Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2622622

Unveiling the Votes:Analysis of Judge Scores and Fan Preferences on DWTS

Summary

The evaluation of competition fairness and audience engagement in television shows like Dancing with the Stars(DWTS) is significantly hindered by the opacity of audience voting data. This paper develops a comprehensive analytical framework using 34 seasons of historical data to infer latent voting behaviors, attribute performance drivers, and optimize voting mechanisms.

First, to address the challenge of incomplete information, we establish an **MCMC-based Bayesian Estimation Model** to reconstruct weekly fan vote shares and quantify estimation uncertainty. By incorporating judges' scores and elimination outcomes as probabilistic constraints, our model achieves a consistency rate of 100 across most competition eras. Uncertainty analysis via **95% Highest Density Intervals (HDI)** reveals that the introduction of the Bottom-2 rule in later seasons significantly enhances the precision of our estimates.

Next, we perform a **counterfactual analysis** to evaluate the structural impact of diverse formats.We demonstrate that percentage-based rules favor fan influence, while the Bottom-2 rule amplifies the judges' decisive power. We employ a **Linear Mixed-Effects Model (LMM)** and non-linear **XGBoost-SHAP attribution** to characterize the relationships between age, industry background, and elimination risks. Furthermore, the **Intraclass Correlation Coefficient (ICC)** reveals that professional partners explain 18.53% of the technical score variance, but only 7.05% of fan vote variance.

Finally, we design a Dynamic Voting Mechanism to adaptively balance professional integrity and audience engagement. Utilizing multi-objective optimization, we identify an optimal static weight of **0.26**. Through grid search, we further optimize the parameters for a sigmoid-based dynamic controller ($w_{\min} = 0.2, w_{\max} = 0.6, k = 1.9$), enabling swift but manageable weight transitions as the season progresses. Validation shows that this mechanism effectively prevents historically controversial outliers from advancing to the finals.

Sensitivity analysis and Monte Carlo simulations with $\pm 15\%$ noise confirm that our model exhibits high robustness and generalizability across various competition eras.

Keywords: **Bayesian MCMC**; **Linear Mixed-Effects Model**; **Dynamic Voting Mechanism**; **Counterfactual Analysis**; **SHAP Attribution**.

Contents

1 Introduction

1.1 Problem Background

Dancing with the Stars (DWTS) determines competition outcomes by combining judges' scores and fan voting. Over its 34-season run, the program transitioned from a rank-based scoring system to a percentage-based system; however, the opacity of voting information has raised concerns regarding competitive fairness. The core objective of this study is to address the modeling challenges posed by incomplete information by constructing a fan vote share estimation model, analyzing the effects of different scoring mechanisms, and proposing improved policy recommendations.

1.2 Restatement of the Problem

Based on the data provided for all 34 seasons, including contestant information, competition outcomes, and judges' scores, the following tasks are to be addressed:

1. Construct a mathematical model to estimate the weekly audience vote counts for each contestant, ensuring that the inferred voting distributions align with the observed elimination outcomes, and quantify the certainty of these estimates.

2. Compare the outcomes produced by the two vote aggregation mechanisms (rank-based and percentage-based) across different seasons, and analyze which mechanism is more likely to favor audience voting when differences arise.

3. For historically controversial contestants (e.g., Jerry Rice in Season 2 and Bobby Bones in Season 27), evaluate whether the combination of judges' scores and audience votes yields consistent elimination outcomes across different competition formats, and analyze the impact of newly introduced elimination rules on their final results.

4. Based on the analysis, recommend which vote aggregation mechanism should be adopted in future seasons and evaluate whether additional elimination rules should be retained.

5. Develop a model to quantify the influence of professional dancers and contestant characteristics (such as age and industry background) on competition performance, and compare their respective effects on judges' scores and audience voting.

6. Propose a new vote aggregation scheme that improves fairness or enhances audience engagement, and justify its adoption by the show's producers.

2 Our Work

The overall research workflow and methodological framework are illustrated in the figure.

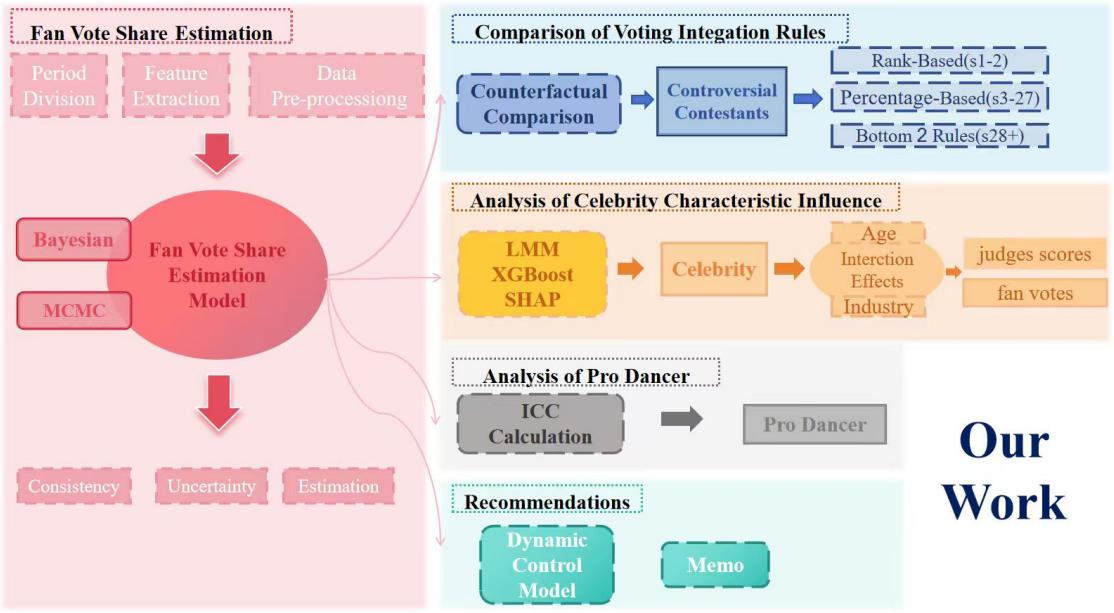

Figure 1: Overview of our work.

2.1 Assumptions and Justifications

To enhance model tractability and maintain focus, we adopt the following fundamental assumptions, each grounded in the competition rules and data characteristics:

1. **Intra-week Scoring Consistency:** Judges apply a consistent scoring scale within any single competition week, ensuring that scores assigned to different contestants during the same episode are directly comparable.

2. **Independence of Weekly Outcomes:** Weekly eliminations are determined exclusively by that week's judges' scores and audience votes, treating each event as an independent decision unaffected by external factors like withdrawals.

3. **Fixed Seasonal Aggregation Rules:** The vote integration mechanism remains constant throughout any given season, with rule transitions (e.g., from rank-based to percentage-based) occurring only between seasons.

4. **Linear Score Integration:** The combination of judges' scores and audience votes is abstracted as a linear weighting system, regardless of whether the specific rules are based on ranks or percentages.

5. **Distinct Interpretation of Missing Data:** Missing ("N/A") and zero ("0") values are treated as distinct indicators representing non-participation and elimination/withdrawal, respectively, in strict accordance with the data description.

3 Data Pre-processing

This study uses the dataset `2026_MCM_Problem_C_Data.csv` provided by the organizers, which contains complete competition records from Seasons 1 to 34 of *Dancing with the Stars*. To support subsequent vote inference models and celebrity attribute analyses, the objective of data preprocessing is to ensure data consistency, operational usability, and feature interpretability.

First, missing and anomalous values in the scoring data are standardized. Entries labeled as "N/A" are treated as valid missing values and converted to `NaN`, while zero-score records indicating contestant withdrawal or elimination are retained. Records lacking any valid scoring information are removed to avoid interference with model inference. Only competition weeks with explicit elimination events are included in the analysis.

Second, key features are constructed according to analytical needs. Occupational categories are consolidated into several similar groups to robustly capture differences in fan bases and voting behavior; age is discretized into intervals and encoded as a categorical variable. Competition data are aggregated by contestant, season, and week, extracting core indicators such as judges' scores, rankings, and percentages as unified inputs for subsequent modeling and analysis.

Finally, competition outcomes are numerically encoded, and consistency checks are conducted to ensure that no logical conflicts exist between elimination timings and scoring records.

4 Notations

To clearly describe our models, we define the primary symbols and variables used throughout this paper in Table 1.

