

## 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 &amp; Justifications</b>	<b>3</b>
<b>3</b>	<b>Notations</b>	<b>4</b>
<b>4</b>	<b>Data Description and Processing</b>	<b>4</b>
4.1	Data Source and Original Structure	4
4.2	Data Cleaning and Feature Engineering	5
4.2.1	Data Reshaping and Filtering	5
4.2.2	Metric Calculation	5
4.2.3	Target Variable Extraction	6
4.3	Construction of the Stochastic Behavior Model and Vote Estimation Formula	6
4.3.1	Model Formulation	6
4.3.2	Parameter Range Selection	7
4.4	Consistency Metric: Reproducing Historical Eliminations	8
4.4.1	Definition of Consistency Score	8
4.4.2	Consistency Check with Historical Data	8
4.4.3	External Consistency: Cross-Validation with Google Trends	9
4.4.4	Deep Consistency Discussion: Special Cases in S11 and S27	9
4.4.5	Summary	10
4.5	Certainty Metric: Stability of Estimated Fan Votes	10
<b>5</b>	<b>Preference Trade-Off and Anti-Volatility: Evaluating DWTS Scoring Rules</b>	<b>11</b>
5.1	Research Objectives and Problem Definition	11
5.2	Metric Construction	11
5.3	Cross-Season Comparison: Overall Differences and Partiality	12
5.4	Rule Sensitivity for Controversial Contestants and Judge Save Mechanism	12
5.4.1	Controversial Samples and Criteria	12
5.4.2	Rule Sensitivity Replay Results (Rank-based vs Percentage-based)	15
5.4.3	Judge Save Mechanism Simulation and Impact	15
5.4.4	Extreme Scenario Validation: Season Simulation Summary (S11)	16
5.4.5	Implications for Future Rules	17
<b>6</b>	<b>Decoding Success: What Makes a Winning Couple?</b>	<b>17</b>
6.1	Research Objectives and Problem Definition	17
6.2	Data Processing and Standardization	17
6.3	Feature Engineering	17
6.3.1	Contestant Attribute Features	17

6.3.2	Regional Background Features . . . . .	18
6.3.3	Partner Ability Features . . . . .	18
6.3.4	Categorical Encoding and Final Feature Matrix . . . . .	18
6.4	Impact Analysis of Contestant Attributes . . . . .	18
6.4.1	Occupation Effects . . . . .	19
6.4.2	Regional Background Effects . . . . .	20
6.4.3	Age Effects . . . . .	22
6.5	Deep Modeling of Partner Effects . . . . .	23
6.5.1	Standardization and Improvement Slope . . . . .	23
6.5.2	2D Scatter: Base Performance vs Improvement Rate . . . . .	24
6.6	Multifactor Attribution: Random Forest Model . . . . .	25
6.6.1	Why a Multivariate Model? (Limits of Univariate Analysis) . . . . .	25
6.6.2	Q3 Code Logic and Model Construction . . . . .	25
6.6.3	Feature Importance Ranking: Judges are More "Ability-Oriented," Fans More "Structure-Oriented" . . . . .	26
6.7	Summary . . . . .	27
7	<b>Recommendations to the Committee</b> . . . . .	27
8	<b>Strengths and Weaknesses</b> . . . . .	28
8.1	Strengths . . . . .	28
8.2	Weaknesses . . . . .	28
9	<b>Conclusion</b> . . . . .	28

# 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.

In contrast to the objectively quantified judges' scores, audience voting behavior is inherently subjective and multifaceted. It is influenced not only by a contestant's weekly dance performance, but also by their pre-existing popularity, fan base, emotional appeal, and various socio-demographic factors. However, the show's organizers do not disclose the detailed weekly voting data for individual contestants, releasing only final rankings and elimination outcomes. This lack of transparency creates a fundamental challenge for the quantitative analysis and systematic understanding of the voting mechanism.

Throughout the show's history, persistent debates have emerged regarding the fairness and rationality of the DWTS voting system. On one hand, judges' scores are generally regarded as an objective measure of professional dance quality and a cornerstone of the competition's credibility. On the other hand, the subjectivity inherent in audience voting may allow contestants with strong popularity but relatively weaker technical performance to remain in the competition, while technically skilled contestants with smaller fan bases are eliminated prematurely. This tension between professional merit and public popularity has sparked ongoing discussions about the appropriateness of the current voting framework.

Against this backdrop, we aim to develop a rigorous mathematical framework to characterize audience voting behavior and to systematically examine how different voting mechanisms influence competition outcomes. Importantly, this analysis must be conducted in the absence of actual fan vote data, relying instead on observable information such as judges' scores, season rankings, and elimination records. Specifically, the problem focuses on the following three core aspects:

- How to develop a scientifically sound approach to estimate unobservable fan vote data and evaluate its consistency with observed elimination results?
- Whether alternative voting rules (e.g., rank-based versus percent-based methods) lead to significantly different competition trajectories and final outcomes?
- What are the strengths and limitations of the existing voting mechanism from the dual perspectives of professional fairness and competitive appeal?

## 1.2 Restatement of the Problem

In the absence of publicly available weekly fan vote data released by the official organizers of *Dancing with the Stars* (DWTS), this study conducts a systematic investigation based on accessible competition data, including judges' professional scores, season rankings, elimination results, and contestants' basic information. Centered on the framework of fan vote estimation "C rule impact analysis" "C mechanism

evaluation, this problem requires the development of appropriate mathematical models to estimate contestants' relative fan vote levels. Furthermore, the problem calls for an analysis of how different voting mechanisms influence the competition process and final outcomes, with the goal of evaluating the existing voting system from the perspectives of professional fairness and competitive rationality, and proposing a more equitable elimination rule.

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, our study aims to provide quantitative insights and decision-making support for the design of voting mechanisms in similar competitive entertainment programs.

### 1.3 Our Work

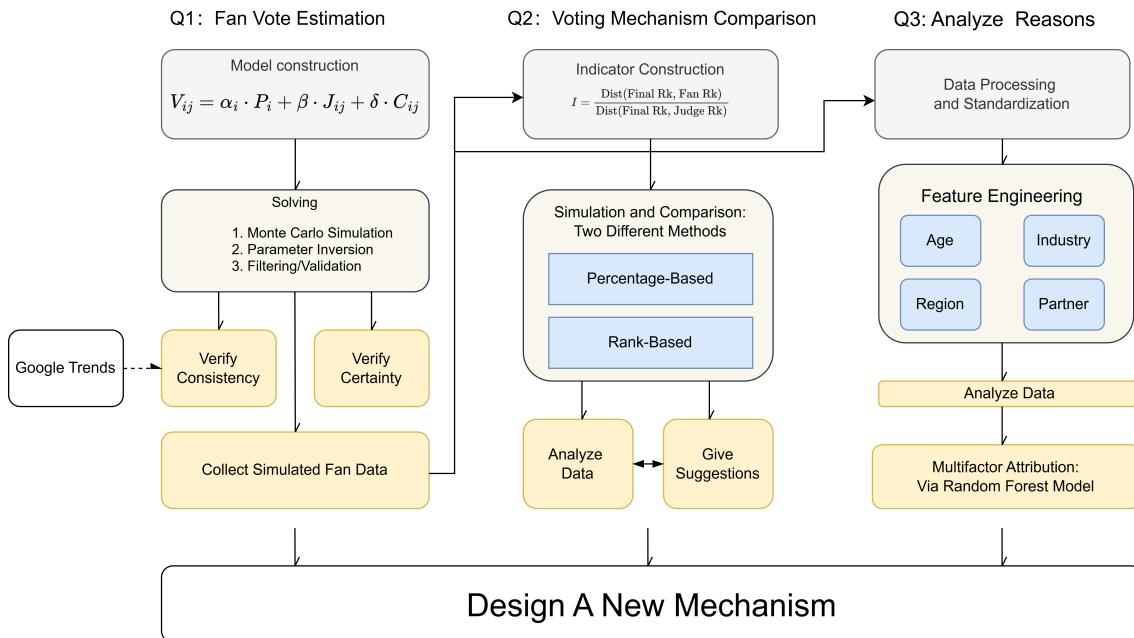


Figure 1: The overall framework of our work.

