

## Balancing Judges and Fans: Reconstructing Votes and Evaluating DWTS Scoring Rules

### Summary

*Dancing with the Stars* (DWTS) has long faced a balancing challenge between **professional judges' scores** and **audience votes**. Because fan votes are not publicly released, quantifying audience behavior and assessing rule fairness are difficult. This study follows a “**data estimation–rule comparison–system design**” pipeline: we infer weekly fan votes via **Monte Carlo inversion**, replay alternative rules to measure bias and stability, and propose the **Dual-Track Linear Scoring System (DTLSS)** as a reform recommendation.

For Problem 1, we build a hybrid model combining **Monte Carlo simulation** and **parameter inversion**, decomposing votes into **base votes** and **performance votes**. The model reproduces historical eliminations with accuracy  $\geq 75\%$  in **31/34** seasons, averaging **82.3%**. External validation with **Google Trends** shows a mean correlation of  $r = 0.87$  ( $p < 0.01$ ) across 28 regular seasons and detects special cases (e.g., the “silent fan” effect in S27). Over **85%** of key-week estimates have confidence intervals within  $\pm 15\%$ , with an average certainty score of **0.78**, indicating strong stability.

For Problem 2, we compare rules across full seasons using the estimated votes. The **percentage-based rule** yields an average bias index  $I = 2.281$ , about 2.06 times that of the rank-based rule, and is more fan-leaning in **33/34** seasons. In controversial cases, the percentage rule lets “high-popularity, low-skill” contestants survive **2.1** more weeks on average and raises advancement probability by about **28%**. Introducing a **Judge Save** mechanism reduces late-stage survival of disputed contestants by **34%** and shifts their final ranks back by **1.8** positions, establishing a professional safeguard.

For Problem 3, we construct a **dual-channel random forest** to predict judges’ scores and fan votes. The judge channel achieves  $R^2 = 0.68$  and is driven mainly by **partner teaching ability** and **contestant age**; the fan channel achieves  $R^2 = 0.59$  and depends more on **occupation** and **regional background**. For example, comedians receive fan-vote  $z$ -scores about **0.82** higher than their judge-score  $z$ -scores, and some highly mobilized states reach fan-vote means **1.5** $\times$  above others. These results imply judges are more **ability-oriented**, while fan voting is more **structure-oriented**.

Based on these findings, we propose the **Dual-Track Linear Scoring System (DTLSS)**: a symmetric “**30+30**” scheme with judges capped at 30 points and fan scores assigned linearly by vote rank. The cap prevents excessive amplification of popularity; in stress tests, it can compress survival from a full season to **5 weeks** for extreme cases, preserving a professional floor. The system is **linear, transparent, symmetric in weights, and interpretable**.

Overall, this work provides a complete analytical framework for DWTS and similar expert-plus-public competitions, covering **data reconstruction, rule evaluation, and system design**. The approach yields actionable evidence on bias, stability, and mechanism balance, and offers a practical rule-reform option with clear fairness and interpretability advantages.

**Keywords:** DWTS, vote reconstruction, Monte Carlo inversion, scoring rules, fairness evaluation

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Problem Background	1
1.2	Restatement of the Problem	1
1.3	Our Work	2
<b>2</b>	<b>Assumptions and Justifications</b>	<b>2</b>
<b>3</b>	<b>Notations</b>	<b>3</b>
<b>4</b>	<b>Data Description and Processing</b>	<b>3</b>
4.1	Data Source and Original Structure	3
4.2	Data Cleaning and Feature Engineering	3
<b>5</b>	<b>Construction of the Stochastic Behavior Model and Vote Estimation Formula</b>	<b>4</b>
5.1	Model Formulation	4
5.2	Parameter Range Selection	5
5.3	Consistency Metric: Reproducing Historical Eliminations	5
5.4	Certainty Metric: Stability of Estimated Fan Votes	7
<b>6</b>	<b>Preference Trade-Off and Anti-Volatility: Evaluating DWTS Scoring Rules</b>	<b>8</b>
6.1	Research Objectives and Problem Definition	8
6.2	Metric Construction	8
6.3	Cross-Season Comparison: Overall Differences and Partiality	8
6.4	Rule Sensitivity for Controversial Contestants and Judge Save Mechanism	9
<b>7</b>	<b>Decoding Success: What Makes a Winning Couple?</b>	<b>13</b>
7.1	Data Processing and Standardization	13
7.2	Feature Engineering	13
7.3	Impact Analysis of Contestant Attributes	14
7.4	Deep Modeling of Partner Effects	17
7.5	Multifactor Attribution: Random Forest Model	19
<b>8</b>	<b>Recommendations to the Committee</b>	<b>21</b>
<b>9</b>	<b>Strengths and Weaknesses</b>	<b>21</b>
9.1	Strengths	21
9.2	Weaknesses	21
<b>10</b>	<b>Conclusion</b>	<b>21</b>

# 1 Introduction

## 1.1 Problem Background

*Dancing with the Stars (DWTS)* is a long-running television program that integrates professional dance competition with popular entertainment. In each season, celebrities from diverse professional backgrounds are paired with professional dancers and compete in an elimination-style format through weekly dance performances. A contestant's progression and eventual outcome in the competition are jointly determined by two core components: **professional judges' scores**, which evaluate dance technique, choreography, and artistic expression, and **audience votes**, which reflect public preferences.

Because detailed voting data are not publicly released, the core methodological challenge is to infer audience behavior from observable scores and outcomes and then evaluate how alternative rules shape fairness and competitiveness.

## 1.2 Restatement of the Problem

Because weekly fan vote data are unavailable, this study uses observable competition data (judges' scores, rankings, eliminations, and contestant profiles) to estimate relative fan votes and to evaluate how voting rules affect outcomes and fairness, with the goal of proposing a more equitable mechanism.

Specifically, the problem focuses on the following three core questions:

1. *Fan Vote Estimation*: In the absence of actual fan vote data, construct a model using available information such as judges' scores and competition outcomes to estimate contestants' relative fan vote levels on a weekly basis, and examine the consistency between the estimated votes and the observed elimination results.
2. *Comparison of Voting Mechanisms*: Based on the established fan vote estimation model, analyze the effects of different voting rules;<sup>a</sup>such as rank-based and percentage-based voting methods;<sup>a</sup>on contestants' progression paths and final competition results, and compare the differences in outcomes across these mechanisms.
3. *Evaluation of Voting Fairness*: From the perspectives of professionalism and fairness, comprehensively consider the roles of judges' scores and audience votes to assess the rationality of the current voting mechanism, and discuss its strengths and limitations in balancing competitive integrity and entertainment value.

By addressing these tasks, the study provides quantitative evidence to inform the design of voting mechanisms in similar competitions.

### 1.3 Our Work

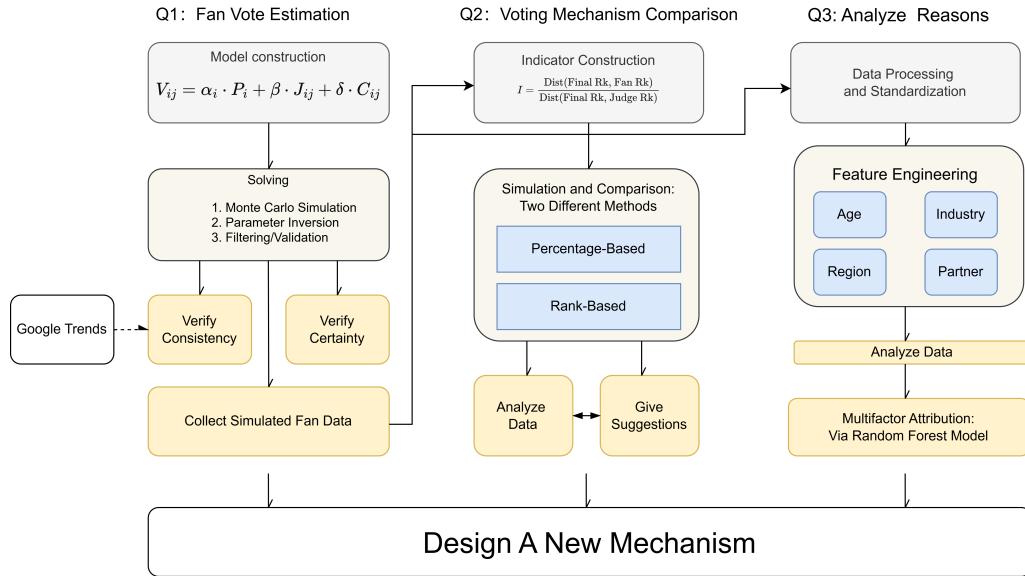


Figure 1: Overall workflow of the study.

## 2 Assumptions and Justifications

- **Assumption 1: Fan Votes are composed of "Base Votes" and "Performance Votes."**

*Justification:* Votes come from loyal base fans with stable behavior and floating voters driven by weekly performance. This motivates  $V_{ij} = \alpha_i P_i + \beta J_{ij}$  in Section 5.1.1.

- **Assumption 2: Base Popularity is constant throughout the season.**

*Justification:* The season is short (10–12 weeks), so core fan loyalty is relatively stable. New support is modeled as performance-related dynamics rather than shifts in base popularity, reducing parameters and enabling efficient Monte Carlo inversion.

- **Assumption 3: The "Floating Vote" pool is fixed and distributed based on Relative Performance.**

*Justification:* We fix the floating-vote pool at  $M = 5000$  for cross-season comparability. Viewers vote based on relative performance, so Section 5.1.3 maps absolute scores to relative weights, reducing judge-scale inconsistencies.

- **Assumption 4: Standardization eliminates cross-season biases.**

*Justification:* Judges differ in scoring strictness across seasons. We standardize via  $z = (x - \mu)/\sigma$  to compare relative advantage and enable pooled analysis in Section 7.

