Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2630964

Data With the Stars: From Vote Inference to Mechanism Design—A Transparent Optimization Framework for DWTS

Summary

This paper develops a mathematical framework to analyze the voting mechanism of Dancing with the Stars(DWTS). The problem lies in inferring weekly fan votes consistent with observed eliminations, and then using the estimates to evaluate scoring methods and design improved evaluation systems.

Fan vote estimation . We reconstruct weekly fan vote shares for all seasons by combining celebrity/pro-dancer fan stocks and performance-driven surges, then enforcing season-era elimination rules (Seasons 1–2 rank-based, Seasons 3–27 percentage-based, Seasons 28–34 bottom-2 lock). Bayesian consistency analysis shows strong reproducible fit, and GA tuning improves overall performance (best fitness $504.95 \to 552.25$; posterior hit $0.704 \to 0.781$ while uncertainty narrows, CI $0.01512 \to 0.01280$).

Scoring method comparison. Across $N = 267$ elimination weeks, rank and percentage rules disagree in 75 weeks (28.09%), confirming that rule choice materially changes eliminations. Against historical outcomes, the percentage rule matches $234/267 = 87.64\%$ while the rank rule matches $166/267 = 62.17\%$; the percentage rule also exhibits larger fan leverage (mean fan-influence 77.4% vs. 31.9%).

Impact factor analysis. Controlling for celebrity age, industry, state, and region, professional-dancer identity remains a first-order driver, adding $\Delta R^2 \approx 0.052$–0.099 across judge and fan outcomes. Age mainly operates through fan channels: per +10 years the implied multiplier is 0.954 on judges but only about 0.66 on fan support.

Evaluation system design. We propose a transparent weekly mechanism with discriminability-based dynamic weighting, dual-threshold elimination, and an optional suspense-triggered judges' save. GA calibration yields three profiles: *Balanced* achieves $\mathrm{MPR}_k = 1.000$ with SI = 0.8686 and zero intervention (BCR = JSR = 0); *Fair* maintains full protection ($\mathrm{MPR}_k = 1.000$) with lower stability (SI = 0.7117); *Show* preserves high suspense (SI = 0.8418) while allowing controlled deviations ($\mathrm{MPR}_k = 0.9525$) with intervention rate 0.1171 (37/316 weeks).

Key Words: elimination consistency; Bayesian posterior; genetic algorithm; scoring-rule comparison; fixed effects; voting system design.

Contents

1 Introduction

1.1 Problem Background

Dancing with the Stars (DWTS) is one of the most successful television shows worldwide, having gained popularity across more than 60 countries. Until now, the program has completed 34 seasons, each pairing celebrity contestants with professional dancers to put on weekly stage performances.[1]

The show adopts a dual-evaluation judging system comprising two parts: (1) expert judges rate each pair on a scale of 1 to 10 points, and (2) the audience cast votes for their favorite pairs. The final rankings are determined by combining these two evaluation components, with the lowest-ranked pair being eliminated each week.

The mathematical relevance of this problem stems from a core information asymmetry: judges' scores are publicly disclosed, yet total fan votes are kept as closely guarded secrets. This dynamic gives rise to a classic inverse problem—one must infer the hidden distribution of votes from indirect observations of weekly elimination results.

1.2 Program's Rating Models

The Ranking Model evaluates each group by the judge scores and fan votes, then sum the ranks:

$$C_i^{Rank} = R_J(i) + R_V(i)[3] \tag{1}$$

The Percentage Model converts the scores to percentage, then sums up:

$$C_i^{Percentage} = \frac{J_i}{\sum_j J_j} + \frac{V_i}{\sum_j V_j}[4] \tag{2}$$

The rule change following Season 2 was prompted by Jerry Rice advancing to the finals despite consistently low judges' scores. However, subsequent controversies involving Bristol Palin (Season 11) and Bobby Bones (Season 27) demonstrated that even the percentage-based scoring method could not prevent technically unskilled contestants from moving forward on the strength of overwhelming fan support. In response, Season 28 introduced a judges' save mechanism as a procedural backstop.

1.3 Paper Goal

The paper addresses questions as follows:

1. **Fan Vote Estimation**: Develop a customized model to estimate plausible vote distributions consistent with elimination outcomes.

2. **Scoring Method Comparison**: Compare the ranking and percentage-based scoring methods using the estimated vote data.

3. **Impact Factor Analysis**: Quantitatively analyze the impact of professional dancers' expertise and celebrity-specific traits on competition outcomes.

4. **Scoring System Design**: Develop an optimized scoring evaluation system that balances technical proficiency and audience popularity.

2 Data Processing and Feature Construction

2.1 Data Structure and Objective

The dataset contains two parts: (i) static "contestant–season" attributes (e.g., celebrity name, ballroom partner, industry, home state/region, age during season, results, and final placement), and (ii) weekly judges' scores recorded by week and judge. Since later modeling requires week-level vote inference and elimination-consistency checks, we reshape the raw data into a "contestant–week" panel and construct judge-side features together with elimination labels.

2.2 Panel Reconstruction

We identify weekly judge-score columns, parse their week and judge indices, and reshape the score table from wide to long format. We then pivot back to a consistent "contestant–week" panel containing four judge-score fields (Judge 1–Judge 4). If a judge field is absent in some seasons, we explicitly create it and fill with missing values to preserve a uniform schema. Static attributes are merged back using the key (season, celebrity name).

2.3 Judge-Side Feature Construction

Valid judges and totals: For each contestant-week, we count the number of non-missing judge scores and sum all available judge scores to obtain the weekly judge total.
Within-week rank and share: For each season-week, we define active contestants as those with a positive weekly judge total, then rank them by this total (with deterministic tie-handling). We also normalize weekly totals by the within-week sum to obtain judge-score shares.

We compute the within-week judge rank by ordering $J_{i,w}$ in descending order within $A_{s,w}$ (minimum-rank ties), and compute the within-week judge share.

$$p^{J}_{i,w} = \frac{J_{i,w}}{\sum_{j \in A_{s,w}} J_{j,w}} \tag{3}$$

2.4 Elimination Labels and Outputs

Elimination week: If the results field explicitly contains "Eliminated Week X", we extract X as the elimination week. Otherwise, we infer the exit timing from participation by defining

$$W_s = \max\{w \mid \exists\, i,\ J_{i,w} > 0\} \tag{4}$$

$$w_i^{\text{last}} = \max\{w \mid J_{i,w} > 0\} \tag{5}$$

If $w_i^{\text{last}} < W_s$, we set the elimination week to w_i^{last}; if $w_i^{\text{last}} = W_s$, we treat the contestant as reaching the final week and leave it empty. The row-level elimination indicator is

$$e_{i,w} = \mathbf{1}(w = \text{elimination week of } i) \tag{6}$$

2.5 Fan Base and Viewership as External Validation

To externally validate the inferred voting outcomes, we additionally incorporate two auxiliary signals: contestants' fan base size and weekly television viewership. Fan base is proxied by publicly observable follower counts prior to each season, capturing ex-ante popularity differences across celebrities. Viewership is measured by season-level average ratings, serving as a proxy for total voting volume. These variables are not used in model estimation, but are employed to assess whether inferred vote shares and elimination probabilities are consistent with broader popularity and audience engagement patterns.

3 Assumptions and Notation

3.1 Key Assumptions

Assumption 1 Fan vote counts consist of two components: the size of the celebrity fan base and the size of the professional dancer fan base.
Assumption 2 Celebrity fan preferences do not exhibit unwarranted abrupt changes across consecutive weeks, i.e., they are temporally smooth.
Assumption 3 The size of professional dancers' fan base increases as the show progresses, satisfying the non-negativity and normalization constraints for the contestants who performed in the previous week; additionally, the growth rate of dancers' fan base size is positively correlated with the rankings of the current week.
Assumption 4 Weekly elimination outcomes are determined by the show's voting rules. To account for unobservable details (e.g., tie scores and similar scenarios), a certain degree of slack error is permitted.
Assumption 5 A subset of non-fan audiences cast votes for the respective dance pairs due to the exceptional technical proficiency of the professional dancers.
Assumption 6 The eliminated contestant must have the lowest score under the known evaluation method.

3.2 Notation

Table 1: Core notation table (selected variables)

Variable	Definition
Indexing and sets	
s	Season index.
w	Week index within a season.
i	Contestant (celebrity–pro pair) index.
$\mathcal{A}_{s,w}$	Active contestants in week (s, w).