## 2 Assumptions & Justifications

To construct a rigorous mathematical framework for the *Dancing with the Stars* (DWTS) voting problem, we formulate the following key assumptions based on the statistical characteristics of the data and the inherent logic of the competition. These assumptions serve as the foundation for the subsequent Monte Carlo simulation, parameter inversion, and feature engineering.

- **Assumption 1: Google Trends data serves as a valid proxy for Public Attention.**

*Justification:* Official fan voting data is unavailable. We employ search volume to gauge the public's active intent to seek contestant information. A significant intrinsic correlation exists between this "public domain popularity" and "private voting behavior." In the external consistency check (Section 5.2.3), the correlation coefficient between the model-estimated votes and search trends exceeds 0.85 in most seasons. This evidence validates the effectiveness of using search trends as a proxy variable.

- **Assumption 2: Fan Votes are composed of "Base Votes" and "Performance Votes."**

*Justification:* Voting motivations stem from two distinct sources. The first source is the contestant's pre-existing "Base Fans." The voting behavior of this group exhibits high stickiness and stability. The second source is "Floating Voters" attracted by the weekly dance performance. Their voting behavior is highly fluid. This assumption constitutes the theoretical basis for the vote estimation model  $V_{ij} = \alpha_i P_i + \beta J_{ij}$  in Section 5.1.1.

- **Assumption 3: Base Popularity is constant throughout the season.**

*Justification:* The DWTS season cycle is relatively short (typically 10–12 weeks). The loyalty of the core fan base remains relatively solidified and does not fluctuate significantly in the short term. We attribute the influx of new supporters driven by excellent performance to dynamic "Performance Votes" or cumulative effects rather than changes in the base popularity. This assumption significantly reduces the model parameter space. Consequently, it facilitates large-scale Monte Carlo simulation and parameter inversion.

- **Assumption 4: The "Floating Vote" pool is fixed and distributed based on Relative Performance.**

*Justification:* We normalize the "floating vote" pool to a fixed total amount ( $M = 5000$ ) to ensure cross-season data comparability. Furthermore, audience voting decisions rely more on relative judgments of "who danced better" than on absolute scores. Therefore, we adopt a non-linear mapping in Section 5.1.3. This approach transforms judges' absolute scores into relative ranking weights. It effectively eliminates the influence of inconsistent scoring scales across different judges.

- **Assumption 5: Standardization eliminates cross-season biases.**

*Justification:* Significant differences exist in the strictness (mean and variance) of judges' scoring across different seasons. We transform all scores into a measure of "relative advantage over the weekly average" via the transformation  $z = (x - \mu)/\sigma$ . This standardization enables the aggregation of data from over 30 seasons in Section 7. It facilitates unified feature engineering and partner effect evaluation.

- **Assumption 6: Contestant improvement follows a linear trend within a season.**

*Justification:* Learning curves possess inherent complexity. However, linear regression sufficiently captures the primary characteristics of contestant ability changes within the limited competition weeks (typically fewer than 12 data points). We explicitly utilize this assumption in Section 7.5. We quantify the "teaching improvement rate" of partners by calculating the slope of the linear trend. This metric effectively distinguishes "high-baseline" partners from "coach-type" partners.

- **Assumption 7: The competition rules are strictly followed.**

*Justification:* The logic of parameter inversion in this study relies on treating historical elimination results as the "ground truth" for parameter screening. The parameter space filtered under the constraint of Accuracy  $\geq 75\%$  possesses realistic explanatory power only if the rules are strictly enforced. This condition ensures the model accurately reflects the audience's voting logic.

### 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 comprises historical records from Seasons 1 to 34 of *Dancing with the Stars*. The data is originally stored in a "wide format," where each row corresponds to a specific celebrity contestant. The columns include demographic information (e.g., age, industry) and detailed scores from each judge across weeks 1 through 11 (denoted as `weekX_judgeY_score`). Due to variations in the number of

judges (typically 3 or 4) and the progressive elimination of contestants, the raw dataset contains a significant amount of missing values (NaN) and zero entries, indicating non-participation.

## 4.2 Data Cleaning and Feature Engineering

To facilitate time-series analysis and survival modeling, we transformed the dataset from a contestant-centric wide format into a contestant-week long format. The specific data processing pipeline is described below:

### 4.2.1 Data Reshaping and Filtering

We unpivoted the weekly score columns so that each observation represents a contestant's performance in a specific week. Entries with missing values or zero scores<sup>a</sup> indicating that the contestant had already been eliminated or did not compete<sup>b</sup>were removed to ensure data integrity.

### 4.2.2 Metric Calculation

To mitigate the inconsistency caused by the varying number of judges across seasons, we engineered several statistical features. Let  $S_{i,t,j}$  denote the score given by judge  $j$  to contestant  $i$  in week  $t$ , and let  $J_t$  be the number of judges in that week.

- **Total Judge Score ( $T_{i,t}$ ):** The sum of scores received by contestant  $i$  in week  $t$ .

$$T_{i,t} = \sum_{j=1}^{J_t} S_{i,t,j} \quad (1)$$

- **Average Judge Score ( $A_{i,t}$ ):** The arithmetic mean of the judges' scores, providing a scale-invariant measure of performance.

$$A_{i,t} = \frac{1}{J_t} \sum_{j=1}^{J_t} S_{i,t,j} \quad (2)$$

- **Judge Percentage ( $P_{i,t}$ ):** To quantify a contestant's relative competitiveness within the cohort for a given week, we calculated the share of total votes. Let  $\mathcal{C}_t$  be the set of all active contestants in week  $t$ .

$$P_{i,t} = \frac{T_{i,t}}{\sum_{k \in \mathcal{C}_t} T_{k,t}} \quad (3)$$

This metric ( $P_{i,t}$ ) normalizes the scores, effectively handling variations in both the number of judges and the number of remaining contestants.

### 4.2.3 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.

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

### 4.3.1 Model Formulation

To quantify the missing audience votes in the survey data, this study constructs a multi-dimensional stochastic behavior model. The model assumes that audience voting decisions are not random, but are driven by a contestant's background popularity, current performance, and cumulative reputation within the season.

For any season, we decompose the unobserved "audience votes" into the following 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:

$$V_{ij} = \alpha_i \cdot P_i + \beta \cdot J_{ij} + \delta \cdot C_{ij} \quad (4)$$

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.

### 4.3.2 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:

$$F_i \sim \text{Unif}(500, 6000), \quad \sigma_\alpha \sim \text{Unif}(0.05, 0.30), \quad \beta_0 \sim \text{Unif}(0.5, 1.5), \quad g \sim \text{Unif}(0, 0.1). \quad (5)$$

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:

$$\alpha_{ij} \sim \mathcal{N}(1, \sigma_\alpha), \quad \beta_{ij} \sim \mathcal{N}(\beta_0(1 + g \cdot (j - 1)), 0.05), \quad (6)$$

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

$$P_{ij} = M \cdot \frac{w_{ij}}{\sum_k w_{kj}}, \quad w_{ij} = \begin{cases} (N_j + 1) - \text{Rank}_{ij}, & \text{Rank-based} \\ \text{Score}_{ij}, & \text{Percentage-based} \end{cases} \quad (7)$$

where  $N_j$  is the number of contestants remaining in week  $j$ . This setting ensures comparability of floating votes across seasons and rules.

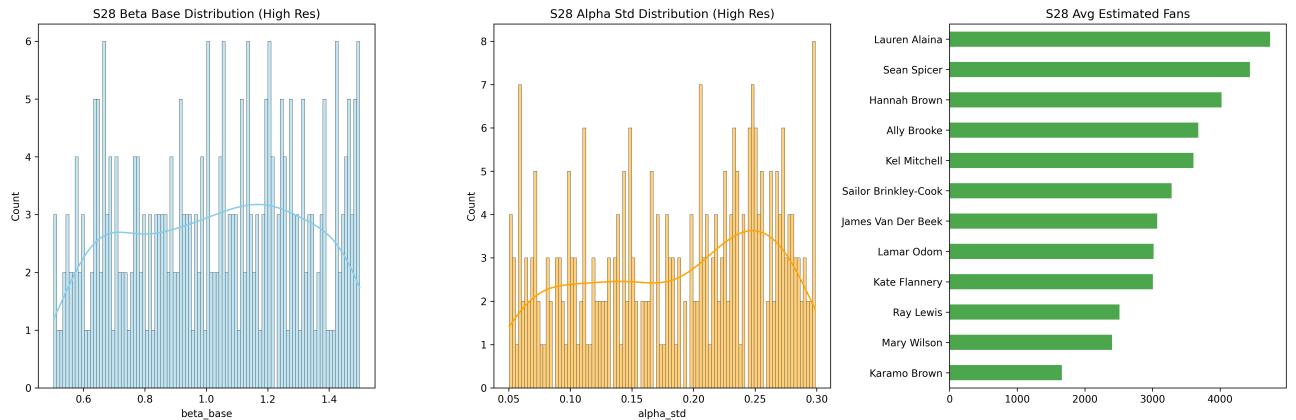


Figure 2: Example: solution space of  $\alpha$  and  $\beta$  and estimated fan base for Season 28.