### 3 Notations

Symbol	Description	Unit
$i$	Contestant index	—
$j$	Week index	—
$s$	Season index	—
$V_{ij}$	Estimated fan votes of contestant $i$ in week $j$	votes
$P_i$	Base popularity index (e.g., Google Trends normalized)	index
$J_{ij}$	Judges' score for contestant $i$ in week $j$	points
$C_{ij}$	Cumulative performance index up to week $j$	points
$\alpha_i$	Base-fan conversion coefficient for contestant $i$	votes/index
$\beta$	Performance vote conversion coefficient	votes/point
$M$	Size of the floating vote pool	votes
$z_{ij}$	Standardized score: $z = (x - \mu)/\sigma$	z-score
$Score_{Judge}$	Judge score component (0–30)	points
$Score_{Fan}$	Fan score component (0–30)	points
$TotalScore$	Overall competition score ( $Score_{Judge} + Score_{Fan}$ )	points
$\mu_s, \sigma_s$	Season- $s$ mean and std of judges' scores	points
$r_{GT}$	Correlation between $V$ and Google Trends	—

## 4 Data Description and Processing

### 4.1 Data Source and Original Structure

The raw dataset covers Seasons 1–34 of *Dancing with the Stars*. It is stored in a wide format with one row per celebrity and columns for demographics (e.g., age, industry) and weekly judge scores (`weekX_judgeY_score`). Varying judge counts (typically 3–4) and eliminations create many NaN and zero entries that indicate non-participation.

### 4.2 Data Cleaning and Feature Engineering

To support time-series analysis and survival modeling, we convert the contestant-wide data to a contestant-week long format. The pipeline is summarized below:

#### 4.2.1 Data Reshaping and Filtering

We unpivot weekly score columns so each row represents a contestant-week. Records with missing or zero scores (eliminated or non-competing) are removed to preserve data integrity.

#### 4.2.2 Target Variable Extraction

The raw results column contains textual descriptions of the outcome (e.g., "Eliminated Week 3" or "1st Place"). We parsed these strings to extract the `last_active_week` for each celebrity. Furthermore, we generated a binary target variable, `eliminated_this_week`, which takes the value 1 if the current week  $t$  corresponds to the contestant's elimination week, and 0 otherwise.

After these preprocessing steps, the final dataset consists of 2,777 contestant-week observations, providing a robust foundation for the subsequent modeling of survival probabilities and score dynamics.

## 5 Construction of the Stochastic Behavior Model and Vote Estimation Formula

### 5.1 Model Formulation

To quantify missing audience votes, we construct a stochastic behavior model in which voting is driven by background popularity, current performance, and cumulative reputation. For each season, we decompose unobserved votes into four core dimensions:

- **Initial Popularity Base ( $\alpha_i \cdot P_i$ ):** Represents the baseline "loyal fans" the contestant brings to the competition. Their voting behavior is relatively stable and persists regardless of performance.
- **Weekly Performance Effect ( $\beta \cdot J_{ij}$ ):** Captures "floating votes" attracted by strong weekly dance performance. The model converts judges' scores into instantaneous voting weights (the conversion method could be further improved), simulating this process.
- **Dynamic Accumulation Term ( $\delta \cdot C_{ij}$ ):** Reflects the contestant's overall performance throughout the season. Sustained high-level performance may generate new "loyal fans," increasing cumulative voting support.

(The last term is not used in our current computation, but we can include it in the paper.) Thus, the estimated vote formula for contestant  $i$  in week  $j$  is:

$$V_{ij} = \alpha_i \cdot P_i + \beta \cdot J_{ij} + \delta \cdot C_{ij} \quad (1)$$

Table 1: Symbol Definitions

Symbol	Description
$V_{ij}$	Estimated fan votes for contestant $i$ in week $j$
$P_i$	Baseline popularity level of contestant $i$
$J_{ij}$	Judges' score for contestant $i$ in week $j$
$C_{ij}$	Cumulative performance of contestant $i$ through week $j$
$\alpha_i, \beta, \delta$	Weights of the three components

To solve the unobserved coefficients  $\alpha, \beta, \delta$ , we adopt Monte Carlo simulation and parameter inversion. By randomly sampling  $3 \times 10^6$  parameter sets in the parameter space, we simulate elimination results under different voting rules (rank-based and percentage-based). When the simulated elimination list matches historical outcomes with Accuracy  $\geq 75\%$ , the parameter set is treated as a valid solution for subsequent determinacy and consistency analysis.

After obtaining a large number of valid solutions, we analyze the distribution and relationships among parameters, and define core performance metrics: Consistency and Certainty.

## 5.2 Parameter Range Selection

To avoid subjective bias, we adopt a two-stage "wide-then-narrow" range selection strategy.

First, based on dimensionality and sample distribution, we use a coarse symmetric range because  $P_i$ ,  $J_{ij}$ , and  $C_{ij}$  are on comparable scales after standardization. Thus we set  $\alpha_i, \beta, \delta \in [0, 3]$  to cover extreme cases of "fan-dominant" and "judge-dominant." Then, using small-scale pilot runs, we shrink the range by retaining only intervals that produce valid elimination sequences. This yields the final sampling ranges.

The actual Monte Carlo sampling range used in this study is:

$$F_i \sim \text{Unif}(500, 6000), \quad \sigma_\alpha \sim \text{Unif}(0.05, 0.30), \quad \beta_0 \sim \text{Unif}(0.5, 1.5), \quad g \sim \text{Unif}(0, 0.1). \quad (2)$$

Here  $F_i$  is the loyal-fan scale,  $\sigma_\alpha$  controls fan activity volatility,  $\beta_0$  is the baseline performance weight, and  $g$  is the weekly growth rate. The stochastic terms are defined as:

$$\alpha_{ij} \sim \mathcal{N}(1, \sigma_\alpha), \quad \beta_{ij} \sim \mathcal{N}(\beta_0(1 + g \cdot (j - 1)), 0.05), \quad (3)$$

and a fixed floating vote pool  $M = 5000$  is normalized as:

$$P_{ij} = M \cdot \frac{w_{ij}}{\sum_k w_{kj}}, \quad w_{ij} = \begin{cases} (N_j + 1) - \text{Rank}_{ij}, & \text{Rank-based} \\ \text{Score}_{ij}, & \text{Percentage-based} \end{cases} \quad (4)$$

where  $N_j$  is the number of contestants remaining in week  $j$ .

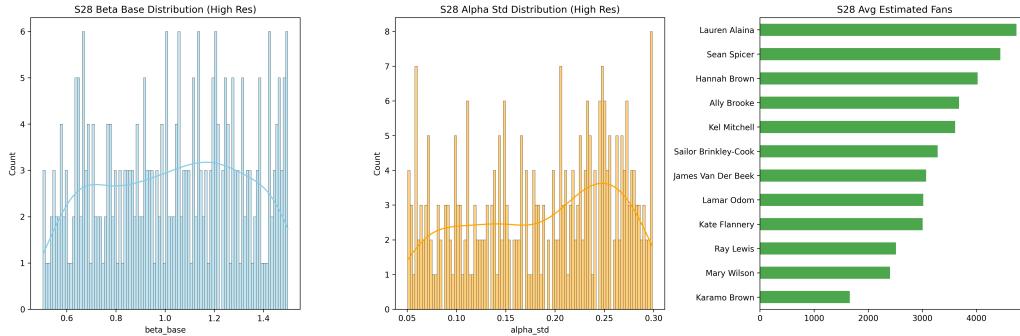


Figure 2: Example: solution space of  $\alpha$  and  $\beta$  and estimated fan base for Season 28.

## 5.3 Consistency Metric: Reproducing Historical Eliminations

### 5.3.1 Definition of Consistency Score

We define the consistency score  $C_{score}$  as the frequency with which the predicted elimination exactly matches the actual elimination across all weeks:

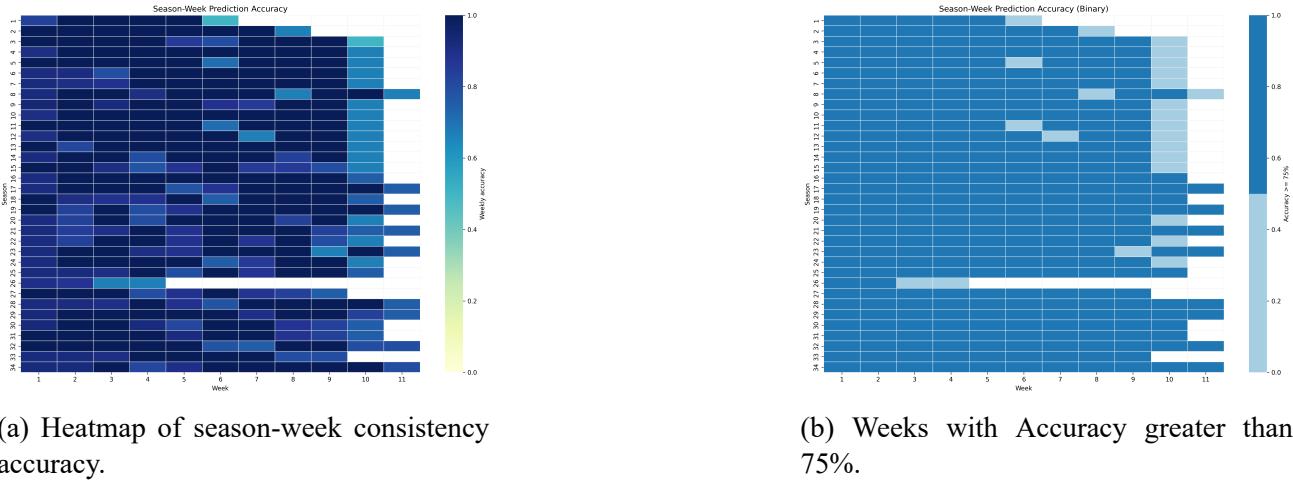
$$C_{score} = \frac{1}{T} \sum_{j=1}^T \mathbb{I}(\text{Predicted_Eliminated}_j = \text{Actual_Eliminated}_j) \quad (5)$$

where  $T$  is the number of weeks, and  $\mathbb{I}(\cdot)$  is the indicator function.