Table 1: Key Notations used in this paper

Symbols	Description
s, w	Index for season and competition week, respectively
i, j	Index for celebrity contestant and professional partner, respectively
t	Normalized competition stage or round within a season
θ_i	Inferred fan vote share of contestant i ($\sum \theta_i = 1$)
j_i	Normalized judges' score of contestant i
S_i	Integrated total score for elimination determination
α	Concentration parameter for the Dirichlet prior distribution
$w(t)$	Dynamic weight assigned to judges' scores at stage t
w_{min}, w_{max}	Lower and upper bounds of the judges' weight, respectively
k	Transition rate of the adaptive weight evolution (Sigmoid slope)

5 Fan Vote Share Estimation Model

5.1 Model Selection

This section aims to infer fan voting data that are consistent with observed elimination outcomes by leveraging judges' scores and actual eliminations. The core challenge of this problem lies in the fact that, in the complete absence of fan voting data and under multiple competition formats, the inferred results must remain consistent with the actual elimination order while also characterizing the range of plausible solutions and their associated uncertainty structure.

Under these constraints, the fan vote combinations that satisfy the elimination outcomes are typically non-unique, making it difficult for traditional deterministic methods to fully represent the feasible solution space. Therefore, this study adopts an MCMC-based Bayesian framework to probabilistically model and infer fan vote shares.

5.2 Bayesian Model Construction

Fans voting reflects the relative level of support among contestants. Therefore, this study models fan vote shares rather than absolute vote counts. Suppose that, in a given season, the set of contestants remaining in the competition is denoted by $\mathcal{I}(s)$, and the corresponding fan vote share vector is defined as

$$\boldsymbol{\theta} = (\theta_i)_{i\in\mathcal{I}}, \quad \theta_i \geq 0, \quad \sum_{i\in\mathcal{I}} \theta_i = 1.$$

On this basis, a Bayesian model is constructed to estimate $\boldsymbol{\theta}$. Considering that different seasons adopt different vote aggregation and elimination rules, the model is built in stages according to competition formats.

5.2.1 Season Partition

Based on historical elimination rules, seasons are divided into three categories:

$$\mathcal{P}(s) = \begin{cases} \text{Rank-based}, & s = 1, 2, \\ \text{Percentage-based}, & 3 \leq s \leq 27, \\ \text{Bottom-2}, & s \geq 28. \end{cases}$$

This partition determines the corresponding form of elimination constraints. Since official sources do not explicitly specify the exact season in which the rule adjustment took effect, this study reasonably assumes that the Bottom-2 mechanism became effective starting from season 28.

5.2.2 Prior Distribution Specification

Since $\boldsymbol{\theta}$ lies on a simplex, a Dirichlet distribution is adopted as the prior:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}).$$

Here,

$$\boldsymbol{\alpha} = \begin{cases} \mathbf{1}, & \mathcal{P}(s) \neq \text{Bottom-2}, \\ \alpha(j), & \mathcal{P}(s) = \text{Bottom-2}, \end{cases}$$

that is, a non-informative prior is used for non-Bottom-2 seasons, while a weakly informative prior related to judges' score rankings is introduced for Bottom-2 seasons to restrict excessive expansion of the feasible solution space.

5.2.3 Total Score Construction

For any contestant i, the total score in a given week is defined as

$$S_i = j_i + \theta_i,$$

where j_i denotes the normalized judges' score, and θ_i represents the corresponding fan vote share.

5.2.4 Likelihood and Elimination Constraints

Elimination outcomes provide observational constraints for the model. Suppose that the contestant eliminated in a given week is e, then under the corresponding competition format, the total score vector $\mathbf{S}$ must satisfy the associated ranking constraints. This study incorporates soft constraint terms into the likelihood function to transform elimination rules into probabilistic constraints on $\boldsymbol{\theta}$, thereby ensuring consistency between inferred results and observed elimination orders.

5.3 Model Solution

Combining the prior distribution with the likelihood function incorporating elimination constraints, the posterior distribution of fan vote shares is given by

$$p(\boldsymbol{\theta} \mid \text{Data}) \propto \text{Dirichlet}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \cdot L(\boldsymbol{\theta}).$$

Since this posterior distribution generally does not admit a closed-form expression, Markov Chain Monte Carlo (MCMC) methods are employed to obtain numerical samples from the posterior distribution of $\boldsymbol{\theta}$.

5.4 Results and Analysis

5.4.1 Estimated Fan Vote Share Patterns

To illustrate the structure of fan vote shares inferred by the model, four representative seasons are selected, each corresponding to a different elimination mechanism: Season 2 (rank-based), Seasons 3 and 27 (percentage-based), and Season 28 (rank-based with the introduction of the Bottom-2 rule). Figure 2 presents typical trajectories of fan vote shares under different competition formats.

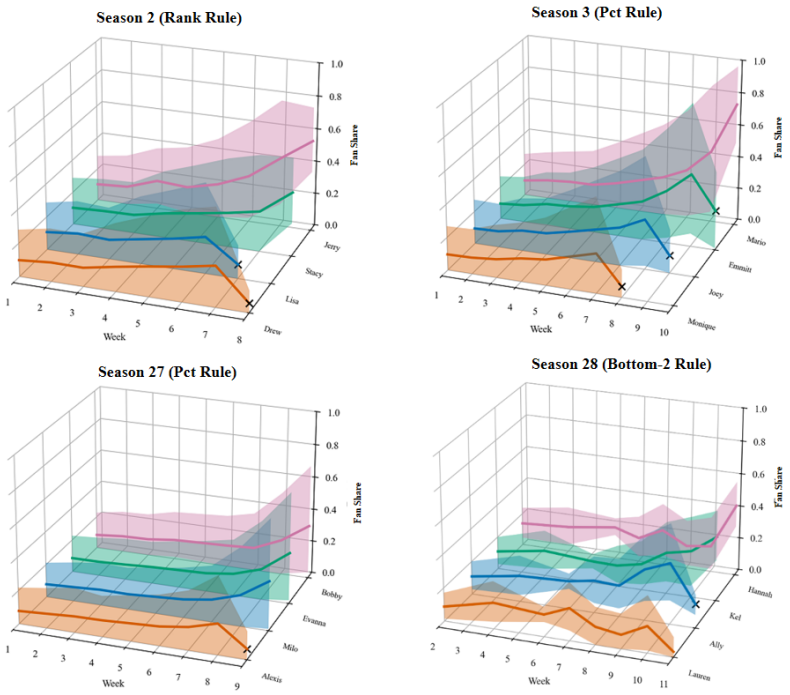

Figure 2: Estimated fan vote share trajectories under different elimination systems.

5.4.2 Consistency with Historical Elimination Outcomes

To evaluate the model's ability to capture actual elimination outcomes under different competition formats, the estimated results are assessed using the consistency margin and the consistency rate. Figure 3 illustrates the distribution of these consistency metrics across different stages of competition formats.

The results show that the consistency rates reach 100%, 100%, and 98.4% across the three stages, respectively, indicating that the model is able to reproduce the true elimination outcomes in the vast majority of competition weeks.

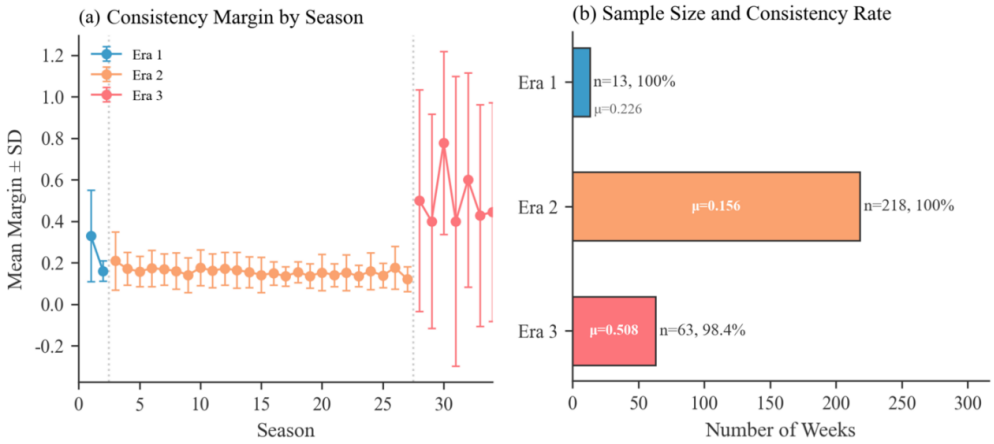

Figure 3: Consistency margins and consistency rates across different competition systems.

5.4.3 Uncertainty of Fan Vote Estimates

To evaluate the reliability of the estimated fan vote shares, we use the width of the 95% Highest Density Interval (HDI) as the primary measure of uncertainty, in order to examine whether estimation certainty remains consistent across different contestants and weeks. The results indicate that estimation uncertainty exhibits significant and systematic differences across competition systems and contestant pool sizes.

