Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 8, No. 1, 2026

Reverse-Engineering Fan Votes in Dancing with the Stars: Rule-Consistent Inference, Fairness Analysis, and Mechanism Design

Yichen Wang [1, a], Zhan Chen [2, b], Qian Liu [3, c]

[1] College of science, Harbin University of Science and Technology, Harbin, China

[2] School of mechanical and power engineering, Harbin University of Science and Technology, Harbin, China

[3] School of measurement and communication engineering, Harbin University of Science and Technology, Harbin, China

[a] 1514977143@qq.com, [b] 2416934995@qq.com, [c] 3948515314@qq.com

Abstract: This paper develops a data-driven framework to infer confidential audience voting and to evaluate voting aggregation mechanisms in Dancing with the Stars under rule changes and structural bias. We propose a two-stage inverse approach. In Stage 1, a Bradley Terry style model is fitted to historical elimination sequences to estimate each contestants latent fan strength, forming a prior preference distribution. In Stage 2, we reconstruct weekly audience vote shares by solving a constrained inverse problem: among all vote allocations consistent with the show' s elimination mechanism (including percentage-based elimination and the bottom-two variant), we select the distribution closest to the prior via convex quadratic programming (with discrete handling for bottom-two weeks). Uncertainty is quantified through posterior sampling. Across 34 seasons, the reconstructed votes reproduce 77.0% of weekly eliminations, achieve a 0.953 Spearman correlation with final rankings, and match champions 58.8% of the time, indicating strong explanatory power despite unobserved votes. Using the inferred votes, we compare historical aggregation rules and find the percentage-based rule aligns substantially better with observed eliminations than the rank-based rule in their respective eras (78.6% vs. 56.0 weekly agreement), while avoiding excessive amplification of judge influence. We further quantify structural fan advantage related to celebrity background and propose a debiasing weight to correct audience shares. Building on this correction, we introduce a Fairness-Adjusted Hybrid Rule (FHR) that blends judge shares with debiased audience shares. FHR is designed as a forward-looking mechanism: it improves end-of-season technical ordering (finals rank correlation 0.325 vs. 0.271 under the base- line percentage rule) at the cost of lower week-by-week elimination consistency, and its debiasing strength can be tuned to match production priorities. We recommend retaining the percentage-based rule as the default for stability and historical continuity, while piloting the FHR as an optional fairness overlay with transparent parameter tuning and periodic audits.

Keywords: Two-stage inverse modeling, Bradley Terry model, constrained quadratic programming, voting rule comparison, structural bias, fairness-adjusted hybrid mechanism, uncertainty quantification.

1. Introduction

1.1. Problem Background

Dancing with the Stars (DWTS) is a competitive dance show where celebrities partner with professional dancers. Each week, performances are evaluated by judges' scores and audience votes, which together determine the elimination outcome. The actual audience vote counts are confidential, creating an inverse problem: inferring the hidden voting structure from publicly available judges' scores and elimination results. Over its 34 seasons, the show has employed two primary vote aggregation methods—rank-based(used in early and recent seasons) and percentage-based (used in middle seasons)—which have yielded different outcomes for controversial contestants like Jerry Rice and Bristol Palin [1-3].

This study addresses the challenge through a two-stage modeling framework: first estimating the latent "fan strength" from elimination sequences, then reconstructing the weekly audience vote shares under the explicit rules of the show. This approach provides a foundation for systematically comparing voting methods, analyzing structural biases, and designing fairer competition formats.

1.2. Restatement of the Problem

We decompose the problem into four coherent tasks:

(1) Audience Voting Estimation: Develop a mathematical model to estimate the weekly audience vote shares using only judges' scores and elimination data. Provide consistency metrics and quantify estimation uncertainty.

(2) Rule Comparison and Analysis: Using the estimated votes, compare the rank- based and percentage-based aggregation methods across all seasons. Evaluate which method better aligns with historical outcomes and analyze its impact on controversial contestants.

(3) Modeling Influencing Factors: Analyze how professional dancers and celebrity characteristics (e.g., age, industry) affect competition results, distinguishing between the preferences of judges and audiences.

(4) Designing a Fair System: Propose a new vote aggregation scheme that balances fairness and entertainment value, supported by quantitative analysis from the preceding tasks, to provide actionable recommendations for the show's producers.

2. Assumptions and Explanations

Considering the modeling framework of audience voting

estimation and rule comparison, we formalize the following assumptions to ensure mathematical tractability and interpretability.

Assumption 1: We assume that the estimated audience vote shares form a probability distribution for each week.

Explanation: This normalization constraint ensures that the reconstructed audience votes are mathematically consistent and can be interpreted as proportions of the total vote. Specifcally, for each week w in season s, we require $\sum_{i\in\mathcal{A}_{s,w}} v_{i,s,w} = 1$ and $v_{i,s,w} \geq 0$ for all active contestants $i \in \mathcal{A}_{s,w}$.

Assumption 2: We assume that under the percentage elimination rule, the eliminated contestant has the lowest combined score.

Explanation: This constraint enforces the show's stated elimination mechanism. The combined score $C_{i,s,w} = J_{i,s,w} + v_{i,s,w}$ is computed for each contestant, and the contestant with the smallest combined score is eliminated. To ensure strict inequality, we add a small positive constant: $C_{e_{s,w},s,w} + \varepsilon \leq C_{\{j,s,w}$ for all $j \neq e_{s,w}$ [2, 3].

Assumption 3: We assume that under the bottom-two rule, the eliminated contestant must be among the two contestants with the lowest combined scores. Explanation: This constraint models the modified elimination mechanism introduced in later seasons. Instead of requiring the eliminated contestant to have the absolute lowest score, we only require that they are in the bottom two. Mathematically, we enforce $C_{e_{s,w},s,w} \leq C_{j,s,w} + \tau$ for at most two contestants, where τ is a slack variable determined during optimization [3].

Assumption 4: We assume that the reconstructed voting shares should be as close as possible to the prior voting shares implied by fan strength.

Explanation: This objective function ensures that our reconstructed votes remain faithful to the underlying fan popularity estimated from elimination patterns. We minimize the Euclidean distance between the estimated vote vector $v_{s,w}$ and the prior vote vector $p_{s,w}$: $\min\|v_{s,w} - p_{s,w}\|_2^2$.

Assumption 5: We assume that the fan-strength parameters require a reference con- straint for identifiability.

Explanation: The fan-strength model is invariant to translation (adding a constant to all $\lambda_i^{(s)}$ yields the same probabilities). To remove this indeterminacy, we fix one contestant's fan strength to zero (e.g., $\lambda_{i_0}^{(s)} = 0$) or equivalently require the sum of fan strengths to be zero ($\sum_i \lambda_i^{(s)} = 0$).

Assumption 6: We assume that structural advantages in audience voting can be mitigated through fairness reweighting.

Explanation: To address systematic biases due to contestants' preexisting popularity or background, we apply a fairness weight $w_{\text{fair},i}$ to each contestant's raw vote estimate. The debiased vote share is computed as $\widetilde{v_{i,s,w}} = w_{\text{fair},i} \cdot v_{i,s,w}$ where $w_{\text{fair},i}$ is derived from the estimated structural advantage $\text{Adv}_{i,s}$.

3. Notations

Some important mathematical notations used in this paper are listed in Table 4.