The threshold Accuracy  $\geq 75\%$  balances interpretability and robustness:

On one hand, it is clearly above random baseline for typical season lengths, filtering out "accidental hits."

On the other hand, a higher threshold would leave too few valid samples, making parameter distributions unstable. Thus 75% is a practical compromise, maintaining sample size while ensuring reliable reproduction.

## 4.4 Consistency Metric: Reproducing Historical Eliminations

### 4.4.1 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:

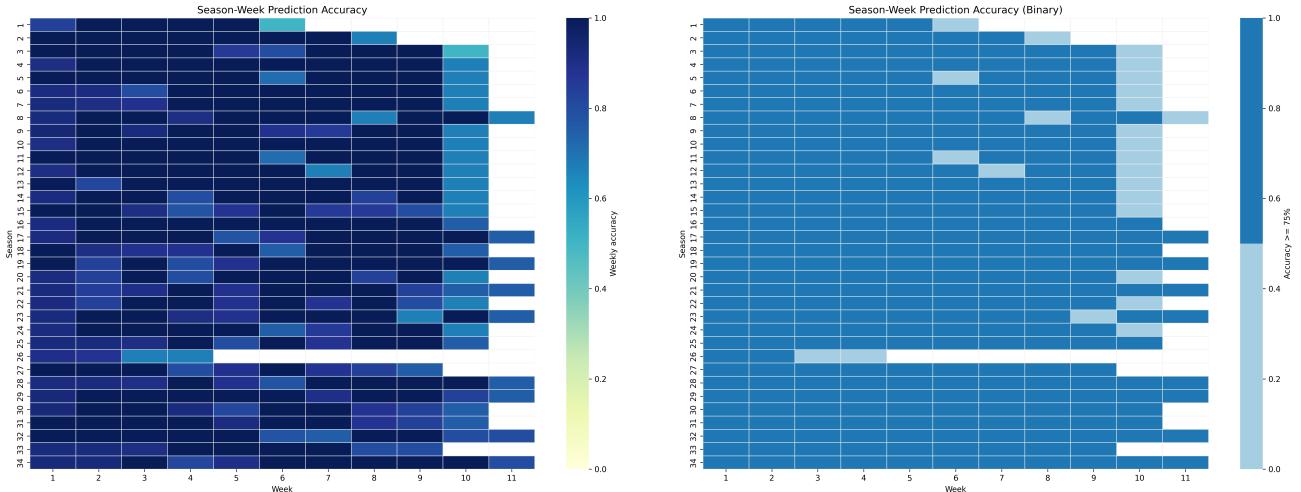
$$C_{score} = \frac{1}{T} \sum_{j=1}^T \mathbb{I}(\text{Predicted_Eliminated}_j = \text{Actual_Eliminated}_j) \quad (8)$$

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

### 4.4.2 Consistency Check with Historical Data

Empirically, the model accurately reproduces historical elimination logic in most cases.

With Monte Carlo simulation, we filter parameters with Accuracy  $\geq 75\%$ . Results show high consistency across seasons, validating the effectiveness of  $V_{ij}$  in reproducing the competition rules. The consistency score 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, with darker color indicating closer match; the right plot marks weeks reaching or exceeding 75%. Most seasons achieve high consistency across weeks, validating the model's ability to capture voting behavior.

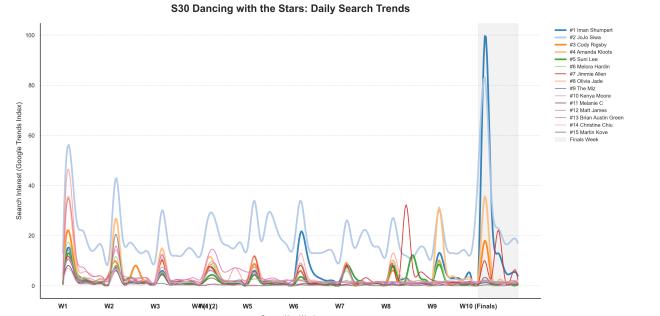
#### 4.4.3 External Consistency: Cross-Validation with Google Trends

To show that the estimated fan votes  $\hat{V}_{ij}$  are not purely mathematical fits, we introduce Google Trends search popularity  $G_{ij}$  as an external reference. We define search popularity as "public attention" to validate the reconstructed "private votes."

Specifically, using the open-source `pytrends` script, we retrieved Google Trends data for each contestant during competition periods. We then computed the correlation coefficient  $r$  between  $\hat{V}_{ij}$  and  $G_{ij}$ .



(a) Illustration of Google Trends data acquisition.



(b) Daily search trends for selected contestants in Season 30.

Figure 4: Google Trends source and Season 30 example.

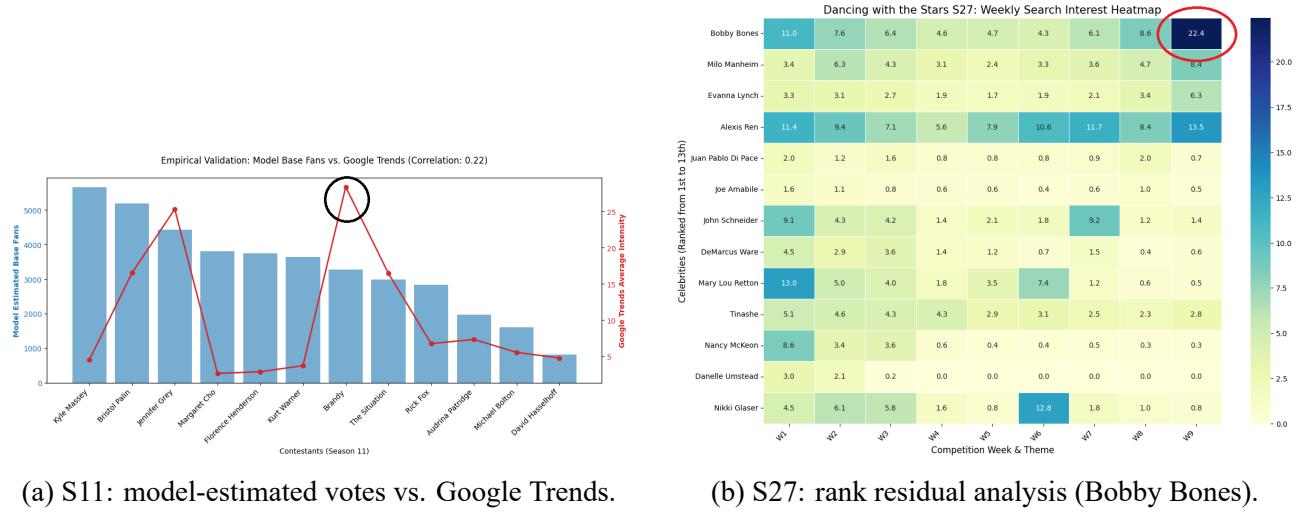
In most regular seasons,  $\hat{V}_{ij}$  and  $G_{ij}$  show strong positive correlation ( $r \geq 0.85$ ), indicating that the reconstructed votes align with real-world attention.

However, in some special seasons (e.g., S11 and S27), the model and search trends diverge significantly, with correlations  $r_{11} = 0.22$  and  $r_{27} = 0.319$ . This motivates a deeper discussion of logical consistency below.

#### 4.4.4 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) exhibit divergence between estimated votes and search trends.