Variable	Definition
$k_{s,w}$	Number of eliminations in week (s, w).
Observed scores and outcomes	
$J_{i,s,w}$	Judges' total score of contestant i in week (s, w).
$M_{s,w}$	Number of valid judges in week (s, w).
`results`	Season outcome label (e.g., eliminated week, finalist, winner).
`placement`	Final placement in the season (1 = champion).
Latent fan votes (core estimation targets)	
$p_{i,s,w}$	Estimated fan-vote share of contestant i in week (s, w).
$V_{i,s,w}$	Estimated fan votes (counts) in week (s, w).
Key smoothness and diagnostics	
λ_s	Temporal smoothness strength for week-to-week vote evolution.
$\xi_{s,w}$	Weekly slack/violation required to satisfy elimination constraints.
R_s	Season-level slack usage rate.
S_s	Season-level total slack magnitude.
Voting-rule quantities (rank vs. percent)	
w_J	Weight of judges' component in the combined score.
w_F	Weight of fan component in the combined score.
$r^J_{i,s,w}$	Within-week rank of judges' scores (1 = best).
$r^F_{i,s,w}$	Within-week rank of fan support (1 = best).
$C^{\%}_{i,s,w}$	Combined standing under the percent-based rule.
$C^{r}_{i,s,w}$	Combined standing under the rank-based rule.
Cross-rule comparison metrics	
$E^{\%}_{s,w}$	Elimination set in week (s, w) under the percent-based method.
$E^{r}_{s,w}$	Elimination set in week (s, w) under the rank-based method.
$D_{s,w}$	Indicator of disagreement between the two methods in week (s, w).
Q_s	Season-level disagreement rate (percent vs. rank).
$e_{s,w}$	Eliminated contestant in week (s, w) when $k_{s,w} = 1$.
$C_{i,s}$	Judge–fan disagreement level (controversy) for contestant i in season s.

4 Fan Vote Estimation Model

4.1 Model Objectives and Observable Inputs

Consider Season s of the show. Let $\mathcal{A}_{s,w}$ denote the set of dance pairs still competing in Week w. For each pair $i \in \mathcal{A}_{s,w}$, our goal is to infer its *fan vote share*

$$p_{i,s,w} \in [0, 1], \qquad \sum_{i \in \mathcal{A}_{s,w}} p_{i,s,w} = 1 \tag{7}$$

Given a weekly voting pool size $T_{s,w}$, the corresponding fan vote count is

$$V_{i,s,w} = p_{i,s,w}\, T_{s,w} \tag{8}$$

In the dataset, weekly judge scores are stored as `weekX_judgeY_score`. Let $M_{s,w}$ be the number of valid judges in Week (s, w) (missing judge scores are ignored). The total judge score for pair i is

$$J_{i,s,w} = \sum_{m=1}^{M_{s,w}} J^m_{i,s,w} \tag{9}$$

and the within-week judge score share is

$$p^{J}_{i,s,w} = \frac{J_{i,s,w}}{\sum_{j \in \mathcal{A}_{s,w}} J_{j,s,w}} \tag{10}$$

Importantly, $\mathcal{A}_{s,w}$ is defined as the set of pairs with valid $J_{i,s,w}$ in that week.

4.2 Voting Pool Size from Viewership

Weekly viewership is provided as `total_viewers_millions`, scraped from the season pages on Wikipedia.[1] We convert it into a weekly voting pool size:

$$T_{s,w} = \text{viewers}_{s,w} \times 10^6 \tag{11}$$

If $\text{viewers}_{s,w}$ is missing, we impute it in three steps:

1. **Within-season interpolation:** interpolate by week index inside the same season, then apply forward/backward fill.
2. **Nearest-season substitution:** if still missing, use the same week index from the closest season number that has data.
3. **Fallback:** if no information exists, use a global median viewership.

In our workflow, $T_{s,w}$ is used *only* to scale shares into vote counts via $V_{i,s,w} = p_{i,s,w} T_{s,w}$. All elimination-consistency constraints are enforced on the shares $p_{i,s,w}$.

4.3 Instagram check

We use weekly TV viewership to scale the voting pool. We also scrape exact Instagram followers for unique celebrities and pros (Seasons 30–34).[2] Plotting followers against the model fan stock on log–log axes shows a clear positive trend for both groups (Fig. 1), supporting that the inferred stock reflects real popularity.

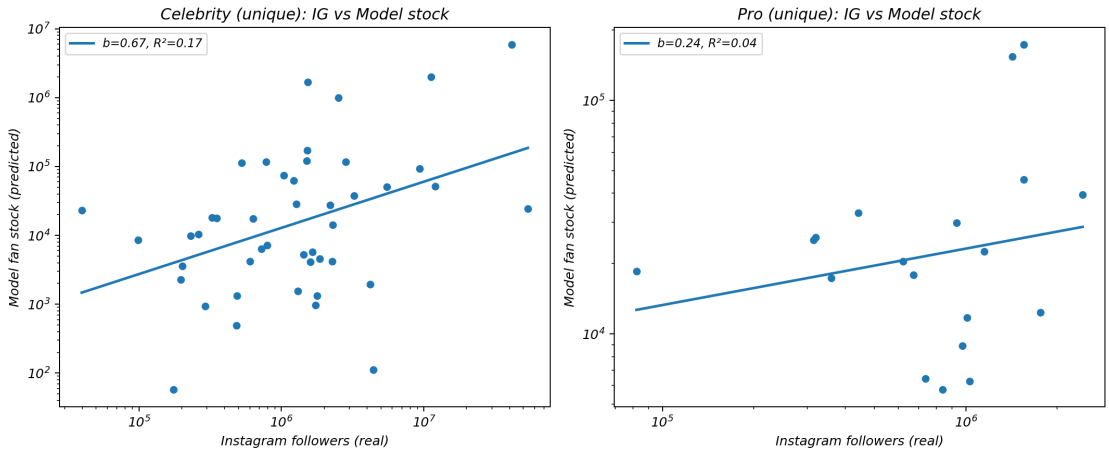

Figure 1: Instagram followers vs. inferred fan stock (Seasons 30–34) on log–log axes for celebrities and professionals.

4.4 Point Estimation of Fan Vote Shares

4.4.1 Attraction intensity and normalization

We first build an interpretable attraction intensity $v_{i,s,w}$, then normalize it into a prior share $\hat{p}_{i,s,w}$:

$$v_{i,s,w} = v^{\text{star}}_{i,s,w} + v^{\text{pro}}_{i,s,w} + v^{\text{perf}}_{i,s,w} + \varepsilon, \qquad \hat{p}_{i,s,w} = \frac{v_{i,s,w}}{\sum_{j\in\mathcal{A}_{s,w}} v_{j,s,w}} \tag{12}$$

where $\varepsilon > 0$ is a small stabilizer.

4.4.2 Celebrity and pro-dancer fan stocks

We maintain two popularity "stocks" in log space:

$$\ell^{\text{star}}_{i,s,w} = \log F^{\text{star}}_{i,s,w}, \qquad \ell^{\text{pro}}_{i,s,w} = \log F^{\text{pro}}_{i,s,w} \tag{13}$$

These stocks represent persistent fan bases that do not reset to zero each week.

Season-start shrinkage mixing. At Week 0, stocks are initialized by mixing: (i) a cross-season baseline g^{star} (by celebrity name) or g^{pro} (by pro dancer), and (ii) a within-season guess derived from Week 1 performance:

$$\ell^{\text{star}}_{i,s,0} = \rho_{\text{star}}\, g^{\text{star}}_{\text{name}(i)} + (1-\rho_{\text{star}})\, \log(\text{star\_guess}_i + 1) \tag{14}$$

$$\ell^{\text{pro}}_{i,s,0} = \rho_{\text{pro}}\, g^{\text{pro}}_{\text{pro}(i)} + (1-\rho_{\text{pro}})\, \log(\text{pro\_guess}_i + 1) \tag{15}$$

where $\rho_{\text{star}}, \rho_{\text{pro}} \in [0,1]$ control how strongly we trust cross-season baselines.

Fixed vote intensities from stocks. Stocks are converted to "fixed" weekly voting intensity:

$$v^{\text{star}}_{i,s,w} = \exp(\ell^{\text{star}}_{i,s,w}), \qquad v^{\text{pro}}_{i,s,w} = \alpha_{\text{pro}} \exp(\ell^{\text{pro}}_{i,s,w}) \tag{16}$$

where $\alpha_{\text{pro}} > 0$ scales the pro-dancer stock to be comparable to the celebrity stock.

4.4.3 Weekly performance features and competition tightness

Let $n_{s,w} = |\mathcal{A}_{s,w}|$. Using the weekly judge totals, we define a judge rank $r^J_{i,s,w} \in \{1,\dots,n_{s,w}\}$ (1 = best), and map it to a smooth strength score:

$$q_{i,s,w} = \frac{n_{s,w} - r^J_{i,s,w} + 1}{n_{s,w}} \in (0,1] \tag{17}$$

Thus, $q_{i,s,w}$ increases with judged performance.

Tightness of a week. We measure how close the week is by the relative spread of judge totals:

$$\text{spread}_{s,w} = \frac{\max(J_{s,w}) - \min(J_{s,w})}{\overline{J_{s,w}} + \varepsilon} \tag{18}$$

then map it to a tightness score with a sigmoid:

$$\text{tight}_{s,w} = \sigma\big(k_{\text{tight}}(\tau_{\text{tight}} - \text{spread}_{s,w})\big), \qquad \sigma(z) = \frac{1}{1+e^{-z}} \tag{19}$$

A larger $\text{tight}_{s,w}$ means the week is closer (harder to separate pairs by judges alone).

4.4.4 Performance-driven weekly votes

Performance-driven voting captures "vote surges" caused by weekly performances. We model two effects:

1. **Excellence bonus:** stronger performances attract extra votes.
2. **Danger-zone mobilization:** when a pair is in the bottom-K by judges, fans may mobilize—especially in tight weeks.

