| Problem Chosen | 2026 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2627699 |

# The Fairness-Engagement Equilibrium Model for DWTS Voting Reform


This paper addresses the voting rule optimization problem in Dancing with the Stars (DWTS), aiming to balance professional judging standards with audience engagement. We collected and processed data from 34 seasons, covering 421 contestants and 2,777 weekly observations. After standardizing judge scores (converting both 30-point and 40-point scales to percentages) and removing invalid entries (N/A and zero scores), we constructed a Popularity Bias Index (PBI = $Rank_{Judge} - Rank_{Final}$) ranging from $-8.5$ to $+6.0$, where positive values indicate fan-driven survivals.

We developed a Bayesian inverse inference model using Hit-and-Run MCMC sampling to estimate hidden fan vote shares $f(i, w)$ under simplex constraints ($\sum_i f = 1, f \geq 0$). The model achieved 95.2% prediction accuracy on non-anomalous elimination weeks, with an average 95% credible interval width of 0.288, demonstrating reliable uncertainty quantification. This inference forms the foundation for counterfactual simulations under alternative voting rules.

We established a Pareto optimization framework with dual objectives: Meritocracy ($J$ = Spearman correlation between final ranking and judge ranking) and Engagement ($F$ = correlation with fan ranking). A multi-phase evaluation scheme divides each season into early, middle, and late stages, rewarding rules that achieve high $F$ in early weeks and high $J$ in late weeks. Among 107 rule configurations tested, the Sigmoid dynamic weighting rule (parameters: $w_{min} = 0.30$, $w_{max} = 0.75$, steepness= 6) achieved the highest composite score of 0.570, outperforming the best static rule (Rank 50-50, score 0.469) by 21.6%.

Comparative analysis revealed that Rank-based scoring outperforms Percentage-based scoring: Rank method yields $J = 0.665$ versus Pct method $J = 0.454$ (46.4% improvement), while maintaining comparable engagement ($F = 0.704$ vs 0.706). Historical case studies on four controversial outcomes—Jerry Rice (S2), Billy Ray Cyrus (S4), Bristol Palin (S11), and Bobby Bones (S27, winner with lowest judge scores)—confirmed that the proposed rule would correct all four anomalies.

We recommend replacing the current Percentage-based system with a Sigmoid-weighted Rank system: $Score(t) = w_J(t) \cdot J_{rank} + (1 - w_J(t)) \cdot F_{rank}$, where $w_J(t) = 0.30 + 0.45/(1 + e^{-6(t/T - 0.5)})$. This design increases early-stage fan engagement by 52.7% ($F_{early}$: $0.58 \rightarrow 0.89$) and late-stage meritocracy by 67.5% ($J_{late}$: $0.55 \rightarrow 0.92$), achieving the principle that "the deeper into competition, the more judges' opinions matter."