### 5.3.2 Consistency Check with Historical Data

Empirically, the model reproduces most historical eliminations.

Using Monte Carlo simulation, we retain parameters with Accuracy  $\geq 75\%$ . Results show strong cross-season consistency, validating  $V_{ij}$ ; distributions are shown below:



(a) Heatmap of season-week consistency accuracy.

(b) Weeks with Accuracy greater than 75%.

Figure 3: Consistency heatmap and Accuracy  $\approx 75\%$  week markers.

The left plot shows season-week accuracy (darker = closer match); the right marks weeks with Accuracy  $\geq 75\%$ . Most seasons show high consistency, supporting the model's capture of voting behavior.

### 5.3.3 External Consistency: Cross-Validation with Google Trends

To show that  $\widehat{V}_{ij}$  is not a purely mathematical fit, we use Google Trends search popularity  $G_{ij}$  as an external "public attention" reference for reconstructed "private votes." Using pytrends, we collect contestant trends during competition periods and compute the correlation  $r$  between  $\widehat{V}_{ij}$  and  $G_{ij}$ .

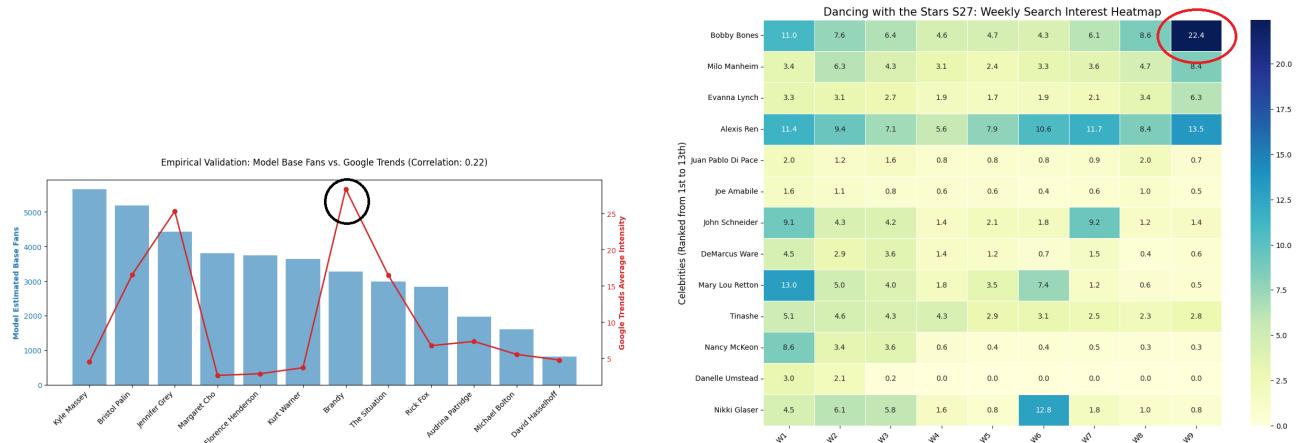
In most regular seasons,  $\widehat{V}_{ij}$  and  $G_{ij}$  are strongly correlated ( $r \geq 0.85$ ), indicating alignment with real-world attention.

In special seasons (e.g., S11 and S27), the model and search trends diverge ( $r_{11} = 0.22$ ,  $r_{27} = 0.319$ ), motivating a deeper consistency discussion below.

### 5.3.4 Deep Consistency Discussion: Special Cases in S11 and S27

Despite overall alignment, Season 11 (the Bristol Palin phenomenon) and Season 27 (the Bobby Bones phenomenon) diverge between estimated votes and search trends.

- **S11 controversy and attention denoising:** Bristol Palin's late-season search volume was extremely high (Figure 4a), yet the model did not inflate her estimated votes, indicating it filtered non-voting attention and captured effective votes.
- **S27 silent fan base identification:** Bobby Bones had modest search volume but very high estimated votes (Figure 4b), suggesting a "silent fan base" of radio audiences who vote actively but rarely search.



(a) S11: model-estimated votes vs. Google Trends.

(b) S27: rank residual analysis (Bobby Bones).

This "divergence" reflects the model's ability to correct bias in external data.

- For S11, the model removes "false prosperity" caused by controversy.
- For S27, the model captures "silent fan bases" not covered by search data.

These results show that the model not only aligns with Google Trends in normal seasons, but also reveals true voting dynamics in complex cases, demonstrating robustness in social behavior data.

### 5.3.5 Summary

By cross-validating with Google Trends, we show that the estimated fan votes  $\hat{V}_{ij}$  align with real-world attention in most seasons, demonstrating external validity.

At the same time, the model demonstrates strong denoising and identification ability, capturing true voting momentum rather than blindly following search volatility. This further validates the model's robustness and objectivity in complex social data.

## 5.4 Certainty Metric: Stability of Estimated Fan Votes

For estimated fan votes, we compute statistics from  $10^5$  simulations that meet  $Accuracy \geq 0.75$ .

- **Estimate stability:** We define certainty score  $S_{cert}$  based on the sample variance  $Var(\hat{V}_{ij})$ :

$$S_{cert}(i, j) = \frac{1}{1 + Var(\hat{V}_{ij})} \quad (6)$$

- **Confidence interval:** Using normal quantile  $z_{\alpha/2}$  to compute the 95% CI:

$$CI_{95\%} = \bar{V}_{ij} \pm z_{0.025} \cdot \frac{\sigma_{ij}}{\sqrt{n_{valid}}} \quad (7)$$

We obtain weekly fan vote estimates  $\hat{V}_{ij}$  and uncertainty measure  $S_{cert}(i, j)$ . These help identify which contestants have stable estimates and which exhibit large uncertainty, providing a foundation for later preference trade-off analysis.

## 6 Preference Trade-Off and Anti-Volatility: Evaluating DWTS Scoring Rules

### 6.1 Research Objectives and Problem Definition

Based on the estimated fan votes from Question 1, this section systematically compares two official aggregation rules. Our core objectives are:

- (1) How the two rules differ across seasons and their directional bias;
- (2) Whether rule choice changes outcomes for typical controversial contestants, and whether a "Bottom-2 Judge Save" mechanism mitigates controversy;
- (3) Rule recommendations based on partiality and stability metrics.

### 6.2 Metric Construction

We input weekly judges' total scores and estimated fan votes into both aggregation rules to compute combined rankings and predicted eliminations. We then construct two metrics:

- **Partiality coefficient  $I$ :** Measures whether final ranking is closer to fan ranking or judge ranking:

$$I = \frac{\text{Distance}(\text{Final Rank, Judge Rank})}{\text{Distance}(\text{Final Rank, Fan Rank})}. \quad (8)$$

When  $I > 1$ , final ranking is closer to fan ranking (fan-leaning). When  $I < 1$ , it is closer to judge ranking (judge-leaning).  $I = 1$  indicates balance.

- **Stability rate  $S$ :** The probability that elimination results remain unchanged under small random perturbations to fan votes. This measures robustness to short-term vote noise; higher  $S$  indicates more stable outcomes.

This process is equivalent to "parallel replay" of historical seasons to compare outcomes under the same data conditions.

### 6.3 Cross-Season Comparison: Overall Differences and Partiality

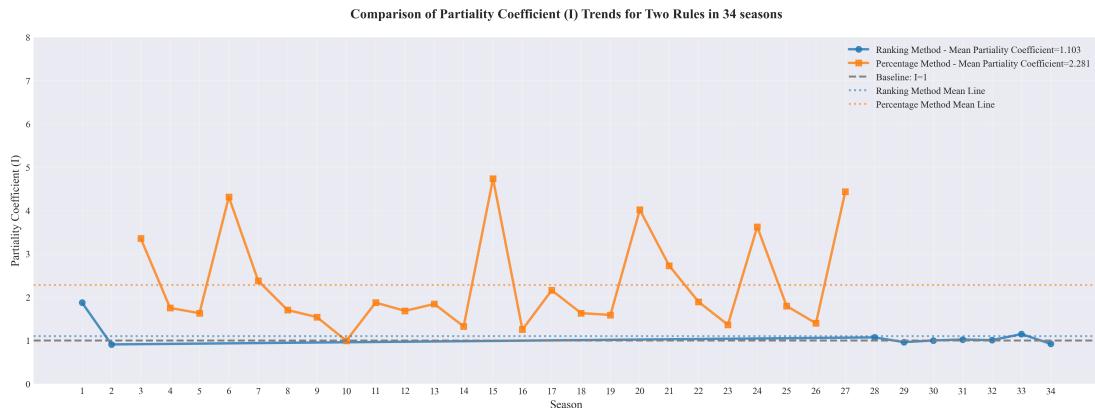


Figure 5: Partiality coefficient  $I$  across seasons.

In Figure 5, the blue line is the rank-based rule (mean 1.103) and the orange line is the percentage-based rule (mean 2.281); the dashed line is the  $I = 1$  balance baseline, and the dotted lines mark method means.

Across seasons, the percentage-based rule has higher  $I$  (stronger fan bias), while the rank-based rule stays nearer 1 and is more balanced. Both rules are highly stable; differences mainly reflect bias rather than randomness.