Let $\mathbf{1}\{i \in \text{bottom-}K\}$ indicate whether i is bottom-K in judge rank. Then

$$v_{i,s,w}^{\text{perf}} = \beta_{\text{perf}} \left(q_{i,s,w}^{a_{\text{exc}}} + \omega_{\text{dang}} \cdot \text{tight}_{s,w} \cdot \mathbf{1}\{i \in \text{bottom-}K\} \cdot (1 - q_{i,s,w})^{a_{\text{dang}}} \right) \tag{20}$$

Here β_{perf} sets the overall scale; a_{exc} shapes how sharply votes concentrate at the top; and $(\omega_{\text{dang}}, a_{\text{dang}})$ control the strength and curvature of the danger-zone effect.

4.5 Elimination-Consistency Correction

The prior $\hat{p}_{i,s,w}$ reflects popularity and performance. However, the show eliminates pairs using a specific rule. Therefore, we *correct* $\hat{p}_{i,s,w}$ into final $p_{i,s,w}$ so that estimated shares are consistent with observed eliminations.

Let $E_{s,w} \subseteq \mathcal{A}_{s,w}$ be the set of eliminated pairs in Week (s, w), with $|E_{s,w}| = k_{s,w}$. Rules differ by season era.

Rank eras (Seasons 1–2 and 28–34). Fans are converted to within-week ranks $r_{i,s,w}^F$ and combined with judges' ranks:

$$C_{i,s,w}^r = w_J r_{i,s,w}^J + w_F r_{i,s,w}^F$$

Season 1–2 requires $E_{s,w} \subseteq \text{Worst-}k_{s,w}(C^r)$, while Seasons 28–34 only require bottom-2 membership $E_{s,w} \subseteq \text{Worst-2}(C^r)$ ("bottom-2 lock"). Because ranks are discrete, we apply a minimum-perturbation loop: if the eliminated pair is outside the required bottom set, slightly decrease its share and redistribute the removed mass to others proportionally to their current shares, until the constraint holds (or a max-iteration cap is reached).

Percentage era (Seasons 3–27). We combine judge share $p_{i,s,w}^J$ and fan share $p_{i,s,w}$:

$$C_{i,s,w}^{\%} = w_J p_{i,s,w}^J + w_F p_{i,s,w}$$

To allow stochasticity in real eliminations, we enforce elimination consistency with slack $\xi \geq 0$ and solve

$$\min_{p \in \Delta,\ \xi \geq 0} \ \lambda_{\text{prior}} \|p - \hat{p}\|_2^2 + \lambda_{\text{slack}} \sum \xi \quad \text{s.t. } C_{e,s,w}^{\%} \leq C_{j,s,w}^{\%} + \xi,\ \forall j \in \mathcal{A}_{s,w} \setminus \{e\}$$

where $\Delta = \{p \geq 0,\ \sum_i p_i = 1\}$. The weekly slack sum $\sum \xi$ quantifies how much inconsistency pressure is needed to reproduce the observed elimination.

4.6 Weekly Stock Updates

After obtaining corrected shares $p_{i,s,w}$, we update the popularity stocks so that next week's fixed voting intensity reflects accumulated momentum.

Log-stock updates. We update celebrity and pro-dancer stocks in log space:

$$\ell^{\text{star}}_{i,s,w+1} = \ell^{\text{star}}_{i,s,w} + \eta^{\text{star}}_{+}\,\text{gain}_{i,s,w} - \eta^{\text{star}}_{-}\,\text{loss}_{i,s,w} - \delta^{\text{star}}, \tag{21}$$

$$\ell^{\text{pro}}_{i,s,w+1} = \ell^{\text{pro}}_{i,s,w} + \eta^{\text{pro}}_{+}\,\text{gain}_{i,s,w} - \eta^{\text{pro}}_{-}\,\text{loss}_{i,s,w} - \delta^{\text{pro}} \tag{22}$$

This captures four intuitive behaviors: (i) strong performance increases popularity, (ii) weak performance reduces it, (iii) popularity decays slowly over time, and (iv) bottom-K mobilization is stronger in tight weeks.

4.7 Cross-Season Outer Loop for Global Baselines

The global baselines g^{star} and g^{pro} should represent "typical" starting popularity for each celebrity and pro dancer across seasons. We estimate them using an outer loop across all seasons:

1. Run the within-season estimation for all seasons using current $(g^{\text{star}}, g^{\text{pro}})$.
2. Collect the season-start stocks $\ell^{\text{star}}_{i,s,0}$ and $\ell^{\text{pro}}_{i,s,0}$ for each celebrity/dancer.
3. Update each baseline by exponential smoothing:

$$g \leftarrow (1 - \rho_{\text{outer}})\, g + \rho_{\text{outer}} \cdot \overline{\ell_0} \tag{23}$$

where $\overline{\ell_0}$ is the current mean season-start log stock for that entity.

Repeating this loop stabilizes season-start initialization and improves cross-season consistency.

4.8 Elimination-Consistency Evaluation Metric

For each elimination week w in season s, define

$$\mathbb{I}_{s,w} = \begin{cases} 1, & \text{the eliminated pair satisfies the season-era bottom-set constraint,} \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

The season-level consistency rate is

$$\text{ConsistencyRate}_s = \frac{1}{|\mathcal{W}_s|} \sum_{w \in \mathcal{W}_s} \mathbb{I}_{s,w} \tag{25}$$

where $\mathcal{W}_s$ is the set of elimination weeks. For Seasons 3–27, we also report the weekly slack sum $\sum_k \xi_{s,w,k}$ as a graded measure of inconsistency pressure.

4.9 Uncertainty Quantification

The model quantifies uncertainty in two complementary ways: a fast feasible-region proxy (mainly for Seasons 3–27), and a Bayesian posterior analysis (for all seasons).As a classical resampling-based baseline, bootstrap methods are widely used for uncertainty assessment.[8]

4.9.1 Feasible-region interval proxy for Seasons 3–27

In the percentage era (Seasons 3–27), elimination consistency constraints are linear in $p_{\cdot,s,w}$, and $p_{\cdot,s,w}$ lies on the simplex. For each pair i we compute

$$p_{i,s,w}^{\min} = \min p_{i,s,w}, \qquad p_{i,s,w}^{\max} = \max p_{i,s,w} \tag{26}$$

subject to the hard constraints and simplex constraint. The interval width

$$U_{i,s,w}^{p} = p_{i,s,w}^{\max} - p_{i,s,w}^{\min} \tag{27}$$

provides a simple "how much room is left" uncertainty proxy. (For rank-based eras, the feasible set is defined by discrete ordering and this proxy becomes unstable.)

4.9.2 Bayesian consistency analysis

For each elimination week (s, w), we construct a posterior for the share vector $p_{s,w} = (p_{i,s,w})_{i \in \mathcal{A}_{s,w}}$, centered at a point estimate $p_{s,w}^{0}$ (taken from our estimator output):

$$p_{s,w} \sim \text{Dirichlet}(\alpha_{s,w}), \qquad \alpha_{s,w} = \kappa_{s,w} p_{s,w}^{0} \tag{28}$$

Here $\kappa_{s,w} > 0$ is a concentration parameter: larger $\kappa_{s,w}$ means the posterior is more tightly concentrated around $p_{s,w}^{0}$.

Hierarchical prior for stability across weeks. To avoid unstable week-by-week estimation, we share information through

$$\log \kappa_{s,w} \sim \mathcal{N}(\mu, \tau^2) \tag{29}$$

and update (μ, τ) by an empirical Bayes outer loop (estimate $\log \kappa_{s,w}$ for all weeks, then refresh the mean and variance).

Likelihood from elimination. Let $C_i(p)$ be the season-era composite score (percentage-based or rank-based). For an eliminated pair e and any survivor j, we model the probability that e is worse than j by

$$\Pr(e \text{ worse than } j \mid p) \approx \sigma\left(\frac{C_j(p) - C_e(p)}{\sigma_{s,w}}\right), \qquad \sigma(z) = \frac{1}{1 + e^{-z}} \tag{30}$$

The noise scale $\sigma_{s,w}$ increases in tighter weeks:

$$\sigma_{s,w} = \sigma_{\text{base}}\big(1 + \sigma_{\text{tight}} \cdot \text{tight}_{s,w}\big) \tag{31}$$

For Seasons 28–34, the likelihood is adapted to match the "bottom-2 lock, then judges choose" mechanism by focusing on bottom-2 membership rather than requiring the eliminated pair to be uniquely worst.

MAP and Laplace approximation. We compute a MAP estimate $p_{s,w}^{\mathrm{MAP}}$ and approximate uncertainty by a Laplace approximation.[5] To optimize without simplex constraints, we use the ALR transform (compositional data standard):

$$y_k = \log \frac{p_k}{p_n}, \qquad k = 1, \ldots, n-1, \quad n = |\mathcal{A}_{s,w}| \tag{32}$$

We approximate the posterior in y-space by a Gaussian around y^{MAP}, sample $y^{(m)}$, map back to $p^{(m)}$, and obtain credible intervals:

$$[p_{i,s,w}^{\mathrm{low}}, p_{i,s,w}^{\mathrm{high}}], \qquad U_{i,s,w}^{p} = p_{i,s,w}^{\mathrm{high}} - p_{i,s,w}^{\mathrm{low}} \tag{33}$$

Using $V_{i,s,w} = p_{i,s,w} T_{s,w}$, we also convert posterior samples to vote counts and compute vote-count intervals.