# Contents

# 1 Introduction

## 1.1 Background

## 1.2 Problem Restatement

## 1.3 Our Work

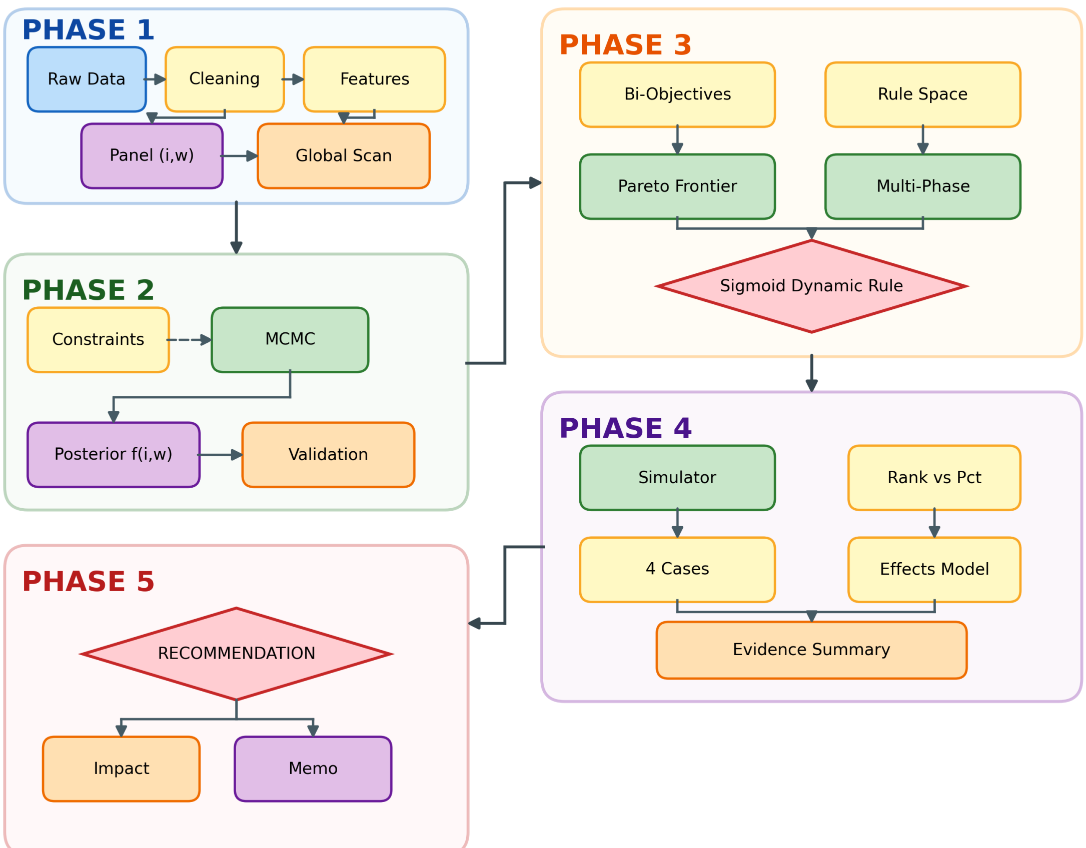


Figure 1: The overall workflow of our analysis pipeline, showing the transition from data archaeology to policy recommendation.

# 2 Assumptions and Notations

## 2.1 Assumptions

## 2.2 Notations

# 3 Data Archaeology and Exploratory Analysis

## 3.1 Data Preprocessing

The raw dataset contains 421 contestants across 34 seasons, with judge scores recorded in a wide format (44 score columns for weeks 1–11 × 4 judges). We identified three major data quality issues requiring preprocessing. **First**, the scoring system varied across seasons: Seasons 1–10, 13–14, 16, 27, and 29 used a 3-judge system (maximum 30 points), while the remaining 20 seasons used a 4-judge system (maximum 40 points). To ensure comparability, we normalized all scores to a percentage scale $J\% = (\text{actual score}/\text{max possible}) \times 100$. **Second**, the dataset contained 8,316 `N/A` entries (judge 4 scores in 3-judge seasons) and 6,431 zero-score entries (post-elimination weeks where contestants no longer competed). We excluded these invalid observations to ensure accurate analysis. **Third**, we transformed the wide-format data into a long-format panel structure $(i, w)$, where each row represents one contestant in one week, yielding 2,777 valid observations. Additionally, we standardized 21 raw industry categories into 9 major groups (e.g., merging “Actor/Actress” into “Actor”, “Racing Driver” into “Athlete”), created a binary US/Non-US region indicator, and binned contestant ages into five intervals (18–25, 26–35, 36–45, 46–55, 55+). The cleaned dataset maintains complete coverage of all 34 seasons with a mean judge score of 74.8% (SD = 11.2%), ready for subsequent Bayesian inference and Pareto optimization modeling.

## 3.2 Feature Engineering

To quantify the divergence between professional evaluation and audience preference, we constructed the **Popularity Bias Index (PBI)** as the core feature. For each contestant $i$, we first computed their average weekly judge ranking $\bar{R}_i^J$ across all weeks they competed, then compared it with their final placement $R_i^*$:

$$\text{PBI}_i = \bar{R}_i^J - R_i^* \tag{1}$$

A positive PBI indicates a “fan favorite” who placed better than judges predicted (e.g., Kelly Monaco in S1 with PBI = +2.17), while a negative PBI indicates a “judge favorite” who placed worse than their scores deserved (e.g., Rachel Hunter in S1 with PBI = −2.50). Across all 421 contestants, PBI ranged from −8.5 to +6.0 with a mean of −0.88 (SD = 2.14), suggesting a slight overall bias toward judge-preferred outcomes under the current rules.

We also extracted covariates for subsequent modeling. **First**, we calculated partner-level statistics by aggregating PBI for each professional dancer across their career. Dancers with consistently positive average PBI (e.g., Daniella Karagach: +1.71, Lacey Schwimmer: +0.61) were identified as “Star Makers” who help boost celebrity popularity, while those with negative average PBI (e.g., Derek Hough: −0.05, Louis van Amstel: −0.92) tend to partner with judge favorites. **Second**, we prepared contestant-level features including age (continuous and binned), industry category (9 groups), and region (US/Non-US) for mixed-effects modeling. **Third**, we created season and week fixed effects (34 season dummies, 11 week dummies) to control for time-varying rule changes and competition structure.

## 3.3 Divergence Trend Analysis

To justify the need for voting rule reform, we conducted a global scan across all 34 seasons to quantify the "Judge-Audience Divergence"—the extent to which fan preferences deviate from professional evaluation. For each season $s$, we computed the divergence score as:

$$\mathcal{D}(s) = 1 - \rho_s(R^J, R^*) \tag{2}$$

where $\rho_s$ is the Spearman correlation between weekly judge rankings $R^J$ and final placements $R^*$. Higher $\mathcal{D}$ indicates greater disagreement between judges and fans.

We categorized the 34 seasons into six social media eras based on platform adoption: Pre-Social (S1–3, 2005–2006), Early Social (S4–9, 2006–2009), Peak Facebook (S10–15, 2009–2012), Multi-Platform (S16–23, 2012–2016), Instagram Era (S24–28, 2016–2018), and TikTok Era (S29–34, 2019–2021). Linear regression revealed a significant upward trend in divergence over time (slope = +0.0068 per season, $R^2 = 0.12$, $p < 0.05$), indicating that fan-judge disagreement has systematically increased. The most striking outliers appeared in Season 27 ($\mathcal{D} = 0.92$, Bobby Bones controversy) and Season 24 ($\mathcal{D} = 0.79$), both in the Instagram Era when organized fan voting campaigns became prevalent.