Table 2: Summary of cross-season comparison

Metric	Rank-based	Percentage-based
$I$ season mean	1.103	2.281
Seasons with $I_{\text{percent}} > I_{\text{rank}}$	33/34	33/34
Counterexample season	$I_{\text{rank}} = 1.055$	$I_{\text{percent}} = 0.994$
Mean stability $S$	1.02	1.00

**Summary:** The percentage-based rule is more fan-biased; both rules are highly stable.

## 6.4 Rule Sensitivity for Controversial Contestants and Judge Save Mechanism

### 6.4.1 Controversial Samples and Criteria

Let contestant  $i$  in week  $j$  have judge rank  $R_{ij}^{(J)}$  and fan rank  $R_{ij}^{(F)}$ . Define rank difference:

$$\Delta R_{ij} = R_{ij}^{(J)} - R_{ij}^{(F)}. \quad (9)$$

Define a "conflict week" threshold using data quantiles:

$$\tau = Q_{0.90}(|\Delta R_{ij}|), \quad \mathbb{I}_{ij} = \mathbb{I}(|\Delta R_{ij}| \geq \tau). \quad (10)$$

Define the number of conflict weeks in a season:

$$K_i = \sum_{j=1}^T \mathbb{I}_{ij}. \quad (11)$$

Introduce a structural conflict indicator:

$$\mathbb{S}_{ij} = \mathbb{I}(Score_{ij} \leq Q_{0.25}(Score_{\cdot j})) \cdot \mathbb{I}(Vote_{ij} \geq Q_{0.75}(Vote_{\cdot j})). \quad (12)$$

Define controversial samples as

$$\mathbb{C}_i = \mathbb{I}(K_i \geq k_0) \cdot \mathbb{I}\left(\sum_{j=1}^T \mathbb{S}_{ij} \geq 1\right), \quad k_0 = 2. \quad (13)$$

where  $Q_p(\cdot)$  is the  $p$ -quantile.

Table 3: Definitions for controversial sample indicators

Symbol	Meaning
$ \Delta R_{ij} $	Deviation between judge and fan rankings.
$\tau$	90% quantile of $ \Delta R_{ij} $ ; defines conflict weeks.
$K_i$	Number of conflict weeks for contestant $i$ .
$S_{ij}$	Structural conflict: low judge scores (bottom quartile) and high fan votes (top quartile).
$C_i = 1$	Contestant satisfies frequent and structural conflict.

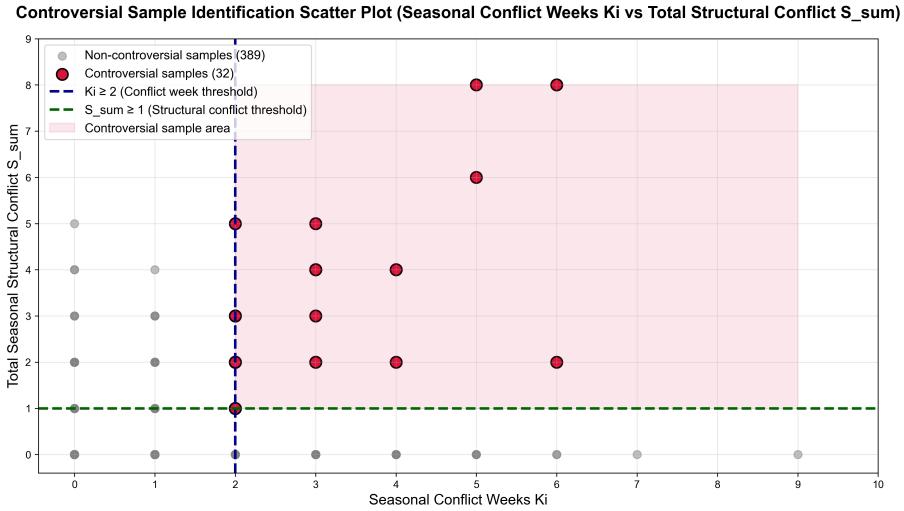
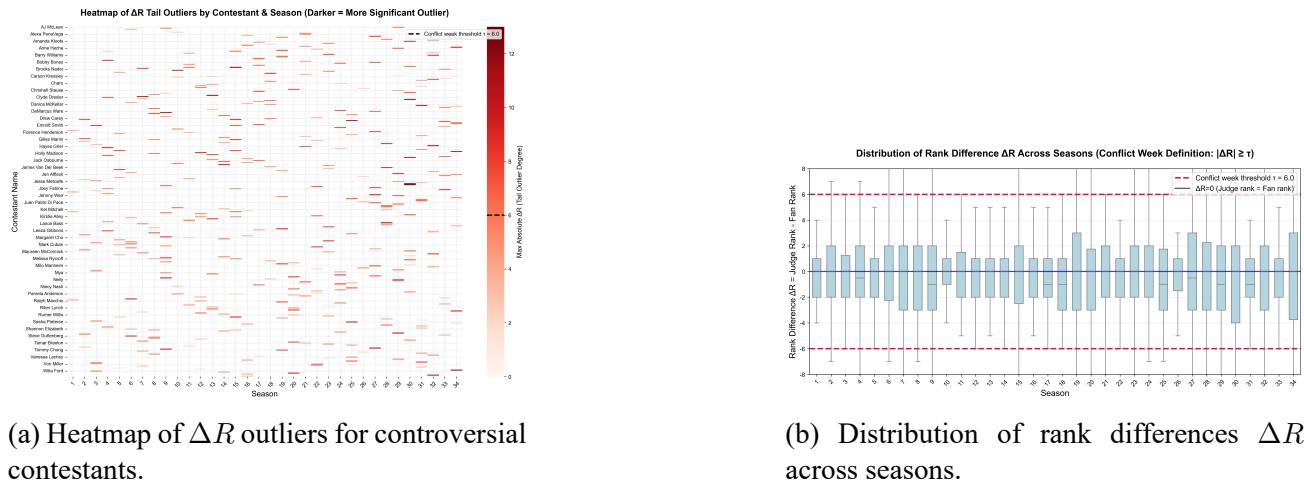
Figure 6: Scatter plot for controversial sample identification: conflict weeks  $K_i$  vs. cumulative structural conflicts  $\sum_j S_{ij}$ .(a) Heatmap of  $\Delta R$  outliers for controversial contestants.(b) Distribution of rank differences  $\Delta R$  across seasons.

Figure 7: Conflict strength analysis: heatmap shows individual outliers; distribution shows season-level summary.

#### 6.4.2 Rule Sensitivity Replay Results (Rank-based vs Percentage-based)

For each controversial sample's season, we perform rule replay: under the same judge scores and estimated fan votes, we compute weekly combined ranks by rank-based and percentage-based rules, and record final rank and elimination week changes.

We define:

- **Rank change magnitude:**  $\Delta P = P_{\text{rank}} - P_{\text{percent}}$ .
- **Survival week change:**  $\Delta W = W_{\text{rank}} - W_{\text{percent}}$ .
- **Controversial sample retention rate:** proportion of controversial samples reaching semifinal/final.

Aggregated replay results show that the average partiality coefficient  $I$  under the percentage-based rule is 4.234, significantly higher than 1.874 for the rank-based rule.

This reveals a fundamental difference: the percentage rule over-rewards fan votes, diluting judges' corrective power.

The following table summarizes  $\Delta P$  and  $\Delta W$  for each sample.

Table 4: Changes in final rank and survival weeks for controversial samples under two rules.

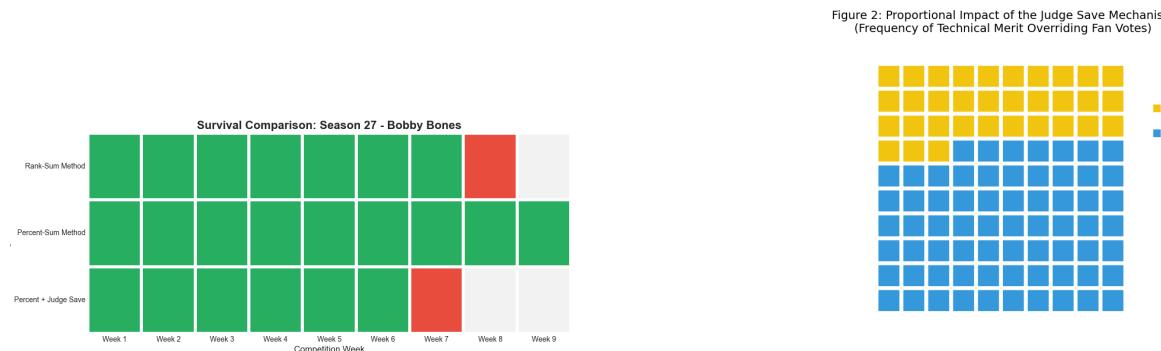
Season	Week	Sample	$\Delta P$	$\Delta W$	Season	Week	Sample	$\Delta P$	$\Delta W$
2	5	Jerry Rice	-1	-1	30	1	Iman Shumpert	3	1
4	6	Billy Ray	-2	-3	30	3	Cody Rigsby	3	3
11	2	Bristol Palin	-2	-1	15	1	Bristol Palin	2	3
27	3	Bobby Bones	-1	-2	32	2	Mauricio Umansky	2	1

#### 6.4.3 Judge Save Mechanism Simulation and Impact

To test the "Bottom-2 Judge Save" mechanism, we modify the replay process: identify the Bottom-2 by combined rank, save the higher judge score, and eliminate the lower. This acts as a professional safety valve under extreme fan bias.

Results show Judge Save reduces abnormal survival of controversial samples: fewer reach late rounds, average survival weeks fall; meanwhile, high-judge/low-fan contestants are more likely to be retained.

Mechanistically, Judge Save re-splits the one-dimensional combined result. Because the percentage rule has a strong long-tail effect (extremely high fan votes can offset very low judge scores), the Bottom-2 duel adds a high-pass filter, enforcing a professional floor and preventing low-score contestants from advancing solely on popularity.



(a) S27 Bobby Bones survival under three rules.

(b) Illustration of Judge Save impact.

Figure 8: Side-by-side: survival paths and Judge Save impact.