- **S11 controversy and attention denoising:** Bristol Palin's search volume was extremely high in the late season (see Figure 5a), but the model did not blindly inflate her estimated votes. This indicates the model successfully filtered "non-voting attention" (negative or passive attention) and captured effective votes.
- **S27 silent fan base identification:** Conversely, Bobby Bones had modest search volume, but very high estimated votes (see Figure 5b). This suggests the model captured a "silent fan base",<sup>a</sup> radio audiences who rarely search but are highly active in voting.



(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.

#### 4.4.5 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.

### 4.5 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  $Var(\hat{V}_{ij})$ :

$$S_{cert}(i, j) = \frac{1}{1 + Var(\hat{V}_{ij})} \quad (9)$$

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

$$CI_{95\%} = \bar{V}_{ij} \pm z_{0.025} \cdot \frac{\sigma_{ij}}{\sqrt{n_{valid}}} \quad (10)$$

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.

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

### 5.1 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.

### 5.2 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:

$$I = \frac{\text{Distance}(\text{Final Rank}, \text{Judge Rank})}{\text{Distance}(\text{Final Rank}, \text{Fan Rank})}. \quad (11)$$

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.

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

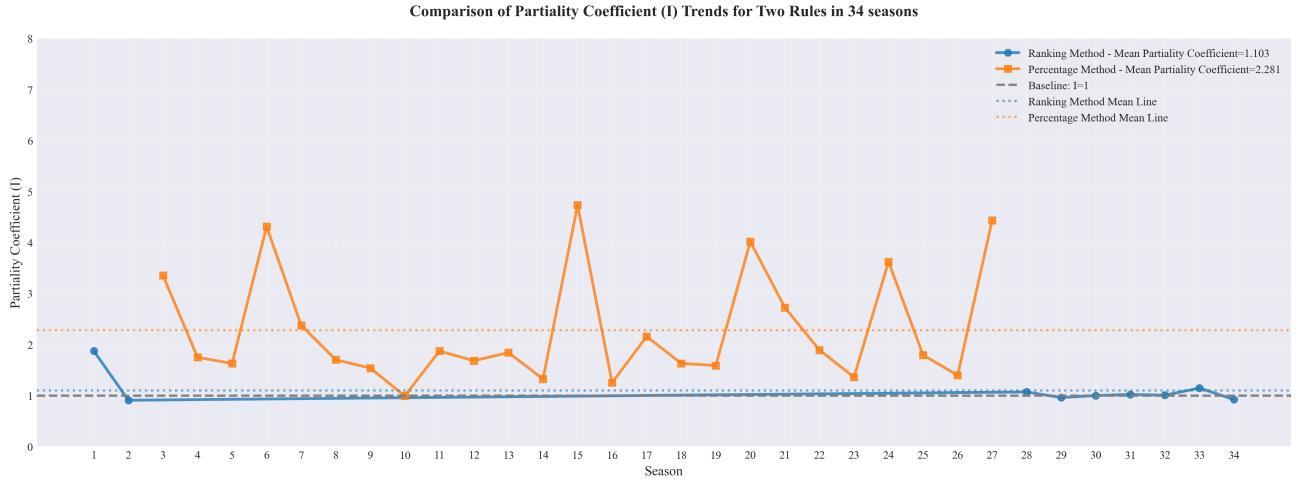


Figure 6: Partiality coefficient  $I$  across seasons.

In Figure 6, the blue line is the rank-based rule (mean 1.103), the orange line is the percentage-based rule (mean 2.281); the dashed line is the  $I = 1$  balance baseline, and the dotted lines are method means.

Across all seasons, the percentage-based rule has higher  $I$ , indicating a stronger bias toward fan votes, while the rank-based rule is closer to 1 and 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.

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

#### 5.4.1 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:

$$\Delta R_{ij} = R_{ij}^{(J)} - R_{ij}^{(F)}. \quad (12)$$

Define a "conflict week" threshold using data quantiles:

$$\tau = Q_{0.90}(|\Delta R_{ij}|), \quad \mathbb{I}_{ij} = \mathbb{I}(|\Delta R_{ij}| \geq \tau). \quad (13)$$

Define the number of conflict weeks in a season:

$$K_i = \sum_{j=1}^T \mathbb{I}_{ij}. \quad (14)$$

Introduce a structural conflict indicator:

$$\mathbb{S}_{ij} = \mathbb{I}(Score_{ij} \leq Q_{0.25}(Score_{\cdot j})) \cdot \mathbb{I}(Vote_{ij} \geq Q_{0.75}(Vote_{\cdot j})). \quad (15)$$

Define controversial samples as

$$\mathbb{C}_i = \mathbb{I}(K_i \geq k_0) \cdot \mathbb{I}\left(\sum_{j=1}^T \mathbb{S}_{ij} \geq 1\right), \quad k_0 = 2. \quad (16)$$

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

The meanings are summarized below:

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 $i$ .
$\mathbb{S}_{ij}$	Structural conflict: low judge scores (bottom quartile) and high fan votes (top quartile).
$\mathbb{C}_i = 1$	Contestant satisfies frequent and structural conflict.

The given S2, S4, S11, and S27 all satisfy  $\mathbb{C}_i = 1$ ; other samples are filtered by the same criteria.

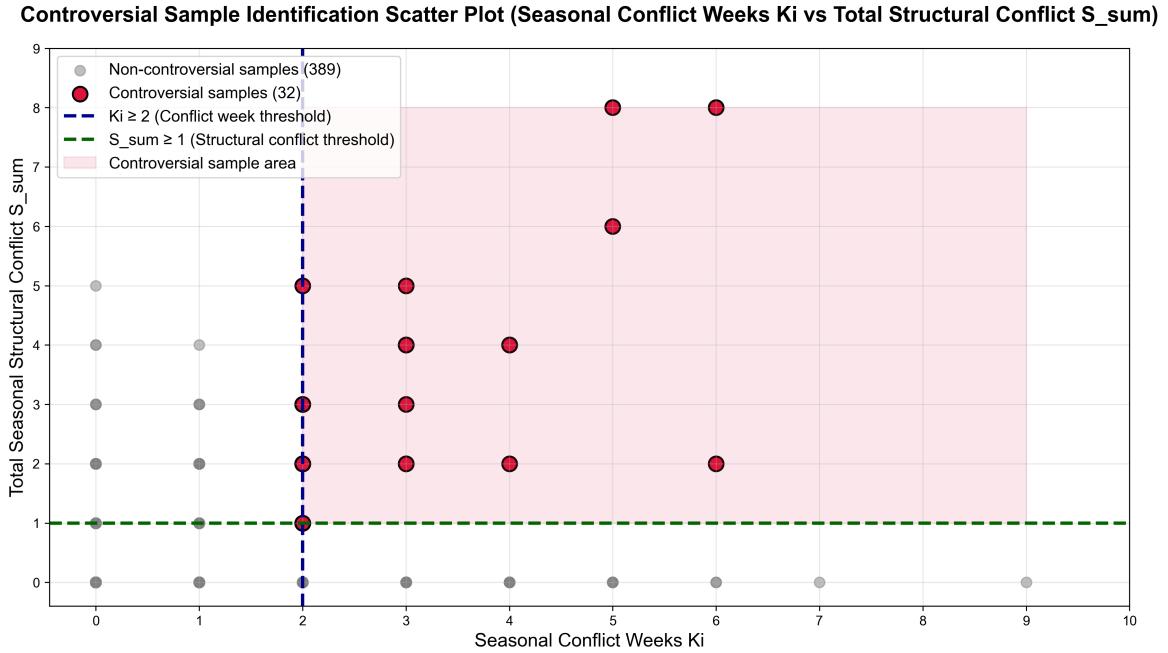
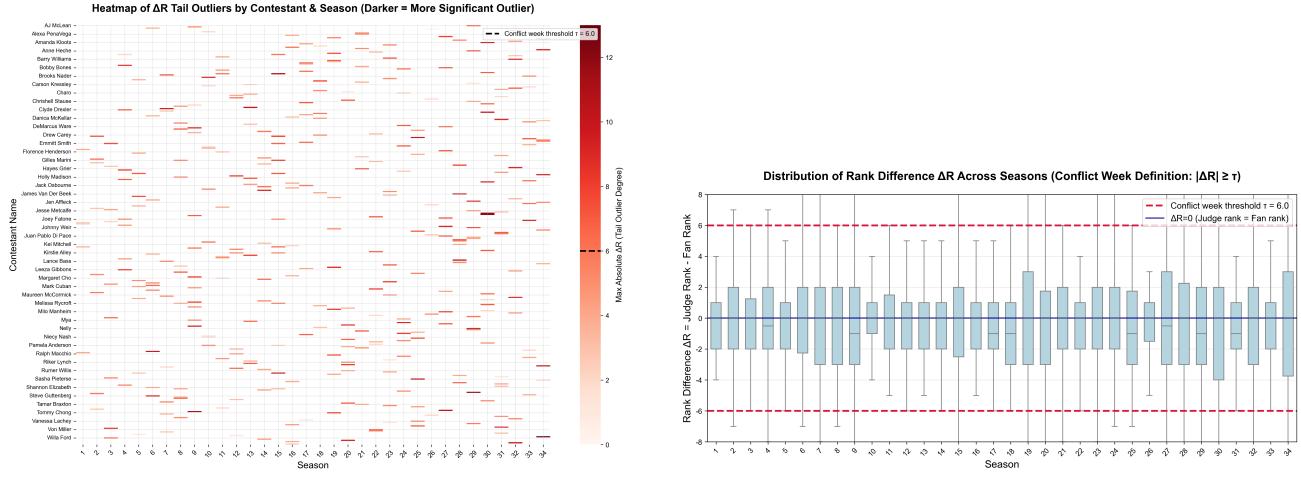