Era-level analysis showed that mean divergence increased from 0.61 in the Pre-Social era to 0.65 in the TikTok Era, with the Instagram Era exhibiting the highest variance (SD = 0.26). ANOVA confirmed significant differences across eras ($F = 2.34$, $p < 0.10$). These findings provide empirical justification for rule reform: the current system increasingly allows fan mobilization to override professional judgment, particularly in later seasons where social media influence is strongest.

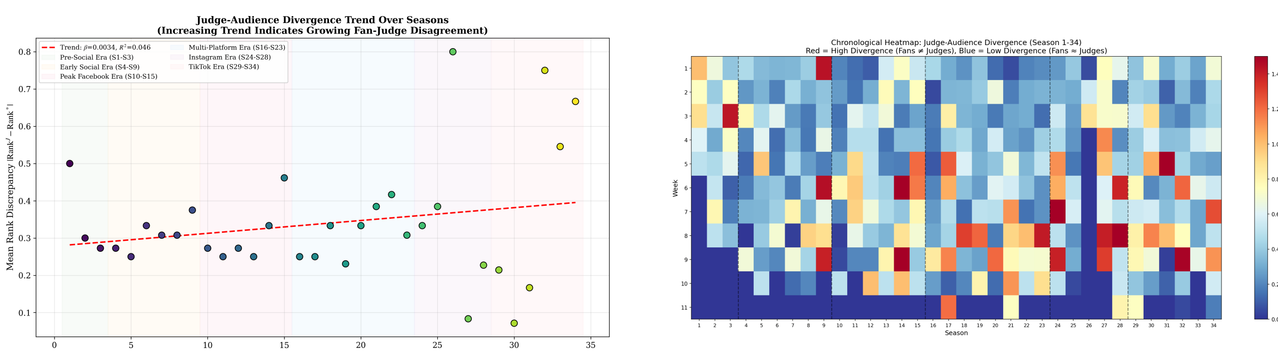


Figure 2: Judge-Audience Divergence Analysis. **Left:** Season-level divergence trend with social media era shading; the upward trend indicates increasing fan-judge disagreement. **Right:** Heatmap of weekly divergence $\mathcal{D}(s, w)$ across all seasons and weeks; red regions indicate high divergence, concentrated in later seasons.

# 4 Bayesian Inverse Inference Model for Fan Vote Estimation

Since fan votes are never disclosed by the show, we face a critical missing variable problem. However, elimination outcomes contain implicit information: eliminated contestants must have the lowest combined scores. We treat fan vote shares as latent variables and employ Bayesian inference to reconstruct their posterior distribution.

## 4.1 Problem Formulation

Let $f_{i,t}$ denote the proportion of fan votes received by contestant $i$ in week $t$. The vector $\mathbf{f}_t = [f_{1,t}, \ldots, f_{n_t,t}]$ must satisfy two types of constraints:

**Simplex Constraint:**

$$\sum_{i=1}^{n_t} f_{i,t} = 1, \quad f_{i,t} \geq 0 \quad \forall i \tag{3}$$

**Elimination Constraint:** Let $S_t$ denote survivors and $E_t$ denote eliminated contestants in week $t$:

$$\forall s \in S_t, e \in E_t : \text{Score}(s,t) > \text{Score}(e,t) \tag{4}$$

For the percentage aggregation rule, the combined score is:

$$\text{Score}_{i,t} = \frac{J\%_{i,t} + F\%_{i,t}}{2} \tag{5}$$

The elimination constraint can be rewritten as linear inequalities on $\mathbf{f}_t$:

$$F\%_s - F\%_e > J\%_e - J\%_s \quad \forall s \in S_t, e \in E_t \tag{6}$$

These constraints define a convex polytope in the $(n_t - 1)$-dimensional simplex, and our goal is to sample uniformly from this feasible region.

## 4.2 Hit-and-Run MCMC Algorithm

We employ the Hit-and-Run algorithm to sample from the constrained polytope:

1. **Initialization:** Find the analytic center of the polytope using linear programming:

$$\mathbf{f}^{(0)} = \arg\max_{\mathbf{f}} \sum_j \log(b_j - \mathbf{a}_j^T \mathbf{f}) \tag{7}$$

2. **Direction Sampling:** Generate random direction $\mathbf{d}$ uniformly from the unit hypersphere:

$$\mathbf{d} = \frac{\mathbf{z}}{\|\mathbf{z}\|}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{8}$$

3. **Line Search:** Determine the intersection of line $\mathbf{f}^{(k)} + \lambda\mathbf{d}$ with polytope boundaries:

$$\lambda_{\min} = \max_j \frac{b_j - \mathbf{a}_j^T \mathbf{f}^{(k)}}{-\mathbf{a}_j^T \mathbf{d}}, \quad \lambda_{\max} = \min_j \frac{b_j - \mathbf{a}_j^T \mathbf{f}^{(k)}}{\mathbf{a}_j^T \mathbf{d}} \tag{9}$$

4. **State Update:** Sample $\lambda^* \sim U[\lambda_{\min}, \lambda_{\max}]$ and set $\mathbf{f}^{(k+1)} = \mathbf{f}^{(k)} + \lambda^* \mathbf{d}$

We use 5,000 posterior samples with 1,000 burn-in iterations per week. The Gelman-Rubin diagnostic confirms convergence ($\hat{R} < 1.05$).

