

# Pairwise, Style-Controlled, Nonparametric Tests for Judge Bias under the ISU Judging System

Michael Allman

[michael.allman1960@gmail.com](mailto:michael.allman1960@gmail.com)

Independent Researcher

---

## Article

**Keywords:** subjective evaluation, judging bias, permutation tests, pairwise comparisons, trimmed mean, figure skating, ISU judging system

**Posted Date:** October 23rd, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-8991583/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Pairwise, Style-Controlled, Nonparametric Tests for Judge Bias under the ISU Judging System

---

Michael Allman  
Independent Researcher

michael.allman1960@gmail.com

## Abstract

Subjectively judged sports require statistical tools that distinguish competitor-specific judge bias from benign judge style differences (leniency/harshness, scale usage, and variance). This paper develops a pairwise, event-local testing framework for the International Skating Union (ISU) Judging System (IJS) that (i) produces an interpretable bias effect size in points for every judge and every pair of competitors, (ii) respects the IJS trimmed-mean scoring mechanism at the element and program-component level, and (iii) assigns nonparametric p-values via permutation tests under null models that retain judge-level scoring behavior. The primary test statistic is a judge's pairwise margin-shift: the change in the published score difference between two competitors attributable to the judge's marks, relative to a median-of-others neutralization. To avoid conflating bias with style, we introduce a residual-label (delta-exchange) permutation null that preserves each judge's realized scoring differentials while permuting their assignment between competitors, removing competitor-specific associations while retaining judge-level scoring behavior. We illustrate the method on the Olympic Winter Games 2026 Ice Dance Free Dance segment using the official 'Judges Details per Skater' sheet. Applying the method to 142 events across 17 competitions (January 2022–February 2026), we identify 20,213 significant judge  $\times$  competitor-pair results at BH  $q \leq 0.05$ . Nine events satisfy a two-part outcome-determinative criterion (significant bias exceeds the official gold–silver margin); the OWG 2026 Ice Dance Free Dance is the most prominent.

Keywords: subjective evaluation; judging bias; permutation tests; pairwise comparisons; trimmed mean; figure skating; ISU judging system

## 1. Introduction

In sports where outcomes depend on human evaluation, the credibility of results hinges on the integrity of judging. Even with formal scoring criteria, individual judges may differ systematically: some are generally lenient, others severe; some avoid outlying marks; others use the full scale. These differences in 'style' need not imply bias toward any particular competitor. The central inferential problem is therefore

not merely whether judges differ, but whether a specific judge's differences are competitor-specific in a way that plausibly changes comparative outcomes.

This paper develops an event-local, pairwise framework for detecting competitor-specific judge bias under the ISU Judging System (IJS). The framework is designed to answer questions of the form: 'Is judge  $j$  unusually favorable to competitor A compared to competitor B, given the rest of the panel?' The output is both an effect size in points (a margin shift) and a nonparametric  $p$ -value that accounts for the competition's scoring mechanics, including trimmed means.

The methodological emphasis is nonparametric: we avoid distributional assumptions about judging noise and instead use permutation tests grounded in label-exchangeability under appropriately specified null hypotheses. A key contribution is the construction of null models that retain judge-level scoring behavior, so that the test targets competitor-specific differential rather than global leniency/harshness or variance.

## 2. Background and related work

The empirical study of bias in judged sports has a substantial literature, including work on nationalistic bias, collusion and vote trading, transparency reforms, and outlier aversion (e.g., Campbell and Galbraith, 1996; Zitzewitz, 2006; Lee, 2008). Most approaches estimate bias using regression or panel models across many events, or by comparing judges' marks to some benchmark such as the panel average. The present work differs in three respects: (i) it produces pairwise competitor-level inference within a single event; (ii) it expresses bias in the native unit of the scoring system (points of margin shift); and (iii) it uses nonparametric permutation inference under style-preserving nulls rather than parametric regression.

Campbell and Galbraith (1996) applied chi-squared tests to judge-level score symmetry, detecting departures from impartial judging at the event level without attributing effects to specific judge-competitor pairs. Zitzewitz (2006, 2014) estimated national-group favoritism using regression across many events, yielding cross-event averages rather than event-local inference. Lee (2008) documented outlier aversion — judges' tendency to avoid extreme marks — a style phenomenon our null explicitly preserves via distributional control, so that residual competitor-specific differentials are not confounded with scale-avoidance behavior. More recently, Braeunig (2025a) finds pervasive patriotic bias among IJS-era figure skating judges across a broad dataset, and Braeunig (2025b) provides a PRISMA-based scoping review of 68 studies documenting nationalistic bias across judged sports.

The ISU's public description of the judging system states: "Out of these scores, the highest and lowest score of each element or program component are ignored and the average will be taken from the remainder, generating the trimmed mean (average score)." This trimming rule is central; an appropriate bias statistic should measure a judge's impact through the trimming mechanism rather than on raw marks alone.

## 3. ISU scoring mechanics and data

3.1. Scoring overview. In IJS, the total segment score is the sum of a Technical Element Score (TES), a Program Component Score (PCS), and deductions. TES is built from element base values (BV) and

Grades of Execution (GOE). Each element has a multiplier  $m_r$  that converts a GOE grade into points. PCS is built from component marks multiplied by component factors. In the Olympic cycle relevant to this paper, panels typically include nine judges; for each element and program component, the highest and lowest marks are discarded and the average of the remaining seven is used (trimmed mean).

3.2. Data source. Our unit of analysis is the ISU ‘Judges Details per Skater’ sheet, which provides judge-by-judge marks for each executed element (GOE grades) and each program component (PCS marks). The worked example in this paper uses the official Olympic Winter Games 2026 Ice Dance Free Dance segment sheet (Milano–Cortina), parsed directly from the published spreadsheet.

3.3. Multi-event database. The broader research program maintains an SQLite database of ISU scoring sheets (2022–2026) spanning 17 competitions across five seasons (2021/22–2025/26) and 144 events in total — 142 analyzed events and 2 excluded events (the ISU Four Continents Championships 2022 Ice Dance Rhythm Dance and Free Dance, which used non-standard 7-judge panels). The database contains 206,682 individual GOE grades and 84,922 individual PCS marks (291,604 total judge marks) from 2,686 competitor entries. Table B provides a complete breakdown; aggregate results are reported in Section 9.

Table B. Database characteristics.

Note: ‘Panel judge-event positions’ counts each judge’s assignment to one event as one entry (one position per judge per event); the same individual may hold positions in multiple events across the database.

| Characteristic                  | Count / Value     |
|---------------------------------|-------------------|
| Seasons                         | 2021/22 – 2025/26 |
| ISU competitions                | 17                |
| Events (total)                  | 144               |
| Main discipline events analyzed | 134               |
| Team event segments analyzed    | 8                 |
| Excluded (7-judge panels)       | 2                 |
| Competitor entries              | 2,686             |
| Panel judge-event positions     | 1,288             |
| Individual judge marks          | 291,604           |
| Pairwise permutation tests (v2) | 271,728           |

## 4. Problem statement and estimands

Fix a single event segment with a panel of  $J$  judges (typically  $J=9$ ) and a set of competitors  $T=1,\dots,N$ . For each competitor  $T$  and each scoring row  $r$  (an element row or PCS-component row), judge  $j$  provides a mark  $x_{\{j,r,T\}}$ . The published scoring rules map the vector of marks for row  $r$  into a points contribution via trimming, scaling, and (for elements) addition of  $BV$ .