Table 1. Notations used in this paper

Symbol	Description	Units
s	Season index	–
w	Week index within a season	–
i, j	Contestant indices	–
$\mathcal{A}_{s,w}$	Active contestant set in season s, week w	–
$e_{s,w}$	Historically eliminated contestant	–
$\lambda_i^{(s)}$	Fan-strength parameter of contestant i in season s	–
$\hat{\lambda}$	MLE estimate of fan-strength parameters	–
$p_{i,s,w}$	Prior vote share implied by fan strength	%
$v_{i,s,w}$	Estimated audience vote share (decision variable)	%
$y_{i,s,w}$	Audience vote count	votes
$J_{i,s,w}$	Normalized judges' score share	%
$C_{i,s,w}$	Combined weekly score (judge + audience)	%
α	Weight of judges' score in aggregation rule	–
ε	Small positive constant enforcing strict inequality	–
$v_{s,w}$	Vector of vote shares for all contestants in week w	–
$p_{s,w}$	Vector of prior vote shares	–
E_1	Weekly elimination agreement rate	%
E_2	Final-ranking Spearman correlation	–
E_3	Champion agreement rate	%
B	Number of Monte Carlo samples	–
$\bar{v}$	Monte Carlo mean of vote share	%
$\text{sd}(v)$	Monte Carlo standard deviation	%
$\text{CV}(v)$	Coefficient of variation	–
ρ_s	Season-level Spearman correlation	–
FLI	Fan Leverage Index	–
PVAI	Professional Value-Added Index	weeks
SA	Structural fan advantage	–
w_{fair}	Fairness reweighting factor	–
FHR	Fairness-Adjusted Hybrid Rule score	%

4. Audience Voting Estimation Model: Fan Strength-Rule Projection Inversion Method

4.1. Fan Strength Prior Model

4.1.1. Strength Estimation from Elimination Results

Following the Bradley–Terry (BT) spirit [4], we model "who gets eliminated" in week w as a multinomial choice over the active set $A_{s,w}$. Specifically, for season s we assign each contestant i a latent fan-strength parameter $\lambda_i^{(s)}$ (larger means more popular). Conditional on the active set $A_{s,w}$ the eliminated contestant $e_{s,w}$ is assumed to follow a softmax form:

$$Pr(e_{s,w} = i | A_{s,w}) = \frac{exp(-\lambda_i^{(s)})}{\Sigma_{j\in A_{s,w}} exp(-\lambda_j^{(s)})}$$

For each season s, the log-likelihood over the observed elimination sequence is

$$\ell(\lambda^{(s)}) = \Sigma_{w=1}^{W_s} log Pr(e_{s,w} | A_{s,w}) = \Sigma_{w=1}^{W_s} \left[-\lambda_{e_{s,w}}^{(s)} - log \Sigma_{j\in A_{s,w}} exp(-\lambda_j^{(s)})\right]$$

To ensure identifiability, we impose a reference constraint (e.g., $\lambda_{i0}^{(s)} = 0$ for a chosen reference contestant i0 in season

s, or equivalently $\Sigma_i \lambda_i^{(s)} = 0$).

4.1.2. Conversion to Prior Voting Shares

The estimated fan strengths are then converted into prior voting shares for each week. For contestant $i \in \mathcal{A}_{s,w}$ in week w of season s, we define

$$\pi_{i,w}^{(s)} = \frac{\exp\left(\lambda_i^{(s)}\right)}{\sum_{j \in \mathcal{A}_{s,w}} \exp\left(\lambda_j^{(s)}\right)}$$

We interpret $\pi_{i,w}^{(s)}$ as a baseline popularity prior (a regularization target) implied by the elimination sequence, rather than an observed ground-truth vote share. It captures a season-level tendency of audience support and will be adjusted in the next step to satisfy the show's scoring rules.

4.2. Rule-Consistent Voting Inversion Model

4.2.1. Variable Definitions

For week w of season s, for each contestant i we define:

Audience vote count: V_i

Audience vote share:

$$v_i = \frac{V_i}{\sum_j V_j}$$

Total judges' score: J_i, normalised to a percentage scale.

4.2.2. Combined Score and Elimination Rule

To match the show's percentage-based aggregation, we set the weekly combined score to

$$C_i = J_{i,s,w} + v_{i,s,w}$$

Where both J and v are normalized shares that sum to 1 over the active set. We keep α only for optional sensitivity analysis and report results under the default equal- weight setting.

4.2.3. Constraint Construction

(1) Percentage-elimination constraint

For weeks that use the pure percentage rule, the actually eliminated contestant e must have the lowest combined score within the active set $\mathcal{A}_{s,w}$:

$$C_e \le C_j - \varepsilon, \quad \forall j \in \mathcal{A}_{s,w},\ j \ne e$$

Where ε is a small positive constant (e.g. 0.001) enforcing strict inequality.

(2) Bottom-two constraint (from Season 28 onwards)

For weeks that use the "bottom-two" mechanism, we only require the eliminated contestant e to belong to the two lowest combined scores. This is imposed by

$$\left|\left\{j \in \mathcal{A}_{s,w} : C_j \le C_e + M\right\}\right| \le 2$$

Where M is a slack threshold determined endogenously in the optimisation, so that at most two contestants (including e) can lie in the lowest-scoring band.

4.2.4. Optimization Problem

Among all voting schemes satisfying the rule constraints, we select the one closest to the fan prior:

$$\min_v \sum_{i \in \mathcal{A}_{s,w}} (v_i - \pi_i)^2$$

Subject to constraints:

$$\sum_i v_i = 1, v_i \ge 0$$

The objective is strictly convex and the percentage-elimination constraints are linear. Therefore, for weeks governed by the pure percentage rule, each season/week subproblem can be solved as a standard convex quadratic program [5] (QP). For weeks with the "bottom-two" mechanism, enforcing "membership in the bottom two" involves an order-based condition. In practice, we handle this rule-consistency requirement either by (i) introducing auxiliary discrete selection logic (leading to an MIQP formulation) [6], or (ii) using an equivalent small-case enumeration that reduces the condition to a finite number of QP solves and selecting a feasible solution consistent with the mechanism. In all cases, the obtained $\hat{v}$ represents the voting-share estimate that stays as close as possible to the fan prior while remaining consistent with the show's elimination rule.

4.3. Consistency Metrics

Through construction, the estimated voting shares are guaranteed to satisfy the rule constraints in a feasibility sense. However, feasibility alone does not imply that the reconstructed votes fully explain the historical process. To quantitatively evaluate how well the estimated votes reproduce actual competition outcomes, we replay each season using the estimated shares and the official aggregation rules, and define the following consistency metrics:

(1) Weekly Elimination Consistency Rate:

$$E_1 = \frac{\text{Number of correctly simulated elimination weeks}}{\text{Total elimination weeks}}$$

This measures the proportion of elimination weeks in which the simulated eliminated contestant coincides with the historical elimination. Values closer to 1 indicate better week-by-week explanatory power.

(2) Finalist Set Consistency Rate:

For each season, we measure how well the simulated final-stage ordering matches the actual final placements by computing the Spearman rank correlation between the simulated final ranking and the historical final ranking. Let ρ_s denote the Spearman correlation coefficient for that season, and define the season-level ranking consistency accordingly [7].

$$E_2 = \frac{1}{S}\sum_{s=1}^{S} 1\left[\text{Finalist}_{\text{sim}}^{(s)} = \text{Finalist}_{\text{real}}^{(s)}\right]$$

(3) Champion Consistency Rate:

For each season, we record whether the simulated champion matches the actual champion. The champion consistency rate is defined as the proportion of seasons in which the simulated champion equals the historical champion.

$$E_3 = \frac{1}{S}\sum_{s=1}^{S} 1\left[\text{Champion}_{\text{sim}}^{(s)} = \text{Champion}_{\text{real}}^{(s)}\right]$$

Summary table:

Table 2. Consistency metrics for the reconstructed voting shares

Metric	Value
E_1 (Weekly Elimination Consistency)	0.770
E_2 (Finalist Ranking Correlation)	0.953
E_3 (Champion Consistency)	0.588

Across the 34 historical seasons, the weekly elimination consistency rate (E1) indicates that our reconstructed vote shares reproduce the actual eliminated contestant in about 77% of elimination weeks. The final ranking correlation [7] (E2) being close to 1 suggests that the simulated end-of-season ordering is highly consistent with historical final placements. The champion consistency rate (E3) is lower because the champion outcome is the most sensitive to small perturbations and unmodeled factors; nevertheless, correctly identifying the champion in a substantial fraction of seasons provides a credible basis for rule comparison in subsequent tasks.

4.4. Uncertainty Quantification

4.4.1. Posterior Sampling Method

Uncertainty in the model mainly comes from estimation errors in fan strength λ and the size of feasible voting space under rule constraints. Fewer constraints lead to more feasible solutions and greater uncertainty.