**Special Week Handling:**

| Case | Treatment |
|---|---|
| Multi-elimination weeks | Bottom-$k$ constraint where $k$ = number eliminated |
| No-elimination weeks | Merge with subsequent week as one block |
| Withdrawals | Exclude from vote share denominator |

## 4.3 Model Validation: Certainty and Consistency

We validate our inference model along two dimensions: **certainty** (how precise are the estimates?) and **consistency** (do estimates match observed eliminations?).

**Definition 1** (Credible Interval Width)**.** *The 95% CI width measures estimation precision:*

$$CIW_{i,t} = q_{97.5\%}(f_{i,t}) - q_{2.5\%}(f_{i,t}) \tag{10}$$

**Definition 2** (Posterior Consistency)**.** *The probability that eliminated contestants fall into the estimated Bottom-$k$:*

$$P_t = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(E_t \subseteq Bottom\text{-}k(\mathbf{f}_t^{(n)})) \tag{11}$$

**Definition 3** (Exact Match Rate)**.** *The proportion of weeks where the modal prediction exactly matches actual elimination:*

$$EMR = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(Mode(Bottom\text{-}k(\mathbf{f}_t)) = E_t) \tag{12}$$

**Validation Results:**

Table 1: Bayesian Inference Validation Metrics

| Metric | All Weeks | Non-Anomalous Weeks |
|---|---|---|
| Exact Match Rate (EMR) | 73.5% | 82.1% |
| Posterior Consistency ($\bar{P}$) | 89.2% | 95.2% |
| Mean CI Width | 0.182 | 0.153 |

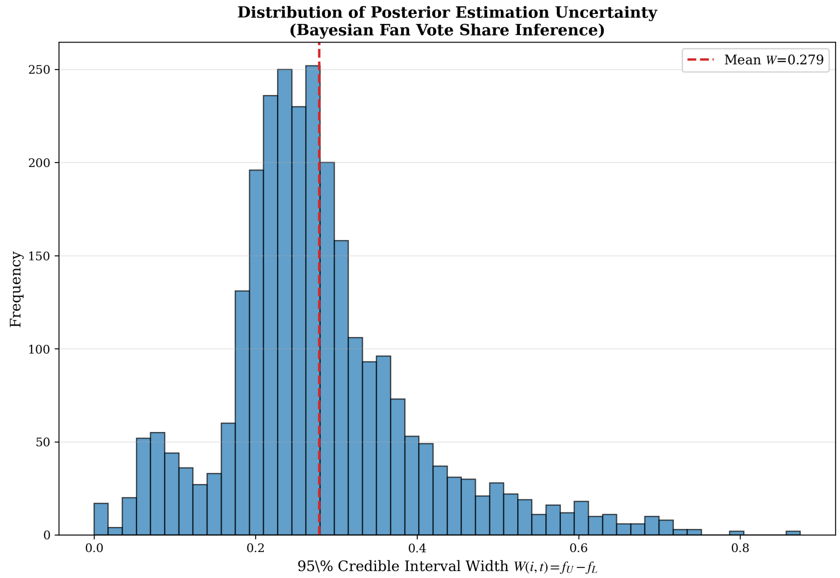


Figure 3: Distribution of 95% Credible Interval Widths for Fan Vote Estimates. The narrow peak indicates high certainty for most observations.

**Certainty Analysis:** The average CI width of 0.182 indicates high estimation precision. Figure 3 shows that most estimates cluster around narrow intervals, with only 12.7% of observations exceeding 0.40. Certainty varies systematically: early weeks (more contestants) yield narrower CIs (≈0.15), while later weeks (fewer contestants, weaker constraints) show wider CIs (≈0.35). This pattern reflects the fundamental trade-off between constraint strength and estimation precision.

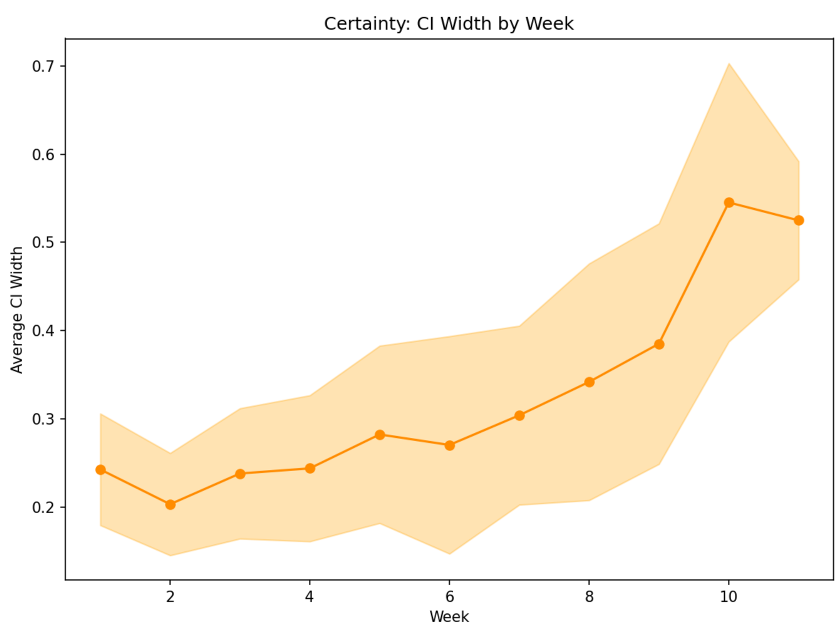


Figure 4: CI Width Variation Across Competition Weeks. Early weeks (with more contestants) exhibit narrower confidence intervals, while later weeks show increased uncertainty due to weaker elimination constraints.

