# From Scores to Trust:
# A Quality Engine for Diving Judges

Miami University Department of Statistics
Rongrong Qian, Jing Jing, Charntel Kazingizi, Sagar Pangeni

Dec 3rd 2025

# Agenda

# How Diving Is Scored

**In competitive diving, even one point can change everything.**

A judge's score directly impacts:

- Rankings
- Advancement
- Recognition
- Athlete opportunities

How scoring works:

- Judges award 0-10 points per dive.
- Highest and lowest scores are removed (if 5+ judges)
- Remaining scores are summed x Degree of Difficulty (DD)

# App Today: What It Already Does Well?



## Strong on Data Capture

- Register judges and divers

- Record scores and compute results

- Store and replay competition videos

# What's Missing: No View of Judge Quality

**What We See**
Individual scores from each judge for every dive performance

**What's Hidden**
Long-term behavioral patterns and consistency metrics over time

| DIVER | | TEAM | | COACH | | AGE 25 | | Final | 1 | 370.90 |
|---|---|---|---|---|---|---|---|---|---|---|
| DATE **March, 18, 2022** | | EVENT **Men's 1M - Finals** | | DIVE ORDER **8** | | | | | | |
| MEET **2022 NCAA Division III (Unofficial)** | | LOCATION **IU Natatorium, 901 West New York Street, Indianapolis,** | | | | | | | | |

| V/O | Rnd | Dive# | Dive Description | DD | J1 | J2 | J3 | J4 | J5 | J6 | J7 | Net Total | Award | Running Total | Rnd Place |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| o | 1 | 405C | Inw 2.5 Som - Tuck | 3.1 | 8.0 | 8.0 | 8.0 | 7.0 | 7.5 | 8.0 | 8.0 | 24.0 | 74.40 | 74.40 | 1 |
| o | 2 | 107C | Fwd 3.5 Som - Tuck | 3.0 | 8.0 | 6.5 | 6.5 | 6.5 | 7.0 | 7.0 | 7.0 | 20.5 | 61.50 | 135.90 | 1 |
| o | 3 | 205C | Back 2.5 Som - Tuck | 3.0 | 5.5 | 5.5 | 6.0 | 5.5 | 5.0 | 5.0 | 5.5 | 16.5 | 49.50 | 185.40 | 5 |
| o | 4 | 305C | Rev 2.5 Som - Tuck | 3.0 | 6.0 | 7.0 | 6.5 | 6.0 | 6.0 | 6.0 | 6.5 | 18.5 | 55.50 | 240.90 | 5 |
| o | 5 | 5152B | Fwd 2.5 Som 1 Tw - Pike | 3.2 | 7.5 | 6.5 | 6.5 | 7.0 | 6.5 | 6.5 | 7.0 | 20.0 | 64.00 | 304.90 | 1 |
| o | 6 | 5335D | Rev 1.5 Som 2.5 Tw - Free | 3.0 | 7.0 | 7.5 | 7.5 | 7.0 | 7.0 | 7.5 | 9.0 | 22.0 | 66.00 | 370.90 | 1 |

DD Totals: Vol: 0 Opt: 18.3 All: 18.3    Rank: Vol: Opt:    **Avg score:** Vol: Opt: 6.79 All: 6.79  Avg counted score: Vol: Opt: 6.75 All: 6.75     FINAL TOTAL  370.90

# What's Missing: No View of Judge Quality

**The Gap**
No structured feedback system to help judges identify and improve their scoring patterns

# From "Feeling" to Data

Subjective, unconscious preferences are inevitable in any judging system.
Today, people judge judges by "*feeling*" rather than objective measures.
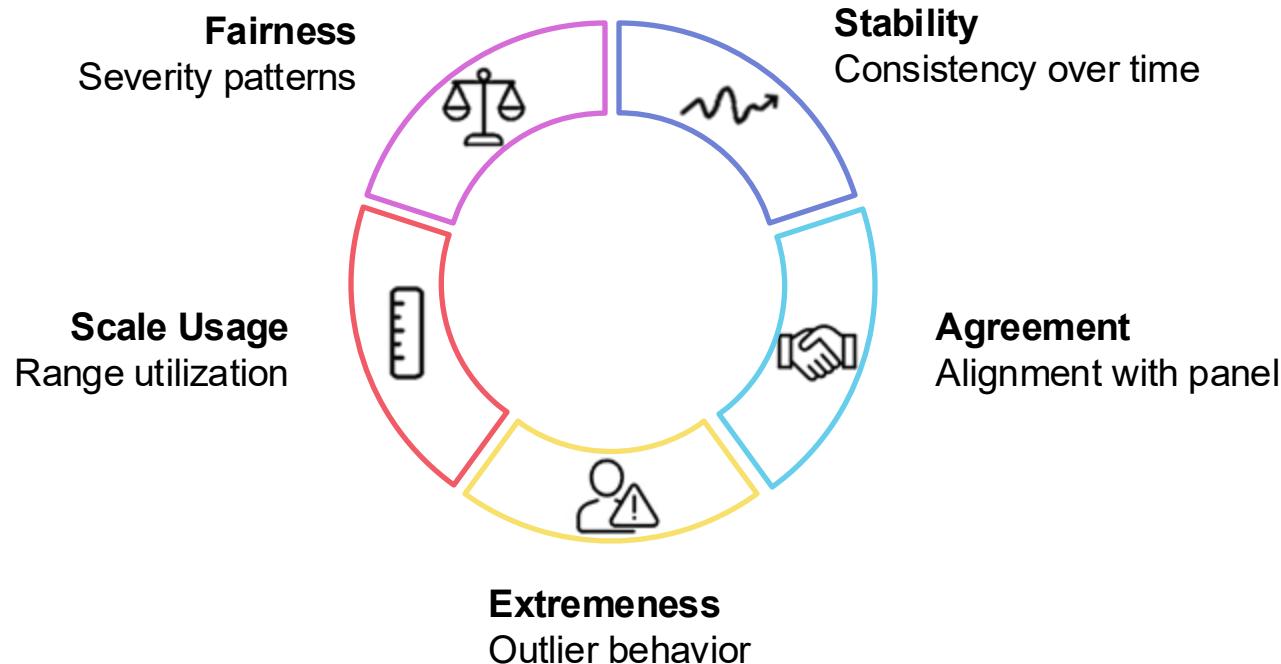
The goal: move to a clear, data-based view of fairness