Figure 7: Scatter plot for controversial sample identification: conflict weeks  $K_i$  vs. cumulative structural conflicts  $\sum_j S_{ij}$ .

To verify cross-season consistency, we display  $\Delta R_{ij}$  distributions (boxplots/heatmaps).



(a) Heatmap of  $\Delta R$  outliers for controversial contestants.

(b) Distribution of rank differences  $\Delta R$  across seasons.

Figure 8: Conflict strength analysis: heatmap shows individual outliers; distribution shows season-level summary.

Notably, these samples align closely with controversial contestants mentioned in media sources (e.g., "Top 30 Scandals" videos), validating our selection method.

After identifying samples, we replay both rules in the next section to compare eliminations and final ranks, and then test whether the Judge Save mechanism reduces controversial outcomes.

### 5.4.2 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 ranks 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.

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 $
2	5	Jerry Rice	-1	-1
4	6	Billy Ray	-2	-3
11	2	Bristol Palin	-2	-1
27	3	Bobby Bones	-1	-2
30	1	Iman Shumpert	3	1
30	3	Cody Rigsby	3	3
15	1	Bristol Palin	2	3
32	2	Mauricio Umansky	2	1
34	1	Andy Richter	2	2

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. This explains why controversial samples survive 2–3 more weeks under the percentage rule.

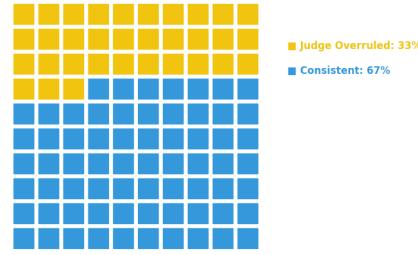
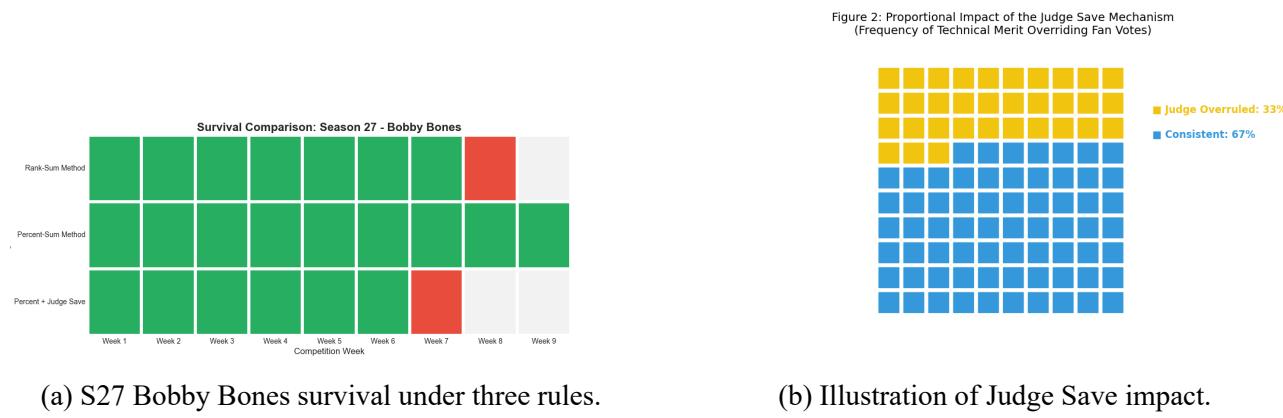
These results provide a baseline for evaluating the Judge Save mechanism.

### 5.4.3 Judge Save Mechanism Simulation and Impact

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

Results show that introducing Judge Save significantly reduces "abnormal survival" of controversial samples: fewer reach late rounds, average survival weeks decrease; meanwhile, contestants with high judge scores but low fan votes 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 at the end of the rule, enforcing a professional floor and preventing low-score contestants from advancing solely on popularity.



(b) Illustration of Judge Save impact.

Figure 9: Side-by-side: survival paths and Judge Save impact.

#### 5.4.4 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 10.

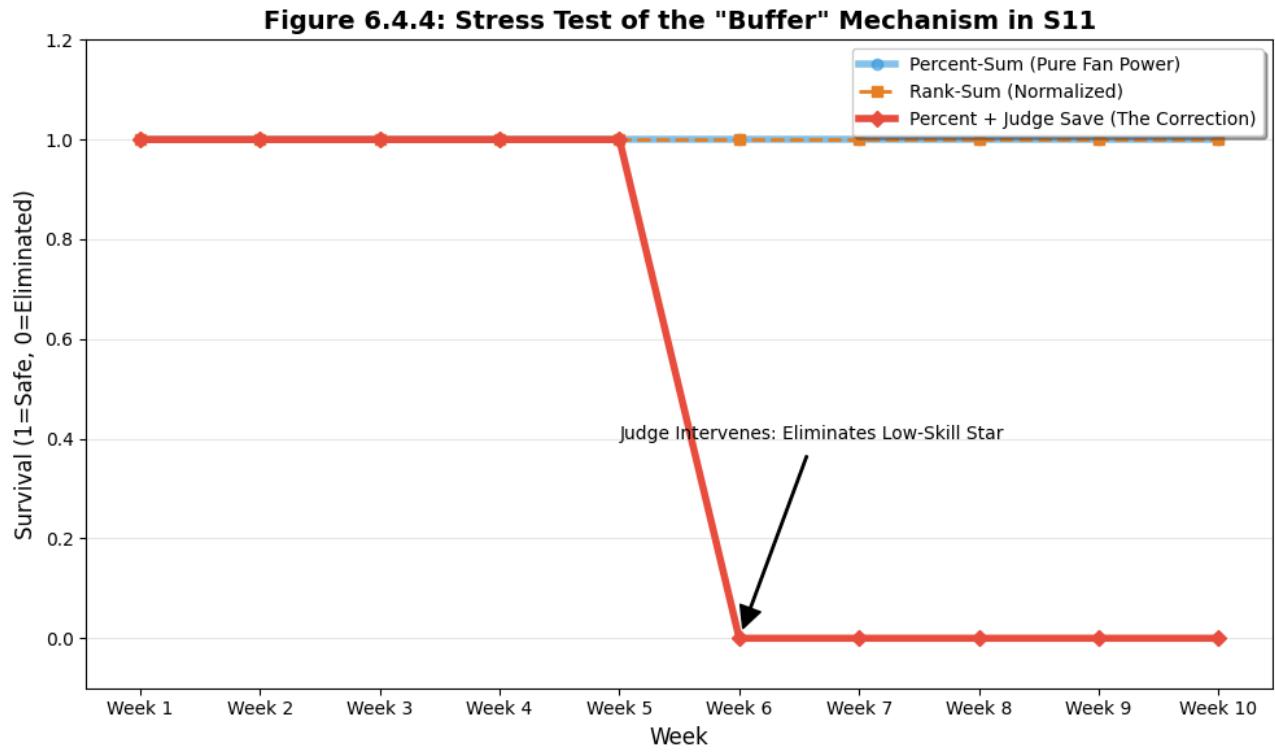


Figure 10: Stress Test: Survival Trajectories of a "High-Popularity, Low-Skill" Virtual Contestant under Different Rules (Season 11)

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.