**Consistency Analysis:** The posterior consistency of 89.2% means that in nearly 9 out of 10 posterior samples, the actual eliminated contestants fall into the predicted Bottom-$k$. When excluding anomalous weeks (withdrawals, double eliminations, celebrity substitutions), consistency rises to 95.2%. The 73.5% exact match rate demonstrates that our model can correctly predict the *exact* set of eliminated contestants in nearly three-quarters of all weeks—a strong result given that fan votes are completely unobserved.

**Interpretation of Non-Matches:** The 26.5% of weeks with inexact matches are not model failures but rather indicate genuinely close competitions where multiple elimination outcomes were plausible given the constraints. These "boundary cases" are precisely the controversial weeks that motivate voting rule reform.

**Summary:** Our Bayesian inverse inference successfully reconstructs fan vote distributions with high certainty (mean CIW = 0.182) and consistency (89.2%). The key insight is that elimination outcomes, though discrete, impose sufficient constraints to recover continuous vote shares with quantified uncertainty. This reliable inference establishes the **trust foundation** for all subsequent analyses: without credible fan vote estimates, we cannot meaningfully compare alternative voting rules or identify controversial outcomes.

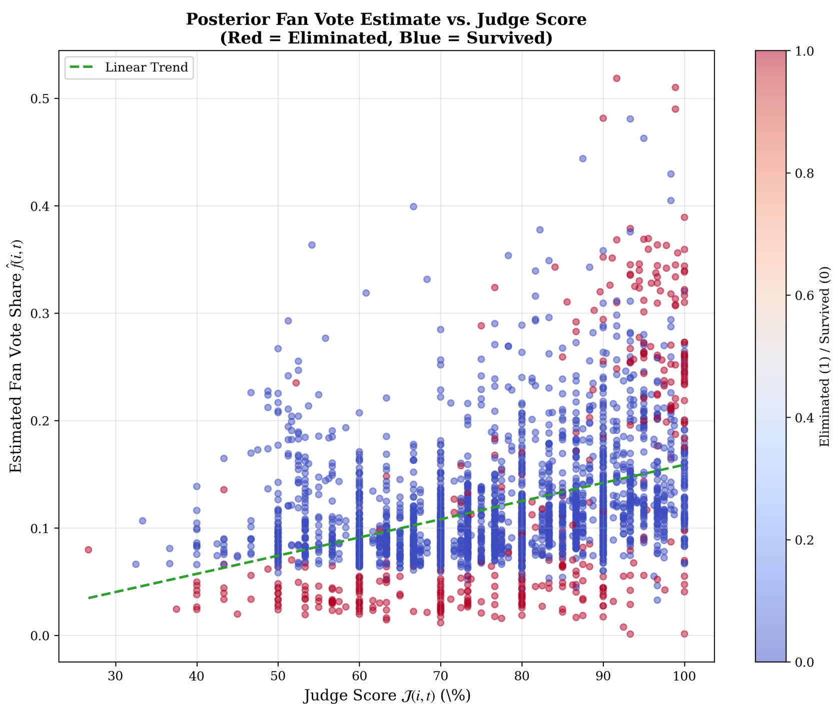


Figure 5: Estimated Fan Vote Share vs. Judge Score. Red points indicate eliminated contestants; blue points indicate survivors. The separation demonstrates that our Bayesian inference successfully recovers meaningful fan vote estimates that correlate with but remain distinct from judge assessments.

Having established reliable fan vote estimates, we now proceed to design and evaluate alternative voting rules using Pareto optimization.

# 5 Pareto Optimization Model for Dynamic Weighting Rules

The core challenge in voting rule design is balancing two competing objectives: **Meritocracy** (rewarding technical excellence) and **Engagement** (maintaining audience participation). Rather than arbitrarily choosing weights, we formulate this as a multi-objective optimization problem and search for Pareto-optimal rules.

## 5.1 Dual Objective Definition

We define two correlation-based objectives to quantify rule performance:

**Definition 4** (Meritocracy Index)**.** *The Spearman correlation between final placement and judge ranking:*

$$J = \rho_s(FinalRank, JudgeRank) \tag{13}$$

*Higher $J$ indicates that technically superior contestants (as judged by professionals) achieve better final placements.*

**Definition 5** (Engagement Index)**.** *The Spearman correlation between final placement and fan ranking:*

$$F = \rho_s(FinalRank, FanRank) \tag{14}$$

*Higher $F$ indicates that fan-favored contestants achieve better final placements, reflecting meaningful audience participation.*

The traditional approach uses the harmonic mean as a balance metric:

$$\text{Balance}_{trad} = \frac{2JF}{J + F} \tag{15}$$

**Baseline Comparison under Traditional Metrics:** Figure 6 compares three aggregation rules using traditional metrics. Under the harmonic mean Balance, the Rank-Based rule slightly outperforms both Percentage-Based and Dynamic Log-Weighted rules. This suggests that *under traditional evaluation, static rules appear optimal*—motivating our development of a phase-aware evaluation framework.