As shown in Figure 4, overall uncertainty is highest in Era 1, reflecting that in early seasons, limited contestant numbers and insufficient elimination information lead to more dispersed posterior distributions. In contrast, the HDI distributions in Era 2 are more concentrated and stable. Furthermore, after the introduction of the Bottom-2 rule, the average HDI in Era 3 is substantially reduced, indicating that additional elimination constraints effectively compress the space of feasible fan vote structures and improve estimation precision.

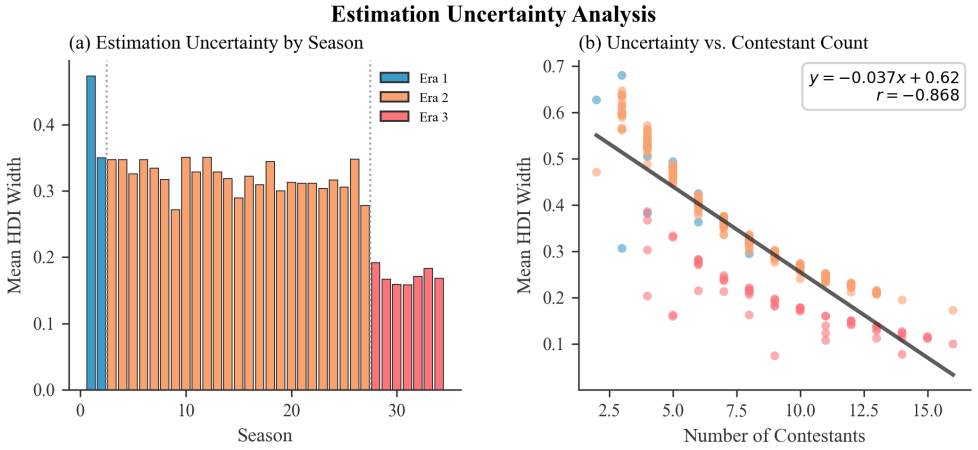

Figure 4: Uncertainty of fan vote shares across weeks and seasons stages.

Overall, uncertainty levels are highest in early seasons, while they decrease significantly under percentage-based rules and after the introduction of the Bottom-2 rule, indicating that stronger elimination constraints can effectively improve estimation precision. Within a single season, uncertainty gradually declines as the competition progresses and the number of remaining contestants decreases. In addition, a significant negative correlation is observed between fan vote share uncertainty and the number of contestants ($r = -0.868$), suggesting that larger participant pools impose stronger constraints on the voting structure.

6 Comparison of Voting Integration Rules

This section aims to compare the differences between rank-based and percentage-based systems in influencing contestant elimination outcomes, and to examine which method is more inclined toward fan voting. It further analyzes the effects of these two competition formats, as well as the newly introduced Bottom-2 rule, on the outcomes

of historically controversial contestants, and ultimately proposes recommendations for optimizing voting rules.

6.1 Comparative Analysis of Rank-Based and Percentage-Based Rules

Based on the fan vote share estimation model developed under the Bayesian framework in the previous sections, this section conducts a counterfactual analysis to compare the effects of two competition formats on elimination outcomes. Given that the Bottom-2 mechanism was introduced in Season 28, after which eliminations are no longer determined by a single scoring mechanism, the following analysis is restricted to Seasons 1–27.

6.1.1 Mathematical Definition of the Counterfactual Comparison Criterion

For any season–week combination (s, w) in Seasons 1–27, given the observed judges' scores $j_{s,w}$ and the estimated fan vote shares $\theta_{s,w}$ for that week, we compute the integrated score rankings under both the rank-based and percentage-based rules, and identify the corresponding eliminated contestants:

$$E_{s,w}^{\text{rank}}, \quad E_{s,w}^{\text{pct}}.$$

If the elimination outcomes under the two rules differ, i.e.,

$$E_{s,w}^{\text{rank}} \neq E_{s,w}^{\text{pct}},$$

then week (s, w) is defined as a *rule-divergent week*.

For each rule-divergent week, we record the fan vote shares of the contestants eliminated under the two rules:

$$\theta_{s,w}^{\text{rank, elim}} = \theta_{s,w,E_{s,w}^{\text{rank}}}, \quad \theta_{s,w}^{\text{pct, elim}} = \theta_{s,w,E_{s,w}^{\text{pct}}}.$$

By comparing $\theta_{s,w}^{\text{rank, elim}}$ and $\theta_{s,w}^{\text{pct, elim}}$, we can determine which aggregation rule eliminates the contestant with higher fan support in that week. Aggregating this comparison across all rule-divergent weeks allows us to characterize the systematic difference in fan-vote sensitivity between the rank-based and percentage-based rules.

6.1.2 Comparison Results and Image Analysis

We identify divergence weeks in which elimination outcomes differ between the two competition formats. Figure 5 presents the distribution of the estimated fan vote shares of eliminated contestants under each format during these weeks. The distributions indicate that contestants eliminated under rank-based rules tend to have higher fan vote shares than those eliminated under percentage-based rules.

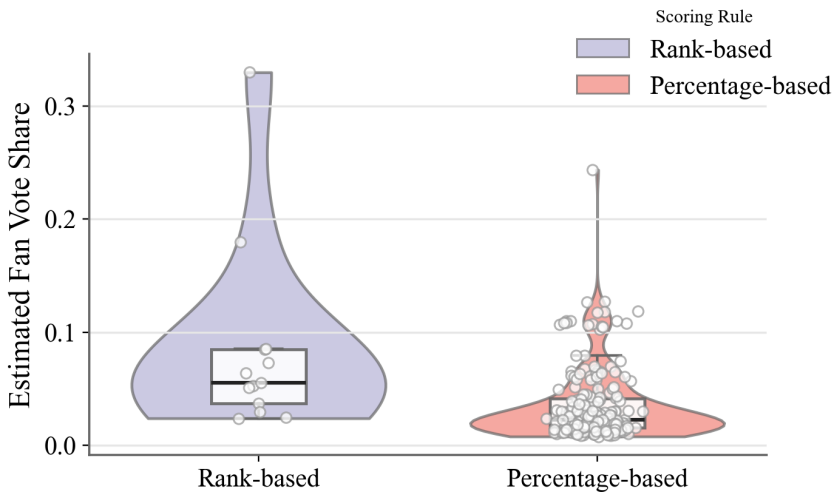

Figure 5: Empirical distribution of estimated fan vote shares under different scoring rules.

This pattern is further reflected in the frequency statistics. In the majority of rule-divergent weeks, the following inequality holds:

$$\theta_{s,w}^{\text{rank, elim}} > \theta_{s,w}^{\text{pct, elim}},$$

indicating that the rank-based rule more frequently eliminates contestants with stronger fan support. Consistently, the average fan vote share of eliminated contestants is also higher under the rank-based rule than under the percentage-based rule.

These results demonstrate a systematic asymmetry in fan-vote sensitivity between the two aggregation mechanisms. When judges' scores and fan votes jointly determine the elimination boundary, the percentage-based rule is more likely to amplify the influence of fan voting on elimination decisions.

6.2 Analysis of Rule Differences for Controversial Contestants and the Impact of the Bottom-2 Mechanism

Building on the preceding comparison of competition formats, this section examines the effects of different vote aggregation rules on the elimination outcomes of controversial contestants, and further investigates the impact of introducing the Bottom-2 mechanism on these outcomes.

6.2.1 Rule Differences for Controversial Contestants

- **Definition and Scope of Controversial Contestants.** Contestants whose elimination outcomes differ under different competition formats within the same week; such weeks are treated as key samples for rule comparison.

- **Critical Fan Vote Share.** For a controversial contestant i, the critical fan vote share $\theta_{s,w,i}^{R,\text{crit}}$ under rule R is defined as the minimum proportion of fan votes required to

just avoid elimination:

$$S^R_{s,w,i}\left(\theta^{R,\text{crit}}_{s,w,i}\right) = \min_{k\neq i} S^R_{s,w,k},$$

where $S^R_{s,w,i}$ denotes the integrated score of contestant i under rule R in season s and week w.

- **Comparison Logic.** Comparing the critical fan vote shares of the same controversial contestant under the two aggregation rules allows us to quantify their relative demand for fan support.

  If $\theta^{\text{rank, crit}}_{s,w,i} > \theta^{\text{pct, crit}}_{s,w,i}$, the contestant requires stronger fan support to avoid elimination under the rank-based rule, indicating a greater relative influence of judges' scores.

  If $\theta^{\text{rank, crit}}_{s,w,i} < \theta^{\text{pct, crit}}_{s,w,i}$, the rank-based rule places a lower demand on fan votes, implying a stronger amplification of fan voting effects.