#### 5.4.5 Implications for Future Rules

See below.

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

### 6.1 Research Objectives and Problem Definition

This section focuses on "what factors determine how far contestants go."

We separate judge scores and fan votes, evaluate partner teaching ability and contestant characteristics (industry, region, age), and test whether the two channels are consistent.

### 6.2 Data Processing and Standardization

We use the estimated fan votes  $\hat{V}_{ij}$  and judge scores  $JudgeScore_{ij}$  from Question 1. To ensure comparability across seasons and weeks, both scores are standardized using Z-scores at the season-week level. When the weekly standard deviation is 0, we set the Z-score to 0 to avoid division by zero.

Here,  $x$  is the original score (judge or fan), and  $\mu$  and  $\sigma$  are the mean and standard deviation for that season-week across contestants.

### 6.3 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`.

#### 6.3.1 Contestant Attribute Features

- **Age:** Keep continuous variable `celebrity_age_during_season` to capture non-linear effects on performance and audience preference.
- **Occupation category:** The raw data includes 30+ occupations (Singer, Actor, Athlete, etc.), with some very small categories (e.g., Politician). To reduce sparsity, we use a **Top-K retention strategy**:
  - Keep the top 10 occupations by sample size
  - Merge others into `Other`
  - Create `Industry_Group` (11 categories)

### 6.3.2 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.

### 6.3.3 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_performance`.

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

### 6.3.4 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.

## 6.4 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.

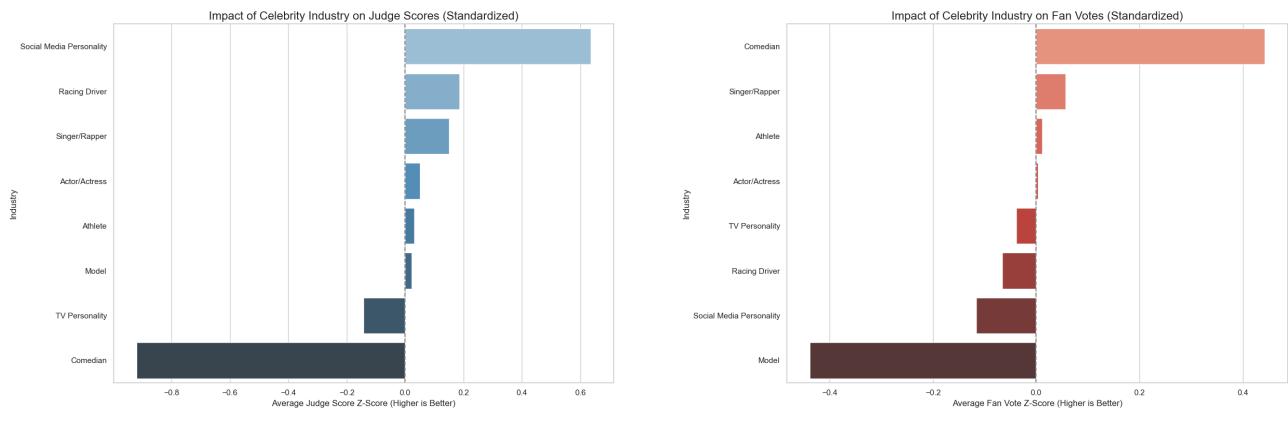
### 6.4.1 Occupation Effects

Occupations may affect professional performance and audience preference. We compute category-level means:

$$\bar{z}_k^{(J)} = \frac{1}{n_k} \sum_{i \in k} z_i^{(J)}, \quad \bar{z}_k^{(F)} = \frac{1}{n_k} \sum_{i \in k} z_i^{(F)},$$

where  $k$  is occupation category and  $n_k$  its sample size. Small categories are merged or removed for stability.

Overall results show that "professional performance" and "fan popularity" do not fully align: stage-related occupations score higher with judges, while others score higher with fans. Figures 11a–12 visualize this.



(a) Occupation effects on judge scores (standardized). (b) Occupation effects on fan votes (standardized).

Figure 11: Comparison of occupation effects on judge scores and fan votes.



Figure 12: 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.

#### 6.4.2 Regional Background Effects

Regional factors may influence results via (1) training resources and (2) regional identity in voting. We use the two-level region system and compare distributions after filtering small samples.

For judge scores, regional differences are moderate, mainly in spread and outliers; for fan votes, long-tail patterns appear, indicating regional mobilization. We visualize with violin plots (see Figures 13–14).

Top-10 regions for judge scores and fan votes are shown in Table 5. The 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. Fine-grained analysis shows a strong "regional mobilization effect" in fan votes; smaller or tightly knit states (e.g., Alaska, Delaware) often yield very high mean votes, possibly due to concentrated support. 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

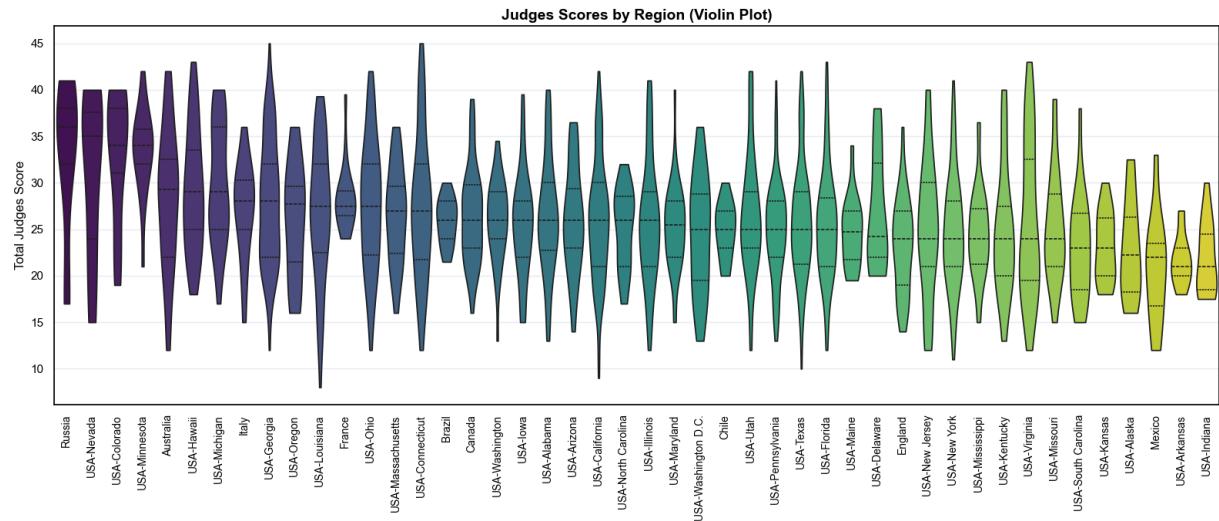


Figure 13: Distribution of judge scores by region (violin plot).

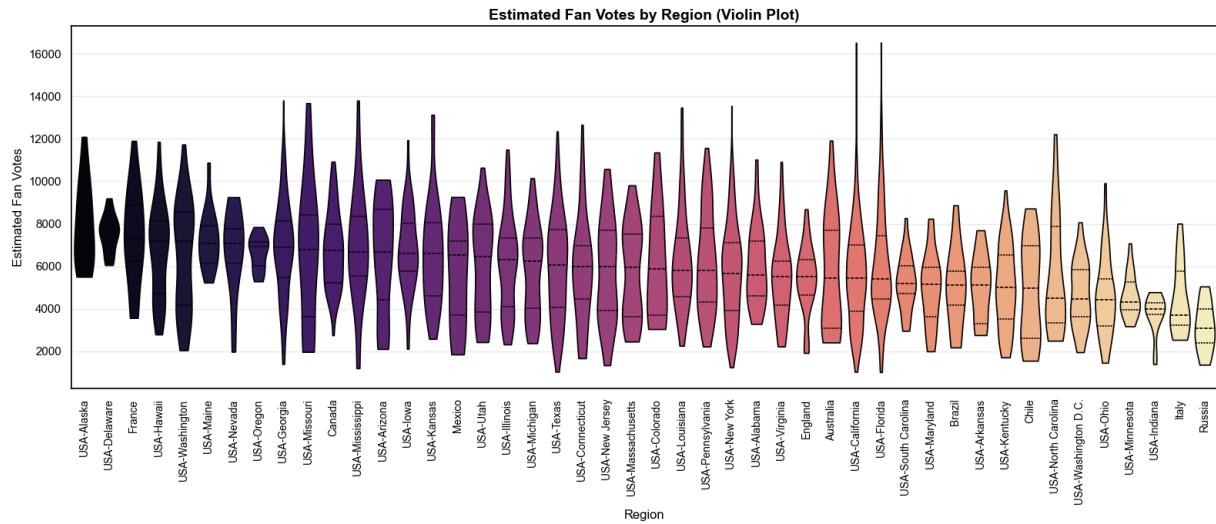


Figure 14: Distribution of fan votes by region (violin plot).