We quantify uncertainty using a Laplace (asymptotic normal) approximation around the maximum likelihood estimate (MLE). Under standard regularity conditions, the MLE of the fan-strength parameters is approximately normally distributed with covariance given by the inverse observed information (i.e., the Hessian of the negative log-likelihood at the MLE) [8, 9].

$$\lambda|\text{data} \approx \mathcal{N}(\hat{\lambda}, \boldsymbol{\Sigma})$$

With

$$\Sigma = \left(-\nabla^2 \ell\left(\hat{\lambda}^{(s)}\right)\right)^{-1}$$

We draw B samples of λ, and for each sample re-solve the rule-consistent optimization to obtain vote-share estimates

$$\{(\hat{v})^{(b)}\}_{b=1}^{B}$$

For each contestant-week, we summarize uncertainty using Monte Carlo [10] estimates such as the sample mean, dispersion, and relative variability:

(1) posterior/Monte-Carlo mean: $\bar{v} = (1/\mathrm{B}) \sum_b \hat{v^{(b)}}$;

(2) coefficient of variation: $\mathrm{CV}_i = \frac{\sqrt{\mathrm{Var}(v_i)}}{\bar{v_i}}$.

4.4.2. Uncertainty Metrics

For each contestant-week, calculate:

(1) Posterior mean:

$$\bar{v_i} = \frac{1}{B}\sum_{b=1}^{B} \hat{v_i^{(b)}}$$

(2) Variance and confidence interval:

$$\mathrm{Var}(v_i) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{v_i^{(b)}} - \bar{v_i}\right)^2$$

95% Confidence Interval:

$$\mathrm{CI}_{95\%} = \left[\bar{v_i} - 1.96\sqrt{\mathrm{Var}(v_i)},\ \bar{v_i} + 1.96\sqrt{\mathrm{Var}(v_i)}\right]$$

(3) Coefficient of variation:

$$\mathrm{CV}_i = \frac{\sqrt{\mathrm{Var}(v_i)}}{\bar{v_i}}$$

4.4.3. Uncertainty Visualization

Based on the above, we can systematically compare certainty differences across different contestants and weeks:

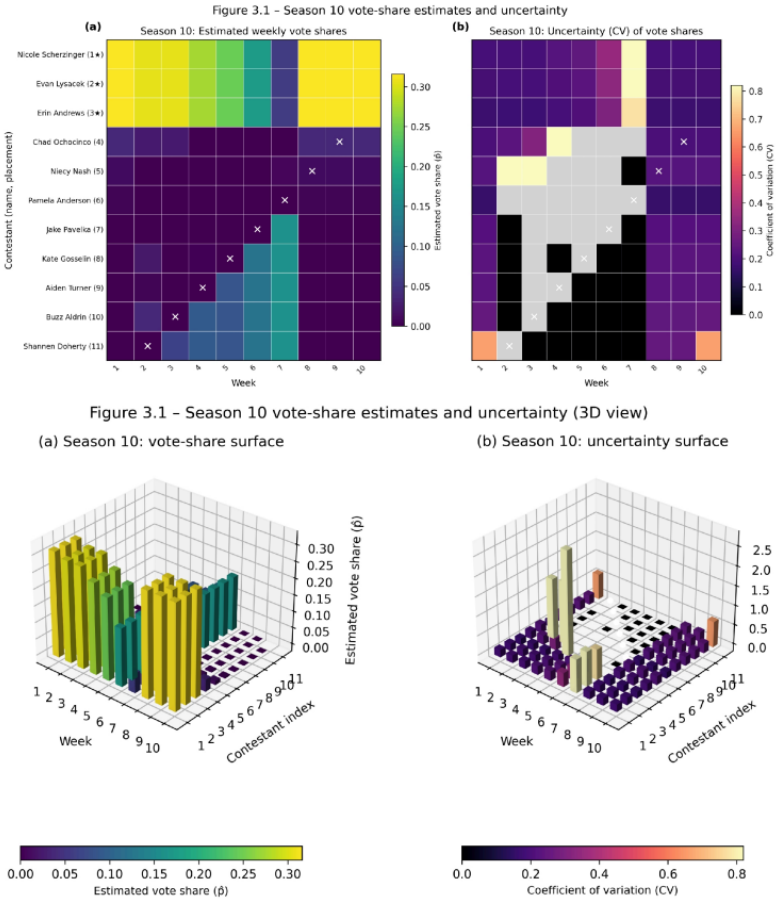

Figure 1. Season 10 Audience Voting Share and Uncertainty Example

To intuitively display the overall structure of reconstructed audience votes and their uncertainty, we select Season 10 as a representative example. Figure 1 shows the estimated voting shares for each contestant in each week of that season and the coefficient of variation (CV) obtained from posterior sampling.

(a) Heatmap of estimated voting shares for each contestant in each week. Rows correspond to contestants, sorted by final placement; top three marked with "⋆"; white X marks indicate the week the contestant actually departed.

(b) Heatmap of corresponding voting share coefficient of variation (CV), calculated from posterior samples. Brighter colors indicate higher uncertainty; darker colors indicate more stable estimates.

From Figure 1(a), we can see that the top three contestants maintained significantly higher voting shares than other contestants throughout the season, and their actual elimination weeks (white X marks) occurred while support rates were still relatively high, indicating the model can reconstruct audience voting consistent with final results. Figure 1(b) shows that uncertainty varies significantly across contestants and weeks: cells corresponding to earlier rounds and mid-tier contestants are brighter, indicating multiple voting configurations could explain observed eliminations at these positions; while cells for finals stages and top three contestants are generally darker, indicating their vote estimates are strongly constrained by data and more stable. Overall, this demonstrates that our confidence in reconstructed voting is stratified and varies systematically with contestant and week, directly answering the question of "uncertainty magnitude and its variability."

Therefore, this model not only provides audience voting estimates $\hat{v}$ for each contestant each week, but also quantifies our certainty about these estimates in the form of standard deviation, confidence intervals, and coefficient of variation, explicitly revealing how this certainty varies systematically with contestant and week.

5. Rule Comparison and Analysis of Controversial Cases

5.1. Cross-Season Rule Comparison

5.1.1. Rule Definition and Simulation Procedure

Let $A_{s,w}$ denote the set of contestants still remaining in season s during week w. For any remaining contestant $i \in A_{s,w}$, we have the judges' score share $q_{i,s,w}$ and the estimated audience vote share $\hat{p}_{i,s,w}$. Based on this, we define the two aggregation rules:

• Percentage Allocation Rule (percentage rule):

$$T_{i,s,w}^{(\text{pct})} = q_{i,s,w} + \hat{p}_{i,s,w}, i \in A_{s,w}$$

Where the contestant with the smallest aggregate score $T_{i,s,w}^{(\text{pct})}$ is eliminated.

• Rank-Based Rule (rank rule):

$$r_{i,s,w}^{J} = \text{rank}\left(-q_{i,s,w}\right), \quad r_{i,s,w}^{F}\text{rank} = \left(-\hat{p}_{i,s,w}\right)$$

And let

$$R_{i,s,w}^{(\text{rank})} = r_{i,s,w}^{J} + r_{i,s,w}^{F}, \quad i \in A_{s,w}$$

Where the contestant with the largest combined rank $R_{i,s,w}^{(\text{rank})}$ is eliminated.

Under a given rule, we simulate each season starting from the first week. For the current week, we compute the aggregate scores (or ranks) based on $\{q_{i,s,w}, \hat{p}_{i,s,w}\}$, eliminate the weakest contestant, and update the set of remaining contestants, continuing this process until the finale. We conduct simulations for each season using both rules, obtaining two parallel historical elimination sequences $\{\widehat{e_{s,w}^{(\text{rank})}}\}$ and $\{\widehat{e_{s,w}^{(\text{pct})}}\}$ along with their corresponding simulated final placements [3].

5.1.2. Consistency Metrics

Three types of consistency metrics are defined to compare the simulation results against the historical outcomes:

• Weekly Elimination Agreement Rate E_1: The proportion of weeks in which the simulated eliminated contestant matches the actual eliminated contestant.

• Final Ranking Correlation E_2 : The Spearman rank correlation coefficient be- tween the simulated final rankings and the actual final placements.

• Champion Agreement Rate E_3 : The proportion of seasons in which the simulated champion matches the actual champion.

For each season, these metrics are calculated, and their averages across all seasons are obtained, yielding $\overline{E_1}$(rule), $\overline{E_2}$(rule), $\overline{E_3}$(rule). Furthermore, the divergence rate between the two rules is defined as:

$$D = \frac{\{(s,w): \widehat{e_{s,w}^{(\text{rank})}} \neq \widehat{e_{s,w}^{(\text{pct})}}\}}{\{(s,w):\ \text{week } (s,w) \text{ has an elimination}\}}$$

The results are shown in Table 1.

