents an unambiguous deviation from the athlete's intended execution, making the frequency and influence of these calls highly quantifiable.

In contrast, evaluating Levels of Difficulty (assigned from Base to Level 4 on non-jump elements like spins, step sequences and lifts) is more complex and requires a statistical baseline. The primary obstacle is that an element's intended level of difficulty is not recorded on the official protocol. For example, a skater who receives a Level 2 on a sit spin may have designed and executed a perfect, planned Level 2 program element, incurring no technical deduction. Alternatively, that same skater may have intended to execute a Level 4 but encountered balance or timing issues on the ice, resulting in an involuntary level reduction. Because these intended designs and on-ice mistakes are unobservable from the scoreboard, this study works with a historical baseline of execution. For every skater and specific element, I calculate their average career Level of Difficulty across all other competitions evaluated by out-of-group (international) panels, establishing an empirical benchmark of the athlete's historical performance quality.

8.1 Methodology for Quantifying Technical Sway

To estimate the global scoring influence of the Technical Panel, this study constructs a deterministic model of point deductions and level adjustments, defined herein as the Net Technical Sway (S_p). This metric represents the aggregate net point-swing introduced by the panel's technical calls and difficulty level assignments relative to a theoretical baseline of clean, expected execution. For any given performance p, the Net Technical Sway is calculated as the sum of three distinct items evaluated at the individual element level:

$$S_p = \sum_{e \in p} (\Delta B\, V_e + \Delta GOE_e + \Delta L_e)$$

Where ΔBV_e is the Base Value Loss (the hard penalty) ,ΔGOE_e is the Grade of Execution loss (the soft penalty), and ΔL_e is the Level Assignment Sway.

To quantify the hard penalty (ΔBV_e), the algorithm isolates the underlying "base element" by parsing and stripping any technical infraction flags (e.g., '<', 'e', 'q') and assigned levels from the protocol description. It then maps the element to a global lookup of median Base Values awarded to clean, unflagged instances of that exact element within the same discipline. The observed Base Value is subtracted from this clean median (bounded at a minimum of zero) to calculate the direct points lost due to technical calls (such as jump under-rotations).

To quantify the soft penalty (ΔGOE_e), the model establishes an empirical penalty matrix that calculates the indirect GOE score depression enforced by technical warnings. This is defined as the difference between the mean GOE awarded to clean instances of an element and the mean GOE awarded to elements carrying a specific infraction flag (e.g., edge warnings, timing violations).

The final parameter, Level Assignment Sway (ΔL_e), measures the point-influence of the panel's difficulty level determinations on non-jump elements (e.g., spins, step sequences, and lifts). Evaluating these elements is complicated by unobservable variables: an assigned "Level 2" on the protocol might reflect a perfectly executed planned Level 2, or it might reflect an error-laden but planned Level 4 element. To address this, the model calculates an Expected Value (EV) baseline representing the element's anticipated difficulty level, derived exclusively from performances evaluated by neutral (out-of-group) technical panels. However, establishing this baseline introduces a data sparsity problem: an athlete may debut a newly learned element or simply lack a sufficient number of historical attempts evaluated by neutral panels. To prevent the exclusion of these observations while maximizing baseline precision, the model employs a tiered hierarchical fallback algorithm. For example, if a skater executes a Flying Camel Spin (FCSp), the algorithm first queries Tier 1: the skater's own historical average level for that specific FCSp (requiring a minimum of 5 neutral observations). If this data threshold is unmet, the algorithm falls back to Tier 2: the global average level awarded to *all* skaters for an FCSp (requiring a minimum of five neutral observations). If the specific element is so rare that Tier 2 also fails, it defaults to Tier 3: the global neutral average for the broader "Spin" category.

The level offset is calculated as the expected level minus the observed assigned level. To translate this numerical offset into an absolute point value, the algorithm calculates the empirical point-value per level ($\approx V_{level}$) for each element, defined as the mean difference in median Base Values between consecutive levels of difficulty. The Level Assignment Sway (ΔL_e) is the numerical level offset multiplied by this category-specific point multiplier.

Crucially, the calculation of ΔL_eremains directionally unclipped. A positive value represents level suppression (points lost relative to expectation), while a negative value represents level boosting (points gained). Therefore, the Net Technical Sway captures the true, aggregate directional influence of the panel. If a Technical Specialist penalizes an athlete with a jump under-rotation but compensates by generously boosting a step sequence level above the athlete's historical baseline, the unclipped formula mathematically offsets these actions, yielding an accurate measure of the net point advantage or disadvantage conferred to the skater.

8.2 Global Analysis of Technical Panel Influence

Applying the deterministic sway model reveals that the Technical Panel exerts a massive, continuous point tax on figure skating performances. The baselined data reveals that the primary driver of this technical tax differs fundamentally between disciplines (Table 7). In the Women's discipline, the impact is driven primarily by the soft penalty of GOE suppression. For example, in the Women's Free Skate, the average performance incurs a direct BV loss of -8.07 points but suffers an even larger average GOE loss of -9.98 points due to the mandatory score caps triggered by technical infraction flags. Conversely, in the Men's discipline, direct Base Value downgrades represent the dominant penalty. In the Men's Free Skate, the average direct BV loss (-10.36 points) substantially eclipses the average GOE loss (-6.45 points), likely reflecting the severe base value penalties associated with downgraded or under-rotated quadruple jumps.

Table 7.

Discipline	Segment	Avg Flags	Avg BV Loss	Avg GOE Loss	Avg Level Loss	Avg Technical Sway	Max Sway
Ice Dance	**Rhythm**	0.1	-1.16 pts	-0.28 pts	+0.02 pts	-1.43 pts	-19.73 pts
Ice Dance	**Free**	0.1	-1.41 pts	-0.37 pts	-0.05 pts	-1.83 pts	-21.18 pts
Men	**Short**	1.0	-4.63 pts	-3.60 pts	+0.07 pts	-8.16 pts	-54.19 pts
Men	**Free**	1.8	-10.36 pts	-6.45 pts	-0.09 pts	-16.90 pts	-66.72 pts
Pairs	**Short**	0.5	-1.92 pts	-2.15 pts	+0.03 pts	-4.04 pts	-18.35 pts
Pairs	**Free**	1.0	-4.37 pts	-4.16 pts	-0.01 pts	-8.54 pts	-37.74 pts
Women	**Short**	1.3	-2.63 pts	-5.29 pts	+0.03 pts	-7.89 pts	-36.14 pts
Women	**Free**	2.6	-8.07 pts	-9.98 pts	-0.07 pts	-18.12 pts	-64.57 pts

Table 7: Global analysis of technical panel influence across segments.
Global analysis of the Technical Panel's scoring influence, measured using the deterministic sway model relative to an empirical baseline of expected, clean execution. For each segment, the table reports the average volume of technical infraction flags called per performance (average flags), the average Base Value loss, the average Grade of Execution loss (the soft penalty of score GOE suppression), and the average Level of Difficulty loss. The Total Technical Sway represents the aggregate point reduction introduced by the panel's determinations, while the maximum observed sway identifies the single largest scoring reduction recorded in the dataset.

The distribution of total technical influence varies dramatically across disciplines, reflecting their underlying rule structures. The highest average Technical Sways occur in the singles disciplines, particularly in the Free Skating segments (Women: -18.12 points, Men: -16.90 points), where the high volume of multi-rotational jumps (averaging 2.6 and 1.8 flags per program, respectively) maximizes the mathematical exposure to edge and under-rotation calls. In contrast, Ice Dance exhibits the lowest average Technical Sway (-1.43 points in the Rhythm Dance, -1.85 points in the Free Dance), reflecting a competitive environment devoid of jumps.

The bidirectional level-assignment metric reveals that the average point influence of Level Loss is remarkably close to neutral across all disciplines. In fact, short segments (such as the Men's Short and Rhythm Dance) exhibit slight aggregate level-boosting (positive values) relative to historical expectations. Because the EV baseline tightly controls for skater ability, the data demonstrates that even in non-jump disciplines like Ice Dance, the technical tax is not driven by arbitrary level suppression, but rather by direct Base Value losses (-1.16 and -1.41 points), capturing discrete errors such as extended lift deductions or invalidated elements.

Finally, the maximum recorded sways in the dataset underscore the immense, leaderboard-altering capacity of these technical decisions. In a sport where major international podium placements are routinely decided by margins of less than one point, the maximum observed technical sways reach staggering magnitudes: -64.57 points in the Women's Free Skate and -66.72 points in the Men's Free Skate. This extreme point-variance provides compelling empirical evidence that any systematic nationalistic bias within the Technical Panel has the potential to alter competition standings far more decisively than the subjective inflation of the primary judging panel, firmly motivating the need for an individualized, baseline-controlled technical bias analysis.

Figure 10.

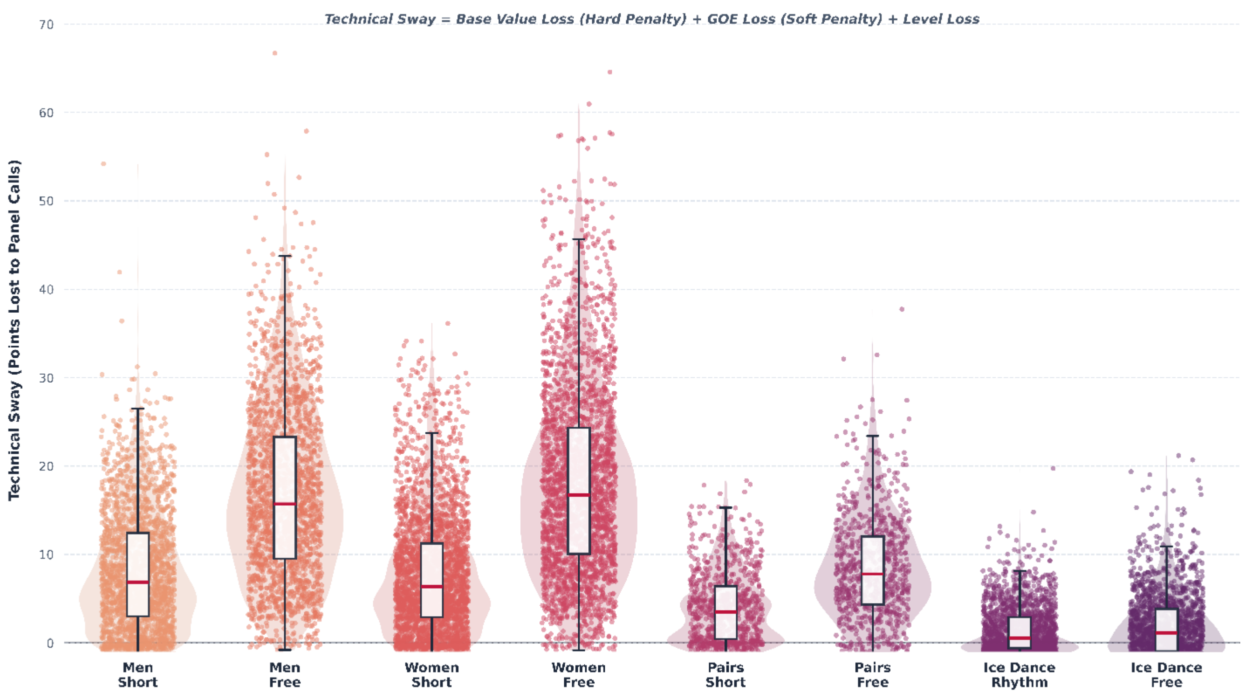


Figure 10: Distribution of technical panel sways across discipline segments.
Hybrid violin and box plots with jittered scatter overlays displaying the distribution of absolute Technical Sway (points) across the eight competitive segments. The vertical axis measures the aggregate point-swing introduced by the panel's infraction calls and Level of Difficulty assignments relative to an empirical baseline of clean execution. The shaded violin contours represent the kernel density estimation of the probability distribution for each segment. Within each violin, the white box plots denote the interquartile range (IQR, 25th to 75th percentiles), the central horizontal line indicates the median, and the whiskers extend to $1.5 \times$ the IQR. The overlaid jittered dots represent individual performances, illustrating the density of the raw data and the presence of extreme outliers.