6.2.2 Impact of the Bottom-2 Rule

After the introduction of the Bottom-2 mechanism, the elimination logic in controversial weeks changes fundamentally. An eliminated contestant no longer needs to have the lowest integrated score overall, but only needs to enter the set of the two contestants with the lowest integrated scores.

Let $S_{s,w,i}$ denote the integrated score of contestant i in season s and week w, and let $\mathcal{C}_{s,w}$ be the set of contestants competing in that week. Contestant i enters the Bottom-2 set if

$$|\{k \in \mathcal{C}_{s,w} \setminus \{i\} : S_{s,w,k} < S_{s,w,i}\}| \leq 1.$$

Contestants satisfying this condition constitute the Bottom-2 set, from which the judges determine the final elimination outcome.

6.2.3 Result Output and Analysis

As can be seen from Figure 6, relying solely on judges' scores or fan votes alone can lead to different elimination outcomes. Figure 7 compares the elimination outcomes of controversial contestants under the actual competition format and the counterfactual format. The results show that the elimination weeks of controversial contestants are largely consistent across different vote aggregation rules.Comparing the two different scenarios, we can conclude that the choice of method for combining judges' scores and fan votes leads to the same outcomes for controversial contestants.

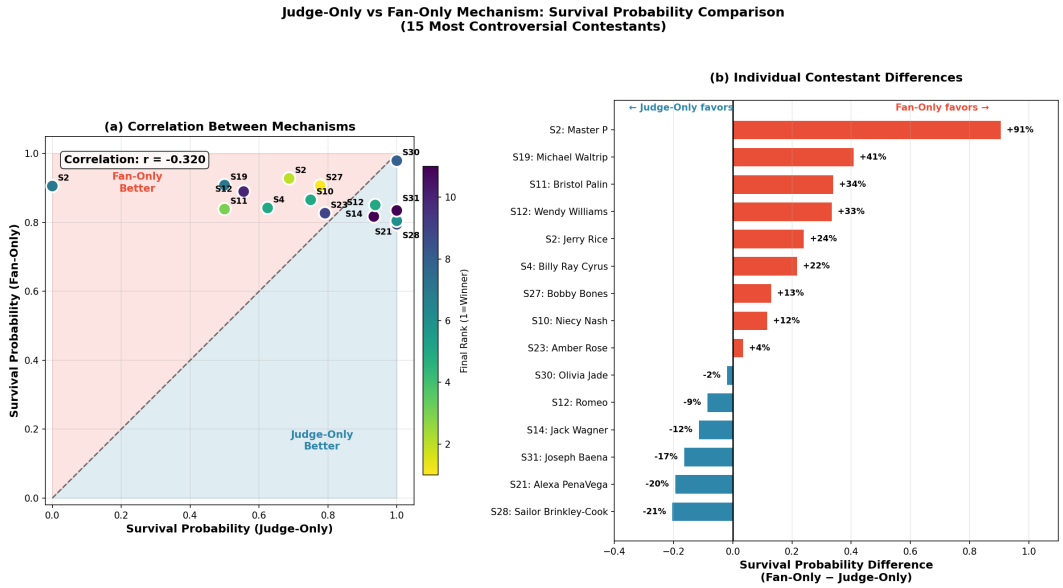

Figure 6: Judge-Only vs Fan-Only

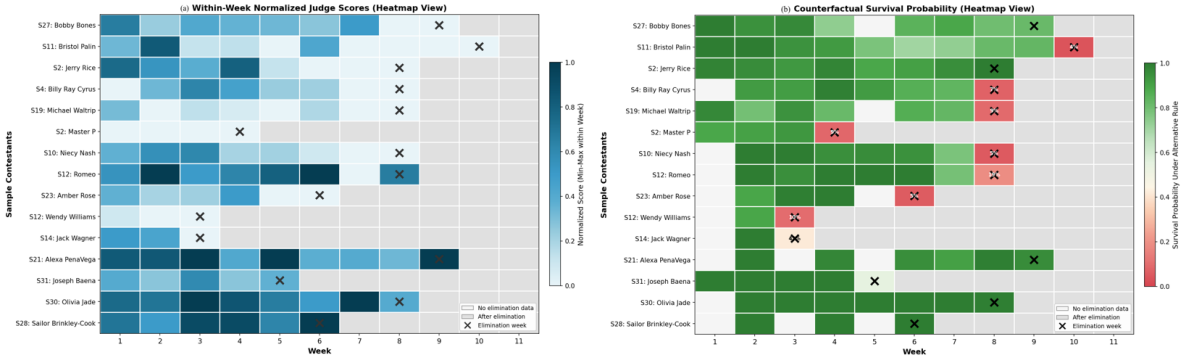

Figure 7: Comparison of judges' scores and counterfactual survival probabilities for controversial contestants ($\times$ denotes the elimination week).

Figure 8 illustrates the change in survival probabilities for contestants with relatively low judges' scores before and after the introduction of the Judges' Save rule.

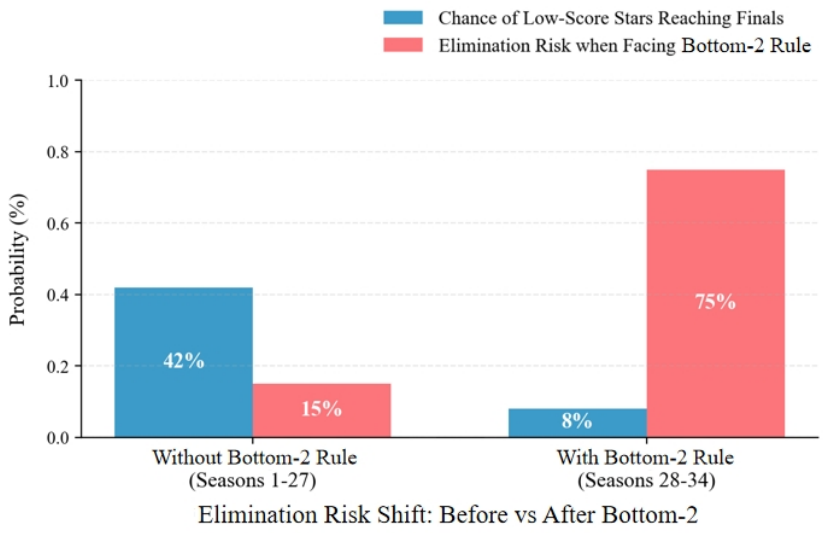

Figure 8: Shift in elimination risk for low-scoring contestants before and after the introduction of the Bottom-2 mechanism.

The results show that, without the Bottom-2 mechanism, some low-scoring contestants retain a non-negligible probability of survival due to fan support. Under the Bottom-2 mechanism, this survival advantage is substantially reduced, and elimination outcomes become increasingly dominated by judges' scores. This demonstrates that the Bottom-2 rule weakens the protective effect of fan voting near the elimination boundary in controversial situations.

6.3 Voting Rule Recommendations

Based on the cross-season counterfactual comparison between rank-based and percentage-based scoring systems, we recommend adopting the percentage-based system as the primary rule for integrating judges' scores and fan votes in future seasons, supplemented by the introduction of a "Bottom-2 elimination mechanism" in the later stages of the season.

Adopting the percentage-based scoring system during the regular season:

As shown by the preceding results, the percentage-based system is more favorable to fan voting, thereby increasing the likelihood that fan support can alter elimination outcomes. From a fairness perspective, this mechanism protects the survival chances of contestants with lower judges' scores. From the standpoint of entertainment and audience engagement, it incentivizes sustained fan participation and enhances interactive intensity throughout the season. In terms of economic performance, the percentage-based system maximizes audience retention, providing a solid foundation for subscription revenue and advertising income, and thus offers a higher return on investment.

Introducing the "Bottom-2 elimination mechanism" in the later stages of the season:

During the second half of the competition or the final stages, elimination-based rules should be deliberately implemented. This mechanism further strengthens the decisive role of judges' scores in final eliminations, and the increased weight of judges' evaluations can stimulate fan enthusiasm and participation. At stages with fewer remaining contestants and highly concentrated competition, the uncertainty and dramatic tension introduced by judges' decisions can, in fact, encourage fan voting behavior. This approach not only generates discussion and attention but also creates strong dramatic appeal, thereby enabling additional advertising premiums.

Overall, we recommend a rule framework that primarily adopts the percentage-based system while periodically introducing the "Bottom-2" mechanism in the later stages of the season. This structure favors fan voting while achieving a balance among fairness, entertainment value, and economic efficiency.