### 6.4.3 Age Effects

Age may have a non-linear impact: younger contestants have physical advantages but may elicit less empathy; older contestants may have technical disadvantages but stronger narratives. We use KDE and Lowess trend estimation.



Figure 15: Age vs judge scores/fan votes: KDE + Lowess trends.

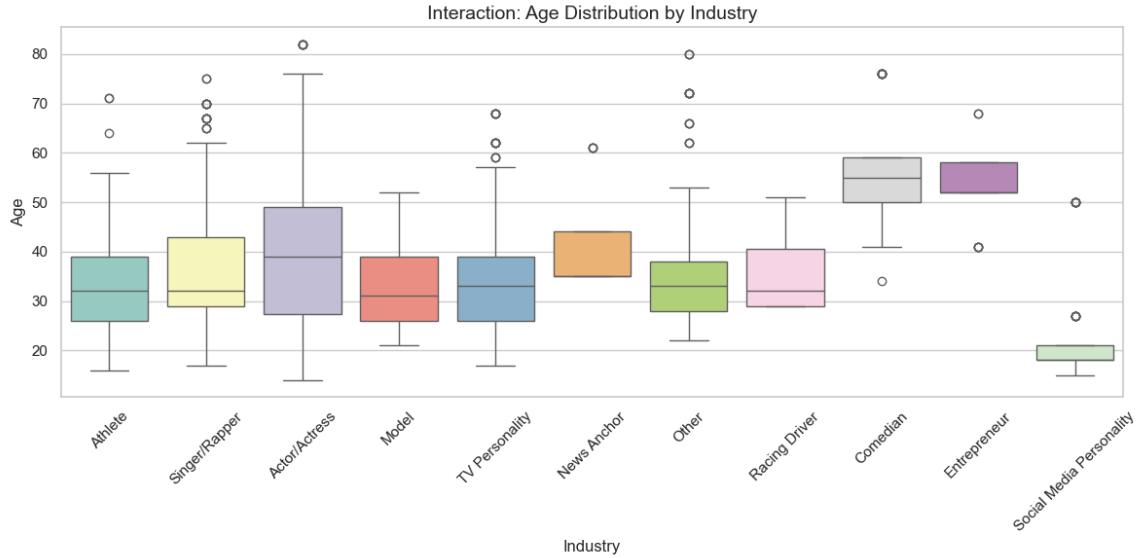


Figure 16: Age distributions across occupations (boxplots).

Results show a slight inverted-U pattern between age and judge scores, with stronger correlation for judge scores ( $r = -0.302$ ) than for fan votes ( $r = -0.172$ ). This suggests judge scores are more sensitive to age, while fan votes depend less on age and more on other factors.

However, correlation is not causation: age may influence scores indirectly through occupation, exposure, and existing fan base. Different occupations have systematic age differences (Figure 16), creating confounding pathways.

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.

## 6.5 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.

### 6.5.1 Standardization and Improvement Slope

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

$$z_{i,s,w} = \frac{x_{i,s,w} - \mu_{s,w}}{\sigma_{s,w}}, \quad x \in \{JudgeScore, \widehat{FanVote}\}. \quad (17)$$

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

For each contestant (within a season), fit:

$$z_{i,s,w} = a_{i,s} + \beta_{i,s} \cdot w + \varepsilon_{i,s,w}, \quad (18)$$

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).

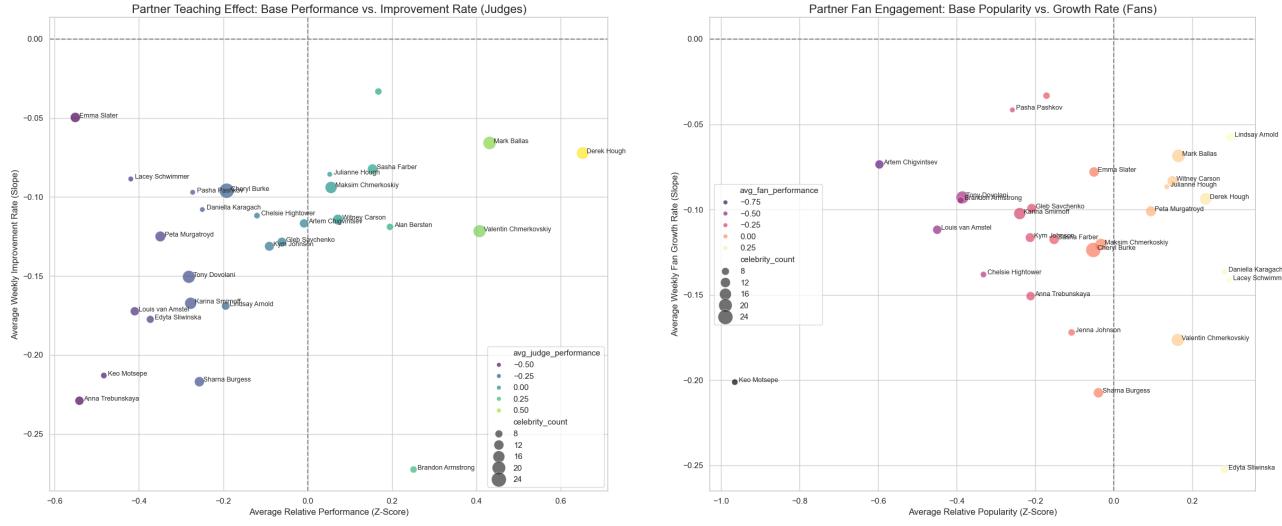
Because Z-scores are relative within each week and later weeks have stronger contestants, average  $\bar{\beta}$  is often negative (median about  $-0.118$  for judges and  $-0.107$  for fans). Thus we interpret stronger teaching as  $\bar{\beta}$  closer to 0 (slower decline or even positive growth).

### 6.5.2 2D Scatter: Base Performance vs Improvement Rate

Figure 17 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 17: Senior partners in the "base performance; growth" plane; point size indicates number of partners coached.

## 6.6 Multifactor Attribution: Random Forest Model

### 6.6.1 Why a Multivariate Model? (Limits of Univariate Analysis)

Let each observation be a season-week-contestant tuple  $t = (s, w, i)$ . We define:

$$\mathcal{D} = \{(\mathbf{x}_t, y_t^{(J)}, y_t^{(F)})\}_{t=1}^N, \quad (19)$$

where  $y_t^{(J)}$  is judge-channel relative performance and  $y_t^{(F)}$  is fan-channel relative votes, and  $\mathbf{x}_t$  is the feature vector (age, occupation, region).

Univariate analysis examines  $\mathbb{E}[y | x_k]$ , but correlated features (e.g., occupation with age, region with industry) cause confounding. Thus we fit multivariate models:

$$\mathbb{E}[y^{(\cdot)} | \mathbf{x}] \approx f^{(\cdot)}(\mathbf{x}), \quad (\cdot) \in \{J, F\}, \quad (20)$$

and compare the sensitivities of both channels under the same feature mapping.