9.1 Methodology for Technical Specialist Bias Analysis

Expanding the empirical analysis to the Technical Panel requires an initial understanding of the structural distribution and certification constraints of the officiating pool. Evaluating the cross-discipline certifications of the 165 Technical Specialists officiating during the 2022–2026 Olympic cycle reveals a highly concentrated intersection of credentials (Figure 11). The largest specialized cohorts are found in Ice Dance (*N* = 38) and Women's Singles (*N* = 31), with a substantial overlap of 49 specialists certified to evaluate both singles disciplines. Conversely, cross-specialization between highly divergent disciplines remains exceedingly rare; only eight specialists in the entire officiating pool maintain active credentials across the distinct technical regimes of both Pairs and a different discipline. This stark concentration of technical expertise underscores the necessity of the segment- and discipline-specific expected value (EV) baselines established in Section 8, as specialists are structurally constrained from operating as a single, homogenous officiating body.

Figure 11.

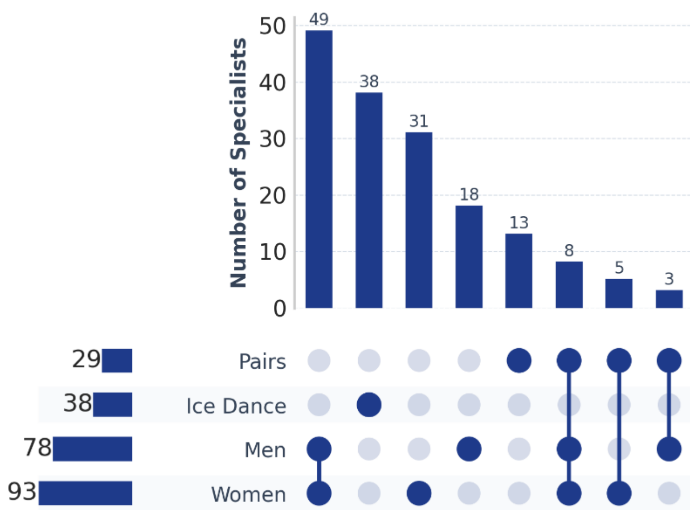


Figure 11: Cross-discipline certification profiles and overlap of international technical specialists. UpSet plot illustrating the intersections and certification profiles of the *N* = 165 unique international technical specialists in the dataset across the four competitive disciplines (Pairs, Ice Dance, Men's Singles, and Women's Singles). The horizontal bar chart on the bottom-left displays the total pool of unique specialists certified to evaluate each individual discipline (set sizes). The main vertical bar chart displays the size of each exclusive intersection subset (the number of specialists sharing that exact combination of credentials). The connected dark-blue circles beneath the vertical bars denote the specific combination of disciplines represented in each intersection.

9.2 Quantifying Nationalistic Bias of the Primary Technical Specialist

To systematically evaluate technical officiating patterns, this study implements a point-based Expected Value (EV) framework designed to detect nationalistic deviations in infraction-calling and level-assignment behaviors. Building upon the element-level metrics defined in Section 8.1, the analysis aggregates the calculated point-sways to the individual *performance* level. This crucial methodological step ensures statistical independence, as multiple elements executed within the same program by the same skater are highly correlated events.

The evaluation is executed at both the aggregate federation level and the individual Technical Specialist level. To ensure statistical robustness, analyses are strictly limited to entities meeting minimum sample size thresholds. Individual specialists are evaluated only if they have assessed a minimum of 4 compatriot performances and 15 international performances. Federation-level aggregations require a minimum of 10 compatriot and 20 international performances.

The framework evaluates Technical Panel bias across two distinct pillars:
1. Infraction Shielding: This pillar measures the panel's propensity to protect compatriots from the hard and soft point deductions associated with technical infractions (e.g., under-rotations, edge calls). For each performance p, the Total Infraction Loss (I_p) is calculated as the sum of the element-level Base Value losses and GOE losses:

$$I_p = \sum \Delta BV_e + \sum \Delta GOE_e$$

2. Level Shielding: This pillar evaluates systematic favoritism in the assignment of difficulty levels. For each performance, the Total Level Loss (L_p) is calculated as the sum of the element-level sway metrics:

$$L_p = \sum \Delta L_e$$

For any officiating entity (a specific specialist or federation), the baseline expected point loss is established exclusively from their evaluation of international (out-of-group) skaters. The Shielding Effect is calculated as the difference between this international baseline point loss and the mean point loss assessed to compatriot skaters. For example, the Infraction Shielding Effect (ΔI_{pts}) is defined as:

$$\Delta I_{pts} = \bar{I}_{international} - \bar{I}_{compatriot}$$

A positive shielding value indicates nationalistic protection (fewer points docked from compatriots compared to the official's own international baseline), while a negative value indicates a punitive approach.

Admittedly, using historical averages to establish execution baselines has limitations; a static model cannot perfectly capture day-of-event physical anomalies or mid-season choreography upgrades. However, the core identification strategy of this study neutralizes these concerns. Because judging panel compositions are determined by randomized lottery, a compatriot specialist is no more likely than an international one to evaluate a skater executing a flawless performance or debuting a choreographed upgrade. Any random execution variance therefore affects compatriot and international evaluations with equal probability. A consistent, systematic point premium awarded by compatriot specialists that exceeds their own international baseline is therefore mathematically isolated from random physical error and natural athletic progression.

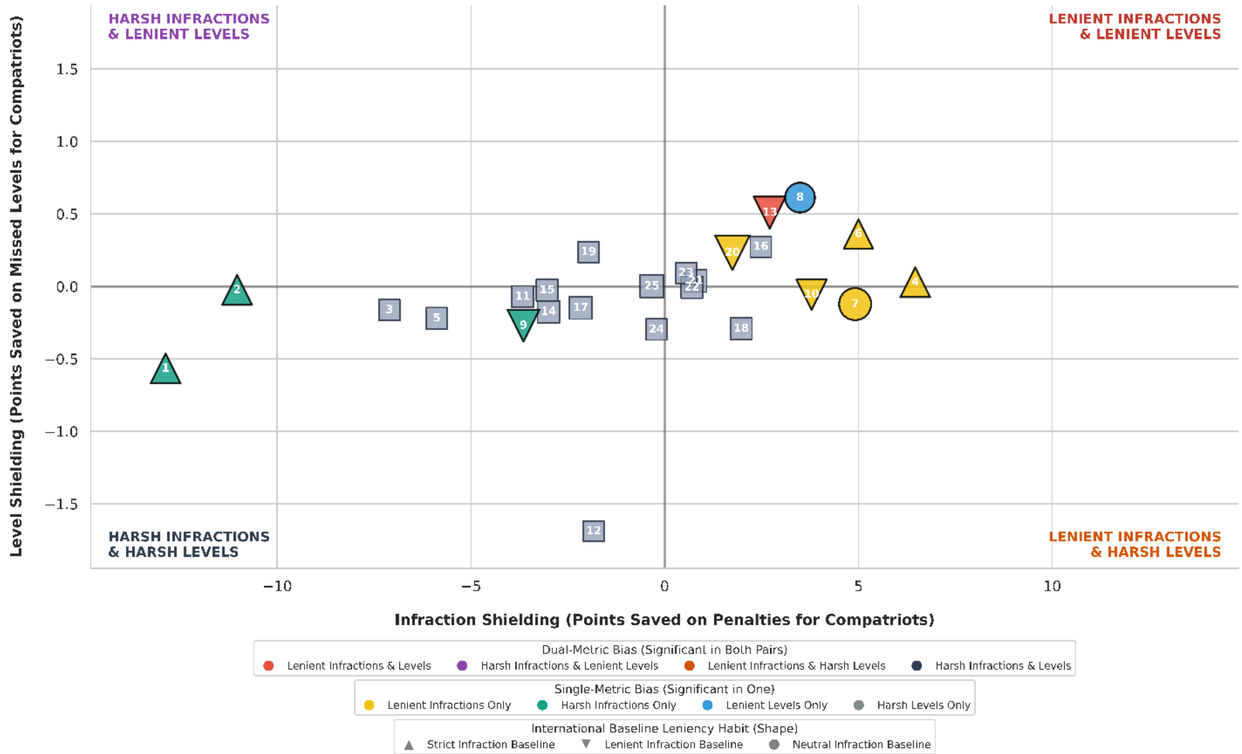
While Section 4 successfully models strategic 'Rival Suppression' (Σ) within the primary judging panel, extending this specific metric to the Technical Panel introduces a critical endogeneity problem regarding skater quality. The judging panel analysis successfully isolates rival suppression by utilizing the Panel Consensus Deviation ($\Delta_{j,p}$). Because the home judge is compared against the concurrent evaluations of the remaining panel members for the *same performance*, objective skater quality is mathematically neutralized. Conversely, due to the single-actor nature of the Technical Specialist, the Expected Value (EV) model must measure bias via *absolute* point-loss relative to an objective baseline of clean execution. Because 'Rivals' are defined strictly by their ordinal proximity to a compatriot, this cohort inherently represents a localized, homogenous band of athletic ability, whether at the top, middle, or bottom of the standings. The 'Non-Rival' cohort, encompassing the remainder of the field, will mathematically possess a drastically divergent mean skill level. Consequently, attempting to quantify Technical Rival Suppression by comparing the absolute EV point-losses between these two cohorts would fatally confound strategic evaluator strictness with the natural variance in athletic execution across different tiers of the leaderboard. To preserve strict mathematical integrity and avoid this selection bias, analysis of the Technical Panel is restricted exclusively to the evaluation of Compatriot Shielding.

9.3 Statistical Testing and Classification

Statistical significance for both technical pillars is determined via two-sample Welch's unequal variances t-tests, comparing the distribution of point losses assessed to compatriot performances against the distribution of point losses assessed to international performances.

Concurrently, to isolate an official's underlying judging philosophy, a one-sample t-test is applied comparing the official's international infraction loss mean ($\bar{I}_{international}$) against the global neutral average ($\bar{I}_{global}$). This establishes whether an entity is systematically "strict" (docking significantly more points globally), "lenient," or "neutral" relative to their peers, ensuring that strict global judging is not misidentified as compatriot bias.

To strictly control for the false positive rate associated with multiple comparisons across hundreds of officials and federations, Benjamini-Hochberg False Discovery Rate (FDR) corrections are applied to all generated p-values. Officiating entities are flagged for systematic technical bias only if their FDR-adjusted p-values fall below the significance threshold ($\alpha = 0.05$). This strict thresholding allows for the objective, data-driven categorization of bias profiles (e.g., "Lenient Infractions & Lenient Levels" vs. "Harsh Infractions Only") based entirely on directional, statistically significant point deviations.



FEDERATION TECHNICAL BIAS KEY (COMPLETE REFINED AUDIT)

Format: Rank. Federation | Baseline Habit | I: Infraction Shielding (Pts) | L: Level Shielding (Pts)

★ Highlighted with Bold Slate Text and Star = Statistically Significant Bias

1. ★ ROU | Strict | I: -12.88 pts | L: -0.57 pts

2. ★ CRO | Strict | I: -11.04 pts | L: -0.02 pts

3. TUR | Strict | I: -7.11 pts | L: -0.16 pts

4. ★ NOR | Strict | I: +6.46 pts | L: +0.03 pts

5. AUS | Lenient | I: -5.88 pts | L: -0.22 pts

6. ★ AUT | Strict | I: +5.00 pts | L: +0.36 pts

7. ★ ITA | Neutral | I: +4.91 pts | L: -0.12 pts

8. ★ KOR | Neutral | I: +3.49 pts | L: +0.61 pts

9. ★ GBR | Lenient | I: -3.65 pts | L: -0.27 pts

10. ★ NED | Lenient | I: +3.79 pts | L: -0.06 pts

11. FIN | Neutral | I: -3.66 pts | L: -0.07 pts

12. AZE | Lenient | I: -1.83 pts | L: -1.69 pts

13. ★ CAN | Lenient | I: +2.71 pts | L: +0.51 pts

14. BUL | Strict | I: -3.00 pts | L: -0.17 pts

15. POL | Lenient | I: -3.03 pts | L: -0.03 pts

16. CHN | Neutral | I: +2.48 pts | L: +0.27 pts

17. SVK | Lenient | I: -2.17 pts | L: -0.15 pts

18. FRA | Neutral | I: +1.97 pts | L: -0.29 pts

19. UKR | Neutral | I: -1.97 pts | L: +0.24 pts

20. ★ USA | Lenient | I: +1.74 pts | L: +0.23 pts

21. JPN | Lenient | I: +0.80 pts | L: +0.04 pts

22. HUN | Strict | I: +0.70 pts | L: -0.01 pts

23. SUI | Strict | I: +0.56 pts | L: +0.09 pts

24. SLO | Strict | I: -0.21 pts | L: -0.30 pts

25. GER | Neutral | I: -0.33 pts | L: +0.00 pts

Figure 12: Federation-level expected value technical assessment mapping infraction shielding against level shielding. Bivariate mapping of federation-level technical evaluation phenotypes derived from the two-pillar Expected Value (EV) model. The horizontal axis measures the Infraction Shielding Effect (ΔI), defined as the percentage of expected technical infractions systematically shielded or overlooked for compatriots relative to the international baseline. The vertical axis measures the Level Shileding Delta (ΔL), representing the mean difficulty level offset systematically awarded to compatriots on non-jump elements relative to their out-of-group career norms. Node shapes denote each federation's baseline officiating tendency when evaluating non-compatriots (upward-pointing triangles represent strict infraction-calling baselines, downward-pointing triangles represent lenient baselines, and circles represent neutral baselines). Node colors distinguish statistically significant dual-metric and single-metric bias profiles after FDR correction (FDR $p < 0.05$). Numbered nodes correspond to the twenty-six national federations meeting the minimum sample thresholds, indexed in the key below.