Posterior hit probability. We summarize probabilistic elimination consistency by

$$\mathrm{PosteriorHitProb}_{s,w} \approx \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{e \in \mathrm{BottomSet}(C(p^m))\} \tag{34}$$

Here, BottomSet$(\cdot)$ is the relevant season-era bottom set (Worst-k in rank eras; bottom-2 in Seasons 28–34).

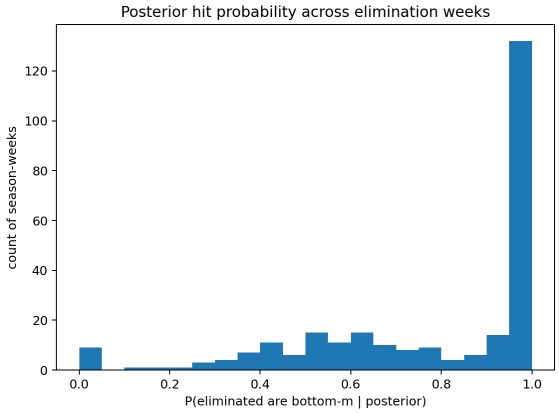

Figure 2: Posterior hit probability across elimination weeks.

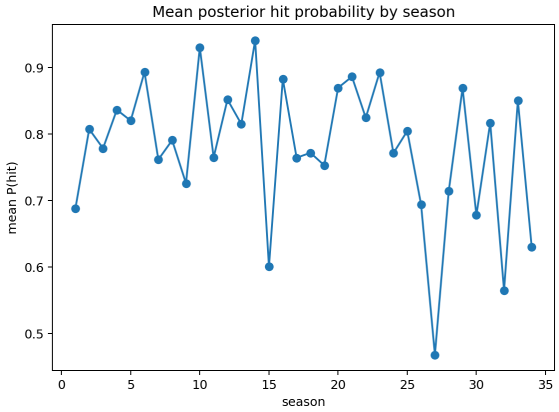

Figure 3: Mean posterior hit probability by season.

4.10 Genetic-Algorithm Tuning

The estimator contains hyperparameters θ (e.g., $w_J, w_F, \alpha_{\mathrm{pro}}, \beta_{\mathrm{perf}}$, decay/gain rates, and tightness parameters). We tune θ via a genetic algorithm (GA) that treats the full pipeline as a black box[6]:

$$\theta \implies \text{estimator outputs } (p_{s,w}^{0}) \implies \text{Bayesian analysis metrics} \implies \mathrm{Fitness}(\theta)$$

We define fitness by combining Bayesian metrics (rewarding consistency and clear separation, penalizing uncertainty and failures):

$$\text{Fitness}(\theta) = a \cdot \overline{\text{PosteriorHitProb}} + b \cdot \overline{\text{AvgPairProb}} + c \cdot \overline{\text{Margin}} - d \cdot \overline{U^p} - e \cdot \text{FailRate} \quad (35)$$

where $\overline{x}$ denotes averages across elimination weeks, and FailRate is the fraction of weeks with failed MAP/Laplace steps. The GA uses selection–crossover–mutation–elitism, and evaluates individuals in parallel to exploit available CPU resources.[7]

4.10.1 Seeded restart strategy and convergence behavior

We run the same GA twice under an unchanged fitness definition. Run-1 starts from a random population for global exploration; Run-2 is a seeded restart initialized around the best individual from Run-1, focusing on local refinement.

Table 2: Run-1 history excerpt (random init).

Gen	best_fitness	mean_fitness	best_hit	best_pairprob	best_ci
1	504.953	459.773	0.704	0.767	0.01512
16	547.459	528.894	0.771	0.813	0.01358
30	552.248	528.604	0.781	0.812	0.01280

Table 3: Run-2 history excerpt (seeded from Run-1 best).

Gen	best_fitness	mean_fitness	best_hit	best_pairprob	best_ci
1	552.851	491.709	0.783	0.810	0.01282
10	553.125	544.789	0.784	0.810	0.02022
19	553.125	546.522	0.784	0.810	0.02022

Model Sensitivity Analysis. Run-1 shows strong global sensitivity: `best_fitness` increases (504.95→552.25) with higher `best_hit`/`best_pairprob` and lower `best_ci`. Run-2 shows local saturation: `best_fitness` and `best_hit` change little near the optimum; variations mainly appear in uncertainty width (`best_ci`) while `best_fail_rate`=0.

5 Scoring Method Comparison

5.1 Cross-Method Application

Let season index be s and week index be w. Denote the set of active pairs (not yet eliminated) by $\mathcal{A}_{s,w}$. For each pair $i \in \mathcal{A}_{s,w}$, we consider two ways to represent judges and fans:

- **Percentage representation:** $P^J_{i,s,w} \in [0,1]$ is the judge score share in week (s,w) (normalized within the week), and $P^V_{i,s,w} \in [0,1]$ is the estimated fan vote share (also normalized within the week).
- **Ranking representation:** $R^J_{i,s,w} \in \{1,\ldots,|\mathcal{A}_{s,w}|\}$ is the rank induced by judges (1 = best), and $R^V_{i,s,w} \in \{1,\ldots,|\mathcal{A}_{s,w}|\}$ is the rank induced by fans (1 = best).

We apply *both* elimination rules to every elimination week, regardless of which rule was historically used. If exactly one pair is eliminated in week (s, w), the counterfactual predicted elimination is:

$$k_{s,w}^{\text{rank}} = \arg\max_{i \in \mathcal{A}_{s,w}} \left(R_{i,s,w}^{J} + R_{i,s,w}^{V} \right), \tag{36}$$

$$k_{s,w}^{\text{pct}} = \arg\min_{i \in \mathcal{A}_{s,w}} \left(P_{i,s,w}^{J} + P_{i,s,w}^{V} \right) \tag{37}$$

For weeks with $k > 1$ eliminations, we take the worst-k set under the corresponding criterion (i.e., the top-k maximizers of $R^J + R^V$ or the bottom-k minimizers of $P^J + P^V$).

Using our estimated P^V and the observed judge scores, we evaluate all elimination weeks contained in the dataset. Across all $N = 267$ elimination weeks, the two scoring methods disagree in 75 weeks, yielding a disagreement rate of $75/267 = 28.09\%$. Moreover, when compared against the actual eliminated pair(s), the percentage method matches the historical outcome far more often than the rank method.

Table 4: Cross-method agreement and historical accuracy over all 267 elimination weeks.

Comparison	Agree	Disagree	Rate	Interpretation
Rank vs. Percentage	192	75	71.91%	Rule choice changes predictions
Rank vs. Actual	166	101	62.17%	Lower historical accuracy
Percentage vs. Actual	234	33	87.64%	Higher historical accuracy

The 28.09% cross-method disagreement is substantial: method choice is consequential, not a minor technicality. Meanwhile, the percentage rule achieves 87.64% agreement with historical eliminations, suggesting that percentage aggregation better explains observed outcomes under our estimated fan votes.

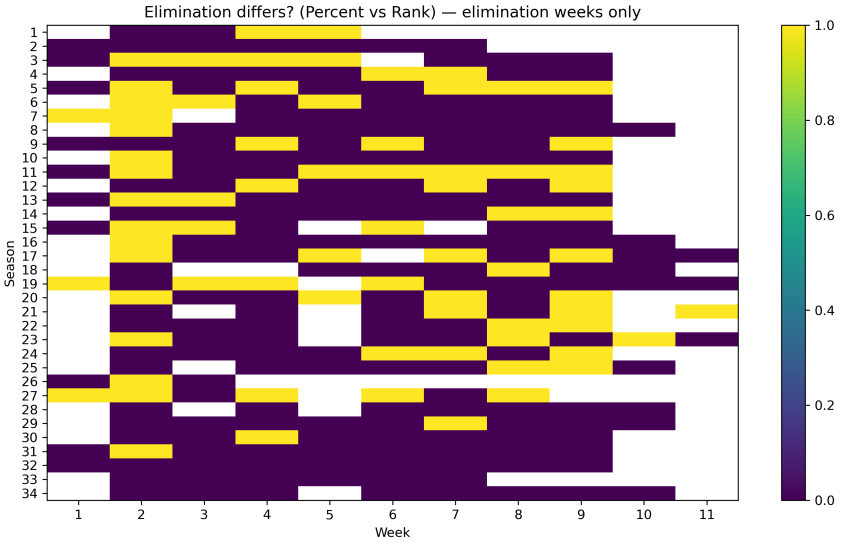

Figure 4: Elimination-week disagreements between percentage and rank rules (1 = different, 0 = same).