We seek, for each judge  $j$  and each unordered competitor pair  $\{A,B\}$ , an interpretable effect size measuring whether judge  $j$  shifts the published margin between  $A$  and  $B$  relative to the rest of the panel — with direction indicated by the sign of  $B_j(A,B)$  — along with a nonparametric  $p$ -value for

extremeness under a null hypothesis of no competitor-specific differential. Throughout this paper, "bias" refers strictly to competitor-specific differential impact — the change in the published score margin attributable to judge  $j$ , conditional on the scoring rules — and carries no implication of intent. Throughout this paper, inference is event-local: p-values and q-values are computed within each event independently, and database-wide figures (Section 9) are aggregates across these independent event-local results.

## 5. Bias statistic: ISU-impact margin shift

Implementation note. In the GOE case, the base value  $BV_r$  cancels in the delta, yielding  $\Delta(j,r,T) = m_r \times (\text{trimmean}(\text{actual}) - \text{trimmean}(\text{neutralized}))$ ; only the judge's marginal contribution through the trimmed-mean mechanism matters. Computations use the raw marks as reported in the judging sheet (integer GOE grades; PCS marks in their native increment), not the ISU-published rounded segment totals. Median-of-others neutralization is preferred over mean-of-others for robustness: the median is less sensitive to any single extreme mark from another judge, and its use aligns with the trimming ethos of the IJS mechanism itself (which discards extreme marks for the same reason).

5.1. Row scoring function. Each scoring row  $r$  has a panel of  $J$  marks  $x_1, \dots, x_J$  and an ISU scaling factor. For element rows (GOE), the scaling factor is the GOE multiplier  $m_r$  and the row has a base value  $BV_r$ . For program component rows (PCS), the scaling factor is the component factor  $f_r$ . Let  $\text{trimmean}(\cdot)$  denote the mean of the middle  $J-2$  marks after discarding the single highest and single lowest.

Define the row points function  $F_r$  as:

- GOE row:  $F_r(x_1, \dots, x_J) = BV_r + \text{trimmean}(x_1, \dots, x_J) * m_r$
- PCS row:  $F_r(x_1, \dots, x_J) = \text{trimmean}(x_1, \dots, x_J) * f_r$

5.2. Median-of-others neutralization. For a focal judge  $j$  on row  $r$ , define a neutralized mark by replacing  $x_j$  with the median of the other  $J-1$  marks (for  $J=9$ , median-of-8 is the average of the 4th and 5th order statistics; in general, the neutralized mark may be non-integer when  $J-1$  is even). For  $J=8$ , the median-of-7 is the 4th order statistic; in general, for a panel of  $J$  judges, the peer median is computed from  $J-1$  values using the standard definition. Let  $x'$  denote the neutralized vector. The row-level ISU-impact delta is:

$$\Delta(j,r,T) = F_r(x \text{ for competitor } T) - F_r(x' \text{ for competitor } T).$$

This delta measures the points moved by judge  $j$  through the published trimmed-mean mechanism, relative to a neutral mark.

5.3. Competitor impact and pairwise margin shift. Aggregate deltas across rows for competitor  $T$ :

$$I_j(T) = \text{sum over rows } r \text{ in } T \text{ of } \Delta(j,r,T).$$

Define the pairwise bias (margin shift) for competitors  $A$  and  $B$  as:

$$B_j(A,B) = I_j(A) - I_j(B).$$

Positive values indicate that judge  $j$  shifts the  $A-B$  published margin toward  $A$  (relative to neutralization); negative values shift toward  $B$ . The statistic is expressed in points of margin shift.

5.4. Pairwise vote mapping. For compatibility with pairwise-comparison frameworks, define a binary vote

$v_j(A,B)=\text{sign}(B_j(A,B))$ , with ties treated as 0. This provides an ordinal reduction of the point-based statistic, useful for comparison with Condorcet-style pairwise-preference methods.

## 6. Null models and permutation inference

6.1. Why style control is required. A judge may be systematically high/low or more/less variable across all competitors. A permutation null that treats judges as identical will flag such style differences as bias, even when they are not competitor-specific. Therefore, the null must preserve judge style.

6.2. Residual-label (delta-exchange) permutation null. For judge  $j$  and competitor pair  $(A,B)$ , let  $d_A = [\Delta_{\{j,r\}} : r \in \text{rows}(A)]$  and  $d_B = [\Delta_{\{j,r\}} : r \in \text{rows}(B)]$  denote judge  $j$ 's delta values for the two entries. Under the null of no directional bias, these  $|A| + |B|$  delta values are exchangeable across the two entries: median-of-8 neutralization substantially attenuates the shared competitor-quality signal row by row, and is intended to leave primarily judge  $j$ 's idiosyncratic deviation. Formally, the exchangeability assumption is that the pooled  $\Delta$  multiset  $\{d_A \cup d_B\}$  is invariant to relabeling rows as belonging to  $A$  or  $B$  under the null. For each of  $M = 10,000$  permutations (seed = 20260223, `numpy.random.default_rng`): pool  $d_A$  and  $d_B$ , randomly split into groups of size  $|A|$  and  $|B|$ , compute  $B_j^{\{\text{perm}\}}(A, B) = \Sigma(\text{group } A) - \Sigma(\text{group } B)$ . The two-sided Monte Carlo p-value is  $p = (1 + \#\{B_j^{\{\text{perm}\}} \geq |B_j^{\{\text{obs}\}}|\}) / (M + 1)$ . This follows Emerson, Seltzer & Lin (2009) and corrects the exchangeability violation in a prior quantile-based null (see Appendix A). Note:  $\Delta$  values from all row types (GOE elements and PCS components) are pooled in the permutation; because each  $\Delta$  is already expressed in points after ISU scaling (GOE factor applied; PCS components scaled to their contribution), pooling treats each row as a points-valued contribution on a common scale. Stratification by row category is a robustness check deferred to future work. While  $\Delta$  values within an entry need not be independent (elements within a program share a competitor-quality context), the residual-label test is exact conditional on the exchangeability of the pooled  $\Delta$  multiset under relabeling. Within-entry dependence is permissible provided the dependence structure is symmetric across  $A$  and  $B$ ; if dependence differs systematically between  $A$  and  $B$  (e.g., due to differing row-type mixes or element leverage), exchangeability may be violated; see Appendix B for step-by-step pseudo-code.

**6.3. Alternative null models considered.** We evaluated four alternative null models before adopting the residual-label permutation test described in Section 6.2. The central methodological challenge is that standard judge-bias tools test for global severity differences — whether a judge scores consistently higher or lower than peers across the full field — whereas the target estimand is directional differential treatment: does judge  $j$  score country  $A$  systematically higher relative to country  $B$ , relative to panel consensus? These are distinct questions requiring distinct instruments.