9.4 Federation-Level Technical Evaluation Phenotypes

Applying the performance-level Expected Value (EV) model to national federations meeting the minimum sample thresholds reveals highly distinct, statistically significant point-evaluation phenotypes (Figure 12). Plotting each federation's mean Infraction Shielding points against its mean Level Shielding points exposes structured, directional biases that are heavily influenced by national origin.

In the upper-right quadrant, Canada (CAN) represents a rare, statistically significant *Dual-Metric Lenient* profile (indicated by the red shape). Canadian specialists exhibit a lenient international baseline overall (downward-pointing triangle), yet they award their compatriots a significant additional dual benefit: saving them an average of +2.71 points on technical infractions while simultaneously boosting them by +0.51 points on non-jump level assignments compared to their already lenient international standard.

A more common phenotype is *Single-Metric Infraction Shielding*, clustered predominantly on the right side of the plot (yellow shapes). Federations such as Norway (NOR, +6.46 pts), Austria (AUT, +5.00 pts), Italy (ITA, +4.91 pts), the Netherlands (NED, +3.79 pts), and the United States (USA, +1.74 pts) demonstrate statistically significant leniency exclusively on technical infraction calls. For these federations, compatriots are systematically shielded from the severe Base Value and GOE point losses associated with jump downgrades and edge calls, while their non-jump levels are evaluated roughly in line with global expectations. Notably, Austria and Norway provide this massive compatriot shielding despite maintaining a "Strict" international baseline (upward-pointing triangles) against the rest of the world.

Conversely, South Korea (KOR) occupies a unique *Single-Metric Level Shielding* profile (blue shape). While Korean specialists evaluate infractions without statistically significant bias, they award compatriots an average premium of +0.61 points strictly through generous Level of Difficulty assignments on spins and step sequences.

Finally, the left side of the plot isolates a *Punitive* or *Harsh Infraction* phenotype (green shapes). Specialists from Romania (ROU, -12.88 pts), Croatia (CRO, -11.04 pts), and Great Britain (GBR, -3.65 pts) dock significantly *more* points from their own compatriots for technical infractions than they do when evaluating international skaters. For Romania and Croatia, this punitive internal standard is compounded by an already strict international evaluation baseline.

These clear, multi-metric variations in point distributions demonstrate that technical evaluation anomalies are highly structured. Rather than representing random evaluation noise, they reflect distinct, quantifiable operational habits and allegiances unique to each national federation's technical specialists.

9.5 Individual Specialist Assessment

Decomposing these technical evaluation patterns to the individual specialist level confirms that the broader federation phenotypes are driven by highly concentrated, statistically significant individual deviations (Figure 13, Supplementary Table 6). At the extreme punitive end of the spectrum, Romanian specialist Grety Marton demonstrates a pronounced dual-metric harsh phenotype (dark slate marker). Operating from an already strict international baseline, she evaluates compatriots with severe domestic rigor, docking them an additional -14.93 points on technical infractions and penalizing them by -0.70 points on level assignments compared to her out-of-group standard.

This stands in direct contrast to officials exhibiting significant lenient phenotypes. For example, United States specialist Judy Blumberg exhibits a statistically significant dual-metric lenient profile (red marker), in which compatriots receive an infraction shielding premium of +1.29 points alongside a +1.29-point boost on non-jump levels, granted on top of a generally lenient international baseline. Extreme point shielding is also visible in single-metric profiles; Norwegian specialist Christine Isaksen demonstrates massive infraction shielding, systematically saving compatriots an average of +9.80 points on technical penalties (yellow marker), while specialists like Sung-Jin Byun (KOR) and Masako Kawai (JPN) systematically boost compatriot scores strictly through generous non-jump Level of Difficulty assignments (blue markers).

The real-world significance of these individual technical anomalies is underscored by their direct intersection with the sport's most prestigious competitive panels. In Figure 13, officials officially selected to the technical panels for the 2026 Milan-Cortina Winter Olympic Games are highlighted, revealing that several of the sport's premier Technical Specialists exhibit statistically significant, non-random deviations from the expected value baseline. For example, Olympic panelist Clarence Ong (AUS) exhibits a highly significant harsh infraction profile (green marker), systematically docking compatriots an average of -7.87 points on technical infractions relative to his neutral global baseline. Conversely, on the lenient end of the Olympic cohort, Canadian specialist Marie Bowness evaluates non-jump elements with a statistically significant compatriot premium, historically boosting Canadian skaters by +0.85 points via level assignments (blue marker) while operating from an already lenient global baseline.

Figure 13.

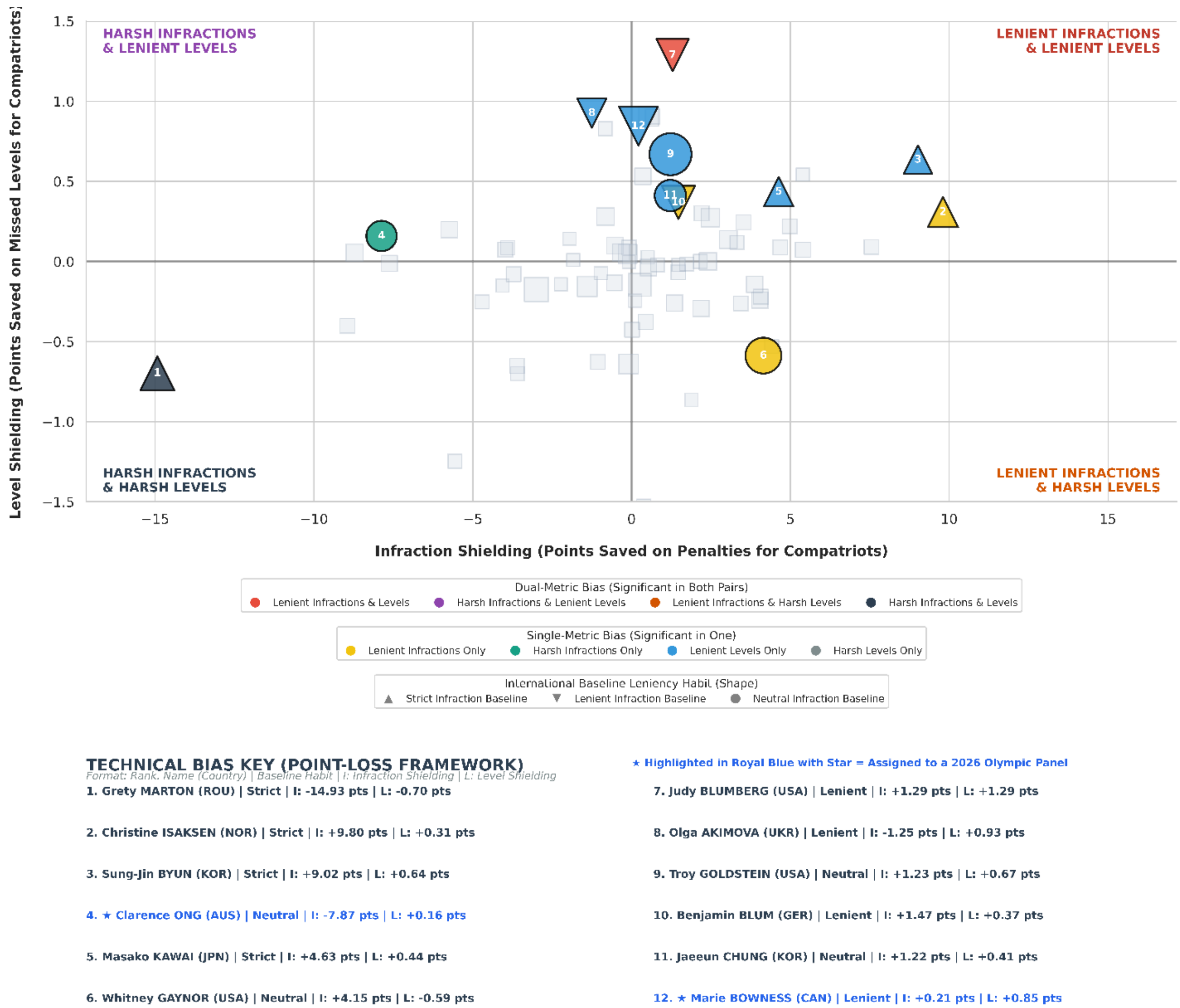


Figure 13: Individual-level expected value technical assessment mapping infraction shielding against level shielding. Bivariate mapping of individual technical specialist phenotypes derived from the two-pillar Expected Value (EV) model. The horizontal axis measures the Infraction Shielding Effect (ΔI), calculated as the percentage of expected technical infractions systematically shielded or overlooked for compatriots relative to the international baseline. The vertical axis measures the Level Shielding Delta (ΔL), representing the mean difficulty level offset systematically awarded to compatriots on non-jump elements relative to their out-of-group career norms. Node shapes denote each specialist's baseline officiating tendency when evaluating non-compatriots (upward-pointing triangles represent strict infraction-calling baselines, downward-pointing triangles represent lenient baselines, and circles represent neutral baselines). Node colors distinguish statistically significant dual-metric and single-metric bias profiles after FDR correction (FDR $p < 0.05$). Numbered nodes correspond to the fifty individual technical specialists meeting the minimum sample thresholds. Specialists highlighted in blue font with asterisk represent those assigned to official technical panels at the 2026 Milan-Cortina Winter Olympic Games.

10. Realized Impact and Leaderboard Simulations

The baseline decomposition models in Section 4 successfully identify the raw point premiums injected by individual judges. However, interpreting the real-world consequences of these values requires a critical mechanical distinction. The metric calculated previously, the Adjusted In-Group Premium ($\pi_{j,p}$), represents the raw mathematical inflation submitted on the judge's scorecard. For clarity in evaluating final outcomes, this metric is conceptually reframed here as the judge's *Attempted Premium*. Crucially, a judge's Attempted Premium is not mathematically equivalent to the *Realized Impact* (R_p) on the athlete's final score.

The relationship between attempted bias and realized impact is not 1:1, nor is it immediately obvious what the precise relationship is, due to the ISU's trimmed mean algorithm. For every evaluated item on the score sheet, the highest and lowest scores awarded by the panel are discarded, and the remaining scores are averaged. Because of this trimming, the actual influence of a biased judge depends entirely on the variance of the rest of the panel.

To illustrate this non-linear dynamic, we consider two theoretical scenarios. In Scenario A **(**Blunt Bias)**,** a highly biased judge awards the absolute highest score for every technical element and program component. Because these scores are consistently the maximum outlier, the algorithm systematically trims them off. The judge's Attempted Premium is large, but the Realized Impact on the skater's final score is exactly zero. In Scenario B (Sophisticated Bias / High Variance) there is considerable variation among the judges about the perceived quality of the elements and components. The biased home-country judge consistently awards the *second-highest* score or lower for every element and component. Because these marks are not the absolute maximum, they survive the trimming process and are incorporated into the average. Here, a smaller Attempted Premium successfully "leaks" through the algorithm, resulting in a positive Realized Impact.