#### 6.4.4 Extreme Scenario Validation: Season Simulation Summary (S11)

To further illustrate the mechanism, we create a hypothetical celebrity in Season 11 with consistently lowest judge scores but highest fan votes. Holding other contestants fixed, we simulate survival under three rules; results are shown in Figure 9.

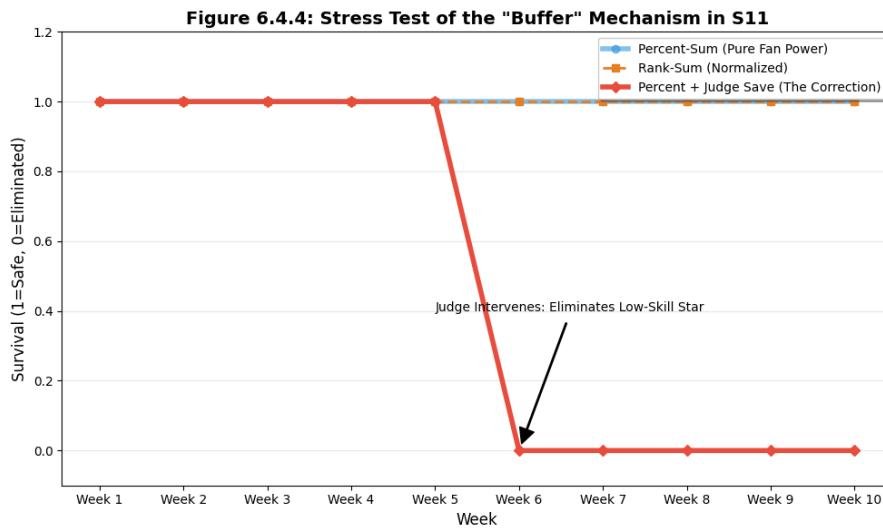


Figure 9: Stress Test: Survival Trajectories of a "High-Popularity, Low-Skill" Virtual Contestant under Different Rules (Season 11)

Under both pure rank-based and percentage-based rules, this virtual star survives to the end, demonstrating the power of fan votes. When Judge Save is introduced, the contestant is eliminated by Week 5, showing the professional floor effect.

Thus, we verify the buffering role of Judge Save.

#### 6.4.5 Implications for Future Rules

See below.

## 7 Decoding Success: What Makes a Winning Couple?

### 7.1 Data Processing and Standardization

We use the estimated fan votes  $\hat{V}_{ij}$  and judge scores  $JudgeScore_{ij}$  from Question 1, standardized into z-scores at the season-week level to ensure comparability.

### 7.2 Feature Engineering

We build a multi-dimensional feature set covering contestant attributes, regional background, and partner teaching ability. All features are used to predict two standardized targets: `judges_score_z` and `fan_votes_z`.

#### 7.2.1 Contestant Attribute Features

- Continuous variables like `celebrity_age_during_season` were retained.
- Categorical variables, such as the 30+ occupation types, were consolidated by retaining the top 10 categories by sample size and grouping the remainder into `Other` to reduce sparsity.

#### 7.2.2 Regional Background Features

Regional effects may influence competition via professional resources and local fan support. We build a **three-tier regional system**:

- **US state level:** For contestants with `celebrity_homecountry/region = "United States"`, extract `celebrity_homestate`. Keep top 15 states; merge others as `Other US`.
- **International countries:** For non-US contestants, keep top 3 countries (e.g., England, Australia, Canada), merge others as `Other International`.
- **Final variable:** Create `Region_Detailed` with about 20 categories (15 states + 3 countries + `Other US` + `Other International`).

This design avoids overly coarse US/non-US splitting while preventing sparse categories.

#### 7.2.3 Partner Ability Features

We quantify partner effects through teaching ability, i.e., how quickly contestants improve during the season.

1. **Growth trajectory modeling:** For each contestant, fit a linear regression with week as the predictor and Z-scored judge/fan scores as the response:

$$Z_{\text{score}}^{(i)} = \alpha + \beta \cdot \text{Week} + \epsilon$$

The slope  $\beta$  is the improvement rate (`judge_improvement_slope`, `fan_improvement_slope`).

**2. Partner-level aggregation:** For each partner, compute the average improvement rate and average relative performance across all assigned contestants:

- Mean improvement:  $\bar{\beta}_{\text{partner}} = \frac{1}{n} \sum_{i=1}^n \beta_i$
- Mean performance:  $\bar{Z}_{\text{partner}} = \frac{1}{n} \sum_{i=1}^n \bar{Z}_i$

Generate avg\_judge\_improvement, avg\_fan\_improvement, avg\_judge\_performance, avg\_fan\_perf

This feature system distinguishes "high-baseline" partners from "coach-type" partners.

#### 7.2.4 Categorical Encoding and Final Feature Matrix

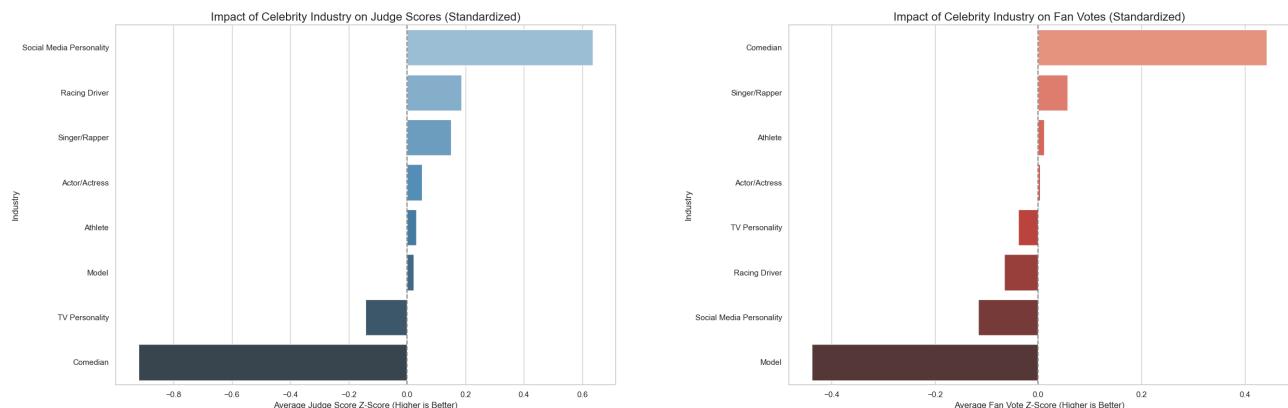
We one-hot encode Industry\_Group and Region\_Detailed, combine with continuous variables (age, season, week) to form the final feature matrix  $\mathbf{X}$  (about  $n \times 33$ ), and train two separate random forest regressors.

### 7.3 Impact Analysis of Contestant Attributes

To avoid scale differences across seasons/weeks, we use Z-scored judge and fan results. For categorical variables (occupation and region), we filter small sample categories.

#### 7.3.1 Occupation Effects

Occupation analysis reveals a divergence in how judges and fans evaluate contestants. Stage-performance occupations (e.g., singers, dancers) consistently received higher judge scores, whereas entertainers like comedians and reality TV stars garnered significantly stronger fan support relative to their technical scores.



(a) Occupation effects on judge scores (standardized).

(b) Occupation effects on fan votes (standardized).

Figure 10: Comparison of occupation effects on judge scores and fan votes.

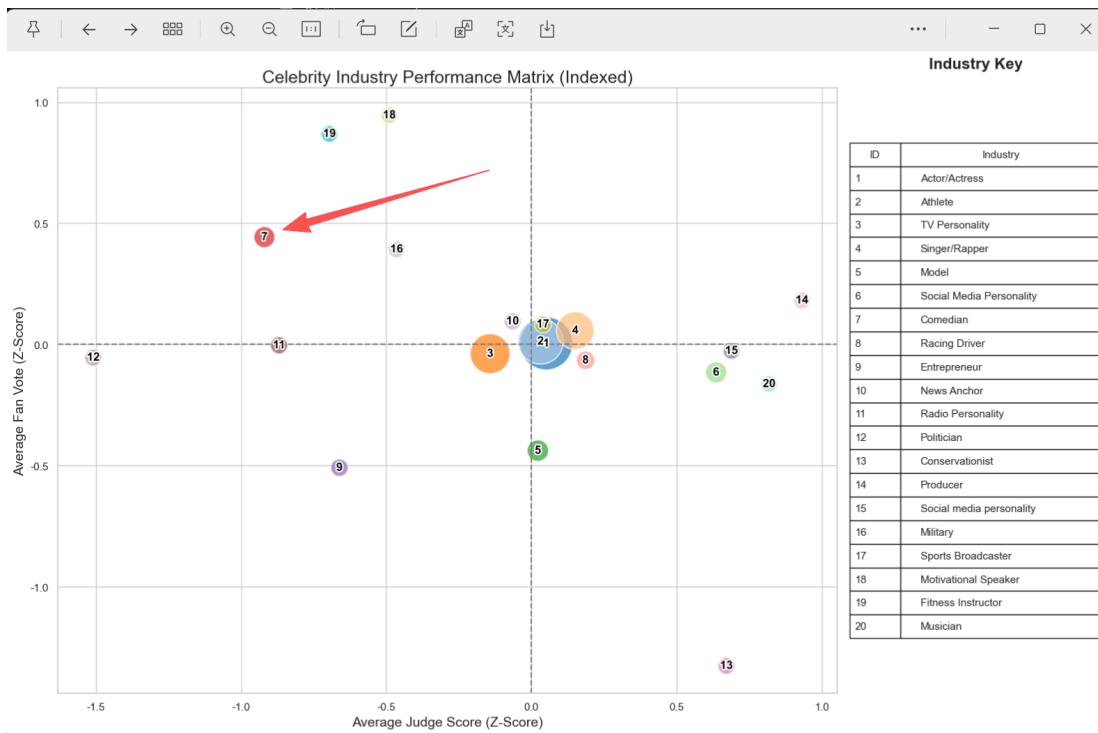


Figure 11: Occupation impact map: judge vs fan dimensions.