**Table A. Alternative null models evaluated.**

| Method                                                                       | Tests for                                                      | Key limitation                                                                                                                         | Status   |
|------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|----------|
| Exact combinatorial $B(j)$                                                   | Pairwise judge deviation, all $C(2k,k)$ element splits         | Exchangeability violation: permuting elements across competitors ignores the shared skater-quality signal, inflating the null variance | Rejected |
| Style-preserving quantile permutation ( <code>isuimpact_quantile_v1</code> ) | Pairwise deviation with judge-style control via empirical CDFs | Same exchangeability violation survives style mapping: skater-quality differences create non-                                          | Rejected |

|                                                                         |                                                                                                           |                                                                                                                                                                                                                                                                                              |                  |
|-------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
|                                                                         |                                                                                                           | exchangeable pools across entries                                                                                                                                                                                                                                                            |                  |
| Friedman rank test + Nemenyi post-hoc                                   | Global judge severity differences across all competitors                                                  | Tests the wrong question: detects whether a judge scores uniformly high/low, not directional national bias. J1 not significant ( $p > 0.10$ ) under Nemenyi despite confirmed OWG finding                                                                                                    | Rejected         |
| Many-Facet Rasch Measurement (MFRM)                                     | Rater severity and fit statistics under a latent-trait model                                              | Requires calibration data not publicly available; assumes a unidimensional latent trait that may not hold across heterogeneous element types                                                                                                                                                 | Not applicable   |
| Residual-label permutation (isuimpact_residual_v1; Emerson et al. 2009) | Directional differential treatment: judge j's residuals (deviation from panel consensus) for entry A vs B | No violations identified in diagnostics to date; exchangeability is plausible given median-of-others neutralization substantially attenuates the shared competitor-quality signal. Exchangeability remains an assumption and is confirmed by upper-tail calibration (Figure 2, Section 9.5). | <b>Adopted ✓</b> |

The Friedman test failure is instructive: applied to the OWG 2026 Ice Dance Free Dance panel, it yields  $\chi^2(8) = 69.98$ ,  $p = 4.96 \times 10^{-12}$ , confirming that judge effects exist globally. Yet the Nemenyi post-hoc test finds J1 not significant (all pairwise adjusted  $p > 0.10$ ), because the Nemenyi test asks whether J1 scores uniformly higher or lower than peers across all competitors — not whether J1 differentially favours one specific nation over another. The residual-label permutation addresses the correct null: given judge j's residuals from panel consensus, are they exchangeable across entries A and B? For J1 in this event, the answer is no ( $p = 0.0003$ , BH  $q = 0.012$ ).

## 7. Effect size and outcome-determinative criterion

$B_j(A,B)$  is expressed in the same units as the published score difference between competitors — points of margin shift. This makes the statistic directly interpretable: a bias of +1.19 points means judge j shifted the A–B published margin by 1.19 points in A's favor, relative to what the panel would have produced with a neutral mark from j.

This unit-consistency enables a natural outcome-determinative criterion: an event is classified as outcome-determinative if a judge j satisfies both (i)  $q \leq 0.05$  under BH-FDR correction within the event, and (ii)  $B_j(A,B) > \text{margin}(A,B)$ , where  $\text{margin}(A,B)$  is the official published segment total difference between A and B (ISU-rounded; the bias  $B_j$  is computed on raw unrounded marks, so differences are small (on the order of hundredths of a point, due to ISU rounding to two decimal places)). Deductions (e.g., falls) are included in the published segment total but are not judge marks; they are unchanged by any judge neutralization, and thus affect the published margin but not the counterfactual adjustment  $B_j$ . The second condition is directional — only bias in favor of the gold medalist can flip the result — so the absolute value is not taken. Events meeting both conditions are those where the pairwise neutralization counterfactual for the gold–silver pair would have changed the winner. Results for all 142 events are reported in Section 9.3.

## 8. Worked example: Olympic Winter Games 2026 Ice Dance Free Dance

We apply the full method to the Olympic Winter Games 2026 (Milano–Cortina) Ice Dance Free Dance segment using the official ‘Judges Details per Skater’ sheet. The segment contains  $N=20$  teams and  $J=9$  judges, yielding 1,710 judge $\times$ pair tests. We compute  $B_j(A,B)$  using both element (GOE) rows and PCS rows, and we estimate p-values using  $M=10,000$  residual-label (delta-exchange) permutations.

Illustrative pair (Gold vs Silver). Let A be the 1st-place team (Fournier Beaudry / Cizeron) and B be the 2nd-place team (Chock / Bates). Under the residual-label (delta-exchange) permutation null, judge J1 shifts the published A–B margin toward A by +1.19 points ( $B_{\{J1\}}(A,B)=+1.190$ ), with  $p=0.0003$  and  $q=0.012$  (BH-FDR within-event). Judge J5 (Janis Engel, GER) also yields a significant margin shift for this pair in the opposite direction ( $B_{\{J5\}}(A,B)=-0.770$ ,  $q=0.035$ ), reflecting a shift toward B (silver); this does not satisfy the directional outcome-determinative criterion and would widen the A–B margin if J5 were removed. No other judge yields  $q \leq 0.05$  for this pair.

Event-level patterns. Across all 190 competitor pairs, the number of FDR-significant pairwise differentials ( $q \leq 0.05$ ) varies by judge in this segment (e.g., J3: 40, J9: 39, J6: 38, J4: 36, J8: 35, J1: 32, J5: 28, J2: 23, J7: 11 (282 of 1,710 pairs, 16.5%); all counts from the residual-label null (isuimpact\_residual\_v1),  $M=10,000$ , seed=20260223). These event-local counts are descriptive and intended to motivate database-scale stability and calibration analyses (Section 9).

Table 1. Gold vs Silver pairwise margin shift by judge (GOE+PCS; residual-label permutation null).

| Judge                                      | BiasPoints<br>$B_j(A,B)$ | Vote | p-value | q-value (BH) |
|--------------------------------------------|--------------------------|------|---------|--------------|
| J1 — Jezabel<br>DABOUI (FRA)               | +1.190                   | A    | 0.0003  | 0.012*       |
| J2 — Elena<br>KHYMYZENKO<br>(UKR)          | −0.240                   | B    | 0.4010  | 0.590        |
| J3 — Marta<br>OLOZAGARRE<br>(ESP)          | +0.710                   | A    | 0.0264  | 0.101        |
| J4 — Virpi<br>KUNNAS-<br>HELMINEN<br>(FIN) | +0.150                   | A    | 0.6410  | 0.778        |
| J5 — Janis<br>ENGEL (GER)                  | −0.770                   | B    | 0.0045  | 0.035*       |
| J6 — Leslie<br>KEEN (USA)                  | +0.020                   | A    | 0.9626  | 0.984        |
| J7 — Isabella<br>MICHELI (ITA)             | −0.130                   | B    | 0.7028  | 0.818        |
| J8 — Feng<br>HUANG (CHN)                   | −0.470                   | B    | 0.1631  | 0.322        |

| Judge                        | BiasPoints<br>B <sub>j</sub> (A,B) | Vote | p-value | q-value (BH) |
|------------------------------|------------------------------------|------|---------|--------------|
| J9 — Richard<br>KOSINA (CZE) | +0.420                             | A    | 0.1465  | 0.304        |

Note: Positive BiasPoints indicate a margin shift toward the 1st-place team. Values are computed via median-of-others neutralization and ISU trimmed means; p-values are two-sided Monte Carlo permutation p-values with M=10,000.