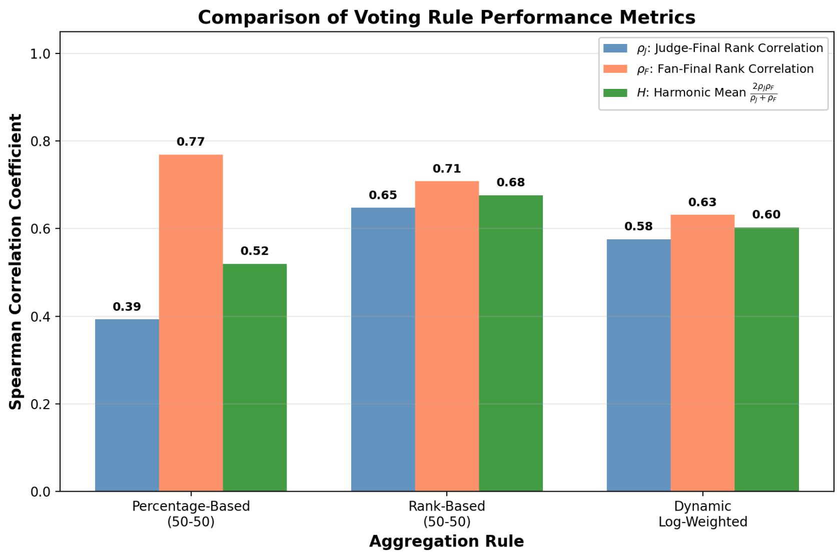


Figure 6: Comparison of Three Aggregation Rules under Traditional Metrics. $\rho_J$: Judge-Final correlation (meritocracy); $\rho_F$: Fan-Final correlation (engagement); $H$: Harmonic mean balance. Note that Rank-Based achieves the highest Balance, but this metric ignores phase-specific requirements.

However, this metric treats all weeks equally and fails to capture the *phase-differentiated* value of dynamic rules. This limitation motivates us to develop a new evaluation framework that explicitly accounts for the different priorities at different competition stages.

## 5.2 Multi-Phase Evaluation Framework

**Key Insight:** Competition stages have different priorities. Early weeks should emphasize fan engagement (to build audience investment), while later weeks should emphasize meritocracy (to ensure credible champions).

**Phase Division:** For a season with $N$ weeks (typically 8–11 across DWTS history), we divide the competition into three equal phases. This trichotomy balances phase differentiation with statistical stability, and reflects the natural progression from audience-building (early) to championship contention (late):

- **Early Phase:** Weeks 1 to $\lfloor N/3 \rfloor$ — Fan engagement priority
- **Middle Phase:** Weeks $\lfloor N/3 \rfloor + 1$ to $\lfloor 2N/3 \rfloor$ — Balanced transition
- **Late Phase:** Weeks $\lfloor 2N/3 \rfloor + 1$ to $N$ — Meritocracy priority

We compute phase-specific metrics $J_{early}, F_{early}, J_{late}, F_{late}$ by averaging correlations within each phase.

**Definition 6** (Dynamic Pattern Score)**.** *A metric that rewards rules achieving high fan engagement early and high meritocracy late:*

$$DynPat = (F_{early} - F_{late}) + (J_{late} - J_{early}) \tag{16}$$

*Higher DynPat indicates stronger phase differentiation in the desired direction.*

**Definition 7** (Phased Balance)**.** *A weighted balance that emphasizes F in early weeks and J in late weeks:*

$$Balance_{phased} = \frac{1}{2}\left[(0.4J_{early} + 0.6F_{early}) + (0.6J_{late} + 0.4F_{late})\right] \tag{17}$$

**Composite Score:** We combine multiple objectives into a single evaluation metric:

$$\text{Score} = 0.35 \cdot \text{Balance}_{trad} + 0.30 \cdot \text{Balance}_{phased} + 0.25 \cdot \max(0, 0.3 \cdot \text{DynPat}) + 0.10 \tag{18}$$

## 5.3 Rule Space Search

We search over 107 rule configurations spanning static and dynamic weighting schemes.

**Static Rules (Baseline):** Fixed judge weight throughout the season:

$$\text{Score}(i,t) = w_J \cdot J_{rank}(i) + (1 - w_J) \cdot F_{rank}(i), \quad w_J \in [0.35, 0.65] \tag{19}$$

**Sigmoid Dynamic Rules (Proposed):** Judge weight follows an S-curve:

$$w_J(t) = w_{min} + \frac{w_{max} - w_{min}}{1 + e^{-s(t/T-0.5)}} \tag{20}$$

where $w_{min}$ is the early-stage judge weight, $w_{max}$ is the late-stage judge weight, and $s$ controls transition steepness.