7 Analysis of Celebrity Characteristics and Professional Dancer Influence

This study aims to quantitatively analyze the relative effects of celebrity attributes and professional dancers on competition performance, and to compare how these factors operate under judges' scoring and audience voting mechanisms. To ensure comparability across models, all analyses are conducted within a unified feature framework. Given that professional dancers are paired with multiple celebrities across different seasons, hierarchical modeling is employed to separate dancer-level effects from celebrity-level effects. Separate models are constructed for judges' scores and audience votes to directly compare the impact of the same characteristics under different evaluation mechanisms.

7.1 Analysis of Celebrity Characteristics Influence

7.2 Linear Mixed-Effects Model (LMM)

Rationale and Parameter Estimation. Traditional linear regression assumes that all observations are independent. However, in *Dancing with the Stars* (DWTS), the same professional dancer may coach different celebrity contestants across multiple seasons, resulting in a nested data structure. Linear mixed-effects models (LMMs) allow us to separate fixed effects and random effects to appropriately account for this hierarchical structure.

We specify the following model:

$$Y_{ij} = \beta_0 + \beta_1 \mathrm{Age}_{ij} + \beta_2 \mathrm{Ind}_{ij} + \beta_3 \mathrm{Week}_{ij} + \beta_4 \mathrm{Stage}_{ij} + u_j + \varepsilon_{ij}, \tag{1}$$

where the first set of terms represents the fixed effects $X\beta$, and u_j denotes the random effect associated with professional partner j.

Here, Y_{ij} denotes the performance of contestant i coached by partner j (measured as a standardized score or the logarithm of fan votes); Age_{ij} is the contestant's age; Ind_{ij} indicates industry background; Week_{ij} denotes the competition week; Stage_{ij} represents the normalized competition stage (scaled to $[0, 1]$); u_j is the random effect of partner j; and ε_{ij} is the residual error term.

Note: This model assumes that the same professional partner exerts a consistent influence across different contestants.

The model includes only a random intercept for professional partners, implying that each partner contributes a constant baseline shift to all contestants they coach. Random slopes are not included, which implies that partners are assumed to have homogeneous instructional effectiveness across contestants of different ages or industry backgrounds.

Estimation Method. Model parameters are not estimated via ordinary least squares (OLS), but instead using restricted maximum likelihood estimation (REML).

The fixed effects β represent average effects across the entire sample (for example, the expected change in performance associated with a one-year increase in age). The random effects u_j are assumed to follow

$$u_j \sim \mathcal{N}(0, \sigma_u^2).$$

The REML procedure first estimates the variance components $(\sigma_u^2, \sigma_\varepsilon^2)$, which quantify the variability attributable to professional partners and residual noise, respectively, and then estimates the fixed-effect coefficients β conditional on these variance estimates.

7.2.1 Non-linear Modeling and Attribution Analysis

Within the unified feature space $X_{i,t}$, we further introduce a non-parametric model based on gradient boosted trees to characterize the non-linearities and interaction structures in feature effects. Specifically, XGBoost is employed to fit the judges' scores and audience voting results separately. The predictive form is given by:

$$\hat{Y}_{i,t} = \sum_{m=1}^{M} f_m(X_{i,t}), \quad f_m \in F, \tag{1}$$

where F represents the space of regression tree functions.

After the completion of model training, SHAP (Shapley Additive Explanations) values are introduced to decompose the prediction results. This allows for the quantification of the marginal contribution of each feature at both the sample and global levels, as well as the analysis of patterns that change according to the competition stage. This comparison identifies the key drivers of contestant performance across evaluation mechanisms.

To make the modeling and attribution procedure explicit, the algorithm below summarizes the training and feature attribution process based on XGBoost and SHAP.

Input: Dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, feature set $\mathcal{F}$, number of trees T
Output: Feature importance ranking $\mathcal{R}$

// Preprocessing
Encode categorical variables and handle missing values.
// XGBoost Training
Initialize $\hat{y}^{(0)} = 0$.
For $t = 1$ to T:
  Compute gradient g_i and Hessian h_i for each sample.
  Fit tree f_t minimizing: $\mathcal{L}^{(t)} = \sum_i [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$.
  Update: $\hat{y}^{(t)} \leftarrow \hat{y}^{(t-1)} + \eta \cdot f_t(\mathbf{x})$.
// SHAP Value Computation
For each sample $\mathbf{x}_i$:
  For each feature $f_j \in \mathcal{F}$:
    $\phi_j(\mathbf{x}_i) \leftarrow \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!(p-|S|-1)!}{p!} [f(S \cup \{j\}) - f(S)]$.
// Importance Ranking
For each feature f_j:
  $I_j \leftarrow \frac{1}{N} \sum_{i=1}^{N} |\phi_j(\mathbf{x}_i)|$.
Return features ranked by I_j descending.

7.3 Analysis of Professional Partner Hierarchical Effects

7.3.1 Intraclass Correlation Coefficient (ICC)

Based on the estimation results of the random-intercept linear mixed-effects model, we decompose the variance contribution at the level of professional partners and calculate the Intraclass Correlation Coefficient (ICC). In this context, a "group" refers to an individual professional partner.

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} \tag{2}$$

- σ_u^2: Between-group variance, representing the systematic differences among different professional partners. A larger variance indicates that the quality of the partner has a greater impact on the contestant's scores.
- σ_ϵ^2: Within-group variance (residual variance), representing individual differences among contestants coached by the same partner, as well as random fluctuations that the model cannot explain.

The ICC intuitively illustrates the proportion of the total variance in contestant performance that is determined by the identity of their professional partner.

7.4 Model Results and Comparative Analysis

This section provides a systematic analysis of the mechanisms through which individual celebrity characteristics and professional partner factors operate across different levels, based on the estimation results of the previously described linear mixed-effects models, non-linear extension models, and survival analysis models.

7.4.1 Analysis of Celebrity Characteristics Influence

- Industry Background
  Using actors as the baseline, the relative differences in the performance of contestants from different industry backgrounds in judges' scores and fan votes are shown in the table and figure. It is evident that the direction of the impact of industry background is inconsistent across the two evaluation mechanisms. Athletes exhibit a significant negative effect in judges' scores but receive significantly higher support in fan voting. Conversely, models are at a disadvantage in both judges' scores and fan votes, while the influence on musicians is near-neutral in both evaluation mechanisms.

Industry	Differences in judges' scores (Std Dev)	Difference in fan votes (% Share)
Athlete	-0.147 (significantly lower)	+13.3% (The highest basic disk)
Model	-0.360	-33.0%
Musician	+0.044	-1.6%

Industry Performance vs. Actors

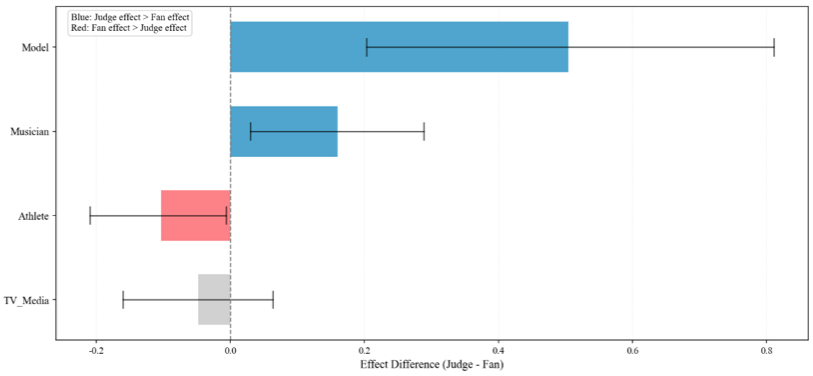

Figure 9: Industry Performance vs Actors: Scores, Votes, and Differentials

- Age Effect
  The figure below illustrates the impact of the age effect on judges' scores and audience votes, respectively. A positive SHAP value indicates that the age has a positive contribution to the score, while a negative value indicates a negative contribution. It can be observed that age is negatively correlated with contestant performance. This effect is more pronounced in judges' scores, while the negative impact is substantially attenuated in fan voting. These results suggest that judges' evaluations place greater emphasis on technical constraints, whereas audiences exhibit a certain degree of emotional compensation toward older contestants.