J1 nationality pattern. Table C presents J1's estimated impact on every team's total segment score in this event. J1 (Jezabel Dabouis, FRA) provided a net positive adjustment to both French-represented teams (+0.26 pts for gold-medal team Fournier Beaudry / Cizeron; +0.08 pts for Lopareva / Brissaud), while all three USA-represented teams received negative adjustments (−0.93 pts for silver-medal Chock / Bates; −0.43 pts for Zingas / Kolesnik; −0.34 pts for Carreira / Ponomarenko). The directional asymmetry between FRA and USA entries is consistent with the significant BiasPoints detected for the gold–silver pair ( $B_{\{J1\}}(FRA, USA) = +1.19, q = 0.012$ ). We did not observe similarly unidirectional patterns among other NOC groupings in this segment. Decomposing J1's net +1.19 pt margin shift by scoring component: element (GOE) rows account for approximately +1.01 pts and Program Component Score (PCS) rows account for approximately +0.20 pts (FRA contribution: +0.16 from GOE, +0.10 from PCS; USA contribution: −0.85 from GOE, −0.10 from PCS; minor discreteness effects from GOE integer grades in the neutralization step).

Table C. J1 estimated impact on each team's TSS (OWG 2026 Ice Dance Free Dance).

| Final rank | Team                       | NOC | J1 impact (pts) |
|------------|----------------------------|-----|-----------------|
| 1          | Fournier Beaudry / Cizeron | FRA | +0.26           |
| 2          | Chock / Bates              | USA | −0.93           |
| 3          | Gilles / Poirier           | CAN | −0.68           |
| 4          | Guignard / Fabbri          | ITA | −0.39           |
| 5          | Zingas / Kolesnik          | USA | −0.43           |
| 6          | Smart / Dieck              | ESP | −0.35           |
| 7          | Reed / Ambrulevicius       | LTU | +0.33           |
| 8          | Lopareva / Brissaud        | FRA | +0.08           |
| 9          | Lajoie / Lagha             | CAN | +0.24           |
| 10         | Carreira / Ponomarenko     | USA | −0.34           |
| 11         | Davis / Smolkin            | GEO | +0.14           |
| 12         | Fear / Gibson              | GBR | −0.05           |
| 13         | Turkkila / Versluis        | FIN | +0.05           |
| 14         | Lauriault / le Gac         | CAN | +0.29           |
| 15         | Taschlerova / Taschler     | CZE | +0.24           |
| 16         | Mrazkova / Mrazek          | CZE | −0.30           |
| 17         | Harris / Chan              | AUS | −0.21           |
| 18         | Bekker / Hernandez         | GBR | +0.04           |
| 19         | Val / Kazimov              | ESP | +0.43           |
| 20         | Reitan / Majorov           | SWE | +0.46           |

Note: Impact points are estimated via median-of-others neutralization (isuimpact\_residual\_v1, M = 10,000, seed = 20260223). FRA entries: mean +0.17 pts. USA entries: mean −0.57 pts.

## 9. Multi-event analysis

Methodological notes. Judge identity across events: judge names are included on post-event ISU scoring sheets; J1–J9 position labels are assigned independently per event and are not linked across events by the ISU. Named identification for specific events is available through official communications (e.g., OWG 2026 named panels). Cross-event identity linking requires name normalization and is deferred to future work (Section 9.5). BH-FDR correction is applied independently within each event (not pooled across the full database), controlling the per-event expected false discovery proportion at 5%. Events with fewer than eight judges are excluded from permutation inference ( $n=2$  events; the ISU Four Continents 2022 Ice Dance segments used non-standard 7-judge panels). Of the 142 analyzed events, 138 have standard nine-judge panels; the four remaining events (ISU Four Continents Championships 2022 Men Single Skating and Pair Skating segments) used eight-judge panels and are included in permutation inference with definitions generalized accordingly. LOJO is computed for all 144 events using the actual panel size for each event.

**9.1. Aggregate prevalence. Across 142 events spanning 17 competitions (January 2022 through February 2026), the ISU-impact method yielded 271,728 pairwise tests ( $J \times C(N,2)$  per event, summed). Of these, 68,914 (25.4%) yielded nominal permutation  $p$ -values  $\leq 0.05$ , approximately 5.1 times the 5% rate expected under a null with no competitor-specific differentials — consistent with a mixture of true alternatives and null tests. After Benjamini–Hochberg FDR correction at  $q \leq 0.05$ , 20,213 pairs (7.4% of all tests) remained significant. Note that BH-FDR is applied independently within each event; the 20,213 figure aggregates per-event corrected counts and does not reflect a single globally corrected procedure.**

Twenty-four of 142 events (16.9%) had zero BH-significant pairs — consistent with the null characterization for those events. Among the 118 events with at least one significant pair, the distribution was highly right-skewed: the median was 120 significant pairs and the mean was 171, with 64 events exceeding 100. This dispersion suggests that detectable competitor-specific differentials are concentrated in specific events and judge panels rather than spread uniformly across the dataset.

**9.2. Variation by discipline and competition. Table 2 presents the breakdown of significant pairs by discipline (excluding team events). Women Single Skating showed the highest proportion (9.0% of tests), followed by Men Single Skating (7.2%), Pair Skating (7.0%), and Ice Dance (6.0%). These differences reflect a combination of scoring structure and the judges assigned to each discipline.**

Significant pairs are not evenly distributed across competitions. The ISU European Championships 2022 alone accounts for 3,884 of the 20,213 significant pairs (19%), despite comprising only 8 of 142 events analyzed. In contrast, the ISU Grand Prix Final 2024/25 produced only 65 significant pairs across 8 events. Competition-level clustering of this magnitude is consistent with shared panel composition across events within a single competition.

*Table 2. Pairwise tests and BH-significant results by discipline (team events excluded). Tests =  $J \times C(N,2)$  per event, summed across all events in each discipline.*

| Discipline | Events | Total tests | Sig. pairs ( $q \leq 0.05$ ) | % significant |
|------------|--------|-------------|------------------------------|---------------|
|------------|--------|-------------|------------------------------|---------------|

|                      |     |         |        |      |
|----------------------|-----|---------|--------|------|
| Women Single Skating | 34  | 87,543  | 7,858  | 9.0% |
| Ice Dance            | 32  | 61,164  | 3,680  | 6.0% |
| Men Single Skating   | 34  | 89,325  | 6,424  | 7.2% |
| Pair Skating         | 34  | 31,797  | 2,210  | 7.0% |
| All main disciplines | 134 | 269,829 | 20,172 | 7.5% |

Note: Totals in this table exclude team event segments (8 events, 1,899 additional tests, 41 additional significant pairs). Full-database totals including team segments are reported in Section 9.1 (271,728 tests; 20,213 significant pairs).