# 100%
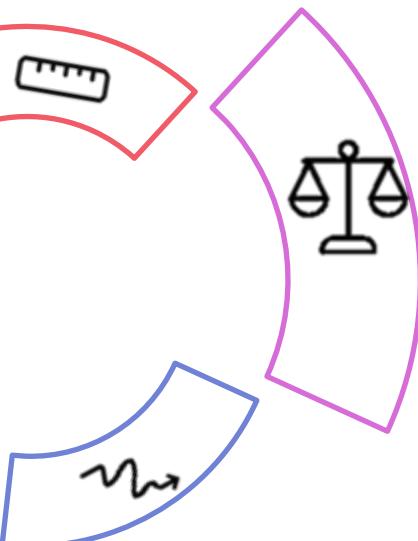
Data-Driven
Objective assessment

# What Makes a Good Judge?

Defining quality in five intuitive dimensions



**Fairness**
Severity patterns

**Stability**
Consistency over time

**Scale Usage**
Range utilization

**Agreement**
Alignment with panel

**Extremeness**
Outlier behavior

# Fairness – Are They Systematically High or Low?



Compare each judge to the **panel consensus baseline**

Above Baseline ➡ Lenient
Below Baseline ➡ Severe

# A Fair Baseline for Each Dive

Baseline Score = Flagged / NA

# A Fair Baseline for Each Dive
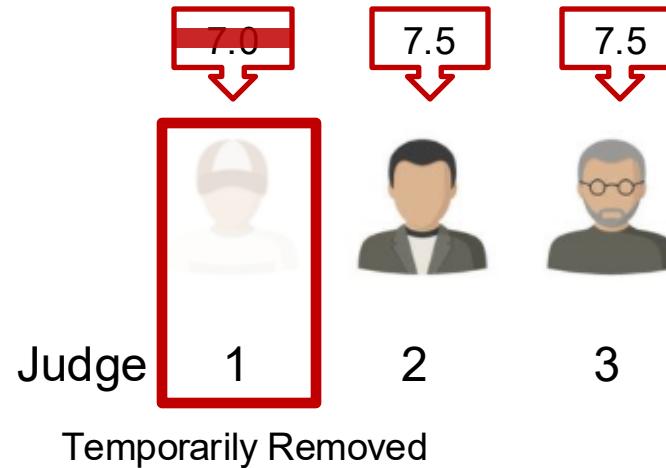
# A Fair Baseline for Each Dive

Step 1 – Leave Judge X Out

| 7.0 | 7.5 | 7.5 |

Judge 1 2 3

Temporarily Removed

Step 2 – Take the Average ➡ Baseline Score

Baseline Score for Judge 1 = Mean Score of 2 and 3 = 7.5

# A Fair Baseline for Each Dive

5 Judges Panel



| 5.5 | 7.5 | 7.5 | 8.0 | 9.0 |

Judge    1      2      3      4      5
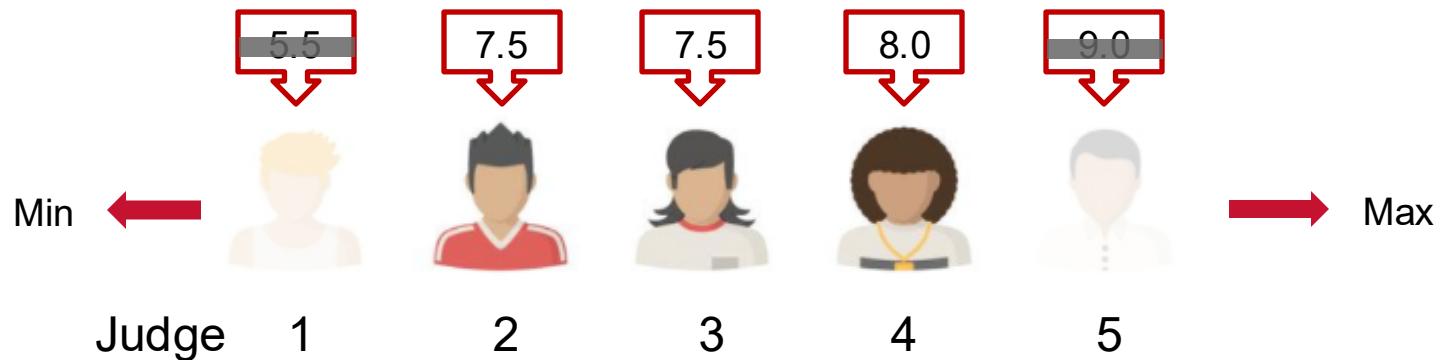
# A Fair Baseline for Each Dive

Step 1 – Remove the Most Extreme Scores
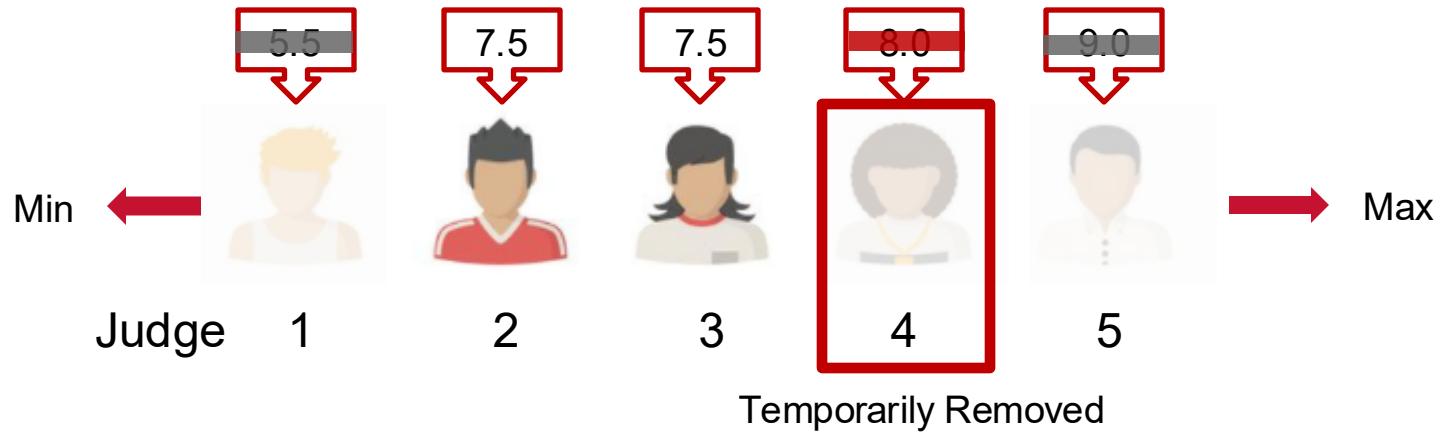
5 Judges Panel