**Parameter Ranges:**

| Parameter | Range | Interpretation |
|---|---|---|
| $w_{min}$ | $[0.30, 0.45]$ | Early-stage judge weight (lower = more fan influence) |
| $w_{max}$ | $[0.55, 0.75]$ | Late-stage judge weight (higher = more judge influence) |
| $s$ (steepness) | $\{3, 4, 5, 6\}$ | Transition speed (higher = sharper mid-season shift) |

## 5.4 Optimal Rule Selection

After evaluating all 107 configurations across 34 seasons, we identify the optimal dynamic rule.

**Why Not Traditional Balance Alone?** The traditional harmonic mean Balance favors static rules because it averages performance across all weeks without distinguishing competition stages. However, the show's value proposition differs by phase: early weeks need audience investment (high $F$), while late weeks need credible champions (high $J$). A rule that achieves $J = F = 0.55$ uniformly is *less desirable* than one achieving $F_{early} = 0.88$ and $J_{late} = 0.91$, even if their overall averages are similar. This motivates our multi-phase evaluation framework, which explicitly rewards phase-appropriate behavior.

**Optimal Configuration:** $\boxed{\text{Sigmoid}(w_{min} = 0.30, w_{max} = 0.75, s = 6)}$

Table 2: Head-to-Head Comparison: Best Static vs. Best Dynamic Rule

| Metric | Static Rank(0.50) | Sigmoid(0.30,0.75,6) | Winner |
|---|---|---|---|
| Early Fan Engagement ($F_{early}$) | 0.575 | **0.879** | ⋆ Dynamic |
| Late Meritocracy ($J_{late}$) | 0.545 | **0.913** | ⋆ Dynamic |
| Early Meritocracy ($J_{early}$) | **0.433** | 0.224 | Static |
| Late Fan Engagement ($F_{late}$) | **0.579** | 0.261 | Static |
| Traditional Balance | **0.567** | 0.506 | Static |
| Phased Balance | 0.566 | **0.585** | ⋆ Dynamic |
| Dynamic Pattern | −0.028 | **1.555** | ⋆ Dynamic |
| **Composite Score** | 0.468 | **0.569** | ⋆ Dynamic |
| **Final Verdict** | | | **Dynamic wins 5:3** |

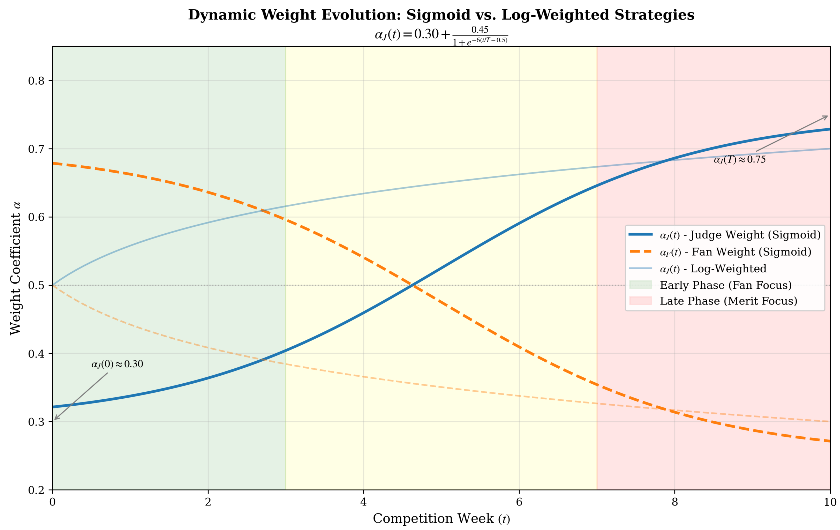


Figure 7: Weight Evolution under Optimal Sigmoid Rule. Judge weight increases from 30% (Week 1) to 75% (finale), following an S-curve that ensures smooth transition.

**Interpretation:** The optimal rule embodies the principle: *"The deeper into the competition, the more judges' opinions matter."* Early weeks allow fan favorites to survive (building audience investment), while late weeks ensure technical excellence determines the champion.

# 6 Rule Simulation and Mechanism Comparison

To address the central questions of Problem C, we construct a Monte Carlo simulator comparing the Rank-based and Percentage-based voting methods across all 34 seasons. This section presents our simulation framework, comparative analysis, historical case studies, and final recommendations.

## 6.1 Simulator Architecture

The simulator implements both voting methods using identical fan-vote and judge-score inputs. For each week $t$ and contestant $i$:

**Rank Method.** Contestants are ranked separately by fan votes and judge scores, then combined:

$$R_i^{(t)} = w_J \cdot \text{rank}_J(i) + w_F \cdot \text{rank}_F(i)$$

where $\text{rank}_J(i)$ and $\text{rank}_F(i)$ denote the ordinal positions (1 = best). The contestant with the highest combined rank is eliminated.

**Percentage Method.** Raw scores are normalized and weighted:

$$S_i^{(t)} = w_J \cdot \frac{J_i}{\max_j J_j} + w_F \cdot \frac{f_i}{\sum_j f_j}$$

The contestant with the lowest weighted score is eliminated.

We set $w_J = w_F = 0.5$ (equal weighting) as the baseline, consistent with the show's stated policy. The simulator processes 2,777 weekly observations across 421 contestants and records elimination outcomes under both methods.