**9.3. Outcome-determinative events. An event is classified as outcome-determinative if at least one judge  $j$  satisfies two conditions simultaneously: (i)  $q \leq 0.05$  under BH-FDR correction, and (ii) the bias  $B_j(A,B)$  between the 1st- and 2nd-place competitors exceeds the observed gold–silver margin (that is, under median-of-others neutralization applied to the gold–silver pair, the implied score-margin shift would change the winner). Nine of 142 events (6.3%) satisfy this criterion (Table 3).**

The nine events span all four main disciplines and seven separate competitions. Margins range from 0.24 to 1.33 points and bias magnitudes from 0.55 to 1.95 points; all are statistically strong ( $p \leq 0.004$ ,  $q \leq 0.050$ ). The OWG 2026 Ice Dance Free Dance (Section 8) is the most prominent given the public significance of the competition, but not the most extreme: the 4CC 2025 Pair Skating Free Skating presents a larger absolute bias (1.95 pts,  $q = 0.011$ ) relative to a margin of 1.33 pts.

One result requires contextual note: the European Championships 2022 Women's Free Skating (Table 3, final row) involves Kamila Valieva, whose competition results were subsequently annulled by the ISU following a doping proceeding. The ISU-impact computation uses the official scoring sheet as published; the sporting standing of that event is subject to the separate ISU ruling.

*Table 3. Outcome-determinative events: judge bias exceeds gold–silver margin at BH  $q \leq 0.05$ .  $BiasPoints = B_j(rank\ 1, rank\ 2)$ .  $M = 10,000$  permutations; seed = 20260223; residual-label (isuimpact\_residual\_v1). † Valieva results subsequently annulled by ISU; see Section 9.3. LOJO: whether removing the judge (replacing marks with the panel median) changes the gold-medal winner. LOJO CF  $\Delta$ : the margin between the top-two finishers in the LOJO counterfactual standings (which may be a different pair of teams than the original gold–silver when LOJO = Yes; CF  $\Delta$  is therefore not directly comparable to the Margin (pts) column in LOJO = Yes rows). CF Winner: the gold-medal winner in the LOJO counterfactual standings (shown for LOJO = Yes rows; em-dash for LOJO = No rows, where the original winner is unchanged). Specific event narratives are in §9.4.*

| Competition        | Discipline   | Segment      | Judge | Bias (pts) | Margin (pts) | p      | q (BH) | LOJO | LOJO CF $\Delta$ (pts) | CF Winner                |
|--------------------|--------------|--------------|-------|------------|--------------|--------|--------|------|------------------------|--------------------------|
| OWG 2026 (Milano)  | Ice Dance    | Free Dance   | J1    | +1.19      | 0.97         | 0.0003 | 0.012  | Yes  | 0.25                   | USA (Chock / Bates)      |
| OWG 2022 (Beijing) | Pair Skating | Free Skating | J4    | +0.55      | 0.47         | 0.0039 | 0.036  | Yes  | 0.77                   | ROC (Tarasova / Morozov) |
| 4CC 2025           | Pair Skating | Free Skating | J9    | +1.95      | 1.33         | 0.0003 | 0.011  | Yes  | 2.47                   | CAN (Stellato-           |

|                      |                            |                      |    |           |      |            |           |     |      | Dudek /<br>Deschamps)       |
|----------------------|----------------------------|----------------------|----|-----------|------|------------|-----------|-----|------|-----------------------------|
| 4CC<br>2025          | Ice<br>Dance               | Free<br>Dance        | J7 | +1.0<br>1 | 0.48 | 0.000<br>4 | 0.00<br>9 | No  | 3.01 | —                           |
| European<br>s 2024   | Women<br>Single<br>Skating | Free<br>Skating      | J1 | +1.2<br>0 | 1.03 | 0.000<br>8 | 0.02<br>1 | No  | 3.53 | —                           |
| GP Final<br>2023/24  | Men<br>Single<br>Skating   | Short<br>Progra<br>m | J6 | +0.9<br>9 | 0.88 | 0.002<br>6 | 0.04<br>9 | Yes | 1.95 | JPN (Shoma<br>Uno)          |
| GP Final<br>2022/23  | Pair<br>Skating            | Short<br>Progra<br>m | J1 | +0.5<br>5 | 0.43 | 0.000<br>6 | 0.04<br>0 | No  | 1.37 | —                           |
| European<br>s 2022   | Men<br>Single<br>Skating   | Short<br>Progra<br>m | J2 | +1.0<br>2 | 0.70 | 0.000<br>9 | 0.01<br>5 | No  | 0.15 | —                           |
| European<br>s 2022 † | Women<br>Single<br>Skating | Free<br>Skating      | J3 | +1.5<br>8 | 0.24 | 0.000<br>3 | 0.00<br>5 | Yes | 6.55 | RUS<br>(Shcherbakov<br>a) † |

**9.4. LOJO counterfactual cross-validation. As a complementary and structurally independent check, we apply the Leave-One-Judge-Out (LOJO) counterfactual: for each judge  $j$ , replace  $j$ 's marks with the median of the remaining  $J-1$  judges' marks and recompute all segment totals and placements. Unlike ISU-impact, LOJO produces no p-value; it answers "would the outcome have changed if this judge had scored at the panel median?"**

Across 1,288 judge-event computations across all 144 events ( $138 \text{ events} \times 9 \text{ judges} + 4 \text{ events} \times 8 \text{ judges} + 2 \text{ events} \times 7 \text{ judges}$ ; the ISU Four Continents 2022 Ice Dance segments used reduced panels; for  $J = 7$  events, LOJO replaces the removed judge with the median of the remaining six judges' marks), LOJO identifies 153 gold-medal changes (11.9% of removals) and 467 podium changes (36.2%). LOJO is a sensitivity analysis, not an inference procedure; high flip rates can occur in close competitions even under benign stylistic dispersion, and should not be interpreted as evidence of misconduct. For the OWG 2026 Ice Dance Free Dance specifically, removing judge J1 and replacing J1's marks with the panel median yields a USA (Chock/Bates) win with a counterfactual margin of 0.25 points; no other single judge removal changes the gold in that event. The LOJO result thus independently identifies the same judge and the same directional effect as the ISU-impact test, providing convergent evidence across two methodologically distinct frameworks. The nine outcome-determinative events (Table 3) include five where LOJO independently confirms a change in gold-medal winner (LOJO = Yes) and four where LOJO does not flip the gold despite the pairwise bias exceeding the margin (LOJO = No). The discordance reflects a structural difference: ISU-impact measures judge  $j$ 's differential on the specific gold–silver pair under median-of-others neutralization, while LOJO replaces  $j$ 's marks for all competitors and recomputes the full trimmed-mean standings. A judge whose bias is concentrated on the gold–silver pair may have countervailing effects on other entries — or may shift the panel's trimming set — such that the full LOJO recomputation does not flip the winner even when the pairwise criterion is satisfied. The counterfactual margins in Table 3 (LOJO CF  $\Delta$ ) show the gold–silver gap under the LOJO scenario; for LOJO = No rows, some values differ markedly from the actual margin (e.g., Europeans 2022 Men Single Skating:

0.15 pts counterfactual vs. 0.70 pts actual), illustrating near-determinative influence even when the gold-medal winner does not formally change.