Because of this non-linear dynamic, we cannot estimate the net effect of nationalistic bias by simply subtracting a judge's mean bias from a skater's total score. To accurately quantify the true competitive distortion, we must determine, performance-by-performance, how much bias survives the trimmed mean.

10.1 Counterfactual Simulation Methodology

To address this, I introduce a theoretical, common-sense simulation. I ask the question: *What would the final scores and standings be if the nationalistic premium was entirely neutralized?*

For every performance evaluated by a home-country judge, the **Item-Level Recused Counterfactual (**E_i^{cf}**)** is calculated:

$$E_i^{cf} = \text{TrimmedMean}(s_1, \ldots, s_{N-1})[s_{home} \text{ removed}]$$

This formula computes the counterfactual score for any evaluated item i (encompassing both technical elements and program components) by entirely removing the home judge's input (s_{home}) from the dataset. This reduces the panel size from *N* to *N*-1 judges. We then apply the official ISU trimmed mean algorithm to each score. By summing these item-level differences across the entire score sheet, we isolate the **Performance Realized Impact (R_p)**:

$$R_p = \sum_{i \in p} (E_i^{actual} - E_i^{cf})$$

This metric represents an estimate of the point bonus that successfully bypassed the trimmed mean to inflate the skater's Total Segment Score. To evaluate the systemic efficiency of the ISU's algorithm at filtering out bias, we calculate the **Bias Conversion Rate (η_x)**:

$$\eta_x = \frac{R_p}{\Pi_j} \times 100\%$$

This quantifies the percentage of the judge's Attempted Premium that successfully reaches the scoreboard. Finally, to measure the macro-level competitive consequences, we conduct a **Leaderboard Simulation**. By subtracting R_p from the scores of all compatriot-judged skaters and re-sorting the standings:

$$S_{neutral} = S_{actual} - R_p$$

we can identify instances where the neutralized, counterfactual scores result in altered final placements or podium compositions.

This recusal-based counterfactual approach builds upon established methodologies in the sports economics literature. For example, Zitzewitz (2006) utilized a similar judge-removal simulation to demonstrate how the 2002 Olympic figure skating standings would have altered under unbiased panels. Similarly, Coupé and Gergaud (2013) applied a recusal counterfactual across a broad spectrum of Olympic judged events, recalculating trimmed means in the absence of national officials to estimate the global volume of altered medal outcomes. By adopting this framework, this study ensures that the simulated leaderboard shifts represent realistic, mathematically grounded competitive outcomes rather than abstract statistical projections.

10.2 The Efficacy of the Trimmed Mean in Reducing Nationalistic Bias

The counterfactual simulation reveals that while the ISU's trimmed mean procedure is very effective at suppressing the *magnitude* of nationalistic bias, it fails to eliminate it entirely. Across all disciplines and segments, the Bias Conversion Rate (η_x) ranges from 13.1.% to 20.8%. For example, in the Women's Short Program, home-country judges inject an average premium of +0.724 points. The trimmed mean successfully filters out approximately 83% of this inflation; however, a mean Realized Impact of +0.126 points still bypasses the algorithm (Table 8).

Crucially, despite this suppression, the Realized Impact remains strictly positive and statistically significant across all eight segments. Utilizing the Wilcoxon signed-rank test with FDR correction, the counterfactual scores are significantly lower than the actual scores ($p < 10^{-26}$ in all segments). This demonstrates that the trimmed mean mitigates, but does not neutralize, the mathematical advantage of having a compatriot on the panel (Table 8).

Table 8.

Segment	Mean Attempted Premium (Π_j)	Mean Realized Impact (Scoreboard)	Conversion Rate	N (Home)	FDR Wilcoxon p-value	Sig?
Ice Dance Free Dance	+2.691	**+0.423**	15.7%	940	5.11e-73	YES
Ice Dance Rhythm Dance	+2.370	**+0.343**	14.5%	978	5.47e-84	YES
Pairs Free Skating	+2.281	**+0.299**	13.1%	578	1.73e-30	YES
Men Free Skating	+2.005	**+0.359**	17.9%	1,174	2.85e-42	YES
Women Free Skating	+1.446	**+0.301**	20.8%	1,433	9.14e-49	YES
Pairs Short Program	+1.195	**+0.196**	16.4%	600	7.13e-31	YES
Men Short Program	+0.858	**+0.139**	16.2%	1,218	4.66e-26	YES
Women Short Program	+0.724	**+0.126**	17.4%	1,514	1.16e-30	YES

Table 8: Efficacy of the trimmed-mean algorithm in suppressing nationalistic judging premiums across segments. Comparison of the mean attempted nationalistic premium (Π_j, the raw deviation submitted on a judge's scorecard) against its mean realized impact (the actual point bonus that successfully bypasses the trimmed-mean algorithm to inflate the final scoreboard). The table reports the total volume of domestic evaluations (N), the resulting bias conversion rate (η_x, representing the percentage of the attempted premium that reaches the scoreboard), and the FDR-adjusted p-values derived from a Wilcoxon signed-rank test comparing actual scores against recusal-based counterfactual scores.

10.3 Competitive Consequences of De-Biased Judging

In elite figure skating, where podium placements are frequently decided by fractions of a point, the leakage of approximately 0.1 to 0.3 realized points may be sufficient to alter competitive outcomes. Indeed, the Leaderboard Simulation (N = 660 competitions) demonstrates that nationalistic bias actively distorts the sport's hierarchy. When the Realized Impact (R_p) is removed and standings are re-sorted, 37.1% of all competitions experience at least one placement change. Consequentially, 4.8% of competitions feature an altered podium makeup (a skater being pushed onto or off the podium), and 5.2% feature a medal color swap whereby the podium athletes remain the same, but their order changes (Table 9a, Supplementary Table 5).

An informative example of the influence of realized nationalistic bias occurred at the 2024/2025 European Figure Skating Championships (Ice Dance Overall Standings). The official results awarded the Bronze medal to the team of Lilah Fear and Lewis Gibson (GBR, 206.02 points). The fourth-place team, Juulia Turkkila and Matthias Versluis (FIN), received 205.69 points. Removing the effect of Realized Impact swaps the third- and fourth-place rankings of these skaters, elevating Turkkila and Versluis to a Bronze medal in this competition (Table 9b). The Realized Impact underlying this shift is particularly interesting because Turkkila and Versluis were slightly punitively evaluated by her compatriot judges, whereas Fear and Gibson received a point bonus. This case study exemplifies the fundamental vulnerability of the current system. The trimmed mean functions to eliminate most of the nationalistic bias from final scores, but the fractional bias that survives is sufficient to alter international medal distributions.

Table 9.

a)

Metric	Segment-Level	Overall Competition-Level
Total Events Analyzed	**1,176 segments**	**660 competitions**
ANY Placement Change	475 (40.4%)	245 (37.1%)
Altered Podium (Top 3 Makeup)	63 (5.4%)	32 (4.8%)
Medal Color Swap (Top 3 Order)	74 (6.3%)	34 (5.2%)

b)

CASE STUDY: RE-RANKED STANDINGS

Example: Season 2425 EUROPEAN FIGURE SKATING CHAMPIONSHIPS 2025 — Dance's Overall Standings (CHAMP Series)

Skater Name	Country	Official Rank	Official Total	Total Bias Impact (R_p)	Neutral Total	De-Biased Rank	If Bias Removed...
Charlene GUIGNARD Marco FABBRI	**ITA**	1	212.12	+0.112	212.01	1	-
Evgeniia LOPAREVA Geoffrey BRISSAUD	**FRA**	2	206.76	+0.150	206.61	2	-
Lilah FEAR Lewis GIBSON	**GBR**	3	206.02	+0.438	205.58	4	⬇ Dropped
Juulia TURKKILA Matthias VERSLUIS	**FIN**	4	205.69	-0.048	205.74	3	⬆ Moved Up
Olivia SMART Tim DIECK	**ESP**	5	198.98	+0.459	198.52	5	-
Allison REED Saulius AMBRULEVICIUS	**LTU**	6	196.66	0.000	196.66	6	-

Table 9: Leaderboard simulation results and case study of judging panel bias neutralization. Leaderboard simulation metrics and a specific event-level case study illustrating the competitive consequences of neutralizing judging panel bias via a recusal-based counterfactual model. **(a)** Aggregate counts and percentages of disrupted competitive outcomes across 801 individual segments and 405 overall competitions. Disruptions are categorized as: any placement changes, altered podium makeups (skaters pushed onto or off the podium), and medal color swaps (changes in the relative ranking order of the top three athletes). **(b)** A case study displaying the re-ranked standings for the Women's overall event at the 2024/2025 Skate Canada Grand Prix. The table reports official ranks and totals alongside the performance realized impact (R_p, the fractional point bias that successfully bypassed the trimmed-mean filter) and the resulting neutral, counterfactual totals and ranks.

10.4 Technical Specialist De-Biasing and Counterfactual Simulation Methodology

While the counterfactual simulation in Section 10.1 demonstrates that the trimmed-mean algorithm successfully suppresses the vast majority of a primary judge's attempted nationalistic premium, the Technical Panel operates under a completely different mechanical regime. The Technical Specialist holds deterministic, single-source control over the underlying Base Value of a performance. Because their infraction calls and level assignments are not averaged or trimmed by a larger panel, any nationalistic deviation they exhibit is injected directly into the final scoreboard without mitigation. Unlike the primary judging panel, the deterministic nature of the Technical Specialist dictates that their bias conversion rate is mathematically absolute ($\eta_T = 100\%$).

To quantify the real-world standings impact of this deterministic bias, this study constructs a point-based Expected Value Substitution (EVS) counterfactual model. For any performance p evaluated by a compatriot Technical Specialist j, the algorithm mathematically extracts the specialist's statistically verified nationalistic premiums: the Infraction Shielding Effect (ΔI_{pts}) and the Level Shielding Effect (ΔL_{pts}), as calculated in Section 8.

We define the Performance Technical Realized Impact ($R_{T,p}$) as the sum of the points artificially saved through infraction leniency and the points gained through difficulty level shielding. Because this simulation specifically tests for the removal of nationalistic favoritism, the calculation isolates positive shielding values (leniency) and bounds negative values at zero:

$$R_{T,p} = \max(0, \Delta I_{pts}) + \max(0, \Delta L_{pts})$$

By subtracting the Performance Technical Realized Impact ($R_{T,p}$) from the skater's actual, officially recorded segment score, the simulation isolates the counterfactual, unbiased technical outcome:

$$\text{Score}_{neutral,p} = \text{Score}_{actual,p} - R_{T,p}$$

Following score neutralization, the algorithm reconstructs the competitive leaderboards at both the individual segment level and the overall total competition level. To quantify the disruptive power of Technical Panel bias, the counterfactual leaderboards are compared against the actual historical ISU standings across three tiers of severity: 1) Any absolute placement change within the field; 2) Altered podium composition (a skater unfairly pushed into or out of the top three); and 3) Medal color swaps (reordering within the top three).

Finally, to accurately measure the density of this threat, the simulation isolates "exposed" leaderboards, defined as any event segment or competition featuring at least one skater evaluated by a compatriot Technical Specialist. Comparing the disruption rates within these exposed competitive environments against the global baseline demonstrates the precise, leaderboard-altering capacity of localized technical alliances.

10.5 Realized Impact of the Technical Panel: Standings and Simulation Results

Applying the Expected Value Substitution (EVS) counterfactual model to the dataset reveals the profound leaderboard consequences of technical specialist nationalistic scoring deviations. Table 10a details the results of this simulation across both individual competitive segments and overall competition standings. However, properly interpreting these results, and comparing them to the judging panel simulations in Section 10.3, requires addressing a critical structural difference between the two officiating bodies, which necessitates the dual-column presentation in Table 10a.