| 5.5 | 7.5 | 7.5 | 8.0 | 9.0 |

Min ← → Max

Judge   1    2    3    4    5

# A Fair Baseline for Each Dive

Step 2 – Leave Judge X Out

| 5.5 | 7.5 | 7.5 | 8.0 | 9.0 |

Min ← Judge  1    2    3    4    5    → Max

Temporarily Removed
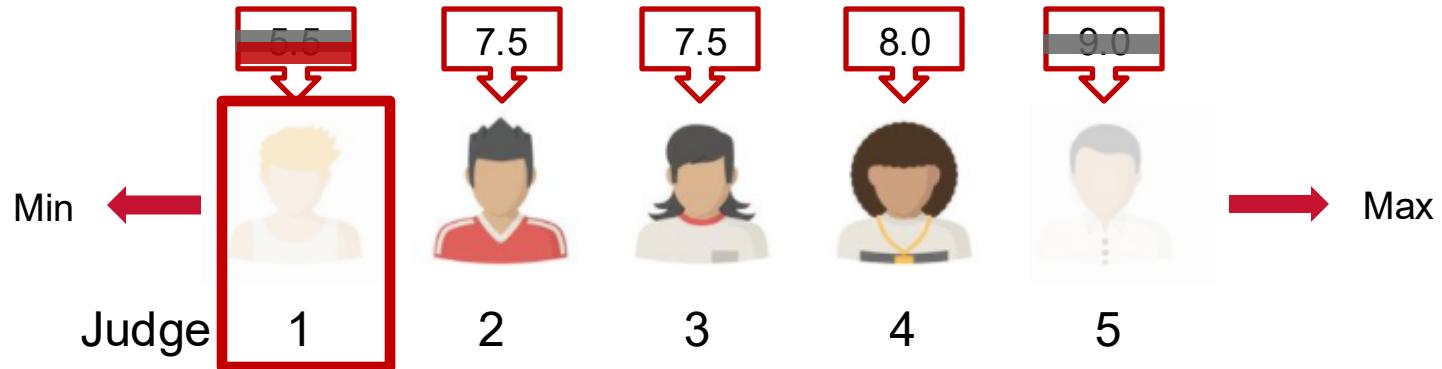
Step 3 – Take the Average  ➡  Baseline Score

Baseline score for Judge 4 = mean score of 2 and 3 = 7.5

# A Fair Baseline for Each Dive

Step 2 – Leave Judge X Out

| 5.5 | 7.5 | 7.5 | 8.0 | 9.0 |

Min ← → Max

Judge 1 2 3 4 5
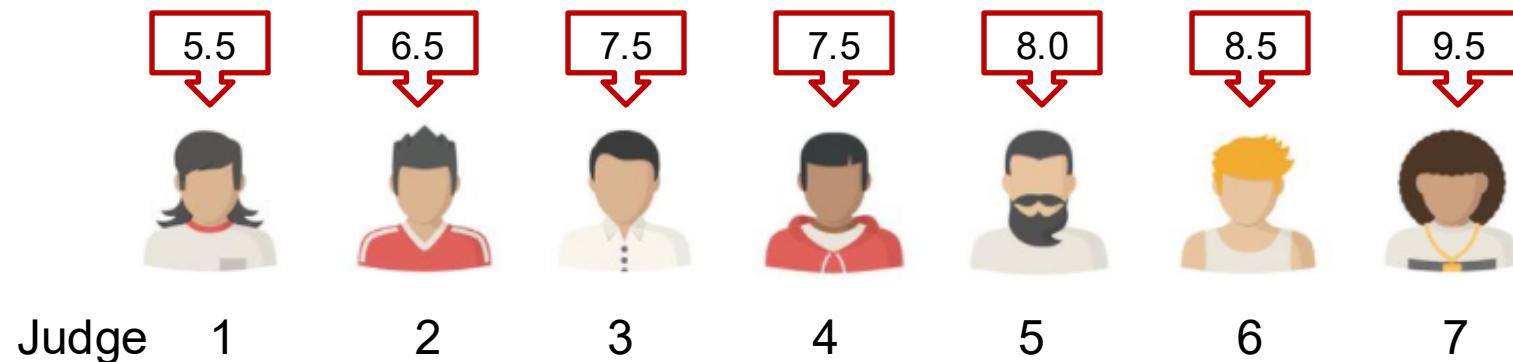
Step 3 – Take the Average ➡ Baseline Score

Baseline Score for Judge 1 = Mean Score of 2, 3 and 4 = 7.67

# A Fair Baseline for Each Dive

7 Judges Panel



| | 5.5 | 6.5 | 7.5 | 7.5 | 8.0 | 8.5 | 9.5 |
|---|---|---|---|---|---|---|---|
| Judge | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
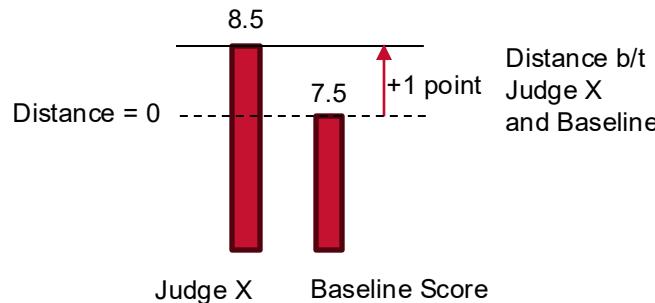
# Fairness – Are They Systematically High or Low?
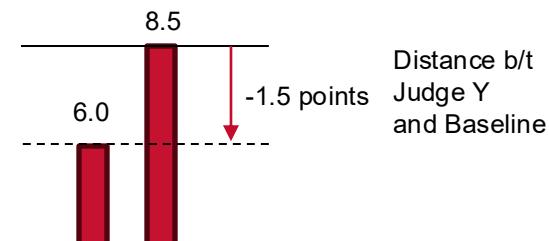