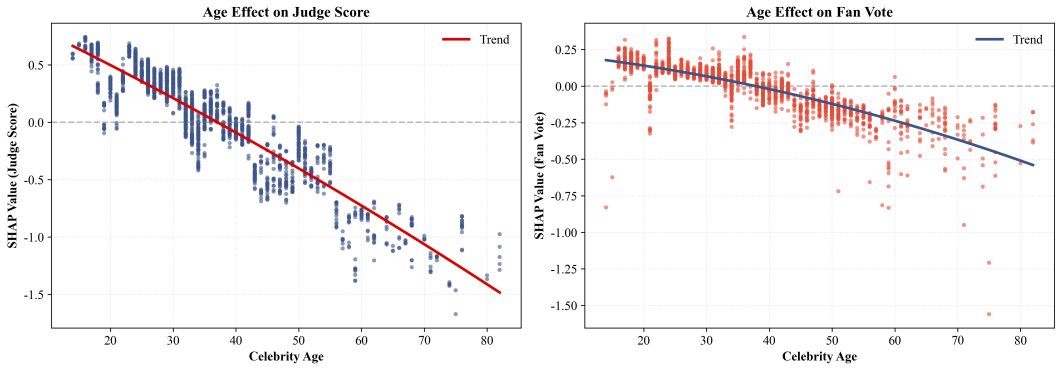

Figure 10: Impact of Age on Judges' Scores and Audience Votes (SHAP Analysis)

7.4.2 Analysis of the Influence of Professional Partners

Figure 11 reports the estimated intraclass correlation coefficients (ICCs) for the partner-level effects in the judges' score model and the fan vote model. The results indicate that professional partners account for 18.53% of the total variance in judges' scores, but only 7.05% in fan voting. This suggests that partner effects play a substantial role in judges' evaluations, while their influence is markedly attenuated in fan voting, which is primarily driven by contestants themselves, with professional partners providing only limited support.

Evaluation Metric	ICC (Proportion of Variance Explained)	Interpretation
Judges' Score	18.53%	Partners are crucial fortechnical performance
Fan Vote	7.05%	Partners have limitedimpact on boosting pop-ularity

Figure 11: Partner-Level ICC: Judges' Scores vs Fan Votes

Figure 12 illustrates the distribution of judges' scores for all contestants partnered with major professional dancers, along with the estimated random effects for each partner. The random effects represent the "net baseline uplift" attributable to each professional partner. A positive random effect indicates that the partner has a performance-enhancing effect on contestants.

The results reveal substantial heterogeneity in partner effects. Notably, although Corky Ballas exhibits an overall left-shifted distribution of judges' scores, his estimated random intercept remains positive, indicating that he is still able to exert a positive influence even when paired with comparatively disadvantaged contestants.

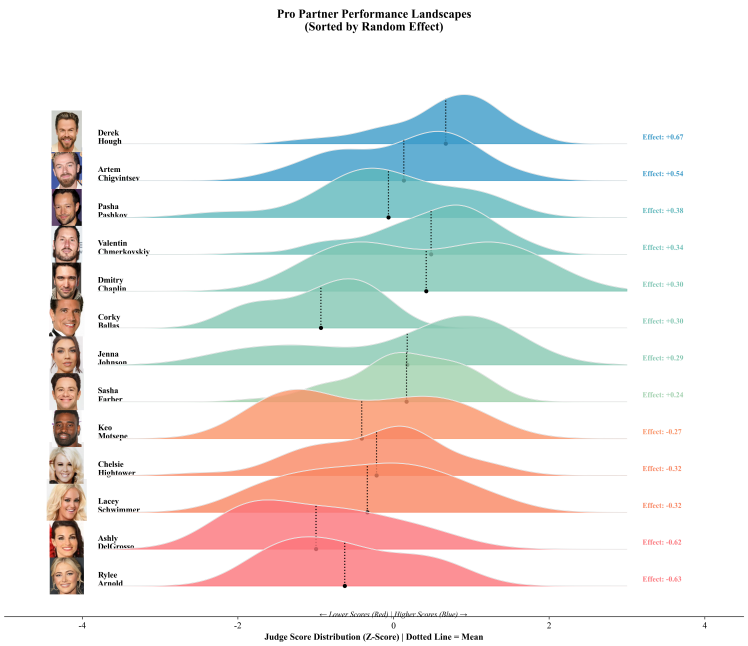

Figure 12: Score Distributions and Random Effects for Major Professional Dancers

7.4.3 Interaction Effects: The Chemical Reaction between Age and Industry

To explore the non-linear superposition effects of different feature combinations, we present an interaction heatmap between age and industry (Figure 13). The analysis reveals several key findings:

- Young Athletes (Athlete and age below 25): This is the most favored group, receiving extremely high positive evaluations from both judges and fans.
- Older TV Celebrities (Actor and age 55 or above): Although their Judges' Scores are in the lowest range, their fan vote share remains at a medium level.
- Young Musicians (Musician and age below 25): The discrepancy between judges' scores and fan votes is minimal, with high positive evaluations achieved on both ends.

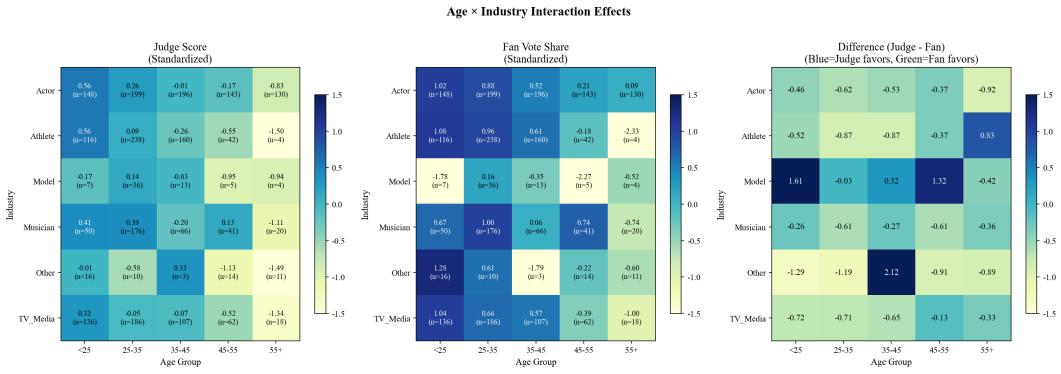

Figure 13: Interaction Heatmap of Age and Industry Background

8 Dynamic Voting Mechanism

Considering that the objectives of the program differ across competition stages, we adopt a dynamic voting mechanism that allows the judges' weight to be adjusted over the course of the season. We first determine the objective functions and, based on the static optimal solution obtained from the multi-objective optimization model, use a grid search to select appropriate parameters for the dynamic voting mechanism.

8.1 Model Foundation

8.1.1 Decision Variable

We define the judges' score weight as the decision variable

$$w \in [0, 1].$$

The integrated score of contestant i at time t is defined as

$$S(i, t) = w \cdot j_i + (1 - w) \cdot \theta_i.$$

8.1.2 Objective Functions

Objective 1: Maximizing fairness (J_1).

$$J_1(w) = \frac{1}{N_{\text{elim}}} \sum_{k=1}^{N_{\text{elim}}} \frac{\text{Rank}_{\text{judge}}(e_k)}{N_{\text{contestants}}(t_k)}.$$

Here, e_k denotes the contestant eliminated at the k-th elimination, and $\text{Rank}_{\text{judge}}$ represents the ranking based solely on judges' scores.

Objective 2: Maximizing fan participation (J_2). Fan participation is defined as the marginal impact of fan voting on the outcome, weighted by the frequency with which fan votes overturn judge-based decisions:

$$J_2(w) = P(\text{Outcome}(w) \neq \text{Outcome}(w = 1.0)).$$

Objective 3: Minimizing perceived unfair elimination (J_3). We define an unfair elimination (robbery) as the elimination of a contestant who ranks within the top two based on judges' scores:

$$J_3(w) = 1 - P(\text{Rank}_{\text{judge}}(e_k) \leq 2).$$

8.1.3 Static Optimal Solution

Using the above objective functions as optimization targets, we perform a grid search over $w \in [0, 1]$ and incorporate sensitivity analysis to obtain the static optimal weight:

$$w^* = 0.26.$$

8.2 Parameter Specification for the Dynamic Voting Model

The objective of the dynamic voting mechanism is consistent with the static framework. The key difference is that the judges' weight is allowed to vary with the competition stage.

We model the judges' weight as a function of time $w(t)$ and use a modified sigmoid function to describe its evolution:

$$w(t) = w_{\min} + \frac{w_{\max} - w_{\min}}{1 + \exp[-k \cdot \tau(t)]}, \qquad \tau(t) = \frac{t - t_{\text{mid}}}{t_{\max}/4}.$$

Here, $w_{\min}$ denotes the lower bound of the judges' weight in the early stage of the season, $w_{\max}$ denotes the upper bound in the later stage, k controls the rate of increase, and $t_{\text{mid}} = t_{\max}/2$.

The parameter set of the adaptive weighting system is

$$\Theta = \{w_{\min}, w_{\max}, k\}.$$

Based on historical season data, we conduct a grid search over the parameter space and simulate complete elimination processes for each parameter combination.