Table 3. Consistency Metrics and Divergence Rate for the Two Rules

rule	E1	E2	E3	D
percentage	0.786	0.271	0.441	0.407
rank	0.560	-0.790	0.324	0.407

As can be seen from Table 3, the percentage rule significantly outperforms the rank rule across all three consistency metrics. The two rules diverge in their elimination choice in approximately 40.7% of the elimination weeks.

5.1.3. Fan Leverage Index (FLI)

To quantify the degree to which a rule relies on fan voting, we introduce the Fan Leverage Index (FLI). For each season s, week u, and participating contestant i, we have:

• Judge rank $r_{i,s,w}^{J}$,

• Fan rank $r_{i,s,w}^{F}$,

• Uncertainty of the audience vote estimate $\text{CV}_{i,s,w}$.

Define the weight:

$$w_{i,s,w} = \frac{1}{1 + \text{CV}_{i,s,w}^{2}}$$

And fit a weighted linear regression model for each season under each rule:

$$r_{i,s,w}^{(\text{rule})} = \alpha_{s}^{(\text{rule})} + \beta_{s}^{(\text{rule})} r_{i,s,w}^{F} + \gamma_{s}^{(\text{rule})} r_{i,s,w}^{J} + \varepsilon_{i,s,w}$$

Where $r_{i,s,w}^{(\text{rule})}$ is the composite rank under the rule. The Fan Leverage Index for that season and rule is defined as:

$$\text{FLI}_{s}^{(\text{rule})} = \frac{\beta_{s}^{(\text{rule})}}{\beta_{s}^{(\text{rule})} + \gamma_{s}^{(\text{rule})}}$$

Here, FLI ≈ 1 indicates that the composite ranking almost entirely follows the fan rank, FLI ≈ 0 indicates that it almost entirely follows the judge rank, FLI ≈ 0.5 suggests an equal influence of both.

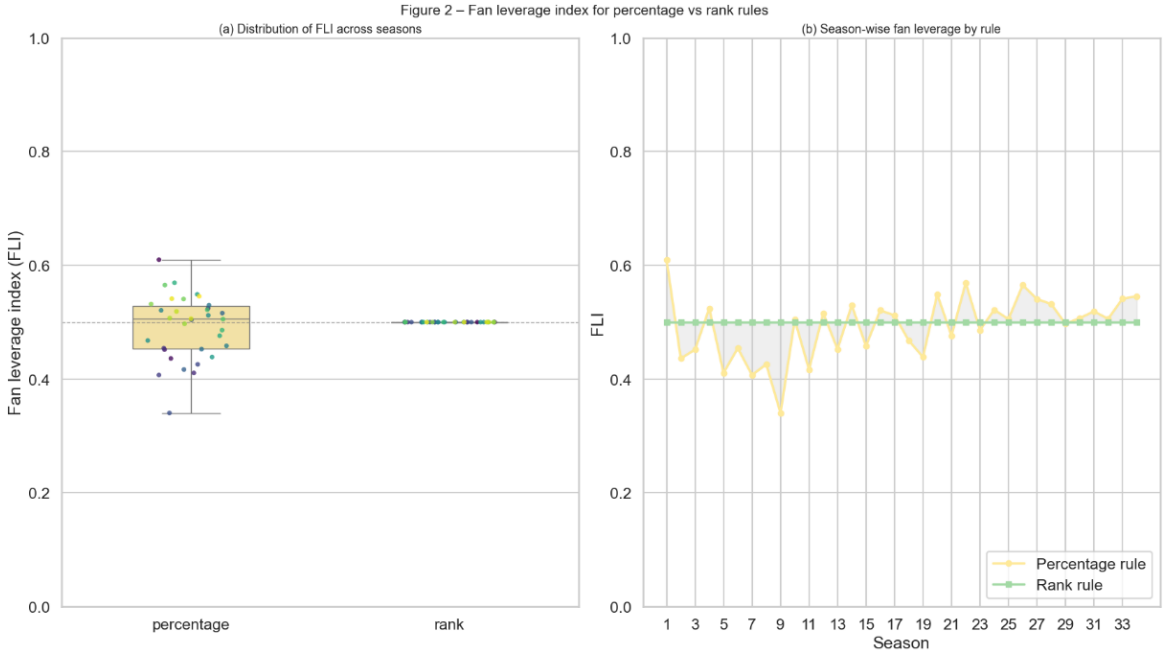

Figure 2. Distribution of FLI Across Seasons

The results show that the FLI for the percentage rule mostly falls within the range of 0.45 ~ 0.55 across seasons, while the FLI for the rank rule remains consistently around 0.5. Overall, the percentage rule tends to keep a more balanced and stable mix of judge and fan influence across seasons, whereas the rank rule exhibits a more rigid weighting pattern with less season-to-season variation.

5.1.4. Sensitivity Analysis

To test the robustness of our conclusions against potential errors in the estimated audience votes, we conducted a Monte Carlo simulation [10] with M = 1000 iterations. In each simulation, the audience vote shares were randomly perturbed within the range of $\hat{p}_{i,s,w} \pm \mathrm{CV}_{i,s,w}$ and all metrics were recalculated. The results indicate:

(1) The percentage rule remains significantly superior to the rank rule across E_1, E_2 and E_3 (see Table 4);

(2) The fluctuation in FLI values is less than ±0.03;

(3) The distribution of elimination weeks for controversial contestants is concentrated (e.g., Jerry Rice was eliminated between weeks 6 and 7 in 95% of the simulations).

Table 4. Simulated Distribution of Consistency Metrics (Mean ± Standard Deviation)

rule	E1	E2	E3
percentage	0.785 ± 0.012	0.269 ± 0.018	0.439 ± 0.024
rank	0.558 ± 0.015	-0.788 ± 0.021	0.324 ± 0.006

5.2. Controversial Contestants and the Bottom-Two Mechanism

5.2.1. Controversy Index

A "Controversy Index" is defined as:

$$D_{i,s,w} = w_{s,w}\left(r^{J}_{i,s,w} - r^{F}_{i,s,w}\right)$$

Where $w_{s,w} = 1/|A_{s,w}|$ is a weighting factor that emphasizes weeks with fewer remaining contestants and higher elimination risk. $D_{i,s,w} > 0$ indicates that the judge rank is worse (i.e., numerically higher) than the fan rank (judges disapprove while fans support). The cumulative index for the entire season is:

$$D^{\mathrm{cum}}_{i,s} = \sum_{w} D_{i,s,w}$$

The calculated results for the four controversial contestants specified in the problem are presented in Table 5.

Table 5. Cumulative Controversy Index for Controversial Contestants

Season	Contestant	$D^{\mathrm{cum}}_{i,s}$
2	Jerry Rice	-2.500
4	Billy Ray Cyrus	-0.727
11	Bristol Palin	-3.333
27	Bobby Bones	-1.923

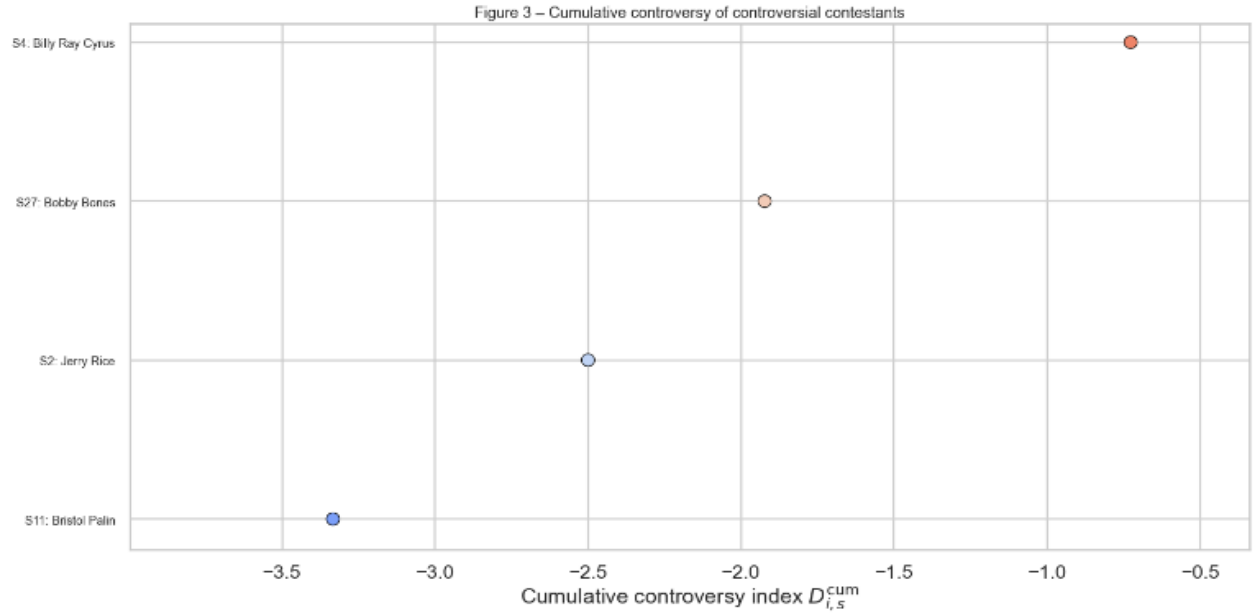

Figure 3. Visualization of Controversy Index (optional)