Severity = Average (Judge score – Baseline)

For One Dive
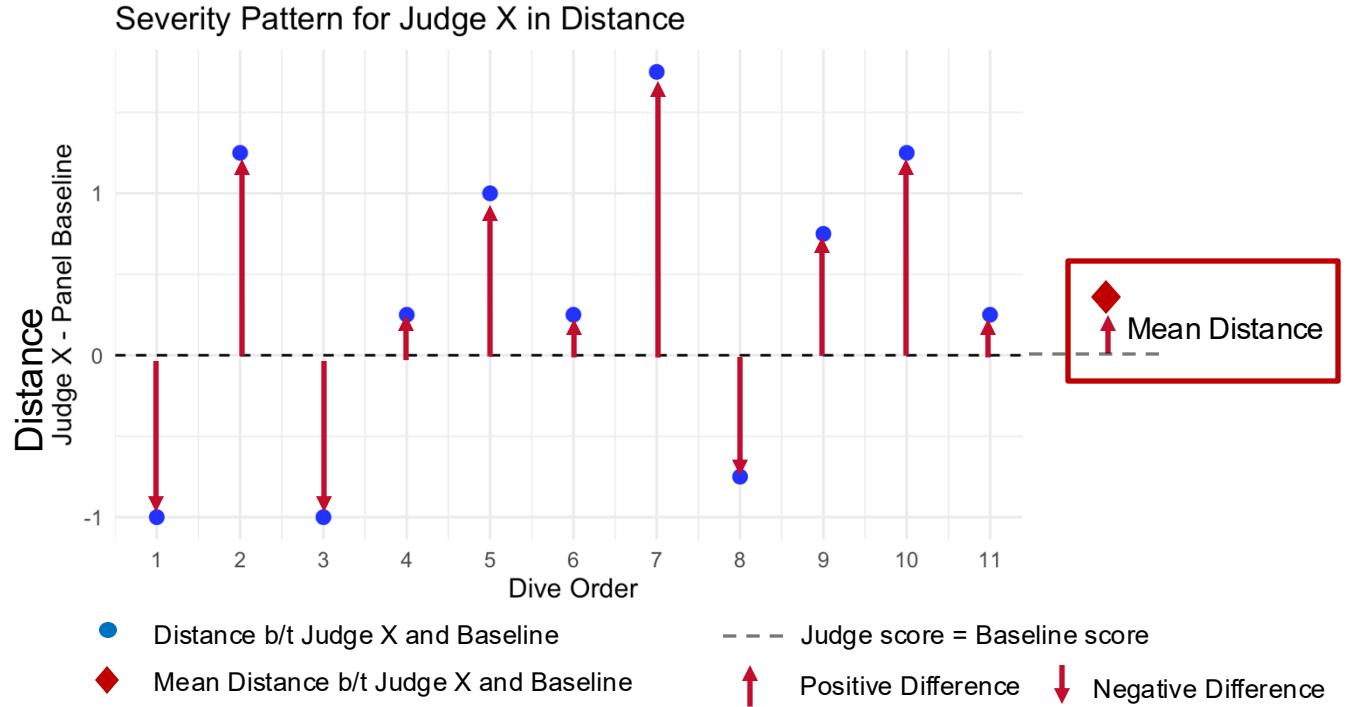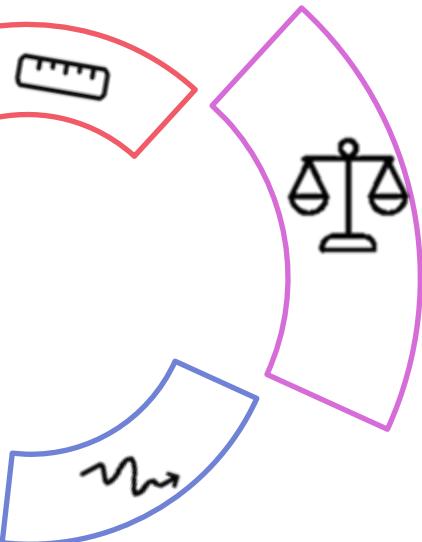
# Fairness – Are They Systematically High or Low?