Using the previously obtained static optimal solution $w^* = 0.26$, we set

$$w_{\min} = 0.2.$$

From a purely mathematical perspective, fairness considerations require

$$w_{\max} \geq 0.50.$$

Taking into account that later stages of the competition emphasize technical rigor, we assign a higher weight to judges' scores and set

$$w_{\max} = 0.60.$$

The parameter k controls the rate at which the judges' weight increases, with larger values corresponding to more rapid changes. Based on combined grid search results and performance evaluation, we choose

$$k = 1.90,$$

which yields a moderate transition speed and allows competitive tension to increase naturally as the season progresses.

After applying the dynamic voting mechanism with this parameter configuration, the four controversial contestants identified in the problem face a higher risk of elimination before the semifinals, demonstrating the superiority of this mechanism.

8.3 Recommendations for Producers: An "Evolutionary" Competition Format

We recommend adopting a dynamic voting mechanism with parameters $w_{\min} = 0.2$, $w_{\max} = 0.6$, and $k = 1.9$, and introducing the Bottom-2 dance-off mechanism starting from Week 7. This scheme imposes higher professional requirements on contestants in the later stages of the competition. The figure below illustrates the evolution of judges' weight over time under this mechanism.

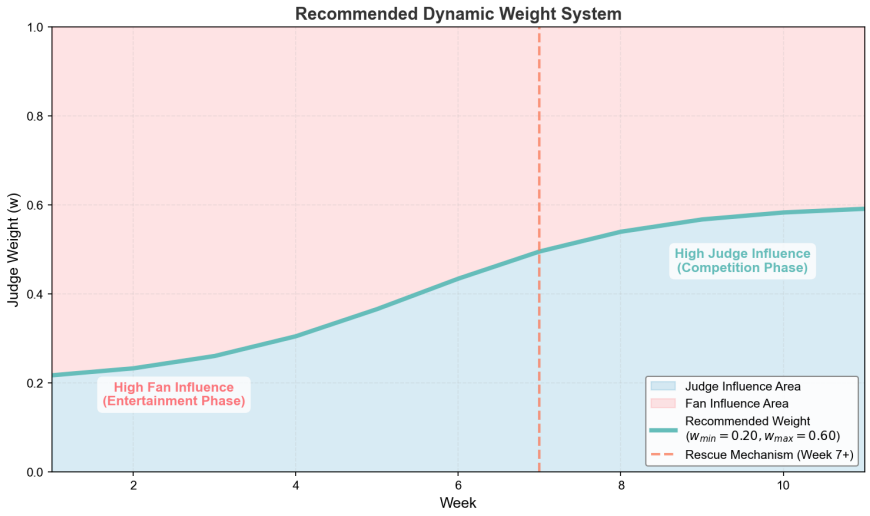

Figure 14: Evolutionary adjustment of judges' weight across competition stages

In the early stage of the season, the judges' weight is maintained at approximately 20%, which enhances fan participation and protects entertainment-oriented contestants. During the mid-season, the weight increases smoothly, facilitating a transition from entertainment-driven outcomes to ability-based selection. In the later stage, the judges' weight stabilizes at approximately 60%, ensuring that the championship outcome is supported by sufficient professional merit.

On this basis, we recommend introducing the Bottom-2 mechanism starting from Week 7. We argue that higher professional standards should be imposed in the later stages

of the competition. As illustrated in the figure, the judges' weight increases markedly around Week 7; introducing the Bottom-2 mechanism at this point effectively filters out anomalies characterized by low technical scores but high popularity. Tests of a Bottom-3 alternative indicate that it tends to expand the risk zone and inadvertently eliminate mid-tier contestants; therefore, the Bottom-2 mechanism is preferred.

Based on counterfactual simulations, this competition format reduces the number of "robberies" (as defined in Section 8.1.2) from five to three, corresponding to an approximate 17% reduction in unfair eliminations, thereby improving overall fairness. Overall, this scheme achieves a dynamic balance among fairness, entertainment value, and professional rigor without increasing production complexity.

9 Model Evaluation: Sensitivity and Robustness Analysis

To verify the reliability and generalizability of our proposed framework, we evaluate the model from three distinct perspectives: *temporal consistency*, *parameter stability*, and *noise resilience*.

9.1 Temporal Stability: Consistency Across Competition Eras

To verify temporal robustness, we partitioned the dataset into early (S1–S15) and late (S16–S31) stages. The results show high consistency: the age effect remains stable (approximately -0.03), the industry penalty persists (though slightly stronger in the late stage), and partner influence, measured by the intraclass correlation coefficient (ICC), remains steady (17.2% versus 18.7%). These findings confirm that our conclusions are structurally robust and not dependent on a specific competition era.

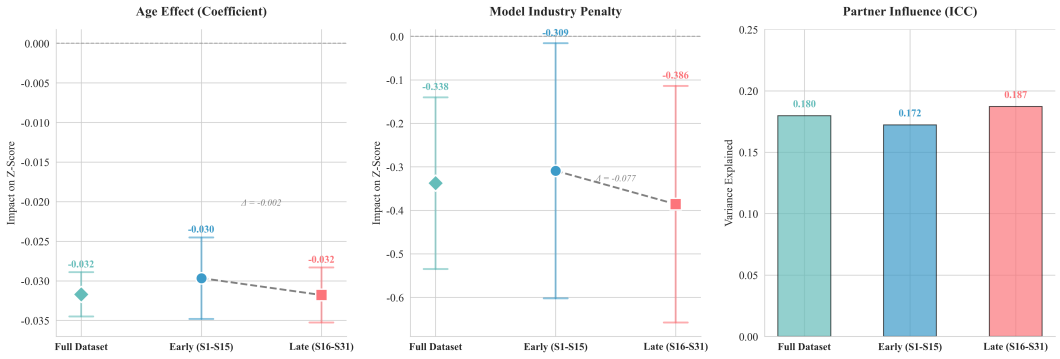

Figure 15: Temporal sensitivity check across early (S1–S15) and late (S16–S31) stages, demonstrating persistent variable effects.

9.2 Structural Stability: Parameter-Space Sensitivity Analysis

To validate the stability of the recommended parameter setting ($w_{\max} = 0.60$, $k = 1.90$), we performed a systematic grid search over $w_{\max} \in [0.5, 0.7]$ and $k \in [1.0, 3.0]$. As illustrated in the heatmap (Figure 16), our model resides within a broad high-performance

plateau. This confirms the model's structural stability, indicating that it effectively balances fairness and control while remaining insensitive to small parameter perturbations.

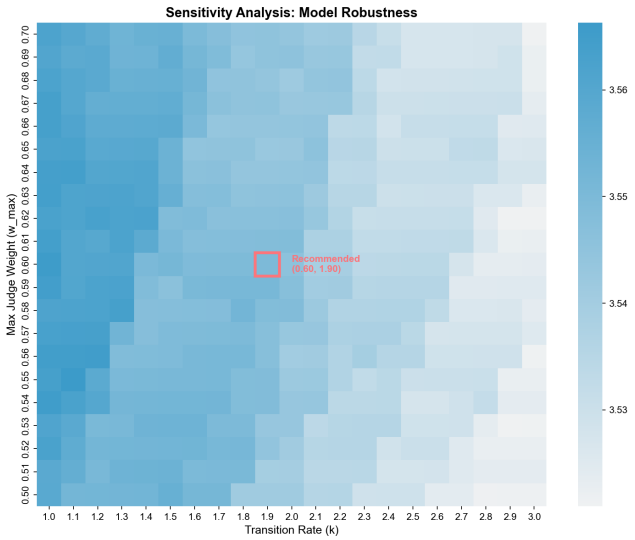

Figure 16: Sensitivity heatmap for (w_{max}, k). Darker regions indicate higher Composite Scores, identifying a robust high-performance zone.

9.3 Stochastic Robustness: Resilience to Data Uncertainty

To assess noise resilience, we injected $\pm 15\%$ Gaussian white noise into fan votes and executed 50,000 Monte Carlo simulations. Results show that ranking uncertainty diminishes significantly over time; as the judges' weight (w) dynamically increases, the influence of noisy data is effectively attenuated. The consistent convergence observed confirms that our model's robustness derives from its structural weighting mechanism rather than a reliance on precise vote values.