Table 10.

a)

Metric	Segment-Level (All Panels: N=1,176)	Segment-Level (Compatriot Seated: N=306)	Competition-Level (All Panels: N=660)	Competition-Level (Compatriot Seated: N=168)
ANY Placement Change	134 (11.4%)	134 (43.8%)	66 (10.0%)	66 (39.3%)
Altered Podium (Top 3 Makeup)	22 (1.9%)	22 (7.2%)	12 (1.8%)	12 (7.1%)
Medal Color Swap (Top 3 Order)	19 (1.6%)	19 (6.2%)	8 (1.2%)	8 (4.8%)

b)

Metric	Segment-Level (All Panels: N=1,176)	Segment-Level (Compatriot Seated: N=1,157)	Competition-Level (All Panels: N=660)	Competition-Level (Compatriot Seated: N=651)
ANY Placement Change	530 (45.1%)	530 (45.8%)	265 (40.2%)	265 (40.7%)
Altered Podium (Top 3 Makeup)	79 (6.7%)	79 (6.8%)	39 (5.9%)	39 (6.0%)
Medal Color Swap (Top 3 Order)	86 (7.3%)	86 (7.4%)	38 (5.8%)	38 (5.8%)

Table 10: Counterfactual leaderboard simulation results for technical specialist and comprehensive dual-panel de-biasing. Aggregate simulation metrics displaying the competitive consequences of neutralizing nationalistic bias within the Technical Panel alone, and within both officiating panels simultaneously. Results are reported across the entire calendar of events ("All Panels") and restricted strictly to the active risk pool where a domestic conflict of interest exists ("Compatriot Seated"). **(a)** Leaderboard disruptions resulting from the Expected Value Substitution (EVS) counterfactual model designed to neutralize technical specialist bias. **(b)** Cumulative leaderboard disruptions resulting from the Comprehensive Dual-Panel Simulation, which simultaneously neutralizes both judging panel and technical specialist bias.

A standard judging panel is composed of up to nine individuals, creating a high statistical probability that at least one skater in any given event will be evaluated by a compatriot judge. Consequently, judging bias is a diffuse, pervasive phenomenon affecting a majority of competitions. In contrast, the Technical Panel features only one primary Technical Specialist responsible for calling the elements and infractions. The probability of a skater drawing a compatriot for this single, highly deterministic seat is significantly lower. Therefore, to accurately measure the potency of technical bias, we must evaluate it not only across the entire calendar of events ("All Panels") but strictly within the concentrated risk pool of events where a conflict of interest actually exists ("Compatriot Seated").

The absolute number of disrupted outcomes remains mathematically identical across both column groupings. For example, exactly 66 competitions experienced a placement change whether measured against the total sample (*N* = 660) or the restricted sample (*N* = 168). This isolates the effect perfectly, confirming that outcome distortions triggered by the EVS counterfactual are exclusively localized to instances where a compatriot technical specialist was seated.

When evaluated within this restricted risk pool, the impact of technical specialist bias is severe and proportionally eclipses the influence of the primary judging panel. In competitions where a compatriot specialist is seated (*N* = 168), correcting for their statistically verified Infraction Shielding and Level Shielding results in at least one placement change in 39.3% of events (Table 10a). This represents a significant volatility rate comparable to the 37.1% alteration rate generated by de-biasing the nine-member judging panel (Table 9a).

The distortion at the top of the leaderboard is also similar to the effect of de-biasing the judging panels. When a compatriot specialist is present, their unmitigated bias alters the physical makeup of the podium (pushing a skater into or out of the top three) in 7.1% of competitions and alters the medal color order in an additional 4.8% of competitions (Table 10a). Both metrics are similar to the equivalent podium disruption rates caused by the primary judging panel (4.8% and 5.2%, respectively; Table 9a).

These comparative results empirically validate the mechanical theory proposed in Section 10.4. While significant nationalistic premiums originating from the primary judging panel are frequent, the ISU's trimmed mean algorithm successfully suppresses the vast majority of its *Attempted Premium*, limiting the *Realized Impact* to fractional point leakages. Conversely, while technical specialist deviations are far less frequent due to panel size constraints, its deterministic nature dictates a mathematically absolute 100% *Bias Conversion Rate*. Technical premiums bypass all regulatory score filters; therefore the technical specialist role represents a highly concentrated point of systemic vulnerability. Any nationalistic variance at this position translates directly to the scoreboard, making it a more potent vector for competitive distortion than a single judge, with the structural capacity to unilaterally alter international medal distributions.

10.6 Comprehensive Dual-Panel Simulation Re-Ranking: Total Systemic Impact

Having established the independent competitive consequences of the judging panel (Section 10.3) and the technical panel (Section 10.5), the final step in quantifying the true footprint of nationalistic bias requires a unified simulation. Because a skater's final score is the combined product of the technical panel and subjective grading, evaluating these officiating bodies in isolation likely underestimates the total systemic distortion. To address this, I construct a Comprehensive Dual-Panel Simulation, subtracting both the judging panel's Performance Realized Impact (R_p) and the technical specialist's Performance Technical Realized Impact ($R_{T,p}$) from the actual scores to yield a fully neutralized, counterfactual leaderboard.

Table 10b presents the results of this combined de-biasing protocol. The first notable observation from this simulation is the ubiquity of domestic influence within the current competitive landscape. When defining "Compatriot Seated" as the presence of *either* a home-country judge or a home-country technical specialist, the risk pool expands to encompass nearly the entire dataset: 651 of the 660 evaluated competitions featured at least one compatriot official evaluating a domestic skater. This confirms that exposure to mathematically verified nationalistic premiums is not an edge case, but a structural constant of the International Judging System.

Within this near-universal risk pool (N = 651), the cumulative effect of judging and technical specialist deviations is striking. When all verified nationalistic premiums are neutralized, 40.7% of all competitions experience at least one placement change in the final overall standings. This demonstrates a highly cumulative effect, where the fractional "leakage" of points through the trimmed mean algorithm operates in tandem with the deterministic base-value inflations of the technical panel to generate significant leaderboard volatility.

The consequences at the absolute pinnacle of the sport are equally severe. In competitions where compatriot officials are seated, the comprehensive simulation reveals that 6.0% feature an altered podium makeup. An additional 5.8% of competitions experience a medal color swap. In aggregate, more than one in ten figure skating competitions yields a medal outcome attributable to domestic favoritism (Table 10b).

Ultimately, the Dual-Panel Simulation demonstrates that the modern IJS does not merely suffer from isolated pockets of subjective leniency. Instead, the aggregate friction of nationalistic bias acts as a persistent, outcome-altering force. While the current regulatory framework actively mitigates the *magnitude* of these individual premiums, it structurally fails to prevent them from routinely deciding international medal distributions.

10.7 Case Study: De-Biasing the 2026 Olympic Games Individual Events

To ground these systemic probabilities in a highly salient real-world context, the Comprehensive Dual-Panel Simulation was applied to the pinnacle events of the 2022–2026 quadrennial: the Individual competitions at the 2026 Milan-Cortina Winter Olympic Games. Tables 11a through 11d present the official standings alongside the neutralized, counterfactual leaderboard for Ice Dance, Men’s Singles, Women’s Singles, and Pairs, respectively.

Analyzing the Olympic standings encapsulates the core econometric dynamics established throughout this study. Most notably, the simulation results directly validate the verifiability hypothesis introduced in Section 5.2. As previously established, evaluators disproportionately demonstrate bias in highly subjective, interpretive scores. The Olympic simulation mirrors this structural vulnerability. Among the four disciplines, Ice Dance (Table 11a) exhibited the highest level of leaderboard volatility upon de-biasing. Removing the verified domestic premiums resulted in four distinct placement swaps affecting eight different teams. In the upper half of the standings, the British team of Lilah Fear and Lewis Gibson artificially lost a placement due to domestic bias. In the neutralized simulation, they move up to 6th place, dropping the Lithuanian team of Allison Reed and Saulius Ambrulevičius to 7th. Further down the leaderboard, the Georgian team of Diana Davis and Gleb Smolkin (a team previously flagged for synchronous reciprocal exchange in Section 7.1) drops from 12th to 13th in the absence of bias, yielding their placement to Finland's Juulia Turkkila and Matthias Versluis. Similar re-rankings occur between the Swedish and Spanish teams, as well as the Korean and Chinese teams.

Conversely, the Olympic Men’s (Table 11b), Women’s (Table 11c), and Pairs (Table 11d) events demonstrate the mitigating effect of wide ordinal margins. The simulation confirms that nationalistic premiums were undeniably present on the Olympic scorecards in these disciplines. For instance, in the Men's event, Canada's Stephen Gogolev received a Realized Impact of +1.129 points, and the USA's Maxim Naumov received +0.961 points. In Pairs, France's Camille Kovalev and Pavel Kovalev received +1.650 points, while the Japanese team of Riku Miura and Ryuichi Kihara were evaluated punitively with a net impact of -1.112 points.

However, despite the mathematical presence of these premiums, the actual performance gaps between the skaters on that specific day were too large for premium realized impacts to successfully bridge. In the Men's and Pairs events, removing the Total Bias Impact (R_p) results in no ordinal changes. In the Women's event, only a single swap occurs, with South Korea's Haein Lee moving up to 7th place and dropping Estonia's Niina Petrokina to 8th.

Table 11.

a)

OLYMPIC ICE DANCE STANDINGS

Skater Name	Country	Official Rank	Official Total	Total Bias Impact (Rp)	Neutral Total	De-Biased Rank	If Bias Removed...
Laurence Fournier Beaudry / Guillaume Cizeron	FRA	1	225.82	-0.717	226.54	1	-
Madison Chock / Evan Bates	USA	2	224.39	-0.357	224.75	2	-
Piper Gilles / Paul Poirier	CAN	3	217.74	+1.016	216.72	3	-
Charlene Guignard / Marco Fabbri	ITA	4	209.58	+0.256	209.32	4	-
Emilea Zingas / Vadym Kolesnik	USA	5	206.72	-0.346	207.07	5	-
Lilah Fear / Lewis Gibson	GBR	7	204.32	-0.354	204.67	6	⬆ Moved Up
Allison Reed / Saulius Ambrulevičius	LTU	6	204.66	+0.000	204.66	7	⬇ Dropped
Evgeniia Lopareva / Geoffrey Brissaud	FRA	8	203.68	-0.146	203.83	8	-
Olivia Smart / Tim Dieck	ESP	9	201.49	-0.489	201.98	9	-
Marjorie Lajoie / Zachary Lagha	CAN	10	199.80	+1.484	198.32	10	-
Christina Carreira / Anthony Ponomarenko	USA	11	197.62	+0.016	197.60	11	-
Diana Davis / Gleb Smolkin	GEO	13	196.02	-0.514	196.53	12	⬆ Moved Up
Juulia Turkkila / Matthias Versluis	FIN	12	196.03	+0.290	195.74	13	⬇ Dropped
Marie-Jade Lauriault / Romain Le Gac	CAN	14	187.18	+0.327	186.85	14	-
Natálie Taschlerová / Filip Taschler	CZE	15	185.00	-0.026	185.03	15	-
Kateřina Mrázková / Daniel Mrázek	CZE	16	181.44	+0.000	181.44	16	-
Phebe Bekker / James Hernandez	GBR	17	179.45	-0.669	180.12	17	-
Holly Harris / Jason Chan	AUS	18	176.39	+0.816	175.57	18	-
Milla Ruud Reitan / Nikolaj Majorov	SWE	20	165.05	+0.000	165.05	19	⬆ Moved Up
Sofía Val / Asaf Kazimov	ESP	19	165.23	+0.371	164.86	20	⬇ Dropped
Hannah Lim / Ye Quan	KOR	22	64.69	+0.000	64.69	21	⬆ Moved Up
Wang Shiyue / Liu Xinyu	CHN	21	64.76	+0.409	64.35	22	⬇ Dropped
Jennifer Janse van Rensburg / Benjamin Steffan	GER	23	63.67	+0.000	63.67	23	-

b)