Severity Pattern for Judge X in Distance
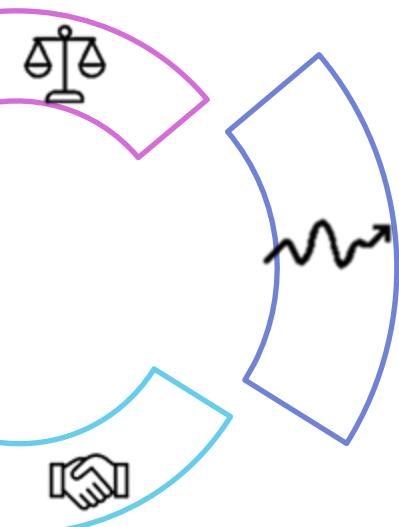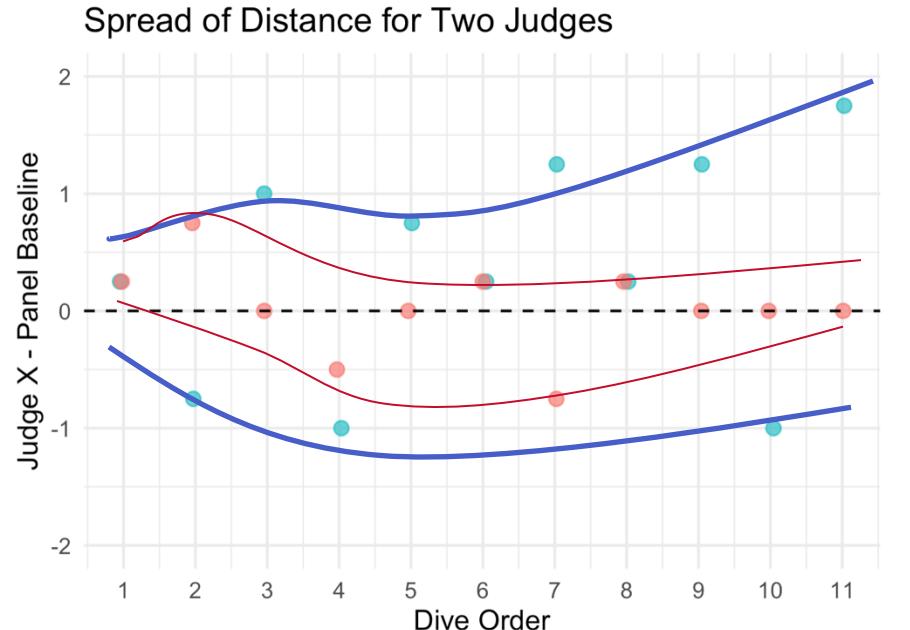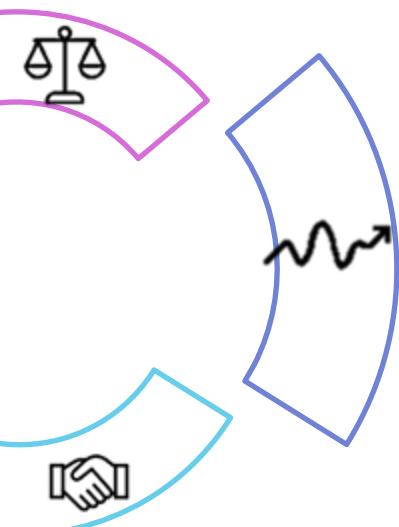
# Stability – Do They Bounce Around?

Inconsistency = Spread of (Judge score – Baseline)

*(How spread out these differences are)*

Small Spread ➡ Stable and Predictable

Large Spread ➡ Noisy and Inconsistent

# Stability – Do They Bounce Around?



Spread of Distance for Two Judges

Dots Stay Close to 0 ➡ Small Spread ➡ More Consistent

Dots Move Up and Down ➡ Large Spread ➡ Less Consistent
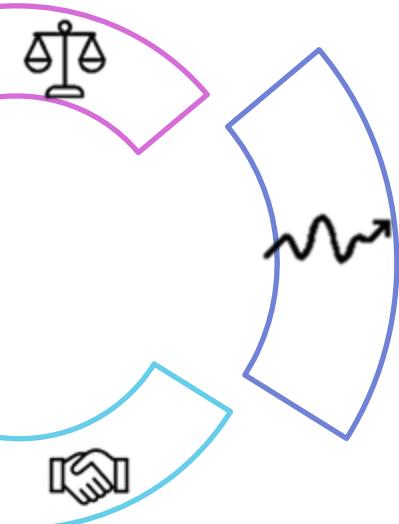
# Stability – Do They Bounce Around?
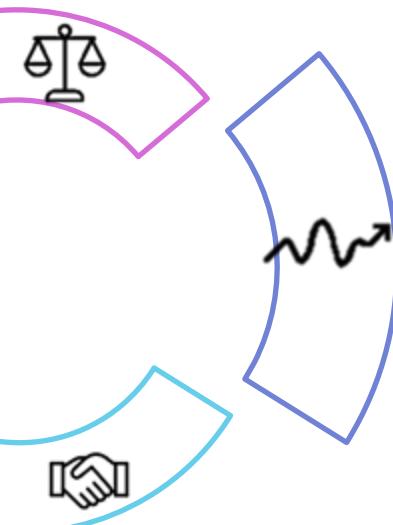
Summary Statistics for Judge Inconsistency (2024&2025)

| Min | 25 %ile | Median | 75 %ile | Max |
|---|---|---|---|---|
| 0.249 | 0.492 | 0.559 | 0.599 | 1.404 |

● **Tier A – Highly Consistent**

   $0 \leq$ Inconsistency $< 0.492$ (25 %ile)

● **Tier B – Typical Consistency**

   $0.492$ (25 %ile) $\leq$ Inconsistency $\leq 0.599$ (75 %ile)

● **Tier C – Needs Review for Consistency**

   Inconsistency $> 0.599$ (75 %ile)

# Stability – Do They Bounce Around?
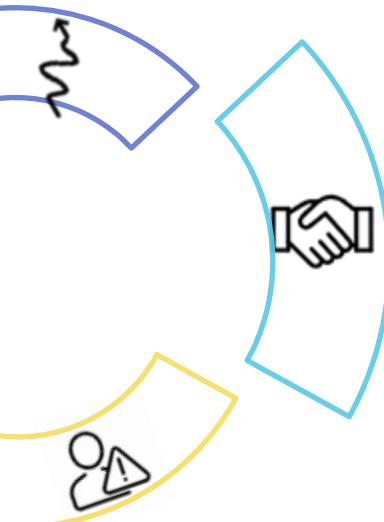
## Summary for selected judge

| Judge ID | Judge | Dives | Score range | Severity | Inconsistency | Agreement | Extremeness | Range use | Overall rank |
|---|---|---|---|---|---|---|---|---|---|
| | | 558 | 8.00 | 0.05 | 0.47 | 0.86 | 0.60 | 3.04 | 31 |

Highly Consistent (Stable and Predictable)