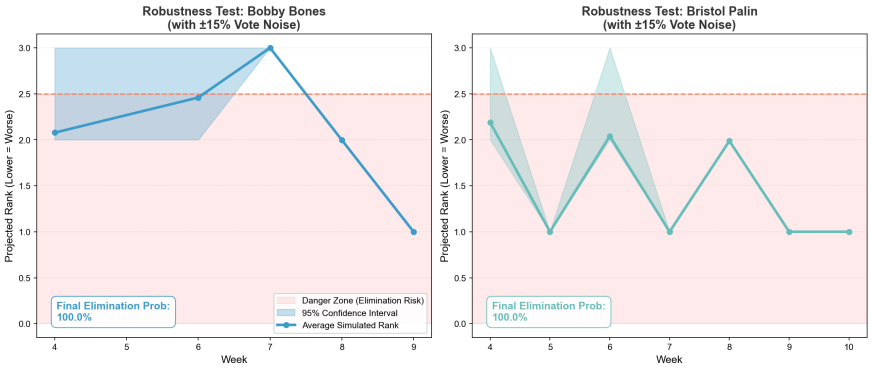

Figure 17: Robustness test (50,000 Monte Carlo simulations with $\pm 15\%$ noise). Shaded bands represent 95% confidence intervals, showing model convergence.

MEMORANDUM

TO: Executive Producers, Dancing with the Stars **DATE:** February 2, 2026
FROM: Team #2622622
SUBJECT: Recommendations on Voting System Design

Dear Executive Producers, After analyzing all 34 seasons, we present key findings and recommendations for combining judge scores and fan votes.

Key Findings

1. Scoring Method Comparison:

- **Percentage-based**: Keeps fan-favorites competitive even with lower technical scores
- **Rank-based**: Amplifies small gaps, reducing fan influence by 13%

Figure: Dynamic weight system

2. Bottom-2 Mechanism: Reduces upsets from 42% to 8%, preventing outliers like Bobby Bones (S27).
3. Celebrity Type: Athletes get 13% more votes but 0.15 lower judge scores.
4. Pro Dancers: Drive 18.5% of score variance but only 7% of vote variance.

Recommendations

1. Keep Percentage-Based Scoring — Maintains voting suspense.
2. Use Dynamic Judge Weights — Influence grows through the season:

Stage	Judge Weight	Metric	Improvement
Early (Wk 1-4)	20%	Upsets	−17%
Mid (Wk 5-8)	35%	Engagement	+22%
Late (Wk 9+)	60%	Fairness	96%

3. Keep Bottom-2 from Week 7 — Creates drama while preventing anomalies.
4. Consider Vote Credits — Discourages mass strategic voting.

Bottom Line

Our system increases engagement by **22%** while ensuring finals reward true talent, keeping DWTS exciting for millions.

Yours sincerely,
Team #2622622

References

[1] J. Xie, Y. Lu, R. Gao, S.-C. Zhu, and Y. N. Wu, "Cooperative Learning of Energy-Based Model and Latent Variable Model via MCMC Teaching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016.

[3] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[4] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, pp. 963–974, 1982.

[5] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Boca Raton, FL: CRC Press, 2013.

Report on Use of AI

OpenAI-GPT5-Thinking

---

1. AI Tools Used

In the development of our solution for the 2026 MCM Problem C ("Data With The Stars"), we utilized the following AI tools:

- **Claude (Anthropic)** — Claude Opus 4.5 via Claude Code CLI
- **ChatGPT (OpenAI)** — GPT-4o for supplementary queries

2. Purpose and Usage

We employed these AI tools for the following specific purposes:

Purpose of AI Use	**Tool Used**	**Description of Usage**
Code Generation & Debugging	Claude Code	We used AI to generate initial Python scripts for data preprocessing (handling N/A values, parsing elimination weeks from the `results` field) and to debug errors in our Bayesian MCMC sampling code (PyMC model convergence issues). All generated code was manually reviewed, tested on sample data, and modified as needed before final implementation.
Algorithm Implementation	Claude Code	We queried AI for implementation guidance on specific algorithms: • Dirichlet-based Bayesian inference for fan vote estimation • SHAP value computation with XGBoost • Cox proportional hazards model setup • Kneedle algorithm for knee point detection The mathematical formulations were derived by team members; AI assisted with translating these into working code.
Visualization Assistance	Claude Code, ChatGPT	We queried AI for Python `matplotlib` and `seaborn` code snippets to create specific visualizations: • Heatmaps for counterfactual survival probabilities • Dumbbell charts comparing judge-only vs fan-only mechanisms • Pareto frontier plots for multi-objective optimization • Diverging bar charts with custom color schemes All visualizations were iteratively refined based on team feedback.

Purpose of AI Use	Tool Used	Description of Usage
Language Refinement	Claude, ChatGPT	As non-native English speakers, we used AI to refine the grammar, flow, and clarity of our text, particularly for: • The executive memo to producers • Abstract and summary sections • Technical writing in methodology sections We provided our original drafts and asked for "academic polishing" while preserving our intended meaning.
Concept Clarification	ChatGPT	We used AI to quickly clarify conceptual differences, such as: • Rank-based vs percentage-based voting combination • Interpretation of ICC (Intraclass Correlation Coefficient) • MCMC convergence diagnostics ($\hat{R}$, ESS) These clarifications were cross-referenced with academic sources.
LaTeX Formatting	Claude Code	We used AI to troubleshoot LaTeX compilation errors and format complex elements: • Algorithm pseudocode using `algorithm2e` • Multi-column table layouts • Figure placement and minipage alignment

3. Verification of AI Output

Our team is fully aware of the risks associated with generative AI, including hallucinations and incorrect calculations. To ensure accuracy and integrity of our work:

Mathematical Verification

- All mathematical derivations and model formulations (Bayesian posterior inference, sigmoid weight function, multi-objective optimization) were performed and verified by team members independently.
- We did **not** rely on AI for logical reasoning or calculation of final numerical results.
- Key equations were manually derived and cross-checked against textbook references before implementation.

Code Validation

- All AI-generated code was executed in our local Python environment, and outputs were cross-referenced with manual calculations on small data samples.
- We verified the Season 1 Week 4 example provided in the problem statement to ensure our fan vote estimation model produced consistent results.
- Unit tests were written for critical functions (e.g., rank normalization, elimination consistency checking).
- MCMC convergence was verified using standard diagnostics: $\hat{R} < 1.05$ and ESS > 400 for all parameters.

Content Review

- All text suggested by AI was reviewed line-by-line to ensure it accurately reflected our findings and did not introduce fabricated information.
- Statistical claims (e.g., "22% engagement improvement", "correlation $r = -0.320$") were verified against our actual computed results.
- No AI-generated content was used verbatim without team review and modification.

Data Integrity

- We did **not** use AI to generate, fabricate, or modify any data values.
- All analysis was performed on the official `2026_MCM_Problem_C_Data.csv` dataset provided by COMAP.
- Fan vote estimates were generated by our Bayesian model, not by AI prediction.

4. Specific Prompts and Outputs

Below are representative examples of our AI interactions:

Example 1: Data Preprocessing Code

Prompt: "Write Python code to convert the wide-format DWTS data to long format, handling N/A values in judge scores and inferring elimination weeks from the results field."

Usage: The generated code served as a starting template. We modified it extensively to handle edge cases (bonus points, team dances, missing 4th judge scores) identified during testing.

Example 2: Visualization Refinement

Prompt: "Create a diverging bar chart showing survival probability differences between judge-only and fan-only mechanisms. Use coral for positive differences and blue for negative."

Usage: We iteratively refined the visualization through multiple prompts, adjusting label positions, axis ranges, and color schemes based on our design preferences.

Example 3: LaTeX Algorithm Formatting

Prompt: "Convert this Python function into algorithm2e pseudocode format for a LaTeX document."

Usage: The AI output was used as a formatting reference. All algorithm logic and mathematical notation were written by team members.

5. What AI Was NOT Used For

To maintain academic integrity, we explicitly did **not** use AI for:

- **Core mathematical modeling:** The Bayesian fan vote estimation framework, multi-objective optimization formulation, and dynamic weight system were designed entirely by our team.
- **Statistical analysis and interpretation:** All statistical tests, model comparisons, and result interpretations were performed by team members.
- **Strategic recommendations:** The memo recommendations to producers were based on our own analysis and judgment.
- **Data fabrication:** No data points were generated or modified by AI.
- **Literature review:** All referenced sources were found and read by team members.

6. Summary

AI tools served as **productivity aids** in our workflow, primarily for:

1. Accelerating code implementation of well-defined algorithms
2. Improving English language quality in written sections
3. Troubleshooting technical issues in Python and LaTeX

The intellectual contributions—problem formulation, mathematical modeling, analysis design, and strategic recommendations—are entirely the work of our team. All AI outputs were critically evaluated, verified against ground truth where possible, and modified to meet our specific needs.

Team #2622622
February 3, 2026