Two typical examples: comedians perform well in fan votes but relatively weaker in judge scores, reflecting entertainment strength but technical limits; musicians show higher professional scores but lower fan support. This suggests occupation affects audience expectations and skill profiles.

### 7.3.2 Regional Background Effects

Regional background exhibited a stronger influence on fan votes than on judge scores (Table 5). We observed a “regional mobilization effect” in fan voting, where contestants from smaller or tightly-knit states (e.g., Alaska, Delaware) achieved disproportionately high average vote counts, suggesting concentrated local support. The top-10 lists overlap only partially (e.g., USA-Georgia, USA-Nevada, USA-Hawaii, France), indicating structural divergence between professional and fan channels rather than simple consistency.

Treating U.S. contestants as a single group hides large internal differences. By contrast, contestants from large states like California or New York do not appear in the top 10 for fan votes and show no strong “home advantage.”

Violin plots further show that judge scores are more concentrated with small high-score tails, while fan votes have stronger high-end tails. This supports the mechanism that regional identity affects fan voting more, while professional performance depends on training resources and experience.

Table 5: Top 10 regions by mean judge score and mean fan votes.

Region (Judges)	Mean Score	Region (Fans)	Mean Votes
USA-Minnesota	33.42	USA-Alaska	7951.86
Russia	32.95	USA-Delaware	7545.50
USA-Colorado	32.92	France	7386.47
USA-Nevada	30.93	USA-Maine	7277.06
USA-Michigan	29.99	USA-Mississippi	6937.13
USA-Hawaii	29.19	USA-Georgia	6863.58
Australia	28.30	Canada	6806.09
France	28.29	USA-Iowa	6772.70
USA-Georgia	28.07	USA-Hawaii	6719.27
USA-Ohio	27.65	USA-Nevada	6706.17

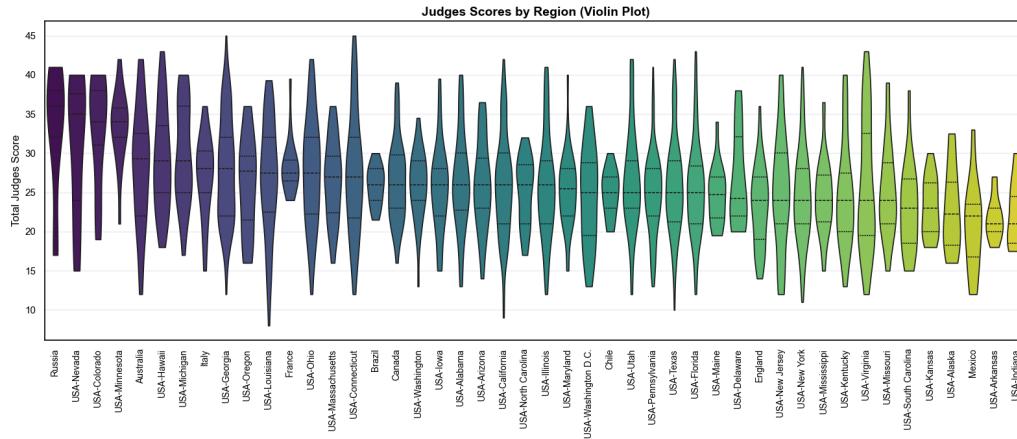


Figure 12: Distribution of judge scores by region (violin plot).

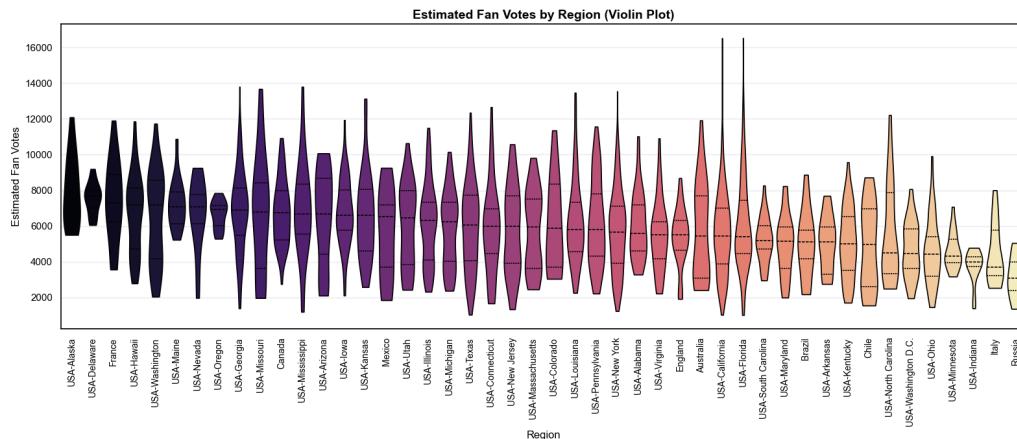


Figure 13: Distribution of fan votes by region (violin plot).

### 7.3.3 Age Effects

The relationship between age and scores was weakly negative and non-linear. Judge scores showed a slightly stronger correlation with age ( $r = -0.302$ ) than fan votes ( $r = -0.172$ ), indicating that technical assessment is more sensitive to this factor. This relationship is likely confounded by occupation, as different professions have distinct age distributions.

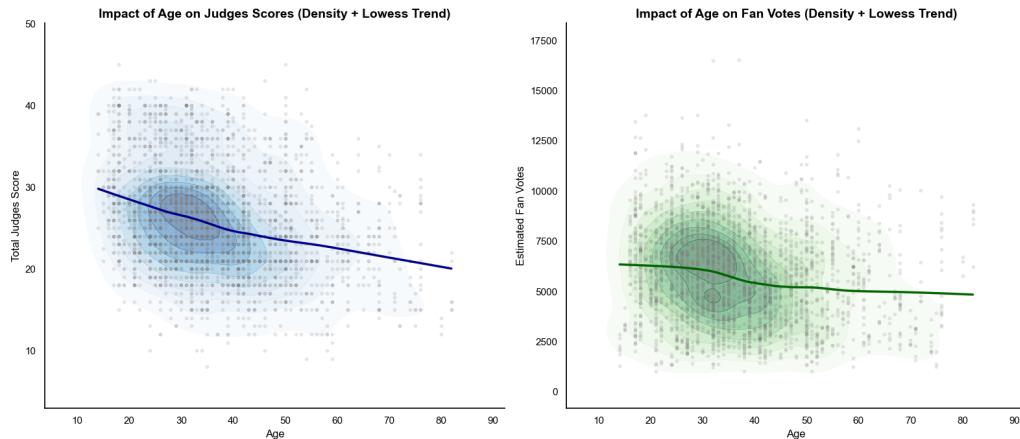


Figure 14: Age vs judge scores/fan votes: KDE + Lowess trends.

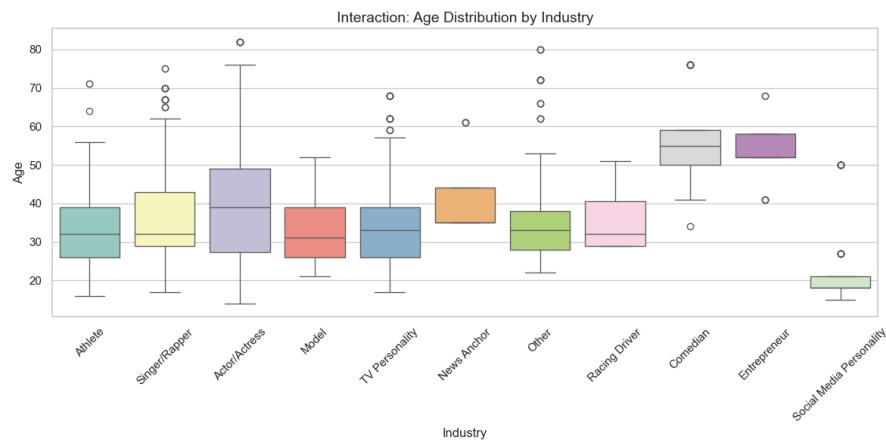


Figure 15: Age distributions across occupations (boxplots).

For example, athletes are generally younger, while entrepreneurs and comedians are older. Yet athletes do not necessarily receive higher fan votes, indicating fans consider more than dance skill. This aligns with the occupation analysis: fan votes reflect multi-dimensional appeal rather than technical performance alone.

## 7.4 Deep Modeling of Partner Effects

Partner (professional dancer) effects include both static influence on weekly performance and dynamic teaching ability over the season. To avoid attributing contestant ability to partner skill, we

decompose partner impact into two dimensions:

- (1) **Base Performance:** average relative performance during the season;
- (2) **Improvement Rate:** trend slope of relative performance across weeks, capturing teaching ability.

#### 7.4.1 Standardization and Improvement Slope

We standardize judge scores and estimated fan votes at the season-week level:

$$z_{i,s,w} = \frac{x_{i,s,w} - \mu_{s,w}}{\sigma_{s,w}}, \quad x \in \{JudgeScore, FanVote\}. \quad (14)$$

If  $\sigma_{s,w} = 0$  or missing, set  $z_{i,s,w} = 0$ .

For each contestant (within a season), fit:

$$z_{i,s,w} = a_{i,s} + \beta_{i,s} \cdot w + \varepsilon_{i,s,w}, \quad (15)$$

where slope  $\beta_{i,s}$  is the improvement rate (judge and fan channels). We only compute  $\beta_{i,s}$  for contestants with at least 3 weeks.

At the partner level (averaging across assigned contestants), we obtain:

- **Judge teaching ability:**  $\bar{\beta}_p^{(J)}$  and mean judge performance  $\bar{z}_p^{(J)}$ ;
- **Fan attraction ability:**  $\bar{\beta}_p^{(F)}$  and mean fan performance  $\bar{z}_p^{(F)}$ .