OLYMPIC MEN'S SINGLE STANDINGS

Skater Name	Country	Official Rank	Official Total	Total Bias Impact (Rp)	Neutral Total	De-Biased Rank	If Bias Removed...
Mikhail Shaidorov	KAZ	1	291.58	-0.722	292.30	1	-
Yuma Kagiyama	JPN	2	280.06	+0.000	280.06	2	-
Shun Sato	JPN	3	274.90	+0.000	274.90	3	-
Cha Junhwan	KOR	4	273.92	+0.140	273.78	4	-
Stephen Gogolev	CAN	5	273.78	+1.129	272.65	5	-
Petr Gumennik	AIN	6	271.21	+0.000	271.21	6	-
Adam Siao Him Fa	FRA	7	269.27	-0.553	269.82	7	-
Ilia Malinin	USA	8	264.49	+0.642	263.85	8	-
Daniel Grassl	ITA	9	263.71	+0.397	263.31	9	-
Nika Egadze	GEO	10	260.27	-0.045	260.31	10	-
Kevin Aymoz	FRA	11	259.94	+0.239	259.70	11	-
Andrew Torgashev	USA	12	259.06	+0.035	259.02	12	-
Kao Miura	JPN	13	246.88	+0.000	246.88	13	-
Lukas Britschgi	SUI	14	246.64	+0.483	246.16	14	-
Matteo Rizzo	ITA	15	243.18	-0.324	243.50	15	-
Aleksandr Selevko	EST	16	236.82	+0.000	236.82	16	-
Boyang Jin	CHN	17	229.08	-0.251	229.33	17	-
Deniss Vasiljevs	LAT	18	226.46	+0.000	226.46	18	-
Kyrylo Marsak	UKR	19	224.17	+0.000	224.17	19	-
Maxim Naumov	USA	20	223.36	+0.961	222.40	20	-
Vladimir Samoilov	POL	21	222.25	+0.000	222.25	21	-
Donovan Carrillo	MEX	22	219.06	+0.000	219.06	22	-
Yu-Hsiang Li	TPE	23	214.33	+0.000	214.33	23	-
Adam Hagara	SVK	24	202.38	+0.000	202.38	24	-
Tomas Guarino Sabate	ESP	25	69.80	+0.000	69.80	25	-
Hyungyeom Kim	KOR	26	69.30	+0.127	69.17	26	-
Andreas Nordeback	SWE	27	67.15	+0.000	67.15	27	-
Fedirs Kuliss	LAT	28	66.86	+0.000	66.86	28	-
Vladimir Litvintsev	AZE	29	63.63	+0.538	63.09	29	-

c)

OLYMPIC WOMEN'S SINGLE STANDINGS

Skater Name	Country	Official Rank	Official Total	Total Bias Impact (Rp)	Neutral Total	De-Biased Rank	If Bias Removed...
Alysa Liu	USA	1	226.79	-0.382	227.17	1	-
Kaori Sakamoto	JPN	2	224.90	+0.423	224.48	2	-
Ami Nakai	JPN	3	219.16	+0.518	218.64	3	-
Mone Chiba	JPN	4	217.88	+0.473	217.41	4	-
Amber Glenn	USA	5	214.91	-0.024	214.93	5	-
Adeliia Petrosian	AIN	6	214.53	+0.000	214.53	6	-
Haein Lee	KOR	8	210.56	-0.875	211.44	7	⬆ Moved Up
Niina Petrokina	EST	7	210.82	+0.000	210.82	8	⬇ Dropped
Anastasiia Gubanova	GEO	9	209.99	+0.000	209.99	9	-
Sofia Samodelkina	KAZ	10	207.46	-0.443	207.90	10	-
Jia Shin	KOR	11	206.68	-0.406	207.09	11	-
Isabeau Levito	USA	12	202.80	-0.438	203.24	12	-
Nina Pinzarrone	BEL	13	200.30	+0.000	200.30	13	-
Loena Hendrickx	BEL	14	199.65	+0.000	199.65	14	-
Lara Naki Gutmann	ITA	15	195.75	+0.000	195.75	15	-
Iida Karhunen	FIN	16	192.79	+0.000	192.79	16	-
Julia Sauter	ROU	17	190.93	+0.434	190.50	17	-
Olga Mikutina	AUT	18	185.59	+0.326	185.26	18	-
Ruiyang Zhang	CHN	19	178.03	+0.000	178.03	19	-
Ekaterina Kurakova	POL	20	173.37	+0.060	173.31	20	-
Livia Kaiser	SUI	21	171.52	+0.191	171.33	21	-
Lorine Schild	FRA	22	167.08	-0.008	167.09	22	-
Kimmy Repond	SUI	23	159.54	+0.127	159.41	23	-
Mariia Seniuk	ISR	24	152.61	+0.366	152.24	24	-
Madeline Schizas	CAN	25	55.38	+0.000	55.38	25	-
Viktoriia Safonova	AIN	26	54.57	+0.000	54.57	26	-
Meda Variakojyte	LTU	27	53.86	+0.000	53.86	27	-
Alexandra Feigin	BUL	28	53.42	+0.002	53.42	28	-
Kristen Spours	GBR	29	45.54	-0.460	46.00	29	-

d)

OLYMPIC PAIR SKATING STANDINGS

Skater Name	Country	Official Rank	Official Total	Total Bias Impact (Rp)	Neutral Total	De-Biased Rank	If Bias Removed...
Riku Miura / Ryuichi Kihara	JPN	1	231.24	-1.112	232.35	1	-
Anastasiia Metelkina / Luka Berulava	GEO	2	221.75	-0.344	222.09	2	-
Minerva Fabienne Hase / Nikita Volodin	GER	3	219.09	+0.236	218.85	3	-
Maria Pavlova / Alexei Sviatchenko	HUN	4	215.26	-0.919	216.18	4	-
Wenjing Sui / Cong Han	CHN	5	208.64	+0.132	208.51	5	-
Sara Conti / Niccolo Macii	ITA	6	203.19	+0.185	203.01	6	-
Emily Chan / Spencer Akira Howe	USA	7	200.31	-0.083	200.39	7	-
Lia Pereira / Trennt Michaud	CAN	8	199.66	-0.433	200.09	8	-
Ellie Kam / Danny O'Shea	USA	9	194.58	+0.043	194.54	9	-
Annika Hocke / Robert Kunkel	GER	10	194.11	+0.065	194.04	10	-
Deanna Stellato-Dudek / Maxime Deschamps	CAN	11	192.61	+0.000	192.61	11	-
Rebecca Ghilardi / Filippo Ambrosini	ITA	12	191.86	+0.428	191.43	12	-
Ioulia Chtchetinina / Michal Wozniak	POL	13	185.86	+0.133	185.73	13	-
Karina Akopova / Nikita Rakhmanin	ARM	14	180.66	+0.000	180.66	14	-
Anastasia Vaipan-Law / Luke Digby	GBR	15	179.06	-0.146	179.21	15	-
Camille Kovalev / Pavel Kovalev	FRA	16	178.43	+1.650	176.78	16	-
Daria Danilova / Michel Tsiba	NED	17	64.07	+0.216	63.85	17	-
Anastasiia Golubeva / Hektor Giotopoulos Moore	AUS	18	60.69	+0.000	60.69	18	-
Yuna Nagaoka / Sumitada Moriguchi	JPN	19	59.62	+0.000	59.62	19	-

Table 11: Counterfactual Olympic leaderboard re-rankings across the four figure skating disciplines. Comprehensive dual-panel de-biasing simulation results applied to the individual events at the 2026 Milan-Cortina Winter Olympic Games: (a) Ice Dance, (b) Men's Singles, (c) Women's Singles, and (d) Pairs. For each competitor, the table reports their official rank and score, the cumulative dual-panel bias impact (R_p, which combines the judging panel's realized impact with the technical specialist's realized impact), and the resulting neutral, counterfactual total and rank.

This Olympic simulation study serves as an empirical illustration of the interaction between evaluator measurement error and ordinal rankings. The data demonstrate that statistically significant domestic point premiums remain measurable even at the sport's highest competitive tier. However, the translation of these premiums into altered competitive placements is contingent upon the underlying variance in athletic execution. The performance margins were sufficiently wide to absorb the documented in-group premiums without disrupting the podium structures. However, the inherently compressed scoring margins characteristic of Ice Dance rendered the standings structurally more sensitive to evaluator variance, demonstrating how realized point premiums can actively alter final Olympic rank orders.

11. Discussion, Recommendations and Future Directions

It is likely that a majority of the officials identified in this study are unaware of their own statistical deviations. Judges and technical specialists in elite figure skating operate in a profound informational vacuum. Even if an official is committed to self-evaluation and objective scoring, the current infrastructure offers no pathway for self-monitoring. For example, the raw materials of figure skating results, individual judge scorecards, segment protocols, and panel compositions, are contained within thousands of unstructured, nested PDF documents. Without advanced programmatic scraping architectures, extracting and analyzing longitudinal scoring patterns is practically impossible for an individual official.

The ISU does not conduct proactive, systemic statistical tracking, thus the responsibility for comprehensive data analyses falls almost exclusively onto independent, external investigators. Consequently, these deep-dive analyses only emerge at decadal intervals (e.g., Zitzewitz 2006; Templon 2018, current study), leaving officials with no ongoing diagnostic feedback.

This institutional blind spot is reflected in the ISU's current disciplinary framework, which remains strictly reactive rather than proactive. During the 2022–2026 Olympic cycle, the ISU Disciplinary Commission did execute high-profile sanctions against officials, such as the lifetime ban of Ukrainian judge Yuri Balkov (International Skating Union 2025a) and the suspension of Hungarian judge Dr. Katalin Balczo (International Skating Union 2025b). However, these cases highlight a critical regulatory limitation: the ISU primarily penalizes "smoking gun" ethical breaches exposed by direct, verbal, or textual whistleblowing, while leaving systemic mathematical bias entirely unaddressed.

Comparing the current findings with the 2018 BuzzFeed News investigation reveals a striking pattern of institutional retention. Of the 27 most statistically biased judges identified in BuzzFeed's 2016–2017 sample (Templon, 2018), exactly one-third (9) were Russian officials who were suspended from international competition during the 2022–2026 cycle due to geopolitical sanctions. However, analyzing the remaining cohort reveals that officials flagged for extreme nationalistic bias nearly a decade ago remain highly active, and highly biased, in the modern officiating pool.

Applying the double-normalized baseline model to this active pool reveals that nine of these historical actors continue to award statistically significant nationalistic premiums during the 2022–

2026 quadrennial. The point premiums awarded by these officials indicate that their evaluation habits have persisted despite changes in the judging system and heightened public scrutiny. Walter Toigo (ITA), who was flagged as the single most biased judge in the world in the 2018 investigation (with a historical preference score of 7.496), remains at the extreme upper bound of the modern judging pool. Under this study's highly conservative mathematical constraints, Toigo exhibits a large, statistically significant Mean Nationalistic Premium of +4.03 points. Similarly, Marta Olozagarre (ESP) and Richard Dalley (USA) maintain statistically significant in-group premiums of +4.06 points and +3.03 points, respectively. Other historically flagged judges, including Cynthia Benson (CAN, +3.23 pts), Nicole LeBlanc-Richard (CAN, +1.85 pts), and Lorrie Parker (USA, +1.27 pts), also successfully cleared the strict False Discovery Rate (FDR) significance thresholds in the current dataset. In fact, among the active non-Russian judges from the 2018 list, only Weiguang Chen (CHN) failed to meet the statistical threshold for significant in-group favoritism in the current cycle.

The real-world consequence of this institutional retention is most evident when examining the sport's premier events. Because the governing body lacks a proactive, data-driven oversight mechanism, historically biased judges are routinely seated at the highest- level competitions. Of the officials flagged by BuzzFeed in 2018 who remain statistically biased in the current quadrennial, four judges, namely Marta Olozagarre (ESP, +4.06 pts), Isabella Micheli (ITA, +3.34 pts), Anna Kantor (ISR, +2.52 pts), and Feng Huang (CHN, +2.27 pts) all served on judging panels at the 2026 Olympic Winter Games (Figure 6).

The promotion of these officials underscores the fundamental vulnerability of the International Judging System. Without a systemic shift toward statistical transparency and proactive judge feedback, officials will continue to be unaware of their historical scoring patterns. Until the ISU implements data-driven monitoring, nationalistic bias will remain not just a persistent feature of the sport, but one that is inadvertently rewarded with Olympic assignments.