**9.5. Stability, calibration, and power .** Two of three validation analyses are deferred to future work. **Stability:** whether a judge exhibits consistent competitor-specific differentials across events requires linking judge identities across events; ISU judging sheets include judge names, but cross-event identity linking requires name normalization across events; stability analysis is deferred to future work. **Calibration:** residual-label permutation p-values are exact conditional on the exchangeability assumption holding — under the null, the permutation p-value is discrete-uniform on  $\{1/(M+1), 2/(M+1), \dots, 1\}$  (converging to continuous uniform(0, 1) as  $M \rightarrow \infty$ ), so deviations from uniformity may reflect a mixture of true alternatives and residual departures from exchangeability; calibration checks are therefore important. Across 271,728 pairwise tests, 25.4% yield  $p \leq 0.05$  (versus 5.0% expected under a pure null with no competitor-specific differentials;  $5.07\times$  enrichment), and 3.78% yield  $p \leq 0.001$  (versus 0.10% under a pure null;  $37.8\times$  enrichment). The p-value distribution shows a pronounced spike at small values and an upper tail broadly consistent with the null above  $p \approx 0.05$ , the two-component signature of a mixture of true-null and true-alternative hypotheses. BH controls the expected false discovery proportion among the rejected hypotheses at 5% within each event; the 7.4% rejection proportion indicates a substantial fraction of tests show competitor-specific differentials under this procedure. To assess null calibration directly, Figure 2 shows a QQ-plot of the upper tail ( $p > 0.5$ ;  $n = 82,062$  pairs), where the null component dominates by construction. The observed upper-tail p-values track the diagonal closely within 95% confidence bands (Figure 2), confirming that the residual-label permutation test is well-calibrated under null conditions. **Power:** characterizing the minimum detectable bias at typical field sizes and score variances is deferred to a companion paper. A complementary quality-gap stratification (Figure 3) confirms that null calibration is stable across all five quintiles of competitor quality gap ( $|TSS\_A - TSS\_B|$ ): upper-tail mean p-values are approximately uniform across strata (range 0.746–0.749; trend  $p = 0.069$ ), consistent with well-specified exchangeability. The overall rejection rate rises modestly from Q1 (23.8%) to Q5 (28.2%) across quintiles (trend  $p = 0.019$ ), reflecting higher signal density in lopsided pairs rather than exchangeability breakdown.

Figure 3. Quality-Gap Stratified Calibration (n = 271,728 pairwise tests)

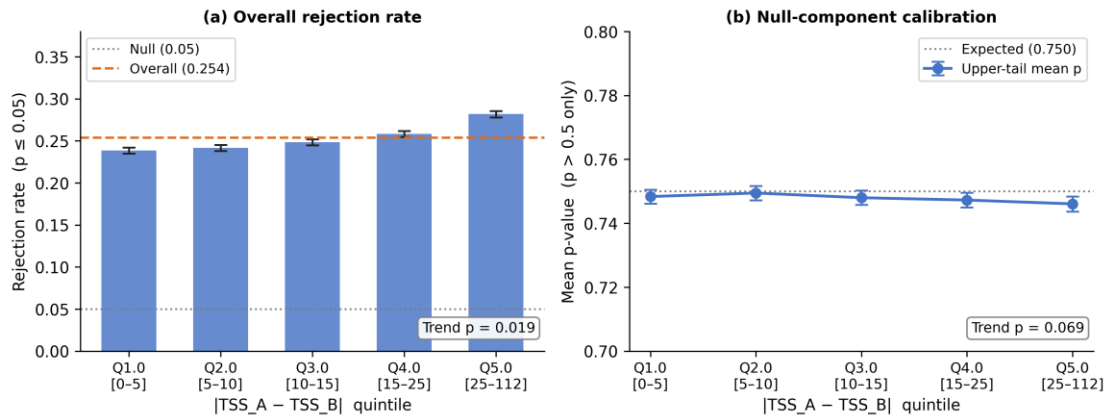


Figure 3. Quality-gap stratified calibration. (a) Overall rejection rate ( $p \leq 0.05$ ) by quintile of competitor quality gap ( $|TSS\_A - TSS\_B|$ ); rates rise from Q1 (23.8%) to Q5 (28.2%), trend  $p = 0.019$ , reflecting

higher signal density in lopsided pairs. (b) Upper-tail mean p-value ( $p > 0.5$ ) by quintile; values are approximately uniform across strata (range 0.746–0.749 vs. expected 0.750), confirming that the null-distribution calibration is stable. Error bars: 95% confidence intervals.  $N = 271,728$  pairwise tests (residual-label permutation,  $M = 10,000$ ).