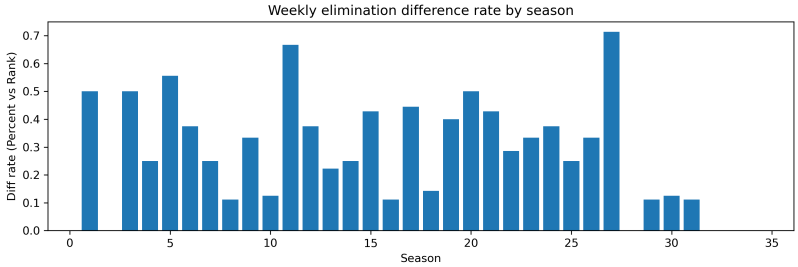

Figure 5: Weekly disagreement rate by season (percentage vs. rank).

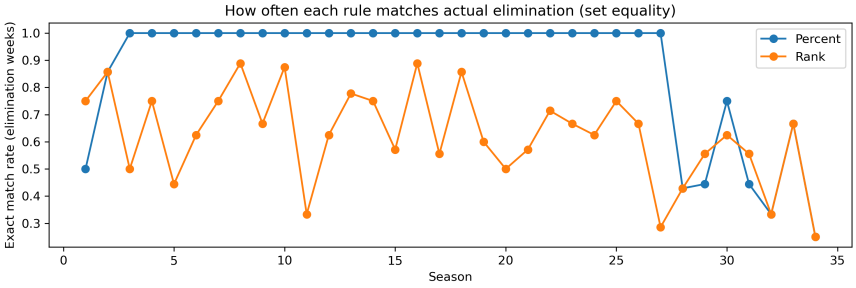

Figure 6: Exact match rate with actual eliminations by season under percentage and rank rules.

5.2 Fan Influence Quantification

Beyond correctness, we quantify *how much* the fan component contributes to the dispersion of the combined score. We define a variance-based fan influence measure (computed *within* each week across active pairs):

Percentage method.

$$\mathrm{FI}_{\mathrm{pct}}(s,w) = \frac{\mathrm{Var}\left(P^V_{i,s,w}\right)}{\mathrm{Var}\left(P^J_{i,s,w} + P^V_{i,s,w}\right)} = \frac{\mathrm{Var}(P^V)}{\mathrm{Var}(P^J) + \mathrm{Var}(P^V) + 2\mathrm{Cov}(P^J, P^V)} i \tag{38}$$

Ranking method.

$$\mathrm{FI}_{\mathrm{rank}}(s,w)i = \frac{\mathrm{Var}\left(R^V_{i,s,w}\right)}{\mathrm{Var}\left(R^J_{i,s,w} + R^V_{i,s,w}\right)} \tag{39}$$

We summarize $\mathrm{FI}_{\mathrm{pct}}$ and $\mathrm{FI}_{\mathrm{rank}}$ over elimination weeks. To avoid rare numerical instabilities (e.g., extremely small denominators), we winsorize the weekly FI values at 1% tails before reporting summary statistics.

Table 5: Fan influence statistics by scoring method over elimination weeks (winsorized at 1% tails for numerical stability).

Method	Mean	Std	Min	Max	Median
Ranking	31.9%	8.8%	24.9%	84.6%	29.1%
Percentage	77.4%	8.0%	61.5%	98.2%	77.6%

In our estimated vote landscape, the percentage aggregation preserves *magnitude* differences in fan vote shares, resulting in a much larger variance contribution from P^V. By contrast, the ranking rule compresses outcomes into ordinal positions, which effectively caps how much a very large fan vote share can separate a contestant from others, thereby reducing the variance share attributable to fans.

5.3 Controversial Case Analysis

To identify contestants with strong judge–fan disagreement, we use the controversy index derived from weekly rank gaps:

$$\mathrm{CI}_{i,s} = \frac{1}{|\mathcal{W}_{i,s}|} \sum_{w \in \mathcal{W}_{i,s}} \left| R^J_{i,s,w} - R^V_{i,s,w} \right| \tag{40}$$

where $\mathcal{W}_{i,s}$ denotes the set of weeks in which contestant i appears in season s.

Below we report three representative cases that have been widely discussed by audiences. For each, we list weekly judge totals (raw), judge ranks, estimated fan vote shares, and fan ranks.(first 5 weeks listed)

Case 1: Jerry Rice (Season 2). This case illustrates persistent judge disadvantage paired with moderate fan support. Under counterfactual scoring, the percentage rule becomes more sensitive to sustained judge-score deficits, and can flag such contestants as vulnerable even when fan ranks are not bottom.

Table 6: Jerry Rice weekly performance summary (Season 2).

Week	JudgeTotal	JudgeRank	EstVoteShare	VoteRank
1	18	6	14.3%	5
2	21	4	12.9%	6
3	22	4	11.8%	7
4	25	3	10.3%	6
5	25	3	10.4%	6

Case 2: Bristol Palin (Season 11). This case exhibits repeated judge-rank weakness with fluctuating fan-rank performance. It emphasizes that week-to-week volatility in R^V (mobilization spikes) can meaningfully alter who becomes "at risk", especially under methods that preserve vote-share magnitudes.

Table 7: Bristol Palin weekly performance summary (Season 11).

Week	JudgeTotal	JudgeRank	EstVoteShare	VoteRank
1	21	7	8.2%	9
2	21	10	10.2%	6
3	18	11	10.3%	7
4	20	9	11.5%	5
5	21	7	12.0%	5

Case 3: Bobby Bones (Season 27). This case demonstrates how a contestant can remain near the bottom in judge ranks while surviving through non-bottom fan ranks over multiple weeks, highlighting the structural tension between technical merit and audience preference.

Table 8: Bobby Bones weekly performance summary (Season 27).

Week	JudgeTotal	JudgeRank	EstVoteShare	VoteRank
1	22	11	8.9%	6
2	23	10	8.0%	9
3	24	10	6.8%	11
4	23	9	7.6%	11
5	24	9	8.7%	9

5.4 Method Recommendation

We synthesize accuracy and influence diagnostics to compare the two rules.

Table 9: Comparative evaluation of the two scoring methods using our estimated fan votes.

Criterion	Ranking	Percentage
Historical accuracy (vs actual)	62.17%	87.64%
Fan influence (variance-based, mean)	31.7%	77.3%
Alignment with judges (mean)	83.8%	65.5%
Alignment with fans (mean)	85.0%	92.7%
Fan overturn rate (mean)	39.7%	50.4%
Sensitivity to score gaps	Low (rank compression)	High (magnitude preserved)
Upset susceptibility	Lower	Higher
Computation	Simple (ranks)	Moderate (percent sums)

Overall, within our reconstructed vote estimates, the percentage method is substantially more consistent with historical eliminations and more responsive to vote-share magnitude, while the rank method is comparatively closer to judge-driven outcomes due to ordinal compression. Which rule is preferred depends on the show objective: prioritizing technical merit favors the rank-based aggregation, whereas emphasizing audience impact and historical faithfulness (under our estimated votes) favors the percentage method.

6 Professional Dancer and Celebrity-Feature Effects on Judges and Fans

6.1 Data and outcomes

Let i index pairs (celebrity–professional), (s, w) index season and week, and let $\mathcal{A}_{s,w}$ be the set of active pairs in week w of season s.

Judges. Let $J_{i,s,w} \in (0, 1)$ be the within-week judge share (`judge_percent`). Define

$$\tilde{J}_{i,s,w} = \operatorname{logit}(J_{i,s,w}), \qquad y^{J}_{i,s,w} = \tilde{J}_{i,s,w} - \frac{1}{|\mathcal{A}_{s,w}|} \sum_{k \in \mathcal{A}_{s,w}} \tilde{J}_{k,s,w} \tag{41}$$

Thus y^J measures relative judge advantage within each (s, w), removing week-level shocks.

Fans. From dwts_fan_vote_estimates_components.csv, we use the estimated vote-share components: P^{tot}, $P^{\star}$ (celebrity fan-base), P^{pro} (pro-fixed), and P^{perf} (performance-induced). For $o \in \{\text{tot}, \star, \text{pro}\}$ we define

$$\tilde{P}^o_{i,s,w} = \text{logit}\Big(P^o_{i,s,w}\Big), \qquad y^o_{i,s,w} = \tilde{P}^o_{i,s,w} - \frac{1}{|\mathcal{A}_{s,w}|} \sum_{k \in \mathcal{A}_{s,w}} \tilde{P}^o_{k,s,w} \tag{42}$$

For the performance component we use a log transform to accommodate very small shares:

$$\tilde{P}^{\text{perf}}_{i,s,w} = \log\Big(P^{\text{perf}}_{i,s,w} + \varepsilon\Big), \qquad y^{\text{perf}}_{i,s,w} = \tilde{P}^{\text{perf}}_{i,s,w} - \frac{1}{|\mathcal{A}_{s,w}|} \sum_{k \in \mathcal{A}_{s,w}} \tilde{P}^{\text{perf}}_{k,s,w} \tag{43}$$

6.2 Explanatory factors

We include all celebrity features present in the panel: `celebrity_age_during_season`, `celebrity_industry`, `celebrity_homestate`, and `celebrity_homecountry_region`. Age is standardized:

$$\text{age\_z}_{i,s} = \frac{\text{age}_{i,s} - \overline{\text{age}}}{\text{sd(age)}} \tag{44}$$

Industries and states are pooled by minimum frequency, and regions are top-N pooled, yielding grouped categories ind_i, state_i, reg_i. The professional dancer identity is pro_i (`ballroom_partner`).