11.1 Effects and Possible Outcomes of Judge Recusal

By providing an unprecedented, empirical analysis of both the judging and technical panels, this study offers a highly granular look at the precise mechanics of evaluation bias. Rather than relying on intuitive or speculative reform concepts, these data-driven findings can directly inform evidence-based policy recommendations designed to safeguard competitive integrity. A primary example of an intuitive but structurally flawed reform is the mandatory recusal of the compatriot judge. Often proposed in the wake of scoring controversies (e.g., Brennan 2014; Zitzewitz 2006), this policy seeks to eliminate nationalistic bias by barring judges from scoring their own compatriots or automatically expunging their evaluations of their own athletes.

From an econometric perspective, however, rational actor theory suggests that a simple recusal rule merely triggers a strategic substitution of bias rather than its elimination. Officials adapt dynamically to institutional rules. If an official is blocked from directly inflating a compatriot's score, they are incentivized to pivot toward alternative, highly damaging strategies, such as the punitive downgrading of that compatriot's closest rivals or engaging in reciprocal point-trading. Because dropping a judge's score for their own country does not restrict them from scoring other competitors, their capacity to monitor the critical point-spread between top contenders remains entirely intact.

The 2026 Olympic Ice Dance event serves as a clear, empirical illustration of this structural limitation. In that competition, where the French team of Laurence Fournier-Beaudry and Guillaume Cizeron won gold over the American team of Madison Chock and Evan Bates by a narrow margin of 1.43 points, controversy centered on the scoring of French judge Jézabel Dabouis (e.g., Hersh 2026; McCarvel 2026). My counterfactual simulation demonstrates that applying a "drop-compatriot" rule, expunging the French judge’s score from the French team and the American judge's score from the American team, completely fails to alter the final standings and actually increases the margin of victory (Table 11a).

This counterintuitive mathematical result occurs because the result-driving vector in that event was not the direct inflation of the French team, but Dabouis' scoring of the American team 5.2 points below the panel average. This dynamic provides striking empirical validation for long-standing warnings in the academic literature. Proposals to systematically exclude compatriot scores (e.g., Fenwick and Chatterjee, 1981) have historically been criticized because they risk introducing new distortions. This case study demonstrates exactly why. Intuitive, recusal-based reforms will probably fail to protect the integrity of the leaderboard because they do not account for the "strategic adaptations" (Scheer & Ansorge, 1987) of the evaluators, nor the mathematical sensitivity of the trimmed-mean algorithm.

11.2 Information Restriction and Leaderboard Blinding

A fundamental principle of market regulation and game theory is that strategic optimization requires information. If a regulatory body can successfully restrict the information set available to an agent, the transaction costs of executing biased or collusive behavior increase dramatically. Empirical research in behavioral economics demonstrates that highly trained, professional officials remain deeply susceptible to immediate, in-arena social and environmental pressures. For instance, Garicano et al. (2005) demonstrated that soccer referees systematically bias their decision-making in response to localized crowd pressure, favoring home teams most aggressively when environmental cues are strongest. In figure skating, an under-appreciated catalyst of bias is the abundance of these real-time, non-essential environmental cues and scoreboard data delivered to the judging panel during a live competition.

This informational exposure is structurally embedded within the competition's design. To maximize spectator entertainment and television ratings, skaters in the Free Skate or Free Dance segment often perform in reverse order of their Short Program standings, ensuring that the highest-ranked medal contenders perform last in a deterministic sequence. While this sequencing is vital for the sport's commercial appeal, it simultaneously exposes officials to a continuous, high-intensity stream of running scores and crowd reactions.

Although the judges' digital touchscreens are technically "blinded" to the running leaderboard, this administrative safeguard is completely neutralized by the public arena environment. Sitting rink-side directly facing the ice, judges have the arena jumbotron in their immediate line of sight.

Furthermore, the arena's public address (PA) system loudly broadcasts each competitor's total score and current ranking to the crowd while the next skater prepares to perform, meaningfully reinforcing the precise point-differential boundaries required for strategic scoring adjustments.

This real-time access facilitates a highly precise mechanism to influence placements. When a home-country judge evaluates a compatriot skating in the final flight, they do not need to guess the required margin to secure a placement over a rival who has already performed because they have heard the rival's exact score and can see the required delta on the arena jumbotron. To clear this threshold, the judge only needs to inflate their Grade of Execution (GOE) and Program Component (PCS) inputs slightly above that delta. This marginal, targeted inflation easily survives the trimmed-mean algorithm because it does not register as a blunt, maximum-outlier score, yet it can successfully leak through to alter the standings.

To mitigate this vulnerability, the ISU could investigate a system of batch-processed, flight-blind scoring. Under this protocol, judges would evaluate a flight of skaters (each "warm-up group") back-to-back without any intermediate scores being calculated, announced, or displayed on the arena jumbotrons. The public scoreboards would remain frozen, and the PA announcer would remain silent regarding placements until the entire flight has concluded.

From a feasibility perspective, this intervention is practical in eliminating a key source of information required for the judges to execute real-time strategic slotting. Physically, it requires no changes to the arena layout or the starting order. Fundamentally, this proposition does not impact the functions of the judging panel in any way; their duties to evaluate the quality of the elements and components presented does not seem to require access to cumulative scores or ordinal placements. The scoring software would simply hold the judges' inputs in a secure queue, releasing the final calculated scores simultaneously once the last skater in the flight finishes. Critics might argue this harms the spectator experience. However, batch-processing might enhance the dramatic tension of the event when standings of the final flight are unveiled in a single, high-suspense climax at the end of the event.

11.3 Stochastic Scoring Mechanisms and Randomized Sub-Panels

As demonstrated by the category-specific decomposition and gamesmanship analyses in this study, modern figure skating officials are not merely biased. They are highly sophisticated, expert navigators of the International Judging System and its trimmed-mean algorithm. Over decades of officiating, judges have developed a deep understanding of how the trimmed mean behaves, allowing an intentionally malicious judge to optimize their inputs so that their *Attempted Premiums* successfully bypass outlier filters to become *Realized Impacts*. They know how to calibrate their scores to survive the trimming process, making coordinated gamesmanship like reciprocal point-trading and strategic slotting highly predictable and low-risk.

To disrupt this strategic certainty, the ISU could experiment with a stochastic scoring mechanism that injects a randomized, mathematical risk into the evaluation process. Currently, more judges are invited to international events than are seated on the active nine-member panel, with the final panel selected via a lottery immediately prior to the segment. A more robust, game-theoretic intervention would be to seat a larger panel of judges (for example, fifteen officials) who all evaluate every performance. However, for each individual skater's performance, a randomized algorithm would select a subset of nine judges whose scores are actually fed into the trimmed-mean calculation. Crucially, this random selection would occur after the scores are submitted or otherwise kept completely blind to the judges during the event.

From an econometric perspective, this stochastic element fundamentally destabilizes not only collusive agreements but also targeted nationalistic bias and strategic point-spread variance. As established by Duggan and Levitt (2002) in their analysis of bilateral sports collusion, corrupt or cooperative behaviors are highly sensitive to certainty. When agents can guarantee that their favors will be successfully delivered, or that a fractional ordinal adjustment will survive the scoring filter, these strategic behaviors remain stable. Introducing strategic uncertainty immediately disrupts these equilibria by dramatically increasing the transaction costs and mathematical risks of manipulation.

Under a standard, deterministic panel, allied judges from Country A and Country B can execute a reciprocal point trade with certainty that both of their scorecards will be integrated into the final trimmed-mean calculation, and if carefully calibrated, a reasonable chance of not being eliminated by the trimmed mean. However, under a stochastic sub-panel model where each judge's card has, for example, a 60% probability of being randomly selected post-performance, the joint probability of both colluding judges having their scores active drops to just 36%. Crucially, this randomized barrier also neutralizes the solitary biased actor. A home-country judge attempting to unilaterally inflate a compatriot or execute a highly precise point-spread adjustment to secure a domestic hierarchy suddenly faces a 40% probability of outright exclusion. By decoupling the act of evaluating from the certainty of scoring, the mathematical viability of both coordinated gamesmanship and targeted unilateral bias is effectively dismantled before the scores are even submitted.

While the random exclusion of judges' scores has been previously proposed as a potential anti-bias measure (Kirkbride, 2013; McFee, 2013), there is an important trade-off to this stochastic design. Introducing random sampling error into the officiating panel may increase scoreboard volatility, making skater placements highly sensitive to the random draw of the sub-panel. Specifically, Emerson (2007) and Emerson and Arnold (2011) warn that if a skater draws a particularly strict or lenient subset of judges by chance, their placement could be disproportionately altered irrespective of their actual performance. Econometrically, there exists a mathematical "sweet spot", an equilibrium where the overall panel size and the randomized subset size are optimized to simultaneously maximize the deterrence of deliberate manipulation while minimizing random sampling variance. Exploring this optimization frontier through predictive modeling, historical re-scoring, and data simulations represents a promising avenue for our future research.

11.4 The Technical Panel: Human Sensory Limits and the Path to Automated Officiating

In theory, the division of labor between the judging panel and the technical panel suggests a clear dichotomy in subjectivity. Under econometric and game-theoretic principles, the duties of the primary judging panel, such as evaluating the aesthetic air position of a jump, or the creative design of a program, or determining if a step sequence expresses the character of the music, are inherently interpretive. Because these tasks have low verifiability, they represent highly fertile ground for nationalistic bias. Conversely, the duties of the technical panel appear objective and physics-based: a skater is either on an inside edge or an outside edge; they either completed three full rotations in the air or they did not; and a spin or death spiral either completed the required number of continuous rotations or fell short. Econometric theory predicts that because these technical calls are highly verifiable, they should be significantly less prone to evaluator bias.

My empirical results, however, do not bear this theory out. The magnitude and consequences of technical panel bias are remarkably high. Through the Expected Value (EV) model, I identified 12 unique technical specialists who significantly alter their baseline infraction calling or difficulty-level assignments when evaluating compatriots (Figure 13). Correcting for these technical deviations alone alters the final standings in 39.3% of competitions where a compatriot specialist is seated and alters the physical makeup of the podium in 11.9% of those events (Table 10a).

When examining the specific metrics of the EV analysis, the magnitude of the individual deviations is surprisingly large. Rather than minor, marginal adjustments, I found that technical specialists systematically shift their evaluations by unexpected margins in both generous and punitive directions. For instance, the Infraction Shielding Effect spans a massive point differential, ranging from Norwegian specialist Christine Isaksen systematically saving compatriots +9.80 points on expected penalties, to an incredibly punitive -14.93 points additionally docked by Romanian specialist Grety Marton. Concurrently, Level Shielding spans a range of approximately two points, ranging from Marton's downgrade of -0.70 points to United States specialist Judy Blumberg's positive level boost of +1.29 points on compatriots' non-jump elements (Figure 13).

This surprising level of deviation from average behavior does not imply that technical specialists are inherently more biased or malicious than judges. Indeed, many specialists frequently cross-train and serve in both roles. Rather, these findings expose an under-appreciated inefficiency, namely the physical limits of human sensory perception. High-speed, multi-rotational athletic movements are exceptionally difficult to evaluate accurately in real-time. Detecting whether a blade takeoff occurred on a flat or a true outside edge or identifying a quarter-under-rotation on a quadruple jump rotating at approximately 300 rotations per minute, requires an ideal camera angle, precise slow-motion capture of where the blade first contacts the ice, and absolute visual clarity.

Here, I highlight a stark contrast between the analyses of nationalistic premiums observed for individual judges and those of individual technical specialists. In the case of the judging panels, the empirical evidence is clear that that statistically significant examples of an official evaluating their own skaters punitively are extremely rare (2/159). For the technical specialists, the situation is fundamentally different, where every possible permutation of statistically significant generosity and strictness toward compatriots is observed in the data. A plausible interpretation of this result is that technical specialists simply lack the visual information required to make these supposedly deterministic evaluations any more accurately than an educated guess.

To resolve this vulnerability, the sport could remove high-speed biomechanical measurements from human evaluation. This recommendation echoes recent calls within the literature for technological interventions as summarized by Braeunig (2025). Electronic scoring and automated recognition systems have been increasingly proposed as essential tools to enhance accuracy and alleviate the cognitive demands placed on human evaluators (Allen et al., 2021; Díaz-Pereira et al., 2014; Sato & Hopper, 2021). The International Skating Union has recently begun exploring this frontier through its ongoing development of an AI-driven Judging Support System (JSS) to assist with technical calling (International Skating Union, 2023). The present research provides robust statistical support for accelerating this technological transition. By delegating strictly physics-based determinations to automated tracking systems, the sport can eliminate the visual ambiguity that currently shields nationalistic bias. Automating these high-speed technical calls would establish a truly objective baseline, relieving human officials of sensory demands they cannot realistically meet and ensuring a level playing field for the athletes.