# Agreement – Do They Rank Dives Like the Panel?

Agreement = Correlation (Judge ranking, Baseline ranking)
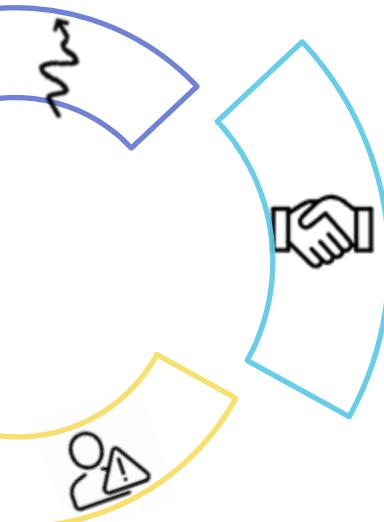
*(Whether the judge's rank-ordering of dives aligns with the panel's. Agreement ranges from −1 to 1)*

Close to 1 ➡ Strong Agreement in Ranking Dives

Close to 0 ➡ Poor Agreement in Ranking Dives

Close to -1 ➡ Actively Reversed Ranking

# Agreement – Do They Rank Dives Like the Panel?

Summary for selected judge

| Judge ID | Judge | Dives | Score range | Severity | Inconsistency | Agreement | Extremeness | Range use | Overall rank |
|---|---|---|---|---|---|---|---|---|---|
| | | 558 | 8.00 | 0.05 | 0.47 | 0.86 | 0.60 | 3.04 | 31 |
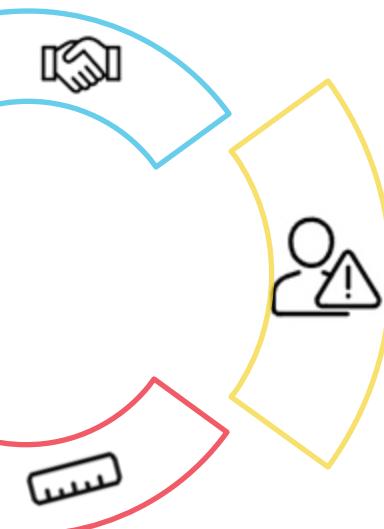
Strong Agreement with the Panel

# Extremeness – Are They Often the Lone Outlier?

- Step 1 – On Each Dive

  Check uniquely highest or uniquely lowest

- Step 2 – Count How Often

  "Extreme dives" = dives where they are the lone outlier

  Extremeness frequency = % of dives where they are the lone outlier

- Step 3 – Measure How Far

  For those dives, measure the gap to the nearest judge

  Extremeness intensity = average gap on those dives

# Extremeness – Are They Often the Lone Outlier?
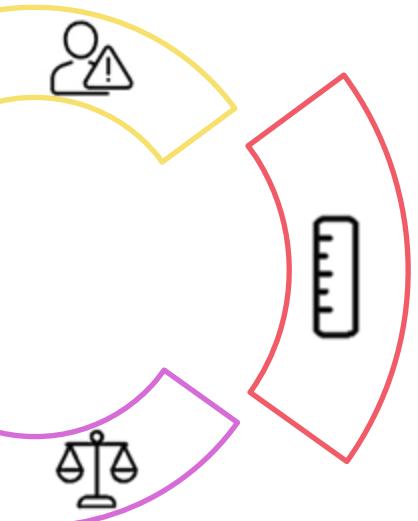
Summary for selected judge

| Judge ID | Judge | Dives | Score range | Severity | Inconsistency | Agreement | Extremeness | Range use | Overall rank |
|---|---|---|---|---|---|---|---|---|---|
| | | 558 | 8.00 | 0.05 | 0.47 | 0.86 | 0.60 | 3.04 | 31 |

Average Gap is 0.60 Points from the Nearest Judge

# Scale Usage – Do They Use the Full Range?

Range ratio = Judge Spread ÷ Panel Average Spread
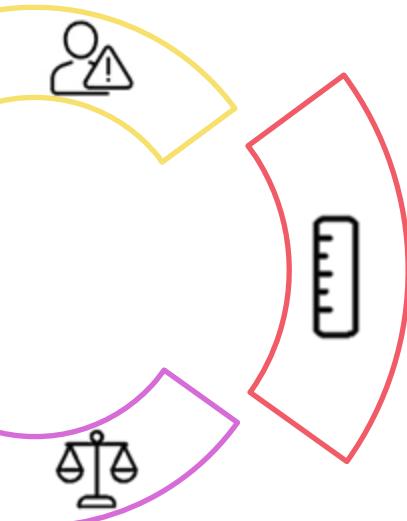
*(How wide their scores spread across dives)*

Range Ratio ≈ 1 ➡ The judge uses the scale similarly to the panel

Range Ratio < 1 ➡ The judge uses a narrower range (stays in the middle)

Range Ratio > 1 ➡ The judge uses a wider range (more extremes)

# Scale Usage – Do They Use the Full Range?

## Summary for selected judge

| Judge ID | Judge | Dives | Score range | Severity | Inconsistency | Agreement | Extremeness | Range use | Overall rank |
|---|---|---|---|---|---|---|---|---|---|
| | | 558 | 8.00 | 0.05 | 0.47 | 0.86 | 0.60 | 3.04 | 31 |

This judge uses a 3 times wider range than the panel average.

# From Metrics to Penalties and Overall Rank

| | |
|---|---|
| Fairness penalty | Grows with the absolute size of Severity |
| Inconsistency penalty | Grows with Inconsistency |
| Agreement penalty | Higher when Agreement is low |
| Extremeness penalty | Higher when the judge is extreme frequently and by a large margin |
| Range-use penalty | Small when range ratio ≈ 1<br>Large when much smaller or larger than 1 |

Note: Each penalty is a non-negative score, where 0 represents ideal behavior
and larger values indicate greater deviation from the panel consensus.