6.3 Model specification

For each outcome $o \in \{J, \text{tot}, \star, \text{pro}, \text{perf}\}$ we fit:

$$y^o_{i,s,w} = \beta^o_{\text{age}}\, \text{age\_z}_{i,s} + \alpha^o_{\text{ind}(i)} + \alpha^o_{\text{state}(i)} + \alpha^o_{\text{reg}(i)} + \gamma^o_{\text{pro}(i)} + \varepsilon^o_{i,s,w}, \tag{45}$$

with category fixed effects α and professional fixed effects γ. Standard errors are clustered at the season–celebrity level (same celebrity across weeks).

6.4 Evaluation metrics

(i) Incremental explanatory power of professionals. Define $R^2_{\text{no\_pro}}(o)$ from the model without $\gamma^o_{\text{pro}(i)}$, and $R^2_{\text{full}}(o)$ from the full model. Then

$$\Delta R^2_{\text{pro}}(o) = R^2_{\text{full}}(o) - R^2_{\text{no\_pro}}(o) \tag{46}$$

(ii) Factor importance via drop-one ΔR^2. For factor $f \in \{\text{age}, \text{industry}, \text{state}, \text{region}, \text{pro}\}$, let $R^2_{-f}(o)$ be the R^2 from the model dropping only factor f, fit on the same complete-case sample as the full model. Define

$$\Delta R^2_{\text{drop}}(o, f) = R^2_{\text{full}}(o) - R^2_{-f}(o) \tag{47}$$

(iii) Cross-outcome alignment (same-way test). For a categorical factor f with category effects $\theta^o_{f,c}$ (centered by category frequency), define

$$\text{Align}_f(o) = \text{Corr}\big(\theta^J_{f,c},\, \theta^o_{f,c}\big) \tag{48}$$

(iv) Age effect as a per-10-year multiplier. Let σ_{age} be the age standard deviation (years). The multiplicative change per +10 years is

$$\text{Mult}^{o}_{\text{age}}(+10) = \exp\left(\beta^{o}_{\text{age}} \cdot \frac{10}{\sigma_{\text{age}}}\right) \tag{49}$$

6.5 Results

6.5.1 Incremental explanatory power of professionals beyond celebrity features

Table 10 reports R^2 from the model without professional fixed effects and from the full model, fitted on the same complete-case sample ($n = 2402$). Define

$$\Delta R^2_{\text{pro}}(o) = R^2_{\text{full}}(o) - R^2_{\text{no\_pro}}(o) \tag{50}$$

Table 10: R^2 comparison and incremental explanatory power of professionals (all celebrity features controlled)

Outcome o	$R^2_{\text{no\_pro}}(o)$	$R^2_{\text{full}}(o)$	$\Delta R^2_{\text{pro}}(o)$
Judges	0.2584	0.3577	0.0994
FansTotal	0.1565	0.2487	0.0922
FansStar	0.1624	0.2547	0.0923
FansPro	0.0655	0.1246	0.0591
FansPerf	0.1624	0.2143	0.0520

Table 10 shows that, after controlling for age, industry, state, and region, professional identity still increases explanatory power by 0.052–0.099 absolute R^2 across outcomes, with the largest gains on Judges and on FansTotal/FansStar.

6.5.2 Marginal contribution (drop-one ΔR^2)

For each factor $f \in \{\text{pro}, \text{industry}, \text{state}, \text{region}, \text{age}\}$, define

$$\Delta R^2_{\text{drop}}(o, f) = R^2_{\text{full}}(o) - R^2_{-f}(o) \tag{51}$$

where $R^2_{-f}(o)$ is obtained by dropping only f and refitting on the same sample as the full model.

Table 11: Factor importance via drop-one ΔR^2 (larger values indicate larger contributions)

Outcome o	$\Delta R^2_{\text{drop}}(o, \text{pro})$	$\Delta R^2_{\text{drop}}(o, \text{industry})$	$\Delta R^2_{\text{drop}}(o, \text{state})$	$\Delta R^2_{\text{drop}}(o, \text{region})$	$\Delta R^2_{\text{drop}}(o, \text{age})$
Judges	0.0994	0.0320	0.0281	-0.0016	0.0885
FansTotal	0.0922	0.0125	0.0330	-0.0001	0.0393
FansStar	0.0923	0.0135	0.0333	-0.0001	0.0412
FansPro	0.0591	0.0065	0.0144	-0.0000	0.0180
FansPerf	0.0520	0.0248	0.0264	-0.0005	0.0585

Professionals are first-order drivers for every outcome, contributing about 0.052–0.099 drop-one R^2; age is the strongest celebrity feature, nearly matching professionals for Judges (0.0885) and remaining large for FansPerf (0.0585); state and industry provide moderate explanatory power; (iv) region contributes essentially zero. The tiny negative values for region are numerical/sample-boundary artifacts and should be interpreted as $\Delta R^2_{\text{drop}}(o, \text{region}) \approx 0$.

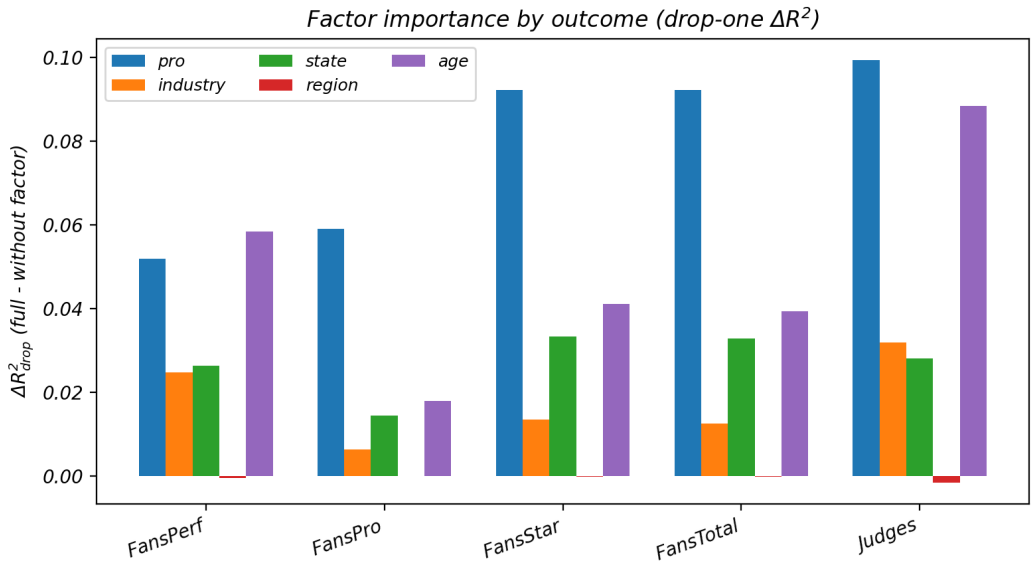

Figure 7: Factor importance across outcomes measured by drop-one ΔR^2 (larger values indicate larger contributions).

6.5.3 Mechanism alignment

For each categorical factor $f \in \{\text{pro}, \text{industry}, \text{state}\}$, let $\theta^o_{f,\cdot}$ denote the centered category-effect vector under outcome o. Define

$$\text{Align}_f(o) = \text{Corr}\big(\theta^J_{f,c},\, \theta^o_{f,c}\big) \tag{52}$$

Table 12: Alignment between judges and fan outcomes (correlation of category effects)

Factor f	$\text{Align}_f(\text{tot})$	$\text{Align}_f(\star)$	$\text{Align}_f(\text{pro})$	$\text{Align}_f(\text{perf})$
pro	0.7546	0.7626	0.5292	0.8933
industry	0.7457	0.7555	0.6036	0.9650
state	0.7181	0.7295	0.4907	0.6901

Table 12 shows moderate alignment on FansTotal/FansStar (≈ 0.75), maximal alignment on the performance component (pro: 0.8933, industry: 0.9650), and minimal alignment on the pro-fixed component (pro: 0.5292, state: 0.4907); thus effects are not transmitted through identical channels across judges and fans.

6.5.4 Age effect across outcomes

Let age enter standardized with standard deviation $\sigma_{\text{age}} = 12.8691$ years. The per-10-year multiplier is

$$\text{Mult}^o_{\text{age}}(+10) = \exp\left(\beta^o_{\text{age}} \cdot \frac{10}{\sigma_{\text{age}}}\right) \tag{53}$$

Table 13 implies a mild judge-side penalty (Judges: 0.954 per +10y, i.e., $\approx 4.6\%$ decrease) but a strong fan-side penalty (FansTotal/FansStar: ≈ 0.66, i.e., ≈ 33–34% decrease), strongest on FansPerf (0.637), indicating age primarily operates through fan-side channels.