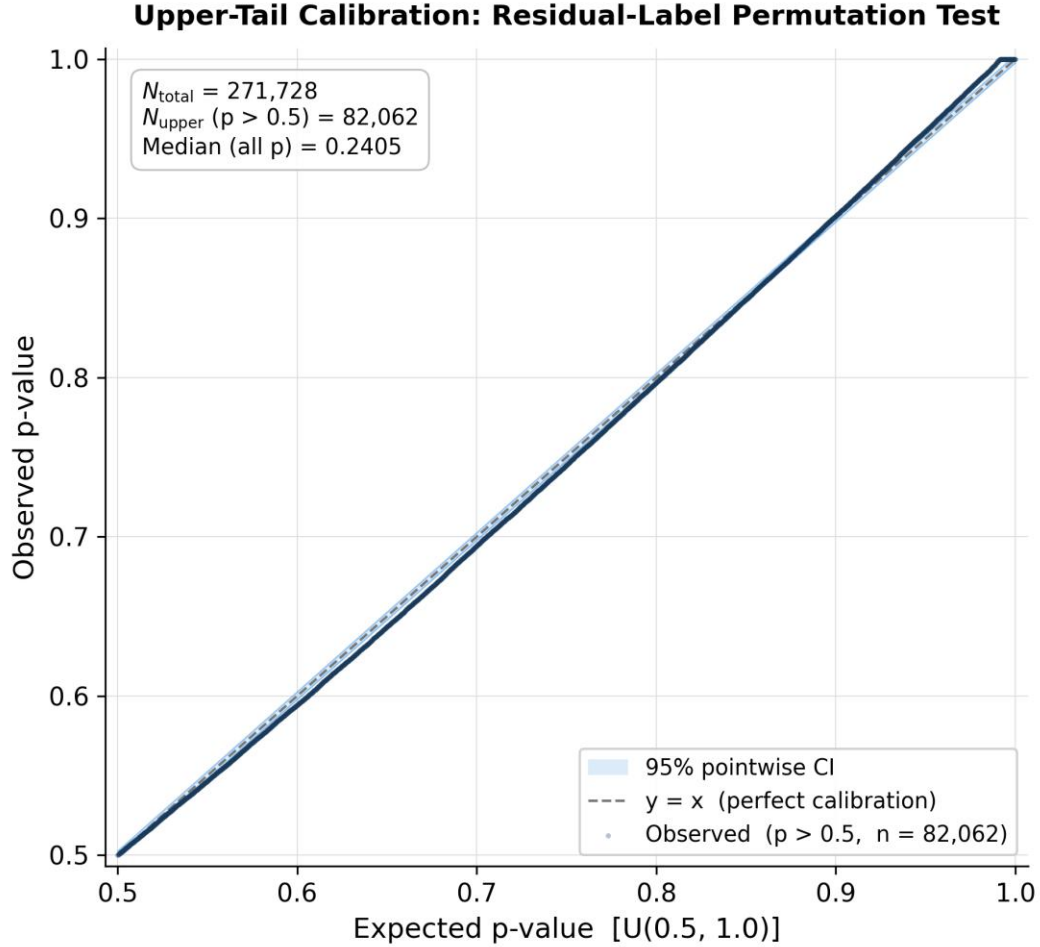
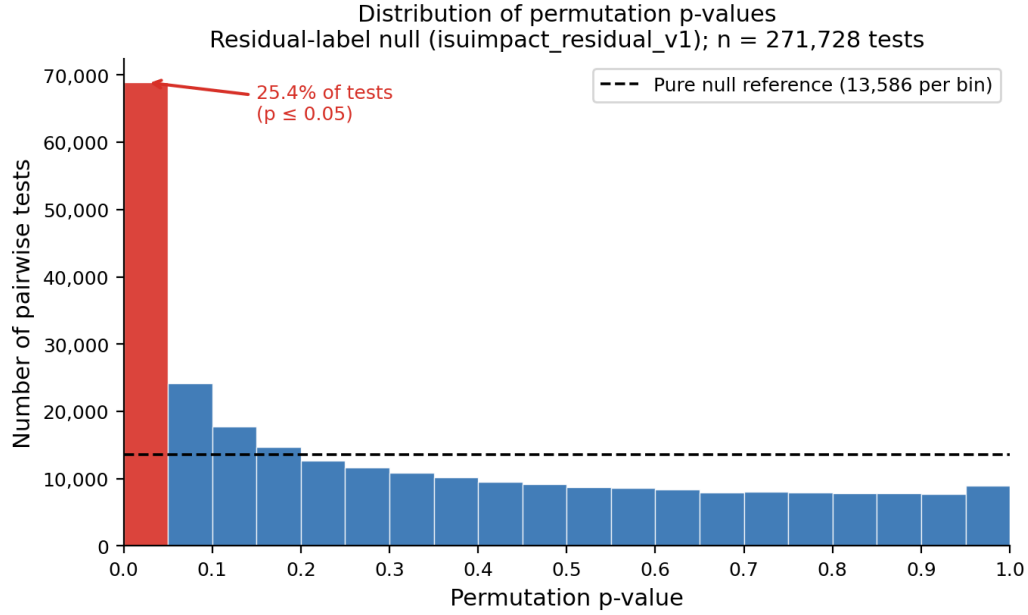


**Figure 2. Upper-tail calibration QQ-plot.** Observed p-values in the upper tail ( $p > 0.5$ ) from 271,728 pairwise tests (residual-label permutation,  $M = 10,000$ ) are plotted against expected uniform(0.5, 1.0) quantiles. The upper tail is the null-dominated component of the null/alternative mixture; departures from the diagonal would indicate mis-calibration. Shaded band: 95% pointwise confidence interval. The minor deviation at  $p = 1.0$  (1.7% of upper-tail pairs;  $n = 1,416$ ) occurs when every permutation is at least as extreme as the observed statistic — arising when judge  $j$ 's differential  $B_j(A,B) \approx 0$ , so all permutations of near-zero deltas equal or exceed the observed value; this is expected null behaviour given  $M = 10,000$  and discrete test statistics.

*Figure 1. Histogram of permutation p-values across 271,728 pairwise tests (residual-label null, `isuimpact_residual_v1`). Red bar: first bin ( $p \leq 0.05$ ), containing 25.4% of tests. Dashed line: pure null reference (13,586 per bin =  $5\% \times n$ ). The pronounced spike at small p-values and upper tail broadly consistent with the null are the two-component signature of a null/alternative mixture.*



Each bar spans 0.05 p-value units. Dashed line = expected count under a pure null (5% x n per bin).

## 10. Discussion and limitations

First, the method is inherently relative: it detects judge-specific differentials versus the remainder of the panel. Bias shared by most or all judges will not be detected.

Second, the residual-label null assumes that pooled delta values are exchangeable across entries for each judge–pair test. This holds when median-of-others neutralization substantially attenuates the shared quality signal. If entry-specific variation persists in the residuals — for example, due to sequencing effects or element difficulty variation not captured by base values — the null could be anti-conservative. The quality-gap stratification (Figure 3) provides empirical evidence against this concern: null calibration is stable across all quality strata (upper-tail trend  $p = 0.069$ ). The approach may also understate collusion effects that manifest as correlated delta patterns across entries rather than single-judge directional differentials.

Third, trimming reduces the influence of a single judge on any row. The proposed statistic measures impact through trimming rather than against it; nevertheless, in very close competitions even small margin shifts can be outcome-determinative.

Finally, the methodology is event-local. Extending to cross-event judge identity linking and causal explanations of bias requires additional assumptions and data (e.g., judge nationality, federation incentives, and assignment mechanisms). A specific variant of this concern is competitor quality mismatch: if entries A and B differ substantially in skill level, their delta distributions may differ structurally even under the null. In practice, within-event quality ranges are narrow — all entries in a given event compete at the same tier (e.g., Olympic final or World Championship) — and median-of-others neutralization already removes the shared quality component from each scoring row before deltas

are pooled. Upper-tail calibration (Section 9.5, Figure 2) confirms that within-event quality variation does not materially inflate the type I error rate.

Planned robustness checks and extensions include: (i) Stratified residual-label permutation — repeating the analysis with GOE rows and PCS rows as separate within-category tests — would provide formal p-values for each score component and directly test whether pooling the two row types drives results. For the OWG 2026 flagship result, J1's +1.19 pt net margin shift decomposes as approximately +1.01 pts (GOE) and +0.20 pts (PCS; Section 8); stratified permutation would confirm or qualify whether each component independently reaches significance. (ii) A database-wide dependence-robust FDR (Benjamini–Yekutieli) sensitivity analysis across all 271,728 tests is planned as a supplement to the event-local BH procedure reported here. (iii) Multi-judge collusion detection — computing cross-judge correlations of impact vectors  $I_j(T)$  within events, clustering correlated judges, and testing whether judge clusters collectively shift margins — addresses forms of bias that single-judge tests are not optimized to detect.

## 11. Conclusion

We introduced a pairwise, point-scale bias statistic for judged competitions scored via trimmed means and developed a nonparametric permutation test that preserves judge style via residual-label (delta-exchange) permutation. The method yields interpretable margin shifts and permutation p-values for every judge×competitor pair within an event. The Olympic Ice Dance Free Dance example demonstrates how a single judge's competitor-specific differential can be quantified and tested without conflating it with overall scoring style. Applied across 142 events spanning 17 competitions (2022–2026), the method identifies 20,213 significant judge × competitor-pair results and nine outcome-determinative events, including a finding that a single judge's competitor-specific differential exceeded the official gold–silver margin at the OWG 2026 Ice Dance Free Dance. The LOJO counterfactual independently corroborates this finding.