- Rank judges on each penalty
  (Rank 1 = best)

- Add the five ranks to get an overall score

Overall rank:

Overall Score = Sum of 5 Penalty Ranks
(Lower is Better)

# From Metrics to Penalties and Overall Rank

## Summary for selected judge

| Row | Judge ID | Judge | Dives | Severity | Inconsistency | Agreement | Extremeness | Range use | Overall rank |
|---|---|---|---|---|---|---|---|---|---|
| Value | | | 558 | 0.047 | 0.468 | 0.859 | 0.597 | 3.038 | 31 |
| Rank (1 = best; of all judges) | | | | 31 / 100 | 18 / 100 | 78 / 100 | 40 / 100 | 37 / 100 | 31 / 100 |

# Judge Leaderboard – Who Is Most Reliable (2024-2025 FL)?

Top 10 judges by composite performance score
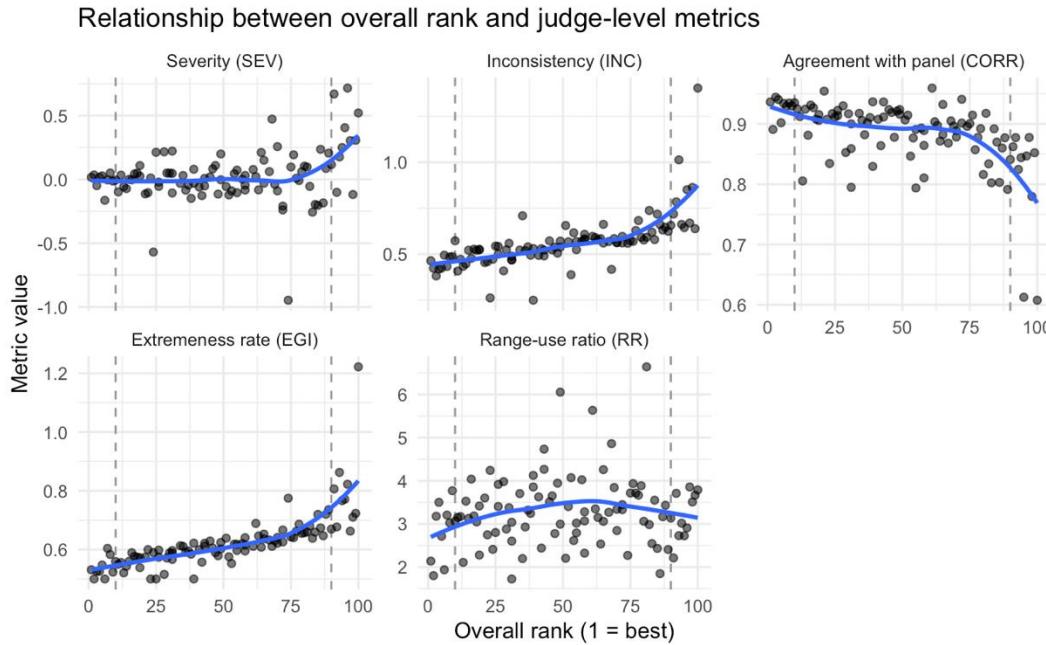
| Judge ID | Judge | Number of dives | Severity | Inconsistency | Agreement with panel | Extremeness | Range-use ratio | Overall rank |
|----------|-------|-----------------|----------|---------------|---------------------|-------------|-----------------|--------------|
|          |       | 274             | 0.016    | 0.463         | 0.937               | 0.531       | 2.139           | 1            |
|          |       | 132             | 0.035    | 0.423         | 0.891               | 0.500       | 1.801           | 2            |
|          |       | 221             | -0.049   | 0.381         | 0.945               | 0.526       | 3.179           | 3            |
|          |       | 209             | 0.017    | 0.418         | 0.941               | 0.526       | 3.501           | 4            |
|          |       | 456             | 0.028    | 0.424         | 0.902               | 0.548       | 2.716           | 5            |
|          |       | 154             | -0.164   | 0.494         | 0.934               | 0.500       | 1.937           | 6            |
|          |       | 1257            | -0.004   | 0.432         | 0.929               | 0.604       | 3.205           | 7            |
|          |       | 55              | 0.050    | 0.484         | 0.935               | 0.583       | 3.025           | 8            |
|          |       | 209             | -0.013   | 0.487         | 0.930               | 0.523       | 3.771           | 9            |
|          |       | 506             | -0.005   | 0.572         | 0.935               | 0.560       | 3.065           | 10           |

## Judge Assessment Card – Example
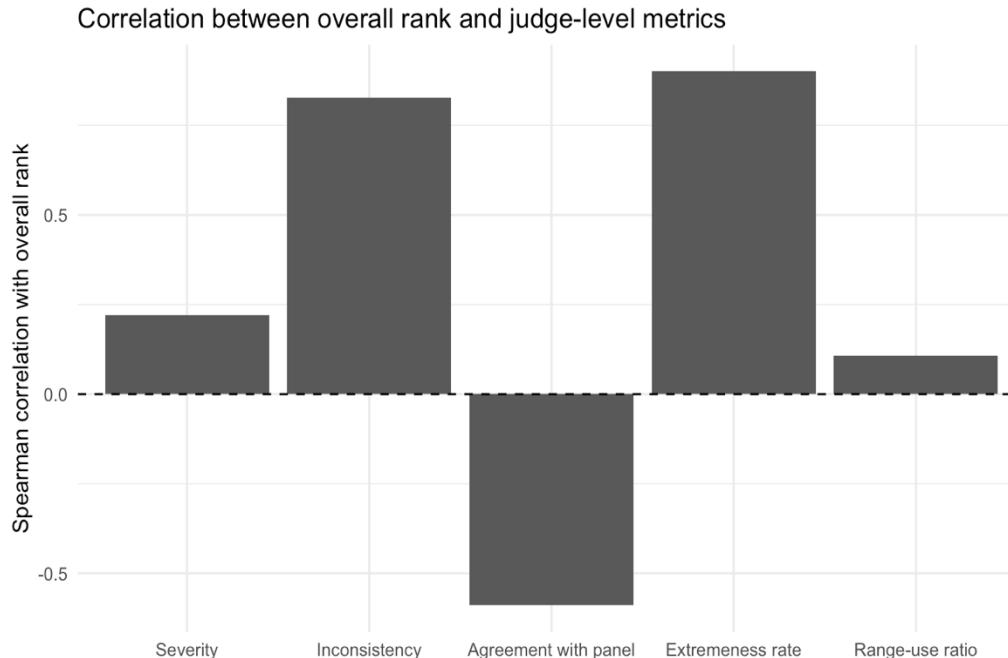
**Live Demo: Choose a Judge ID, See Their Profile**