12. Conclusions

This study has presented comprehensive analysis of elite-level figure skating evaluation across the complete 2022–2026 Olympic cycle. By leveraging a fully de-anonymized dataset, I introduced a series of baseline-controlled models designed to separate true nationalistic bias from general evaluator variance. The findings demonstrate that subjective judging remains highly sensitive to strategic adjustments under the principle of verifiability. Strikingly, the deterministic gatekeeping of the technical panel, previously unexamined in the literature, represents an equally potent and unmitigated vector of outcome distortion. Through counterfactual simulations, I mapped how these dual-panel biases bypass standard scoring filters to alter competitive standings in over 40% of competitions.

Ultimately, the objective of this research is not to assign blame or focus scrutiny on individual actors, but to provide a clear, data-driven diagnostic of the sport's institutional and structural vulnerabilities. In subjectively judged sports, maintaining competitive integrity requires continuous, proactive self-monitoring and a willingness to adapt scoring mechanics to the strategic behavior of the evaluators. It is my sincere hope that these findings find their way to the International Skating Union, national federations, officials, coaches, and the athletes themselves. By empowering the global figure skating community with objective statistical diagnostics, this research seeks to catalyze a collaborative push toward greater transparency. These insights can inform future policy reforms and scoring system designs, helping figure skating ultimately achieve the truly fair and objective playing field that its dedicated athletes and fans deserve.

13. Data Availability

The supplementary data tables referenced in this manuscript are hosted and available without restriction at https://phenologic.net/assets/FS_Bias_Supplementary_Tables.zip. The raw data underlying this study consists of official competition protocols published by the International Skating Union (ISU). However, the aggregated, row-level dataset utilized for this econometric analysis was compiled using proprietary scraping and parsing pipelines developed by PhenoLogic Co. and is therefore not publicly available.

14. Declaration of Competing Interests

The author is affiliated with PhenoLogic Co., a private data analytics firm. PhenoLogic Co. holds no financial contracts or consulting agreements with the International Skating Union (ISU) or any national figure skating federations. The author declares no personal relationships with any officials, athletes, or governing bodies evaluated in this manuscript. The views and findings expressed are strictly those of the author based on mathematical modeling.

15. Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. It was conducted as independent research by PhenoLogic Co.

16. Declaration of Generative AI in Research Methodology

During the data collection phase of this research, the author used the Google GenAI SDK (Gemini-3.1-Flash-Lite) solely for the purpose of parsing and extracting structured data from unstructured ISU protocol PDF documents. After using this tool, the author conducted mathematical reconciliations and manual verifications to ensure data integrity. No generative AI technologies were used in the original drafting or conceptualization of this manuscript, nor in the generation of the statistical conclusions. Gemini 3.1 Pro was used exclusively for proofreading and final editing of the manuscript.

17. Acknowledgements

The author wishes to thank Jeroen Prins for his valuable feedback on early drafts of this manuscript, and Abraham Wyner for discussions regarding the econometric modeling.

References

P.A. Brauenig. 2025. “National bias in international sports judging: a scoping review”. *Cogent Social Sciences* 11 (1): https://doi.org/10.1080/23311886.2025.2594841.

Ansorge, C. J., and J. K. Scheer. 1988. “International bias detected in judging gymnastic competition at the 1984 Olympic Games." *Research Quarterly for Exercise and Sport*, *59* (2), 103–107. https://doi.org/10.1080/02701367.1988.1 0605486.

Heiniger, S., and Mercier, H. (2019). “*National bias of International Gymnastics Judges during the 2013-2016 Olympic cycle”*. ArXiv. https://arxiv.org/abs/1807.10033.

Heiniger, S., and Mercier, H. (2021). “Judging the judges: Evaluating the accuracy and national bias of international gymnastics judges”. *Journal of Quantitative Analysis in Sports*, *17*(4), 289–305. https://doi.org/10.1515/jqas-2019-0113

Sandberg, A. (2018). “Competing identities: A field study of in-group bias among professional evaluators”. *Economic Journal*, *128* (613), 2131–2159. https://doi.org/10.1111/ecoj.12513

Wolframm, I. (2023). “Let them be the judge of that: Bias cascade in elite dressage judging”. *Animals: An Open Access Journal from MDPI*, *13* (17), 2797. https://doi.org/10.3390/ani13172797

Emerson, J. W., Seltzer, M., and Lin, D. (2009). “Assessing judging bias: An example from the 2000 Olympic games”. *American Statistician*, *63* (2), 124–131. https://doi.org/10.1198/tast.2009.0026

Emerson, J. W., and Meredith, S. (2011). “Nationalistic judging bias in the 2000 Olympic diving competition”. *Math Horizons*, *18* (3), 8–11. https://doi.org/10.4169/194762111X12954578042812

Krumer, A., Otto, F., and Pawlowski, T. (2022). “Nationalistic bias among international experts: Evidence from professional ski jumping”. *Scandinavian Journal of Economics*, *124* (1), 278–300. https://doi.org/10.1111/sjoe.12451

Zitzewitz, E. 2006. "Nationalistic Bias in the 2002 Olympic Figure Skating Judging." *Journal of Economic Perspectives* 20 (2): 209–222. https://doi.org/10.1257/jep.20.2.209.

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review* 101 (4): 1410–35**.** https://doi.org/10.1257/aer.101.4.1410

Lee, J. (2008). “Outlier aversion in subjective evaluation: Evidence from world figure skating championships”. *Journal of Sports Economics*, *9*(2), 141–159. https://doi.org/10.1177/1527002507299203

Prendergast, C. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37 (1): 7–63. https://doi.org/10.1257/jel.37.1.7.

Prendergast, C. 2002. "The Uncertainty of Subjective Judgments." *American Economic Review* 92 (2): 215–221. https://doi.org/10.1257/000282802320189151.

Campbell, B., and Galbraith, J. (1996). “Nonparametric tests of the unbiasedness of Olympic figure-skating judgments”. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *45* (4), 521–526. https://doi.org/10.2307/2988550

Duggan, M., and S. D. Levitt. 2002. "Winning Isn't Everything: Corruption in Sumo Wrestling." *American Economic Review* 92 (5): 1594–1605. https://doi.org/10.1257/000282802762024665.

Emerson, J. W., and Arnold, T. B. 2011. “Statistical sleuthing by leveraging human nature: A study of Olympic figure skating”. American Statistician, 65 (3), 143–148. https://doi.org/10.1198/tast.2011.10165

Templon, J. 2018. "Top-Level Figure Skating Judges Consistently Favor Skaters From Their Home Countries. Now Many Of Those Judges Are At The Olympics” *BuzzFeed News*, February 19, 2018. https://www.buzzfeednews.com/article/johntemplon/the-edge

Zitzewitz, E. (2014). “Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging”. *Journal of Sports Economics*, *15*(1), 3–30. https://doi.org/10.1177/1527002512441479

Courty, P., and G. Marschke. 2004. "An Empirical Investigation of Gaming Responses to Explicit Performance Incentives." *Journal of Labor Economics* 22 (1): 23–49. https://doi.org/10.1086/380612.

Lyngstad, T. H., Härkönen, J., & Rønneberg, L. T. S. (2020). “Nationalistic bias in sport performance evaluations: An example from the ski jumping world cup”. *European Journal for Sport and Society*, *17*(3), 250–264. https://doi.org/1 0.1080/16138171.2020.1792628.

Plessner, H., and Haar, T. (2006). “Sports performance judgments from a social cognitive perspective”. *Psychology of Sport and Exercise*, *7*(6), 555–575. https://doi.org/10.1016/j.psychsport.2006.03.007.

Ste-Marie, D. M. (1996). “International bias in gymnastic judging: Conscious or unconscious influences?” *Perceptual and Motor Skills*, *83*(3), 963–975. https://doi.org/10.2466/pms.1996.83.3.963.

Ste-Marie, D. M., & Lee, T. D. (1991). “Prior processing effect on gymnastic judging.” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 126–136. https://doi.org/10.1037/0278-7393.17.1.126

Clerides, S., and T. Stengos. 2012. "Love Thy Neighbor, Love Thy Neighbor's Music: National Bias in the Eurovision Song Contest." *Journal of Cultural Economics* 36 (2): 125–151. https://doi.org/10.1007/s10824-012-9162-4.

Fenn, D. J., M. A. Porter, P. J. Mucha, M. J. McDonald, and N. S. Jones. 2006. "How the Eurovision Song Contest Works." *Physica A: Statistical Mechanics and its Applications* 360 (2): 576–598. https://doi.org/10.1016/j.physa.2005.06.050.

International Skating Union. 2025a. "Decision of the ISU Disciplinary Commission: Case No. 2024-02 (Yuri Balkov)." Lausanne, Switzerland: International Skating Union, March 2025. https://www.isu.org/inside-isu/rules-regulations/disciplinary-commission-decisions.

International Skating Union. 2025b. "Decision of the ISU Disciplinary Commission: Case No. 2025-02 (Dr. Katalin Balczo)." Lausanne, Switzerland: International Skating Union, June 2025. https://www.isu.org/inside-isu/rules-regulations/disciplinary-commission-decisions.

Brennan, C. 2014. "Sochi Olympics: Figure Skating Judging Scandals Demand Serious Reform." *USA Today*, February 21, 2014. https://www.usatoday.com/story/sports/columnist/brennan/2014/02/20/sochi-olympics-figure-skating-judging-scandal-adelina-sotnikova-yuna-kim/5657801/.

Hersh, P. 2026. "Controversy Clouds Olympic Ice Dance Scoring as French Duet Edges Americans." *NBC Sports*, February 16, 2026. https://www.nbcsports.com/olympics/figure-skating/2026-milan-ice-dance-scoring-controversy.

McCarvel, N. 2026. "Scrutiny Falls on French Judge's Scorecard in Tight Olympic Ice Dance Battle." *USA Today*, February 17, 2026. https://www.usatoday.com/story/sports/olympics/2026/02/17/milan-olympics-figure-skating-ice-dance-judging-controversy/5657801/.

Fenwick, I., and Chatterjee, S. (1981). "Perception, preference, and patriotism: An exploratory analysis of the 1980 winter olympics". *American Statistician*, *35* (3), 170–173. https://doi.org/10.2307/2683998

Garicano, L., I. Palacios-Huerta, and C. Prendergast. 2005. "Favoritism under Social Pressure." *Review of Economics and Statistics* 87 (2): 208-216. http://www.mitpressjournals.org/doi/pdf/10.1162/0034653053970267

Kirkbride, A. N. (2013). Scoring/judging applications. In T. McGarry, P. O'Donoghue, & J. Sampaio (Eds.). *Routledge handbook of sports performance analysis* (pp. 140–152). Routledge.

McFee, G. (2013). "Officiating in aesthetic sports". Journal of the Philosophy of Sport, 40 (1), 1–17. https://doi.org/10.1080/00948705.2012.725910

Emerson, J. W. (2007). "Chance, on and off the ice". *CHANCE*, *20* (2), 19–21. https://doi.org/10.1080/09332480.2007.10722841

Allen, E., Fenton, A., & Parry, K. (2021). "Computerized gymnastics judging scoring system implementation – An exploration of stakeholders' perceptions". *Science of Gymnastics Journal*, *13* (3), 357–370. https://doi.org/10.52165/ sgj.13.3.357-370

Díaz-Pereira, M. P., Gómez-Conde, I., Escalona, M., & Olivieri, D. N. (2014). "Automatic recognition and scoring of olympic rhythmic gymnastic movements". *Human Movement Science*, *34*, 63–80. https://doi.org/10.1016/j.humov.2014.01.001

Sato, N., & Hopper, L. S. (2021). "Judges' evaluation reliability changes between identifiable and anonymous performance of hip-hop dance movements". *PloS One*, *16*(1), e0245861. https://doi.org/10.1371/journal.pone.0245861.

International Skating Union. (2023). *ISU and Fujitsu agree to continue the development of the Judging Support System.* Lausanne, Switzerland: International Skating Union.