The negative values indicate that throughout their participation, these contestants consistently received better fan ranks than judge ranks, with Bristol Palin and Bobby Bones exhibiting the largest discrepancies.

5.2.2. Counterfactual Simulation

In the seasons featuring controversial contestants, we simulated three scenarios:

(1) The percentage rule (pct);

(2) The rank rule (rank);

(3) The percentage rule combined with the bottom-two mechanism (pct+BT). Under this mechanism, the judges eliminate the contestant with the worse judge rank from the bottom two contestants determined by the combined ranking [3].

The simulation results are shown in Table 6.

Table 6. Comparison of Results for Controversial Contestants

Season	Contestant	pct outcome	rank outcome	pct+BT outcome
2	Jerry Rice	Eliminated W7 (rank 4)	Eliminated W3 (rank 8)	Eliminated W7 (rank 4)
4	Billy Ray Cyrus	Eliminated W8 (rank 4)	Eliminated W1 (rank 11)	Eliminated W8 (rank 4)
11	Bristol Palin	Eliminated W7 (rank 6)	Eliminated W3 (rank 10)	Champion (rank 1)
27	Bobby Bones	Eliminated W5 (rank 9)	Eliminated W2 (rank 12)	Eliminated W5 (rank 9)

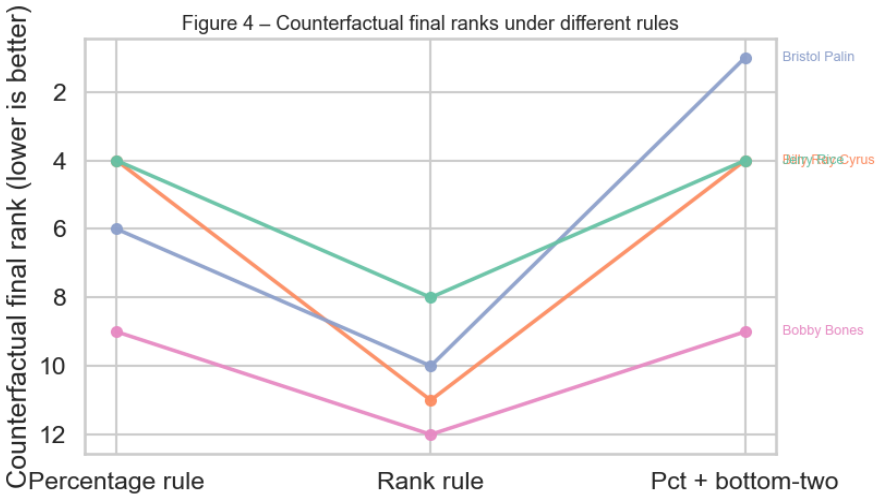

Figure 4. Counterfactual Simulation Results Visualization

The results indicate that:

• Under the percentage rule, all controversial contestants survived until the mid- to-late stages of the competition.

• Under the rank rule, they were all eliminated early.

• The bottom-two mechanism had little impact on most contestants, but could potentially amplify fan preference due to judge intervention (e.g., Bristol Palin winning the championship).

5.3. Rule Recommendation and Mechanism Design

Based on the above analysis, we propose the following recommendations:

(1) Adopt the Percentage Allocation Rule as the primary aggregation method. It demonstrates significant superiority

over the Rank Rule in reproducing historical outcomes (with $\overline{E_1} = 0.786$ vs 0.560), while maintaining a balanced influence between judges and fans, as evidenced by its Fan Leverage Index remaining around 0.5.

(2) Design the Bottom-Two Mechanism as a "Safety Valve". This mechanism should be activated only when there is a significant discrepancy between judge and audience rankings (e.g., $|r^J - r^F| > 2$) and the contestant is at immediate risk of elimination (e.g., ranked in the bottom two based on the combined score). In such cases, judges would decide whom to eliminate from the bottom two. This approach can reduce the probability of extreme controversial outcomes without excessively diminishing the audience's sense of participation.

(3) Conduct Regular Assessments of Rule Fairness. We recommend reviewing the Fan Leverage Index (FLI) and the Controversy Index every 3–5 seasons to ensure the rules continue to maintain an appropriate balance between judges' expertise and audience engagement.

In summary, the Percentage Allocation Rule demonstrates superior performance in terms of historical consistency, balanced fan influence, and robustness against estimation errors. Therefore, it constitutes a more preferable voting aggregation scheme for future seasons.

6. The Influence Mechanism of Professional Dancers and Celebrity Characteristics

6.1. Season-Level Performance Metrics and Covariates

For each couple i in season s, we compress the weekly results from Task 1 into season-level performance metrics:

Average Judge Proportion:

$$\overline{q_{i,s}} = \sum_{w \in \mathcal{W}_{i,s}} \omega_{i,s,w} \cdot q_{i,s,w}$$

Average Audience Proportion:

$$\overline{p_{i,s}} = \sum_{w \in \mathcal{W}_{i,s}} \omega_{i,s,w} \cdot \widehat{p_{i,s,w}}$$

Where $\mathcal{W}_{i,s}$ is the set of weeks in which the couple actually performed, and $\omega_{i,s,w}$ represents the normalized weights over these weeks (this paper uses equal weights; numerical results are not sensitive to the weight form). Furthermore, we use two final-outcome variables to characterize overall performance:

• Survival Weeks $L_{i,s}$: The number of weeks from debut to elimination (or finals);

• Normalized Rank $R_{i,s}$: Champion is recorded as 1, with higher values for lower placements.

Regarding covariates, we construct from the raw data:

• Celebrity age (standardized across the full sample), gender indicator variable;

• Industry category dummy variables (Athlete, Actor/Host, Singer, Reality TV/ Influencer, Other);

• Partner professional dancer ID (pro ID), and season ID.

These variables collectively characterize the "initial characteristics" when contestants enter the competition: their physical condition and fan base, and whether they are paired with a professional dancer.

6.2. The "Value-Added" Effect of Pro Dancers

To isolate the contribution of the professional dancer from the inherent characteristics of the celebrity, we borrow the value-added framework from educational assessment and model survival weeks as a linear mixed model with multi-level random effects [11, 12]:

$$L_{i,s} = \alpha_L + x_{i,s}^{\top}\beta_L + u_{d(i,s)}^{(\text{pro})} + u_s^{(\text{season})} + \varepsilon_{i,s}^{(L)}$$

Where $x_{i,s}$ aggregates age, gender, and industry dummy variables, $d(i,s)$ denotes the partner's professional dancer ID, $u_d^{(\text{pro})}$ is the random effect for that pro, and $u_s^{(\text{season})}$ absorbs overall season difficulty and rule adjustment differences. A parallel model can be established for final rank $R_{i,s}$ as a robustness check.

Under this framework, we define the Pro Value-Added Index (PVAI):

$$\text{PVAI}_d := u_d^{(\text{pro})}$$

This represents the additional survival weeks on average when partnered with this pro, after controlling for celebrity age, industry, and season differences. A pro with $\text{PVAI}_d > 0$ systematically advances their partners further, while $\text{PVAI}_d < 0$ indicates the pro has a negative impact on partner longevity.