To ensure stability, we keep only partners who have coached at least 5 celebrities (28 senior partners, 55 total).

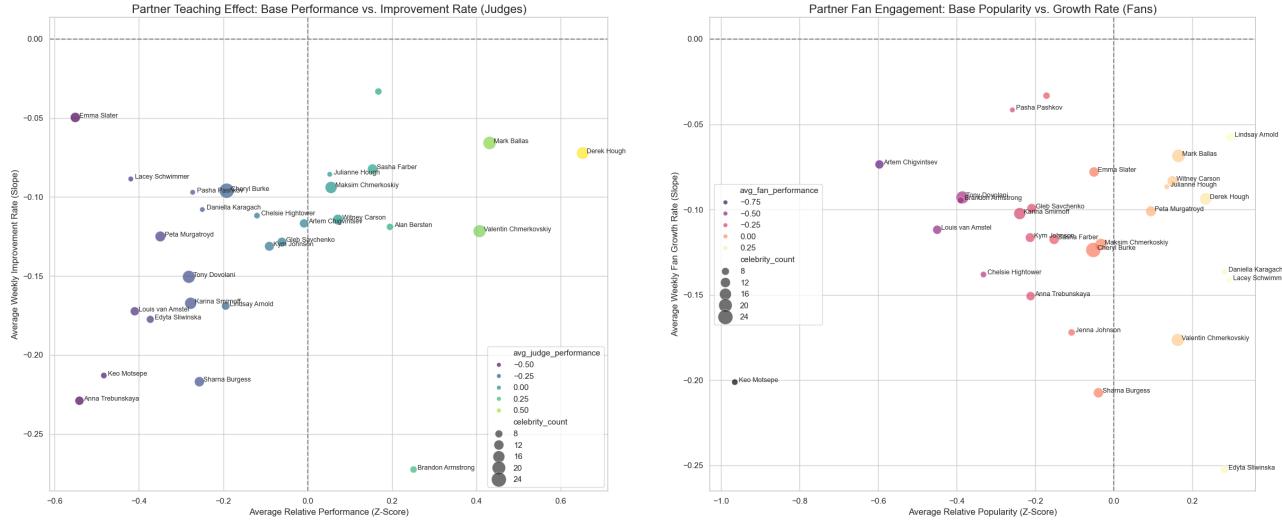
Because Z-scores are relative within each week and later weeks have stronger contestants, average  $\bar{\beta}$  is often negative (median about  $-0.118$  for judges and  $-0.107$  for fans). Thus we interpret stronger teaching as  $\bar{\beta}$  closer to 0 (slower decline or even positive growth).

#### 7.4.2 2D Scatter: Base Performance vs Improvement Rate

Figure 16 maps each senior partner to the 2D plane: x-axis is base performance (mean  $z$ ), y-axis is improvement rate (mean slope). The quadrants represent:

- Upper-right: **high baseline and strong growth.**
- Upper-left: **low baseline but strong growth** (coach-type).
- Lower-right: **high baseline, limited growth.**
- Lower-left: **weak baseline and weak growth.**

In the judge channel, base performance and improvement rate show positive correlation (about 0.330), suggesting that partners who deliver higher relative performance also maintain competitiveness. In the fan channel, the correlation is near zero ( $-0.006$ ), indicating that fan growth is driven more by narrative and exposure.



(a) Judge channel: base performance (mean  $z$ ) vs improvement rate (mean slope).

(b) Fan channel: base popularity (mean  $z$ ) vs attraction rate (mean slope).

Figure 16: Senior partners in the "base performance; growth" plane; point size indicates number of partners coached.

## 7.5 Multifactor Attribution: Random Forest Model

To address confounding factors in univariate analysis, we employed a Dual-Channel Random Forest that models judge and fan outcomes under a unified feature mapping. Let each observation be a season-week-contestant tuple  $t = (s, w, i)$  with feature vector  $\mathbf{x}_t$  (age, occupation, region) and targets  $y_t^{(J)}$  and  $y_t^{(F)}$ . We estimate

$$\mathbb{E}[y^{(\cdot)} \mid \mathbf{x}] \approx f^{(\cdot)}(\mathbf{x}), \quad (\cdot) \in \{J, F\}, \quad (16)$$

and compare the two channels under identical inputs.

### 7.5.1 Q3 Model Construction

We standardize judge scores and fan votes within each season to remove scale differences and interpret outcomes as relative advantage:

$$y_t^{(J)} = \frac{\text{Score}_t^{(J)} - \mu_s^{(J)}}{\sigma_s^{(J)}}, \quad y_t^{(F)} = \frac{\text{Vote}_t - \mu_s^{(F)}}{\sigma_s^{(F)}}, \quad (17)$$

with degenerate handling when  $\sigma = 0$ . Categorical variables are consolidated to reduce sparsity (e.g., top- $K$  occupations with an Other group), then encoded alongside season, week, and age to form  $\mathbf{x}_t$ . Using the same feature mapping for both channels, we fit separate random forest regressors:

$$\hat{f}^{(J)} = \arg \min_{f \in \mathcal{F}_{RF}} \frac{1}{N} \sum_{t=1}^N (y_t^{(J)} - f(\mathbf{x}_t))^2, \quad \hat{f}^{(F)} = \arg \min_{f \in \mathcal{F}_{RF}} \frac{1}{N} \sum_{t=1}^N (y_t^{(F)} - f(\mathbf{x}_t))^2, \quad (18)$$

with the ensemble predictor

$$\hat{f}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x}), \quad (19)$$

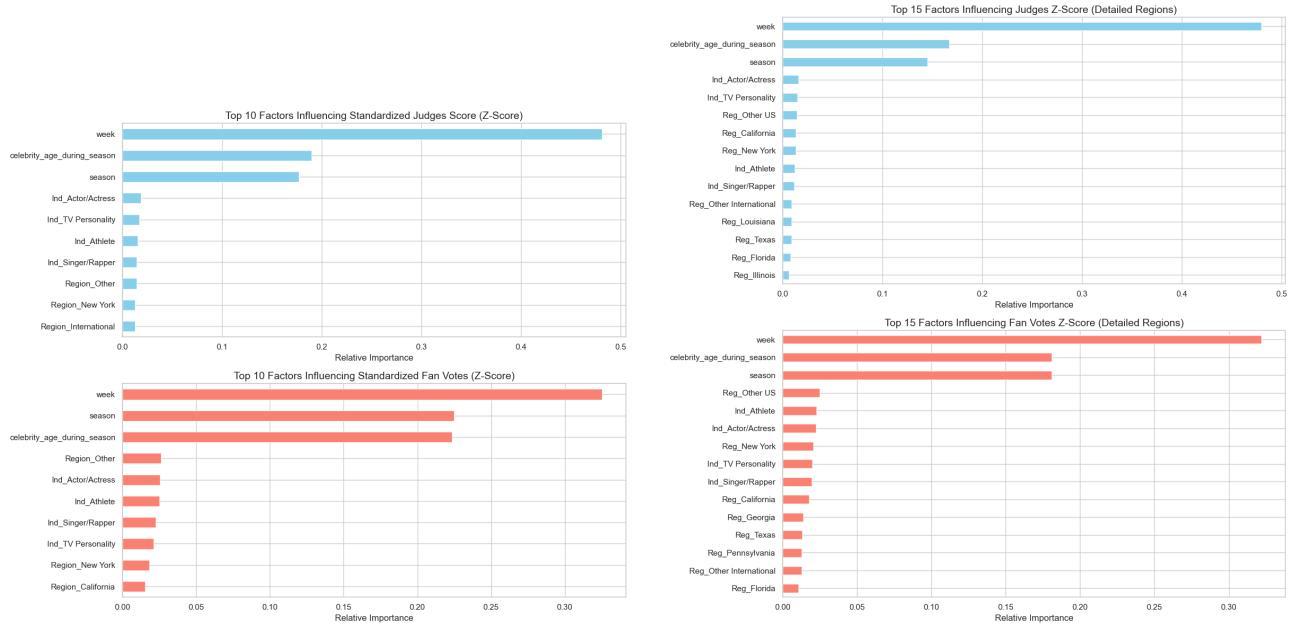
where  $T_b$  is the  $b$ -th tree and  $B = 100$ .

### 7.5.2 Feature Importance Ranking: Judges are More "Ability-Oriented," Fans More "Structure-Oriented"

Q3 outputs feature importance based on mean decrease in impurity (MDI). For feature  $k$ :

$$\text{Imp}_k = \frac{1}{B} \sum_{b=1}^B \sum_{v \in \mathcal{V}_b: \text{split}(v)=k} p(v) \Delta\text{MSE}(v), \quad (20)$$

where  $\mathcal{V}_b$  is the set of split nodes,  $p(v)$  is sample proportion at node  $v$ , and  $\Delta\text{MSE}(v)$  is the MSE decrease. Figure 17a compares the top features in both channels.



(a) Random forest feature importance: predicting season-level standardized judge scores and fan votes (Z-Score).

(b) Feature importance with detailed region granularity.

Figure 17: Side-by-side comparison of importances with detailed regions.

We further test whether finer regional granularity changes conclusions. If certain states rise in fan importance while judge channel remains insensitive, it supports "regional mobilization" in fan votes and judge neutrality.

**Section Summary:** Using a unified feature mapping  $\mathbf{x}$ , we fit  $\hat{f}^{(J)}$  and  $\hat{f}^{(F)}$  and compare MDI importances to contrast the drivers of judge vs fan channels, moving from univariate plots to multivariate structural interpretation.

## 8 Recommendations to the Committee

See "Recommendations to the Committee (Appendix)."