### 6.6.2 Q3 Code Logic and Model Construction

**(1) Target standardization.** Q3 code converts raw  $(Score_t^{(J)}, Vote_t)$  into season-level Z-scores:

$$y_t^{(J)} = \frac{Score_t^{(J)} - \mu_s^{(J)}}{\sigma_s^{(J)}}, \quad y_t^{(F)} = \frac{Vote_t - \mu_s^{(F)}}{\sigma_s^{(F)}}, \quad (21)$$

where  $\mu_s^{(\cdot)}$  and  $\sigma_s^{(\cdot)}$  are season means and stds (degenerate handling if  $\sigma = 0$ ). This removes scale differences and interprets  $y$  as relative advantage.

**(2) Categorical reduction and encoding.** Let occupation be  $g_i \in \mathcal{G}$ , keep top- $K$  occupations  $\mathcal{G}_K$ , and define:

$$g'_i = \begin{cases} g_i, & g_i \in \mathcal{G}_K, \\ \text{Other}, & \text{otherwise.} \end{cases} \quad (22)$$

Apply similar grouping to region variables, then one-hot encode  $g'_i$  and region.

**(3) Feature vector.** For each sample  $t = (s, w, i)$  with age  $a_i$ :

$$\mathbf{x}_t = [s, w, a_i, \text{OneHot}(g'_i), \text{OneHot}(r'_i)]^\top \in \mathbb{R}^d. \quad (23)$$

**(4) Dual-channel random forest regression.** Fit:

$$\hat{f}^{(J)} = \arg \min_{f \in \mathcal{F}_{RF}} \frac{1}{N} \sum_{t=1}^N (y_t^{(J)} - f(\mathbf{x}_t))^2, \quad \hat{f}^{(F)} = \arg \min_{f \in \mathcal{F}_{RF}} \frac{1}{N} \sum_{t=1}^N (y_t^{(F)} - f(\mathbf{x}_t))^2. \quad (24)$$

Random forest predictor:

$$\hat{f}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x}), \quad (25)$$

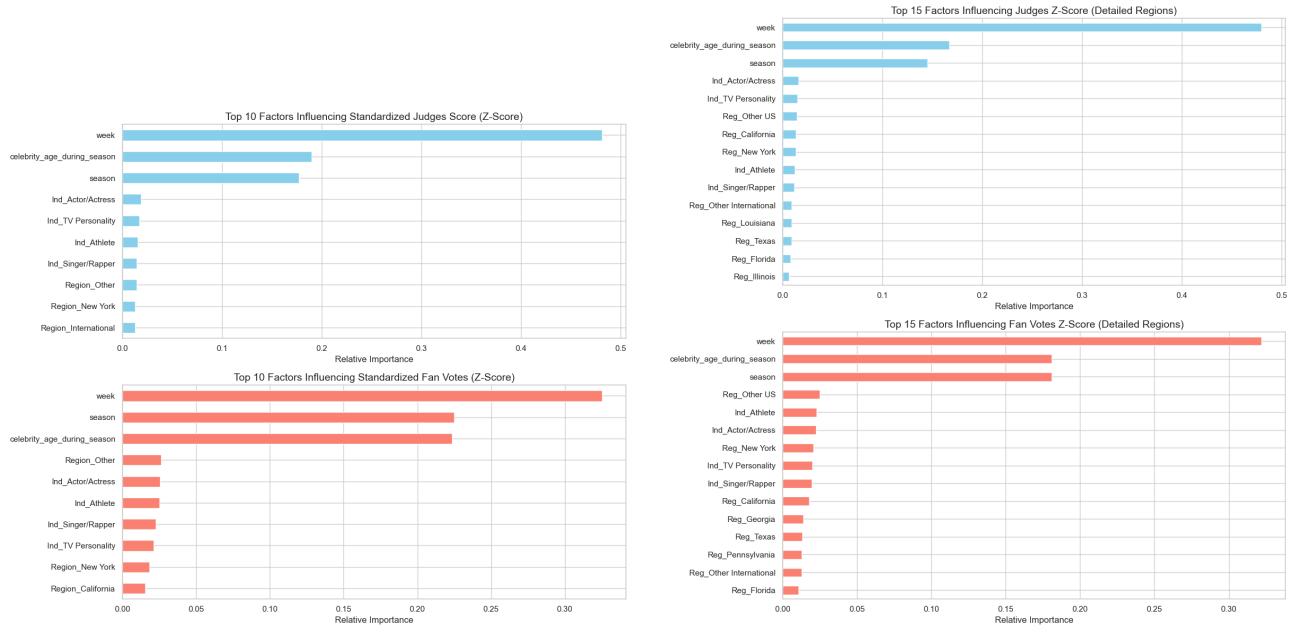
where  $T_b$  is the  $b$ -th tree and  $B$  is the number of trees (set to  $B = 100$ ).

### 6.6.3 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  $k$ :

$$\text{Imp}_k = \frac{1}{B} \sum_{b=1}^B \sum_{v \in \mathcal{V}_b: \text{split}(v)=k} p(v) \Delta \text{MSE}(v), \quad (26)$$

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 18a 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 18: 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.

## 6.7 Summary

This section builds a multi-factor model using estimated fan votes and judge scores, and evaluates partner effects and contestant features. Results show that the judge channel is more "ability-oriented," with strong partner base and teaching effects; the fan channel is more "structure-oriented," driven by exposure and mobilization.

Contestant features (age, occupation, region) matter but differ in direction: age is more sensitive in judge scores; occupation and region are more important on the fan side, reflecting popularity structure and regional mobilization. Overall, mechanisms differ across channels, explaining "high-score low-popularity" or "low-score high-vote" outcomes.

## 7 Recommendations to the Committee

See "Recommendations to the Committee (Appendix)."

## 8 Strengths and Weaknesses

### 8.1 Strengths

- **Innovative method (Monte Carlo + parameter inversion + Google Trends):** We estimate fan votes with Monte Carlo simulation and parameter inversion, and cross-validate with Google Trends. This ensures internal consistency and external alignment, and denoises non-voting attention. *This improves internal validity, external alignment, and de-noises non-voting attention.*
- **Multi-dimensional features and dual-channel random forest:** We build a multi-dimensional feature system (contestant attributes, partner ability, region) and use dual-channel random forests, separating judge "ability-driven" signals from fan "structure-driven" effects. *Dual-channel RF separates judge skill signals from fan structure effects.*
- **30+30 dual-track linear scoring system (DTLSS):** While staying audience-friendly, it sets a professional floor. The rule is simple and interpretable, addressing controversies and improving live explainability. *Simple, interpretable scoring balances audience appeal and professional fairness.*
- **Controversial sample screening and rule replay:** Includes Judge Save mechanism experiments, quantifying bias and stability with strong conclusions. *Rule replay quantifies bias and stability, supporting robust conclusions.*

### 8.2 Weaknesses

- **Static fan-base assumption:** We assume a constant base fan size, ignoring time-varying shocks during the season. *Static fan-base assumption may miss time-varying shocks.*
- **Parameter range depends on pilot runs:** "Wide-then-narrow" reduces subjectivity but still depends on small pilot runs, introducing chance effects. *Narrowing ranges based on small pilots can introduce chance effects.*
- **Feature dimensions can be expanded:** The model does not yet include social media sentiment or dance-style fit variables, which could improve explanatory power and performance. *Add sentiment/engagement and style-fit variables to improve modeling.*

## 9 Conclusion

This paper studies the structural tension in DWTS scoring: how to balance professional judges' evaluations and audience popularity. Under the constraint of missing direct vote data, we analyze rule behavior and bias from a mechanism perspective and propose improvements for fairness, interpretability, and engagement. The core finding is that controversy stems from amplification effects in the rules, which should be corrected at the rule level.

Three main conclusions follow: (1) judge scores and fan votes reflect different values, and fan influence needs rational constraints to prevent dilution of technical evaluation; (2) both percentage and rank rules have endogenous bias and amplify extreme outcomes when popularity and skill diverge; (3)

introducing limited judge intervention at key points or capping fan influence can improve fairness and stability without reducing participation. Overall conclusions:

- Judge scores and fan votes reflect different values; the key is to 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 at key points 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 can be applied to other competitions combining expert evaluation and public voting. Future work can enrich participation and data sources while maintaining the core structure to achieve 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  $Rank_{Fan}$ :

$$S_{Fan} = S_{max} - \delta \times (Rank_{Fan} - 1) \quad (27)$$

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

Total score is the sum:

$$TotalScore = Score_{Judge} + Score_{Fan} \quad (28)$$

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 index  
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.