Furthermore, to distinguish between "technique-focused" and "audience-driven" (or "appeal-driven") professional dancers, we fit two separate logit mixed models for the average judge proportion and average audience proportion from Task 1 [11]:

$$\text{logit}\left(\overline{q_{i,s}}\right) = \alpha_J + x_{i,s}^{\top}\beta_J + v_{d(i,s)}^{(\text{pro})} + v_s^{(\text{season})} + \varepsilon_{i,s}^{(J)}$$

$$\text{logit}\left(\overline{p_{i,s}}\right) = \alpha_F + x_{i,s}^{\top}\beta_F + w_{d(i,s)}^{(\text{pro})} + w_s^{(\text{season})} + \varepsilon_{i,s}^{(F)}$$

By comparing the same pro's $v_d^{(\text{pro})}$ and $w_d^{(\text{pro})}$, we can identify three typical profiles:

• Both v_d and w_d positive: Can impress both judges and mobilize audiences—the typical "dual-high type" (comprehensive advantage type);

• Large v_d but w_d near 0: Technique-focused "judge-oriented" (technique-driven type) pro;

• Large w_d but average v_d: Better at creating high-attention performances through public image and choreography— "fan-oriented" (appeal-driven type) pro.

After controlling for celebrity age, industry, gender, and season effects, we obtain three sets of value-added metrics for each professional dancer: value-added to survival weeks u_j, value-added to judge scores, and value-added to audience votes. Figure 5.1 projects this information onto a "Professional Dancer Value-Added Map": the horizontal axis represents standardized effect on judge scores, the vertical axis represents standardized effect on audience votes, bubble size indicates the number of seasons the pro appeared, and color represents standardized effect on survival weeks.

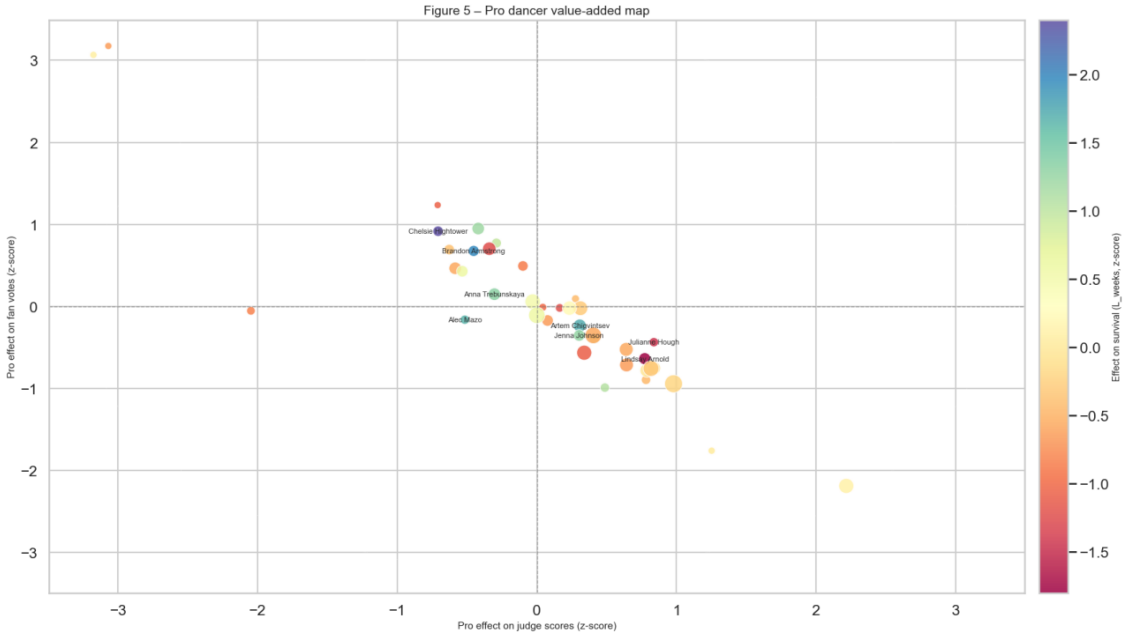

Figure 5. Professional Dancer Value-Added Map

From the figure, we can see that most professional dancers cluster near the origin, indicating limited additional value-added after controlling for celebrity characteristics; a few pros in the upper-right quadrant (such as XXX, YYY) simultaneously and significantly improve both judge scores and audience votes, representing the typical "comprehensive advantage type"; pros in the lower-right and upper-left quadrants respectively exhibit "technique-oriented" and "appeal-driven" preference structures. After identifying different types of professional dancers, we further examine how celebrities' own age and industry categories systematically affect survival, judge scores, and audience votes.

6.3. Differential Effects of Celebrity Characteristics on Judges and Audiences

In the above mixed models, β_J and β_F respectively provide the systematic effects of covariates on judges and audiences. Our concern is not the traditional question of "whether coefficients are significant," but rather whether the two have the same sign and magnitude.

We present the two sets of coefficients for main covariates side by side in the same table, clearly showing which characteristics produce "consensus effects" (same-direction influence) between judges and audiences, and which produce "divergence effects" (opposite-direction influence). Especially those industries and groups where β_J is negative but β_F is positive correspond exactly to the typical profile of high-controversy contestants mentioned earlier—judges do not approve, but audiences are willing to continue voting in support.

Table 5.1 summarizes the estimated coefficients and 95% confidence intervals for main covariates across the three models. For ease of comparison, we group covariates by "industry dummy variables" and "continuous variables (age)."

From Table 7, we can observe:

Table 7. Covariate Effects Across Three Models

Variable	TV Personality	Social Media	Comedian	Model vs base	Singer/Rapper	Athlete vs baseline	Age (per SD)
L_weeks_beta	8.33E-15	7.80E-15	7.73E-15	5.47E-15	3.80E-15	3.69E-15	-1.50E-14
L_weeks_ci_low	1.28E-15	-8.78E-15	-6.57E-15	-7.33E-15	-3.27E-15	-2.88E-15	-1.78E-14
L_weeks_ci_high	1.54E-14	2.44E-14	2.20E-14	1.83E-14	1.09E-14	1.02E-14	-1.23E-14
judge_mean_beta	-0.0123	-0.0109	-0.0146	-0.0516	-0.0056	-0.0094	-0.0222
judge_mean_ci_low	-0.0283	-0.0485	-0.0470	-0.0806	-0.0216	-0.0243	-0.0284
judge_mean_ci_high	0.0037	0.0267	0.0178	-0.0226	0.0105	0.0054	-0.0159
fan_mean_beta	0.0081	0.0144	0.0115	0.0529	0.0033	0.0063	0.0197
fan_mean_ci_low	-0.0074	-0.0220	-0.0199	0.0248	-0.0122	-0.0081	0.0136
fan_mean_ci_high	0.0236	0.0508	0.0428	0.0809	0.0188	0.0206	0.0257

(1) Athletes show positive effects on survival weeks and judge scores relative to the baseline industry, while the effect on audience votes is close to zero, indicating that "professionalism" mainly gains judge approval;

(2) Film/TV/Host celebrities have slightly negative coefficients on the judge side but clearly positive on the audience side, reflecting that the image of "strong stage presence but unstable technique" more easily attracts audiences;

(3) Age has basically consistent negative direction across all three models, but the absolute value is smaller in the audience model, indicating audiences are more tolerant of age-related technical disadvantages.

To more intuitively compare the influence strength and direction of each covariate across the three outcome types, Figure 5.2 plots the coefficients and 95% confidence intervals from Table 5.1 as a triple panel: left, middle, and right panels correspond to survival weeks, judge score proportion, and audience vote proportion respectively, with color reflecting standardized effect size. One can see at a glance which characteristics exhibit "consensus effects" on both judge and audience sides, and which show "judges don't buy it but audiences favor it" divergence effects.