# Relationship between overall rank and judge-level metrics



Relationship between overall rank and judge-level metrics
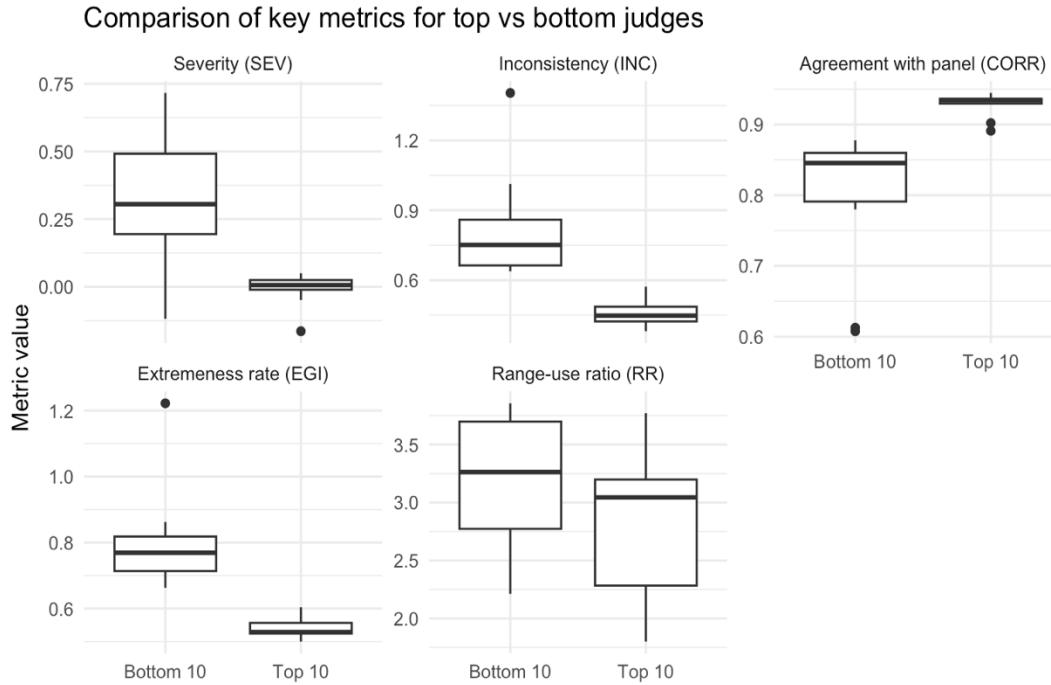
- **SEV**: Severity stays near zero, then turns clearly positive for the lowest-ranked judges (consistent high scorers).
- **INC**: Inconsistency rises steadily with worse rank and spikes for the bottom judges.
- **CORR**: Agreement with the panel declines with rank and drops sharply in the worst decile.
- **EGI**: Extremeness intensity stays modest until it rises noticeably near the worst ranks.
- **RR**: Range-use ratio is slightly higher in the middle ranks but does not clearly separate the extremes.

# Correlation between overall rank and judge-level metrics



Correlation between overall rank and judge-level metrics

- Overall rank is most strongly and positively associated with inconsistency (INC) and most strongly and negatively associated with agreement with the panel (CORR).

- Severity (SEV) shows only a modest association with rank, and the range-use ratio (RR) is only weakly related.

# Comparison of key metrics for top vs bottom judges



Comparison of key metrics for top vs bottom judges

- **INC & SEV**: Bottom 10 have higher inconsistency and more positive severity; top 10 have low INC and SEV ≈ 0.
- **CORR level**: Agreement is much stronger for the top judges.
- **CORR spread**: Top group has tight, high CORR; bottom group shows lower medians and wider spread.
- **EGI**: EGI is clearly higher and more variable among the bottom judges
- **RR**: Bottom judges tend to use slightly more of the scale (higher RR)

# Questions?

# Reference

- Maronna, R. A., Martin, R. D., & Yohai, V. J. (2019). Robust statistics: theory and methods (with R) (2nd ed.). Wiley. https://www.wiley.com/en-us/Robust+Statistics%3A+Theory+and+Methods+(with+R)%2C+2nd+Edition-p-9781119214687
- Zitzewitz, E. (2006). Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making. Journal of Economics & Management Strategy, 15(1), 67–99. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1530-9134.2006.00092.x
- Emerson, J. W., Seltzer, M., & Lin, D. (2009). Assessing Judging Bias: An Example From the 2000 Olympic Games. The American Statistician, 63(2), 124–131. http://www.jstor.org/stable/25652240
- Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982 Dec;38(4):963-74. PMID: 7168798.

# Judge Leaderboard – Who Is Least Reliable (2024-2025 FL)?

Bottom 10 judges by composite performance score

| Judge ID | Judge | Number of dives | Severity | Inconsistency | Agreement with panel | Extremeness | Range-use ratio | Overall rank |
|---|---|---|---|---|---|---|---|---|
| | | 176 | 0.670 | 0.718 | 0.862 | 0.806 | 2.212 | 91 |
| | | 506 | -0.101 | 0.785 | 0.877 | 0.677 | 3.709 | 92 |
| | | 156 | 0.176 | 1.013 | 0.825 | 0.863 | 2.728 | 93 |
| | | 77 | 0.250 | 0.661 | 0.844 | 0.767 | 3.017 | 94 |
| | | 66 | 0.405 | 0.644 | 0.613 | 0.771 | 2.725 | 95 |
| | | 310 | 0.716 | 0.851 | 0.847 | 0.822 | 2.909 | 96 |
| | | 132 | 0.304 | 0.669 | 0.878 | 0.663 | 3.856 | 97 |
| | | 36 | -0.118 | 0.863 | 0.780 | 0.711 | 3.509 | 98 |
| | | 88 | 0.307 | 0.638 | 0.852 | 0.723 | 3.666 | 99 |
| | | 12 | 0.521 | 1.404 | 0.608 | 1.222 | 3.789 | 100 |

Rongrong Qian qianr5@miamioh.edu

Jing Jing jingj4@miamioh.edu

Charntel Kazingizi kazingcr@miamioh.edu

Sagar Pangeni pangens@miamioh.edu

Michael Hughes hughesmr@miamioh.edu

# Thank you!