Table 13: Age effects: coefficients and per +10-year multipliers ($\sigma_{\text{age}} = 12.8691$ years)

Outcome o	β^o_{age} (on age_z)	$\text{Mult}^o_{\text{age}}(+10)$
Judges	-0.060852	0.953815
FansTotal	-0.519122	0.668055
FansStar	-0.535738	0.659485
FansPro	-0.214542	0.846444
FansPerf	-0.580932	0.636727

6.6 Evaluation

Professionals and celebrity features jointly shape performance: professionals add $\Delta R^2_{\text{pro}}(o) \in [0.052, 0.099]$ beyond celebrity features (Table 10) and are among the most important drivers across outcomes (Table 11); among celebrity features, age is strongest and affects fans far more than judges (Table 13), while industry and state are moderate contributors. Mechanistically, effects are not identical across judges and fans: alignment peaks on the performance component and is weakest on the pro-fixed component (Table 12), implying judges track performance-driven variation most consistently whereas fan totals also reflect fixed/support-base channels.

7 New Elimination Mechanism

DWTS combines two signals: judges' evaluations (technical merit) and audience popularity. Building on Problem 1, we assume week-w reconstructed vote-share components $s^{\star}_{i,w}, s^{\text{pro}}_{i,w}, s^{\text{perf}}_{i,w}$ are available for each active pair $i \in C_w$, together with judges' totals $\text{JudgeScore}_{i,w}$. We design a *transparent, rule-based* weekly elimination mechanism that (i) protects technical fairness, (ii) preserves suspense, and (iii) avoids opaque manual intervention.

7.1 Inputs and Hyperparameters

In week w, let C_w be the active set, $N_w = |C_w|$. Define judges' share

$$J_{i,w} = \frac{\text{JudgeScore}_{i,w}}{\sum_{j \in C_w} \text{JudgeScore}_{j,w}} \tag{54}$$

This mechanism uses a small hyperparameter set

$$\Theta = \{\phi,\ \beta,\ k,\ \lambda_\alpha,\ \alpha_{\min}, \alpha_{\max},\ \mathbb{I}_{\text{tight}},\ \tau,\ \mathbb{I}_{\text{tie}}\}$$

and $\varepsilon, \varepsilon > 0$ for numerical stability. We use $\text{BottomK}(X_{\cdot,w}, k)$ for the bottom-k set (with deterministic tie-handling stated in Algorithm 1).

7.2 Mechanism

(1) Reconstruct an effective vote share from heterogeneous components. We aggregate reconstructed components by a weighted power transform:

$$r_{i,w} = w_\star (s^{\star}_{i,w} + \varepsilon)^{\phi} + w_{\text{pro}} (s^{\text{pro}}_{i,w} + \varepsilon)^{\phi} + w_{\text{perf}} (s^{\text{perf}}_{i,w} + \varepsilon)^{\phi}, \qquad \sum_{\ell \in \{\star, \text{pro}, \text{perf}\}} w_\ell = 1 \tag{55}$$

and normalize to obtain an effective audience vote share:

$$V_{i,w} = \frac{r_{i,w}}{\sum_{j \in C_w} r_{j,w}} \tag{56}$$

To damp extreme mobilization, apply a second power transform and renormalize:

$$\tilde{V}_{i,w} = \frac{V_{i,w}^{\phi}}{\sum_{j \in C_w} V_{j,w}^{\phi}}, \qquad \phi > 0 \tag{57}$$

(2) Adjust judges' signal and compute discriminability-based dynamic weights. Because judges may partially reflect show appeal, we blend judges with $\tilde{V}$:

$$J_{i,w}^{*} = \beta J_{i,w} + (1-\beta)\tilde{V}_{i,w}, \qquad \beta \in (0,1) \tag{58}$$

For any share vector $X_w = \{X_{i,w}\}_{i \in C_w}$ define normalized entropy and discriminability:

$$H(X_w) = -\sum_{i \in C_w} X_{i,w} \ln X_{i,w}, \qquad D(X_w) = 1 - \frac{H(X_w)}{\ln N_w} \tag{59}$$

Let $D_J(w) = D(J_{\cdot,w}^{*})$, $D_V(w) = D(\tilde{V}_{\cdot,w})$, and define

$$\alpha_w^{\text{raw}} = \frac{D_J(w)}{D_J(w) + D_V(w)}, \quad \tilde{\alpha}_w = \text{clip}(\alpha_w^{\text{raw}}, \alpha_{\min}, \alpha_{\max}), \quad \alpha_w = (1-\lambda_\alpha)\tilde{\alpha}_w + \lambda_\alpha \alpha_{w-1} \tag{60}$$

(3) Combined score and dual-threshold elimination. Compute the unified weekly score

$$S_{i,w} = \alpha_w J_{i,w}^{*} + (1-\alpha_w)\tilde{V}_{i,w} \tag{61}$$

Define bottom-k sets under each dimension and the *AND-pool*:

$$B_w^J = \text{BottomK}(J_{\cdot,w}^{*}, k), \qquad B_w^V = \text{BottomK}(\tilde{V}_{\cdot,w}, k), \qquad \mathcal{E}_w = B_w^J \cap B_w^V \tag{62}$$

If $\mathcal{E}_w \neq \varnothing$, eliminate the weakest in the pool: $\text{Elim}(w) = \arg\min_{i \in \mathcal{E}_w} S_{i,w}$. If $\mathcal{E}_w = \varnothing$, the two signals structurally disagree; we then use a transparent fallback rule driven by *suspense*.

(4) Suspense-triggered judges' save. Let $S_{(1),w} \leq S_{(2),w} \leq \cdots \leq S_{(N_w),w}$ be the order statistics of $\{S_{i,w}\}$ and define

$$\text{SI}(w) = 1 - \frac{S_{(2),w} - S_{(1),w}}{S_{(N_w),w} - S_{(1),w} + \varepsilon} \tag{63}$$

When $\mathbb{I}_{\text{tight}} = 1$ and $\text{SI}(w) \geq \tau$, we consider $\mathcal{B}_w = \text{BottomK}(S_{\cdot,w}, 2)$ (expanded if $\mathbb{I}_{\text{tie}} = 1$) and eliminate the contestant in $\mathcal{B}_w$ with the smaller $J_{i,w}^{*}$. Otherwise, eliminate $\arg\min_{i \in C_w} S_{i,w}$.

Algorithm 1 Weekly elimination mechanism (deterministic ties)

Input: C_w, $\{\text{JudgeScore}_{i,w}\}$, $\{s^{\star}_{i,w}, s^{\text{pro}}_{i,w}, s^{\text{perf}}_{i,w}\}$, α_{w-1}, hyperparameters Θ
Compute $J_{i,w}$ and reconstruct $V_{i,w}$ via Eqs. (55)–(56)
Compute $\tilde{V}_{i,w}$ via Eq. (57) and $J^{*}_{i,w}$ via Eq. (58)
Compute α_w via Eqs. (59)–(60)
Compute $S_{i,w}$ via Eq. (61)
Form $\mathcal{E}_w$ via Eq. (62)
if $\mathcal{E}_w \neq \varnothing$ **then**
  Eliminate $\arg\min_{i\in\mathcal{E}_w} S_{i,w}$
else
  Compute $\text{SI}(w)$ via Eq. (63)
  if $\mathbb{I}_{\text{tight}} = 1$ and $\text{SI}(w) \geq \tau$ **then**
    Let $\mathcal{B}_w = \text{BottomK}(S_{\cdot,w}, 2)$ (expand if $\mathbb{I}_{\text{tie}} = 1$); eliminate $\arg\min_{i\in\mathcal{B}_w} J^{*}_{i,w}$
  else
    Eliminate $\arg\min_{i\in C_w} S_{i,w}$
  end if
end if
Tie-handling: for any arg min, break ties by smaller $J^{*}_{i,w}$, then smaller $\tilde{V}_{i,w}$, then a fixed pair-key order.

7.3 Evaluation Metrics

We evaluate counterfactual replays across seasons using:

Technical Fairness: Modified Protection Rate (MPR). Let $\mathcal{B}^{J}_{w}(k) = \text{BottomK}(J^{*}_{\cdot,w}, k)$. Then

$$\text{MPR}_k = \frac{1}{W}\sum_{w=1}^{W} \mathbf{1}\big(\text{Elim}(w) \in \mathcal{B}^{J}_{w}(k)\big) \tag{64}$$

Entertainment: Suspense Index.

$$\text{SI} = \frac{1}{W}\sum_{w=1}^{W} \text{SI}(w) \tag{65}$$

Operational robustness: disagreement and intervention rates.

$$\text{BCR} = \frac{1}{W}\sum_{w=1}^{W} \mathbf{1}\{\mathcal{E}_w = \varnothing\}, \qquad \text{JSR} = \frac{1}{W}\sum_{w=1}^{W} \mathbf{1}\{\mathcal{E}_w = \varnothing,\ \mathbb{I}_{\text{tight}} = 1,\ \text{SI}(w) \geq \tau\} \tag{66}$$

7.4 Calibration and Results

We calibrate Θ by GA-based counterfactual replay on Seasons 1–34 under three objective profiles: *Balanced*, *Fair*, and *Show*. All stochastic tie-handling is seeded for reproducibility.