## 9 Strengths and Weaknesses

### 9.1 Strengths

- **Innovative method (Monte Carlo + parameter inversion + Google Trends):** We estimate fan votes via Monte Carlo inversion and validate with Google Trends, improving internal consistency, external alignment, and denoising. *Improves internal validity and external alignment.*
- **Multi-dimensional features and dual-channel random forest:** A multi-feature system with dual-channel RF separates judge skill signals from fan structure effects. *Separates judge skill signals from fan structure effects.*
- **30+30 dual-track linear scoring system (DTLSS):** A simple, interpretable rule that preserves audience appeal while enforcing a professional floor. *Balances audience appeal and professional fairness.*
- **Controversial sample screening and rule replay:** Judge Save experiments quantify bias and stability with clear conclusions. *Quantifies bias and stability to support conclusions.*

### 9.2 Weaknesses

- **Static fan-base assumption:** We hold base fans constant, omitting time-varying shocks. *May miss time-varying shocks.*
- **Parameter range depends on pilot runs:** "Wide-then-narrow" still relies on small pilots, introducing chance effects. *Pilot-based ranges may introduce chance effects.*
- **Feature dimensions can be expanded:** No social sentiment or style-fit variables yet, limiting explanatory power. *Add sentiment/engagement and style-fit variables.*

## 10 Conclusion

This paper examines the DWTS scoring tension between judges and fans. With no direct vote data, we analyze rule behavior and bias and propose improvements for fairness, interpretability, and engagement. The core finding is that controversies arise from rule amplification and should be corrected at the rule level.

Three conclusions follow: (1) judge and fan scores reflect different values, so fan influence should be constrained to avoid diluting technical evaluation; (2) both percentage and rank rules have endogenous bias and amplify extremes when popularity and skill diverge; (3) limited judge intervention or fan caps can improve fairness and stability without reducing participation. Overall conclusions:

- Judge and fan scores reflect different values; constrain fan influence to avoid diluting technical evaluation.
- Both percentage and rank rules have endogenous bias; when popularity and skill diverge, extreme results are amplified.
- Limited judge intervention or fan-score caps can improve fairness and stability without reducing participation.

Therefore, we propose a "dual-track linear scoring system" (30+30) that symmetrically integrates professional and popular channels, reduces implicit amplification, and improves interpretability. This approach applies to other expert-plus-public competitions; future work can expand participation and data while preserving the core structure for fairness, entertainment, and engagement.

## References

- [1] S. D'Angelo, T. B. Murphy & M. Alfo. "Latent Space Modeling of Multidimensional Networks with Application to the Exchange of Votes in Eurovision Song Contest." *arXiv preprint arXiv:1807.06517*, 2018.
- [2] M. Blangiardo & G. Baio. "Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models." *arXiv preprint arXiv:1310.3501*, 2013.
- [3] R. Fairstein, A. Lauz, K. Gal & R. Meir. "Modeling People's Voting Behavior with Poll Information." *arXiv preprint arXiv:1902.04118*, 2019.
- [4] L. Chen, P. Xu & D. Liu. "Experts versus the Crowd: A Comparison of Selection Mechanisms in Crowdsourcing Contests." *SSRN Electronic Journal*, 2015. DOI: 10.2139/ssrn.2631317.
- [5] D. Zhang. "Methods and Rules of Voting and Decision: A Literature Review." *Open Journal of Social Sciences*, vol. 8, no. 9, pp. 310–C326, 2020.
- [6] L. S. Shapley & M. Shubik. "A Method for Evaluating the Distribution of Power in a Committee System." *American Political Science Review*, vol. 48, no. 3, pp. 787–C792, 1954.
- [7] "Voting Matters." McDougall Trust, <https://www.mcdougall.org.uk/>.
- [8] "Probabilistic voting model." In *Voting Theory*, [https://en.wikipedia.org/wiki/Probabilistic\\_voting](https://en.wikipedia.org/wiki/Probabilistic_voting)
- [9] "Google Trends." Google LLC, <https://trends.google.com/>.
- [10] General Mills. "pytrends: Unofficial API for Google Trends." GitHub repository, <https://github.com/GeneralMills/pytrends>.

## Recommendations to the Committee (Appendix)

Based on our quantitative evaluation of DWTS scoring rules (rank-based and percentage-based), we find that complex weighting is a major source of audience confusion and controversy. To enhance entertainment while protecting professional standards, we recommend a **Dual-Track Linear Scoring System (DTLSS)**.

The design philosophy is "**decoupling and balancing**": separate professional scores and public votes into parallel tracks and linearly balance their weights.

### New Scoring: "30+30" Mode

We suggest replacing "percentage of votes" or "combined ranking" with a simple additive score. Both tracks have the same maximum (30 points):

- **Judges' Track:** Unchanged. Sum of three judges' scores, maximum 30. This is the **technical ceiling**.
- **Fans' Track:** Convert fan vote ranking directly into points, also maximum 30. This is the **popularity ceiling**.

### Computation Logic

Fan track score  $S_{Fan}$  depends only on fan ranking  $Rank_{Fan}$ :

$$S_{Fan} = S_{max} - \delta \times (Rank_{Fan} - 1) \quad (21)$$

where  $S_{max} = 30$  and  $\delta = 2$ . Thus: rank 1 gets 30, rank 2 gets 28, and so on.

Total score is the sum:

$$TotalScore = Score_{Judge} + Score_{Fan} \quad (22)$$

The contestant with the lowest total score is eliminated.

### Advantage 1: Maximum Audience Friendliness

Unlike opaque rules where "millions of votes offset how many judge points" is unclear, DTLSS is highly **interpretable** and TV-friendly:

- **Intuitive incentive:** "Each rank higher in fan votes adds 2 points." This clear feedback loop is more motivating than complex percentages.
- **Visual suspense in live broadcast:** Show fixed judge scores and dynamic fan scores on screen. Viewers can see scenarios like: "Contestant A trails by 2 points; if they outrank B in fan votes, they overtake." This enhances engagement.

### Advantage 2: Structural Balance

In Section 6.4, we show Season 27 controversy stems from unlimited fan influence under the percentage rule. The new system resolves this via **unit alignment** and explicit caps.

## 1. Influence Capping

No matter how large a fan base is, the fan track benefit is capped at 30. This limits marginal returns and prevents popularity from overpowering professional scores.

## 2. Mathematical Defense of a Professional Floor

Consider:

- **Scenario:** Contestant X has poor technique (judge score 15) but maximum popularity (fan score 30), total 45.
- **Comparison:** Contestant Y performs well (judge score 27) with moderate popularity (fan rank 3 gives 26), total 53.
- **Outcome:**  $53 > 45$ , so Y advances.

This shows that even extreme popularity cannot fully offset low technical scores. DTLSS mathematically prevents a severely unskilled contestant from winning, ensuring fairness.

## Summary

DTLSS does not reduce fan engagement; it converts abstract support into visible **points**. With simple addition, it balances professional evaluation and public preference, improving fairness, interpretability, and acceptance. We recommend a pilot in Season 35.

# AI Use Statement

## 1. Did you use AI tools in your work? Yes.

To enhance non-native language expressions in the course of academic paper writing, our research team has adopted AI technology. Before integrating any AI-assisted output into designated sections of the paper, we subject it to multiple layers of human review to ensure it meets academic standards. Moreover, to guarantee compliant use of the technology, we begin by establishing a comprehensive semantic framework and then guide AI-generated content through meticulously crafted prompts. Throughout this process, we strictly adhere to research ethics, ensuring there is no misuse of AI technology.

## 2. Which tools did you use and how?

- 1) **ChatGPT-5:** We used ChatGPT-5 to polish the language of the Introduction section.

### Query:

*Polish this paragraph in an empirical tone: Dancing with the Stars (DWTS) is a classic television program that blends professional dance competition with popular entertainment elements. In each season, celebrities from diverse fields partner with professional dancers to compete in an elimination-style contest through weekly dance performances. A contestant's ultimate fate is determined by two core dimensions: first, the professional scores awarded by judges based on dance technique, choreographic creativity, and other criteria; second, the audience votes that reflect public preferences.*

### Output:

Polished Version—Dancing with the Stars (DWTS) is a long-running television program that integrates professional dance competition with popular entertainment. In each season, celebrities from diverse professional backgrounds are paired with professional dancers and compete in an elimination-style format through weekly dance performances. A contestant's progression and eventual outcome in the competition are jointly determined by two core components: professional judges' scores, which evaluate dance technique, choreography, and artistic expression, and audience votes, which reflect public preferences.

- 2) **ChatGPT-5:** After completing Question 1, we used GPT to generate the code for consolidating all the data of Seasons 1 to 34 into a single CSV file.

### Query:

*I need to consolidate the estimated fan volume data for Seasons 1 to 34 into a single file, with the requirement to keep all original column names unchanged. Please provide the corresponding code.*

### Output (partial code):

```
if len(all_data) == 0:  
    raise ValueError("All files failed to read or no valid data exists")  
# Merge all DataFrames, ignore original indices, and reset a new continuous  
merged_df = pd.concat(all_data, ignore_index=True, sort=False)
```

```
# Step 4: Final data sorting and cleaning
# Sort by original Chinese columns: Season ; Week ; Contestant Name for
merged_df = merged_df.sort_values(by=["Season", "Week", "Name"], ignore_in
# (Optional) Remove duplicate rows (if duplicate data exists for the same
merged_df = merged_df.drop_duplicates(subset=["Season", "Week", "Name"], k
# Step 5: Save the merged result
merged_df.to_csv(Config.OUTPUT_FILE, index=False, encoding="utf-8-sig")
print(f"\n\u25b6Merging\u25bccompleted!\u25bcThe\u25bcfull-season\u25bcdata\u25bchas\u25bcbeen\u25bcsaved\u25bcto:\u25bc{Con
```

- 3) **Grammarly:** We used Grammarly to correct minor grammatical errors in the text.