## References

1. Campbell, B., and Galbraith, J. W. (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.
2. International Skating Union (ISU). ISU Judging System: Panel of Judges and trimmed mean. <https://www.isu.org/figure-skating-rules/> (accessed 2026-02-26).
3. International Skating Union (ISU). (2024). Special Regulations & Technical Rules: Single & Pair Skating and Ice Dance. International Skating Union (27 Nov 2024). <https://www.isu.org/figure-skating-rules/>
4. Lee, J. (2008). Outlier Aversion in Subjective Evaluation: Evidence From World Figure Skating Championships. *Journal of Sports Economics*, 9(2), 141–159.
5. Emerson, J. D., Seltzer, M., and Lin, D. (2009). Assessing Judging Bias: An Example From the 2000 Olympic Games. *The American Statistician*, 63(2), 124–131.
6. Zitzewitz, E. (2006). Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making. *Journal of Economics & Management Strategy*, 15(1), 67–99.

7. 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.

Braeunig, P. A. (2025a). Patriotic bias in sports judging: A matter of juror and athlete socialization at the national level. *Cogent Social Sciences*, 11(1), Article 2557014.  
<https://doi.org/10.1080/23311886.2025.2557014>

Braeunig, P. A. (2025b). National bias in international sports judging: A scoping review. *Cogent Social Sciences*, 11(1), Article 2594841. <https://doi.org/10.1080/23311886.2025.2594841>

### **Data availability**

The ISU ‘Judges Details per Skater’ scoring sheets used in this analysis are publicly available from the International Skating Union website ([isu.org](http://isu.org)). The compiled SQLite database of 291,604 individual judge marks across 142 events (2022–2026) is openly available on Zenodo at <https://doi.org/10.5281/zenodo.18804555> (CC BY 4.0).

### **Code availability**

The ISU-impact analysis pipeline, permutation test implementation, and all scripts used to produce the results in this paper are available at <https://github.com/michaelallman1960-max/figure-skating-judge-bias>. The implementation uses Python 3 with NumPy (PCG64 PRNG, seed=20260223).

### **Author contributions**

M.A. conceived and designed the study, developed the statistical methodology, assembled the database, implemented the computational pipeline, performed all analyses, and wrote the manuscript.

### **Funding**

The author received no funding for this work.

### **Competing interests**

The author declares no competing interests.

## **Appendix A. Why the residual-label null was adopted over the quantile permutation null.**

The quantile permutation null (`isuimpact_quantile_v1`) mapped each judge’s marks to percentiles within that judge’s empirical CDF, permuted percentile labels across judges within each scoring row, then mapped back via each judge’s inverse CDF. The flaw is a misalignment of the permutation axis. The quantile null permutes across judges within a row — asking whether judge  $j$ ’s rank among judges for this

element is unusual — but the target null is within judge, across competitors: whether judge  $j$ 's deltas for competitor A versus competitor B are exchangeable. Although mapping through each judge's own inverse CDF does preserve judge  $j$ 's marginal distribution, the exchangeability violation is structural: the null tests the wrong comparison object. The v1 null therefore does not represent a credible 'no directional bias' baseline for this estimand. Empirical diagnostics comparing both methods: under the quantile null, the mean permutation p-value across all tests was 0.249 (expected 0.500 under a well-specified null) and 2,812 of 272,072 tests reached  $p \leq 0.001$  — approximately  $10\times$  the expected false-positive count — with 86 of 144 events showing  $\geq 2$  judges flagged simultaneously; these diagnostics are reported in full in methodology\_diagnosis\_v1.md (available in the code repository; see Code Availability).

The residual-label null (isuimpact\_residual\_v1) operates on  $\Delta_{\{j,r\}}$  values directly. Because median-of-8 neutralization substantially attenuates the shared panel consensus signal, the remaining deltas represent judge  $j$ 's idiosyncratic contribution. Under the null of no directional bias for pair (A,B), these deltas are exchangeable across A's and B's rows — no CDF mapping required, no style assumption imposed. See Section 6.2 and methodology\_diagnosis\_v1.md for the full root-cause analysis.

## Appendix B. Algorithm 1: Residual-Label Permutation Test

| Algorithm 1 — Residual-Label Permutation Test (isuimpact_residual_v1) |                                                                                                                                                  |
|-----------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                       | <b>Input:</b>                                                                                                                                    |
|                                                                       | Scoring rows for competitors A, B; judge marks $x_{\{j,r\}}$ ; scaling factors $m_r$                                                             |
|                                                                       | Judge position $j$ ; permutation count $M$ ; RNG seed                                                                                            |
|                                                                       | <b>Output: bias statistic <math>B_{\text{obs}}</math> (points); one-sided p-value <math>p</math></b>                                             |
| 1                                                                     | For each row $r \in \text{rows}(A) \cup \text{rows}(B)$ :                                                                                        |
| 2                                                                     | $\Delta_{\{j,r\}} \leftarrow m_r \times (x_{\{j,r\}} - \text{median}\{x_{\{k,r\}} : k \neq j\})$                                                 |
| 3                                                                     | $B_{\text{obs}} \leftarrow \sum_{\{r \in \text{rows}(A)\}} \Delta_{\{j,r\}} - \sum_{\{r \in \text{rows}(B)\}} \Delta_{\{j,r\}}$                  |
| 4                                                                     | $\text{Pool} \leftarrow [\Delta_{\{j,r\}} : r \in \text{rows}(A) \cup \text{rows}(B)] \triangleright  \text{rows}(A)  +  \text{rows}(B) $ values |
| 5                                                                     | $\text{count} \leftarrow 0$                                                                                                                      |
| 6                                                                     | For $m = 1$ to $M$ :                                                                                                                             |
| 7                                                                     | $(\Delta_A, \Delta_B) \leftarrow \text{random partition of Pool into }  \text{rows}(A)  \text{ and }  \text{rows}(B)  \text{ values}$            |
| 8                                                                     | $B^{(m)} \leftarrow \sum \Delta_A - \sum \Delta_B$                                                                                               |
| 9                                                                     | If $B^{(m)} \geq B_{\text{obs}}$ : $\text{count} \leftarrow \text{count} + 1$                                                                    |
| 10                                                                    | $p \leftarrow (1 + \text{count}) / (M + 1)$                                                                                                      |
| 11                                                                    | Return $B_{\text{obs}}, p$                                                                                                                       |
|                                                                       | <b>Notes:</b>                                                                                                                                    |
|                                                                       | $m_r$ = ISU GOE multiplier for element rows; $m_r = 1$ for PCS rows.                                                                             |
|                                                                       | The one-sided test assesses whether judge $j$ systematically favours A over B.                                                                   |
|                                                                       | Run for all unordered pairs $\{A, B\}$ and all $j$ ; apply BH-FDR correction                                                                     |
|                                                                       | within each event across all (pair, judge) combinations.                                                                                         |