Table 14: GA-tuned hyperparameters and aggregate performance (across evaluated weeks N).

Profile	ϕ	β	$\alpha_{\min}$	$\alpha_{\max}$	λ_α	k	$\mathbb{I}_{\text{tight}}$	τ	$\mathbb{I}_{\text{tie}}$	N	MPR_k	SI	BCR/ JSR
Balanced	1.30	0.797	0.517	0.745	0.646	4	0	0.95	0	328	1.0000	0.8686	0.0000/0.0000
Fair	0.30	0.990	0.800	0.950	0.582	6	1	0.637	1	312	1.0000	0.7117	0.0000/0.0000
Show	1.30	0.990	0.472	0.502	0.014	2	1	0.634	1	316	0.9525	0.8418	0.1171/0.1171

Balanced yields the strongest credibility (perfect protection, high suspense, no intervention). Fair enforces technical protection but trades off stability (lower SI). Show maintains high suspense while allowing *rule-triggered* interventions (nonzero BCR/JSR), offering controllable narrative tension without ad-hoc manual overrides.

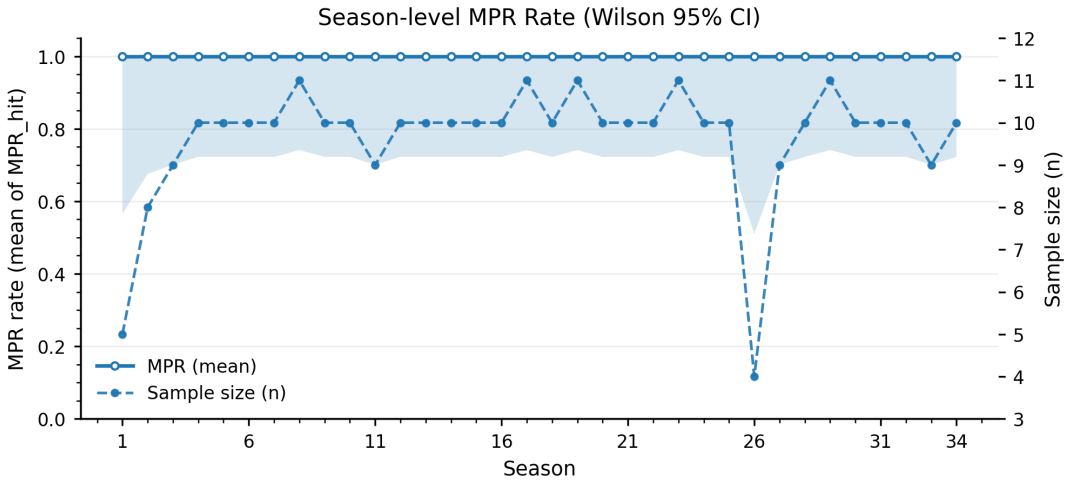

Figure 8: Season-level MPR hit rate with Wilson 95% confidence intervals; the dashed curve reports the number of evaluated elimination weeks per season.

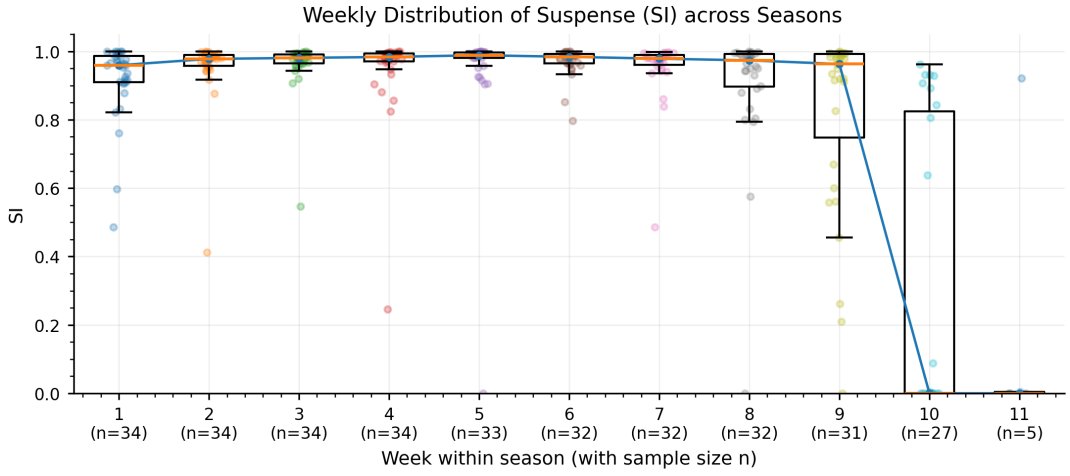

Figure 9: Weekly distribution of suspense index (SI) pooled across seasons; boxplots summarize seasons within each week and sample sizes are shown under the x-axis.

References

[1] Wikipedia contributors. (n.d.). *Dancing with the Stars (American TV series)*. Accessed 2026-02-02. https://en.wikipedia.org/wiki/Dancing_with_the_Stars_%28American_TV_series%29

[2] Instagram Help Center. (n.d.). *About Instagram insights*. Accessed 2026-02-02. https://www.instagram.com

[3] 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

[4] Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2), 193–202. https://doi.org/10.2307/2346567

[5] 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

[6] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197. https://doi.org/10.1109/4235.996017

[7] Eshelman, L. J., & Schaffer, J. D. (1993). Real-coded genetic algorithms and interval-schemata. In *Foundations of Genetic Algorithms 2*, 187–202. https://doi.org/10.1016/B978-0-08-094832-4.50018-0

[8] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Report on Use of AI Tools

1. AI tools used

- DeepSeek
  Model: DeepSeek-v3.2
  Use: Language polishing and structural refinement of the written report.
- Doubao
  Model: Doubao-Seed-1.8
  Version: Latest web release
  Use: Translation support between Chinese and English during drafting.

2. Uses of AI tools in this work

AI assistance was limited to writing support.

- Translation of short text segments for drafting.
- Grammar and syntax correction to improve fluency and readability.
- Academic tone and terminology adjustment without changing technical meaning.
- Rewriting for clarity, cohesion, and consistent terminology across sections.
- Concise reformulation to reduce redundancy while preserving meaning.
- Formatting-oriented edits of narrative text, including reordering sentences for logical flow.

3. Uses excluded from AI assistance

AI tools were not used for scientific or technical decision making.

- Designing or modifying mathematical models, assumptions, or algorithms.
- Deriving formulas, performing calculations, running simulations, or producing final numerical results.
- Selecting parameter values, interpreting outcomes, or drawing conclusions on behalf of the team.
- Generating, fabricating, or inventing citations, references, or factual claims.
- Producing any content that substitutes for the team's independent reasoning.

4. Verification and integrity

All AI-assisted edits were reviewed by the team before inclusion.

- Technical statements were checked against our own derivations, code outputs, and data tables.
- All references were manually verified to ensure they exist and are correctly cited.
- Final wording was edited by the team to preserve authorship and avoid plagiarism risk.

5. Representative record

Model: DeepSeek-v3.2
Platform: Web
Query:

> In the *Judge-Side Feature Construction / Elimination Labels and Outputs* stage, key research decisions include the definition of the active set, ranking rules, tie-handling, and the logic for inferring elimination weeks. How should I clearly document the validation procedures I used—such as spot-check sampling, targeted boundary-case testing, and cross-referencing against the *results* field—to ensure these rules and outputs are correct?

Output:

1. **Key research decisions and rationale.**

   - **Active set $\mathcal{A}_{s,w}$:** A pair is treated as active in week w if its weekly judges' total score is strictly positive, which excludes eliminated pairs and pre-entry records in accordance with the competition format.
   - **Ranking and tie handling:** We adopt the minimum-rank convention: tied pairs receive the same rank and subsequent ranks are skipped accordingly (e.g., $1, 1, 3$), avoiding arbitrary tie-breaking.
   - **Elimination-week inference:** When the *results* field explicitly reports "Eliminated Week X", we extract X directly; otherwise we infer the elimination week from the last scored week w_i^{last} relative to the season length W_s, treating $w_i^{\text{last}} = W_s$ as finalists.

2. **Systematic validation and quality assurance.**

   - **Level 1 (spot checks):** We randomly sample elimination weeks across seasons and manually reconcile raw tables with the derived active sets, ranks, and inferred elimination weeks.
   - **Level 2 (boundary cases):** We perform targeted checks for tie weeks, premiere/finale weeks, and irregular or missing *results* entries to ensure rule-consistent behavior.
   - **Level 3 (full cross-reference):** We compare the program-generated elimination-week list against the elimination-week list parsed directly from *results* and verify zero discrepancies.
   - **Level 4 (downstream consistency):** We confirm that the derived elimination indicator $e_{i,w}$ yields feasible solutions in the downstream elimination-consistency constraint model, providing an additional coherence check for the labels.

3. **Final statement (transparency and auditability).** We state that the rules were specified by the team based on competition-format analysis; implementation relies on standard tooling, while correctness is supported by manual review and programmatic cross-comparisons; validation artifacts (sampling lists, boundary-case inventories, and full comparison tables) are retained to ensure auditability.