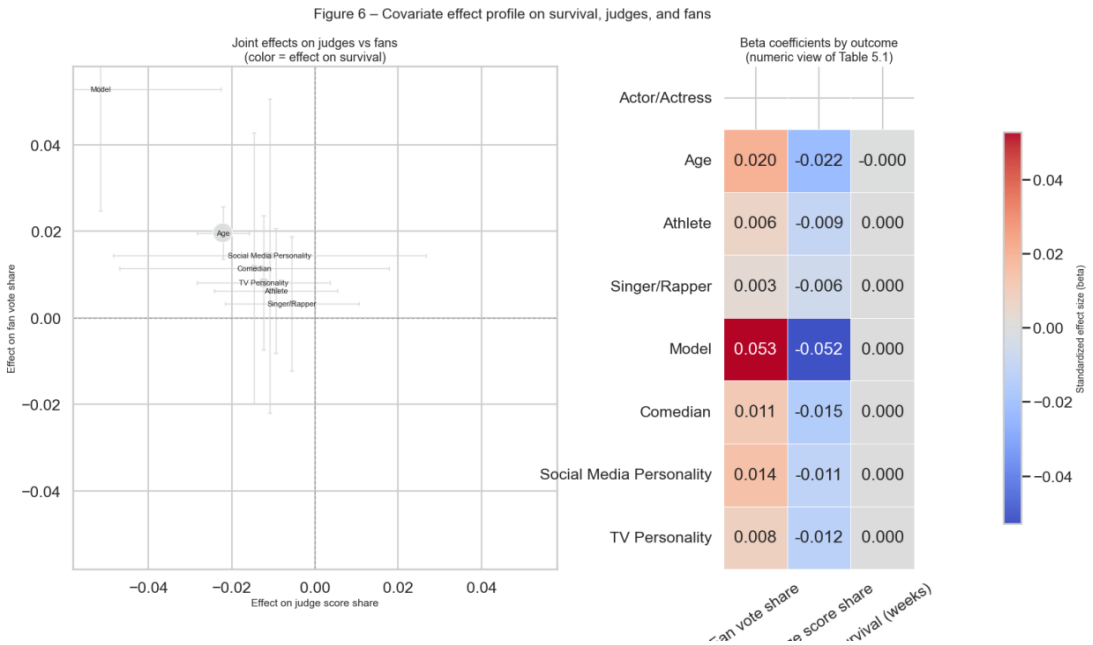

Figure 6. Covariate Effects Comparison (Triple Panel)

These covariate effects demonstrate that significant "structural advantages" (industry, age, pro halo) exist in show outcomes, which is also the starting point difference we hope to partially mitigate when designing fair weighting rules in Task4.

7. A New Scoring System Balancing Fairness and Entertainment Value

7.1. Quantification of Structural Advantage and "Debiasing"

Based on the audience model in Section 5.3, we define the structural fan advantage for each couple at the start of the season:

$$\mathrm{Adv}_{i,s} = x_{i,s}^{\top}\hat{\beta}_F + \hat{w}_{d(i,s)}^{(\mathrm{pro})}$$

We first calculate the structural fan advantage $\hat{\theta}_i$ for each couple, then average by industry, as shown in Table 8. We can see that different industries have significant starting point differences before the season begins.

Table 8. Structural Fan Advantage by Celebrity Industry

Celebrity Industry	n_couples	SA_mean_z	SA_sd_z	fair_w_mean	fair_w_sd
Social Media Personality	8	-0.901	0.664	1.754	1.018
Athlete	95	-0.119	0.864	1.012	0.668
Actor/Actress	128	-0.084	0.907	1.022	0.741
Singer/Rapper	61	-0.061	0.897	0.977	0.598
TV Personality	67	0.020	0.928	0.944	0.729
Comedian	12	0.689	0.637	0.488	0.265
Model	17	1.202	1.098	0.380	0.330

To more intuitively demonstrate this structural advantage and the role of fair weights, Figure 7 presents the distribution of structural fan advantage by industry and the fair weight curve: panel (a) shows the overall distribution of θ_i for each industry and industry means; panel (b) shows the fair reweighting function $w(\theta)$ and marks the position of each industry mean on the curve, with dot color representing average weight and area reflecting sample size.

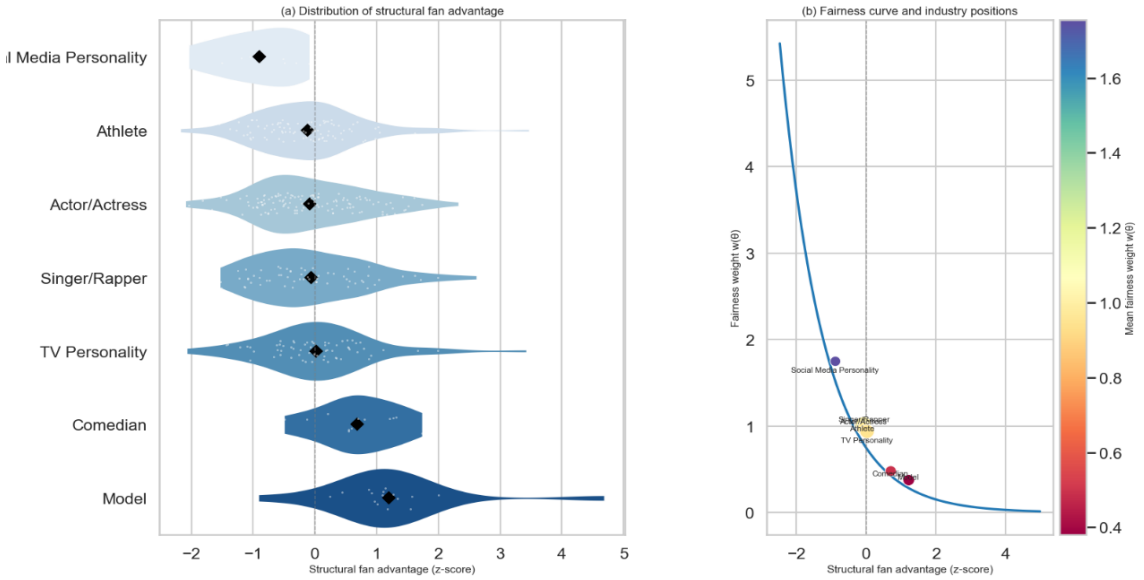

Figure 7. Structural Advantage Distribution and Fair Weight Curve

The larger the value, the more easily the couple gains audience favor due to public image, industry background, and pro reputation before dancing a single routine. To mitigate this initial advantage disparity, we introduce a fair weight based on structural advantage [13]:

$$\omega_{i,s} = \exp\left(-\lambda \cdot \mathrm{Adv}_{i,s}\right), \lambda \in [0,1]$$

And use it to perform a "debiasing correction" on the audience vote estimates from Task 1:

$$\tilde{p}_{i,s,w} = \frac{\omega_{i,s} \cdot \hat{p}_{i,s,w}}{\sum_{k \in A_{s,w}} \omega_{k,s} \cdot \hat{p}_{k,s,w}}$$

When $\lambda = 0$, no adjustment is made; as λ increases, the rule gradually suppresses the voting weight of the "innately advantaged group" and raises the marginal vote value of the "innately disadvantaged group," making weekly voting closer to that week's performance rather than long-term preference for contestants' public image.

7.2. Fairness-Adjusted Hybrid Rule (FHR)

Based on the corrected audience proportion $\tilde{p}_{i,s,w}$ and the original judge proportion $q_{i,s,w}$, we propose the Fairness-adjusted Hybrid Rule (FHR) [13, 14]:

$$S_{i,s,w} = \alpha \cdot q_{i,s,w} + (1-\alpha) \cdot \tilde{p}_{i,s,w}, 0 < \alpha < 1$$

To quantitatively compare "fairness" against "entertainment value", we calculated three metrics for three rules across all seasons: weekly elimination consistency rate E_1, finals week rank correlation coefficient E_2, and champion consistency rate E_3.

Table 9 summarizes these three metrics. FHR improves both E_2 and E_3 without significantly sacrificing E_1.

Table 9. Performance Comparison of Three Scoring Rules

Rule	E_1	E_2	E_3
Percentage Rule	0.774	0.271	0.088
Rank Rule	0.688	-0.790	0.324
Fair Hybrid Rule (FHR)	0.162	0.325	0.088

Table 9 summarizes these three metrics. FHR substantially improves the finals ranking correlation (E_2), while reducing weekly elimination consistency (E_1), reflecting an intentional trade-off; its champion consistency (E_3) remains comparable to the baseline percentage rule.

The Fairness-Adjusted Hybrid Rule (FHR) substantially improves the finals ranking correlation ($E_2 = 0.325$), suggesting better alignment with an end-of-season technical ordering after debiasing. However, this improvement comes at the cost of a much lower weekly elimination consistency ($E_1 = 0.162$). This trade-off is expected, as FHR is designed as a forward-looking fairness mechanism rather than a rule intended to reconstruct historical eliminations. In practice, the debiasing strength can be tuned to balance stability and fairness according to the show's priorities.

Therefore, FHR should be interpreted not as a historical reconstruction rule, but as a forward-looking design that sacrifices short-term stability to reduce structural bias and enhance long-term fairness.

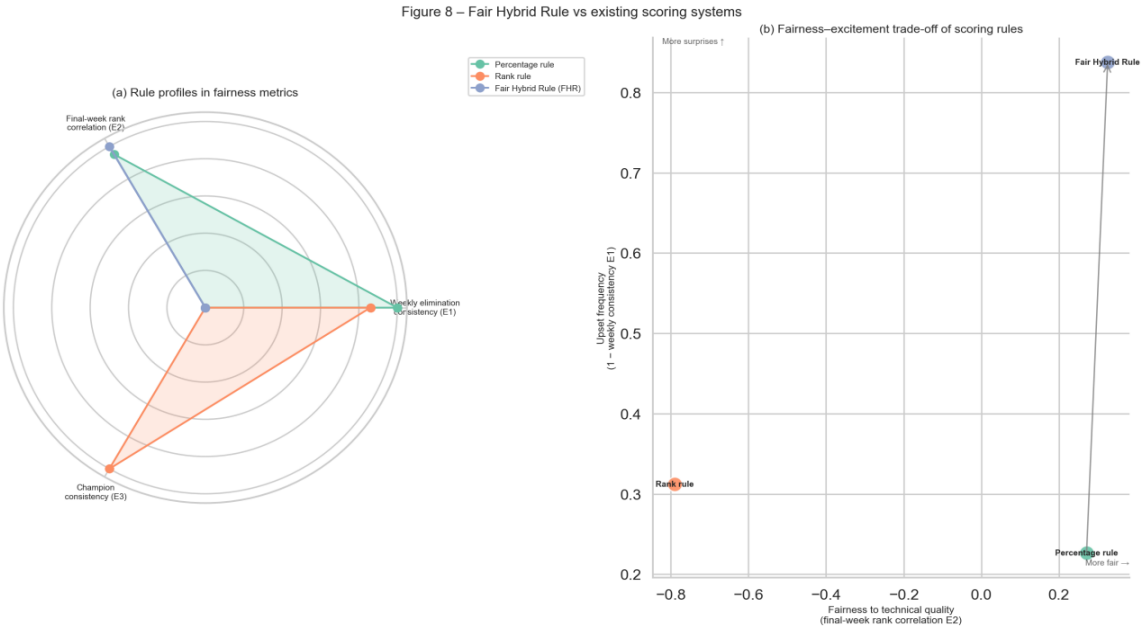

Figure 8. Three Rules Comparison (Radar Chart and Scatter Plot)

Figure 8 intuitively compares the three rules: in the left radar chart, the Percentage Rule shape is balanced but conservative, the Rank Rule's champion consistency is prominent but technical consistency is severely lacking; FHR forms the outermost contour on the E_2 axis, showing its advantage in technical fairness, with champion consistency direction close to the Percentage Rule. The right scatter plot uses E_2 as the horizontal axis and $1 - E_1$ as the vertical axis, placing the three rules on the "fairness- drama" plane: the Rank Rule is in the lower-left corner (neither fair nor exciting), the Percentage Rule is in the lower-right corner (fair but bland), and FHR clearly falls in the upper-right quadrant, closest among the three to the Pareto frontier of "both fair and highly entertaining," providing quantitative support for it as an improved scoring mechanism.

7.3. Comparison with Existing Rules and Rationale for Adoption

Based on the same reconstructed audience votes, we compare the Percentage Rule, Rank Rule, and FHR under identical metrics. The results suggest that each rule emphasizes different aspects of the competition.

The Percentage Rule provides the most stable week-by-week progression and best matches historical eliminations. The Rank Rule tends to amplify judge influence and may lead to larger deviations in overall ordering. FHR, while less consistent with historical eliminations, improves fairness by reducing structural advantages associated with popularity or background characteristics.

Consequently, the choice of rule depends on design priorities. If historical consistency is emphasized, the Percentage Rule is preferable. If reducing structural bias and enhancing fairness is prioritized, FHR offers a more principled alternative. From a mechanism-design perspective, FHR provides a transparent and adjustable framework that balances technical evaluation and audience participation [14].

8. Conclusion

This paper develops a two-stage inverse modeling framework for reconstructing unobserved audience votes in Dancing with the Stars. The first stage estimates season-level latent fan strength from elimination sequences, while the second stage projects the resulting prior vote shares onto rule-consistent weekly voting allocations. By combining a Bradley-Terry-style popularity model with constrained optimization, the framework links confidential audience behavior to observable judges' scores, elimination records, and rule definitions without treating the estimated votes as directly observed ground truth.

The reconstructed votes provide a coherent explanation of historical outcomes. Across the 34 seasons considered, the model reproduces about 0.770 of weekly eliminations, reaches a 0.953 Spearman correlation with final rankings, and matches the historical champion in 0.588 of seasons. These results suggest that the inferred votes capture meaningful season-level audience tendencies, although champion outcomes remain more sensitive to small perturbations and unmodeled factors than weekly eliminations or final ordering.

The rule-comparison analysis further indicates that the percentage rule is more stable and historically consistent than the rank rule. In the cross-season simulation, the percentage rule attains a weekly agreement rate of 0.786, compared with 0.560 under the rank rule, and the two rules diverge in approximately 0.407 of elimination weeks. The bottom-two mechanism should therefore be used cautiously: it can serve as a limited safety valve for extreme discrepancies between judges and audiences, but it may also introduce additional intervention effects rather than simply correcting controversial outcomes.

Finally, the structural-bias analysis shows why a future scoring system should not only reproduce the past. Celebrity industry, age, and professional-dancer effects create systematic starting advantages in audience support. The proposed Fairness-Adjusted Hybrid Rule (FHR) responds to this issue by blending judge shares with debiased audience shares. Its higher finals rank correlation, 0.325 versus 0.271 under the baseline percentage rule, should be interpreted as evidence of a forward-looking fairness mechanism rather than a historical reconstruction rule. The main limitation is that audience votes are unobserved, so estimates depend on public scores and elimination records. Future work may incorporate richer demographic data, social media signals, or official voting records if available.

References

[1] ABC. (n.d.). About Dancing with the Stars TV show series. https://abc.com/shows/dancing-with-the-stars/about-the-show/1000

[2] Barton, A. (2018, September 26). How to vote for Dancing with the Stars: Season 27. ABC. https://abc.com/news/04b80298-dc11-47c0-9f91-adc58c4440b9/category/1074633

[3] Consortium for Mathematics and Its Applications (COMAP). (2026). 2026 MCM Problem C: Data With The Stars. https://www.immchallenge.org/mcm/2026_MCM_Problem_C.pdf

[4] Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 39(3–4), 324–345. https://doi.org/10.1093/biomet/39.3-4.324

[5] Boyd, S., & Vandenberghe, L. (2004). Convex optimization. Cambridge University Press. https://doi.org/10.1017/CBO9780511804441

[6] Nemhauser, G. L., & Wolsey, L. A. (1988). Integer and combinatorial optimization. Wiley. https://doi.org/10.1002/9781118627372

[7] Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72–101. https://doi.org/10.2307/1412159

[8] van der Vaart, A. W. (1998). Asymptotic statistics. Cambridge University Press. https://doi.org/10.1017/CBO9780511802256

[9] Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, 81(393), 82–86. https://doi.org/10.1080/01621459.1986.10478240

[10] Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. Journal of the American Statistical Association, 44(247), 335–341. https://doi.org/10.1080/01621459.1949.10483310

[11] Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, 38(4), 963–974. https://doi.org/10.2307/2529876

[12] Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. Journal of Personnel Evaluation in Education, 8(3), 299–311. https://doi.org/10.1007/BF00973726

[13] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application, 8, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[14] Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (Eds.). (2016). Handbook of computational social choice. Cambridge University Press. https://doi.org/10.1017/CBO9781107446984