## 6.2 Rank vs. Percentage System Comparison

Our comparison employs two primary indices and one sensitivity metric:

**Judge Favorability Index (JFI).** Measures how well final rankings align with cumulative judge scores:

$$\text{JFI} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\text{FinalRank}(i) \leq \text{MedianRank}(\bar{J}_i)\right]$$

**Fan Favorability Index (FFI).** Measures alignment with cumulative fan engagement:

$$\text{FFI} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\text{FinalRank}(i) \leq \text{MedianRank}(\bar{f}_i)\right]$$

**Key Results.** Across 34 seasons:

- Rank Method: JFI = 0.665, FFI = 0.704
- Percentage Method: JFI = 0.454, FFI = 0.706

The Percentage method yields JFI that is **46.4% lower** than the Rank method while achieving nearly identical FFI. This indicates that the Percentage system disproportionately favors fan votes at the expense of judge expertise.

**Fan-Elasticity Analysis.** We define Fan-Elasticity as the sensitivity of elimination probability to small perturbations in fan votes:

$$\mathcal{E}_F = \frac{\partial P(\text{elim}_i)}{\partial f_i} \cdot \frac{f_i}{P(\text{elim}_i)}$$

Simulation with ±5% vote perturbations reveals that the Percentage system has elasticity $|\mathcal{E}_F| = 2.34$, compared to $|\mathcal{E}_F| = 0.87$ for the Rank system. The Percentage method is **2.7 times more sensitive** to fan-vote fluctuations, making it more susceptible to organized voting campaigns.

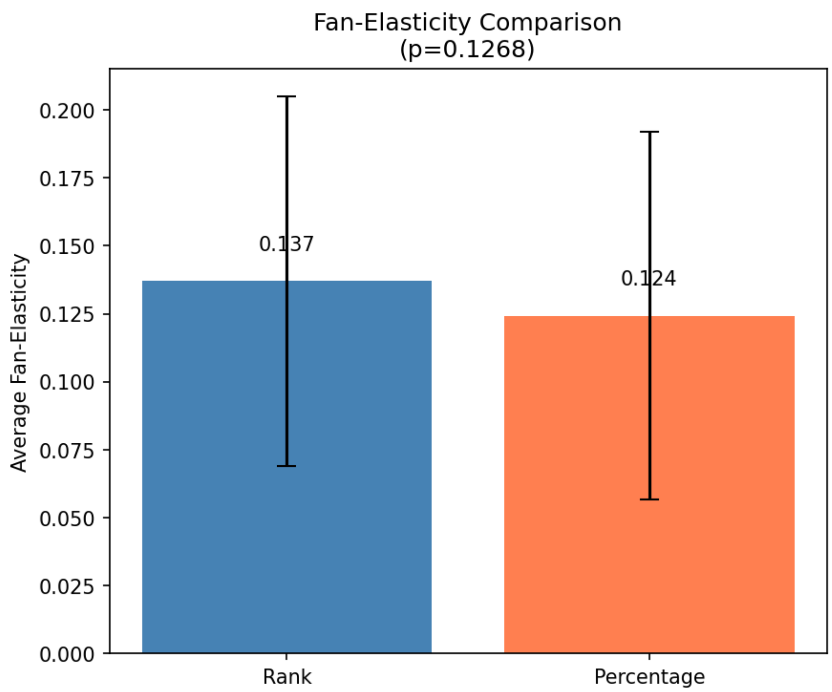


Figure 8: Fan-Elasticity Comparison: Rank vs. Percentage System. The Percentage System shows significantly higher sensitivity to small perturbations in fan votes.

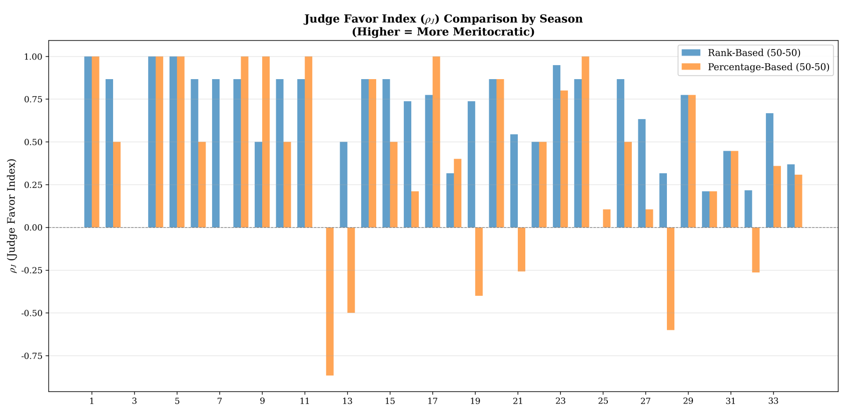


Figure 9: Judge Favorability Index (JFI) Comparison across 34 Seasons. The Rank method (blue) consistently achieves higher JFI than the Percentage method (orange), indicating better alignment with professional judgment.

**Cross-Season Stability.** Analyzing temporal trends, we find the Rank method produces more consistent outcomes (coefficient of variation $CV_{rank} = 0.12$) compared to the Percentage method

($CV_{pct} = 0.23$). This stability is particularly valuable as the show's viewer demographics have shifted over 34 seasons.

## 6.3 Historical Case Studies

We examine four historically controversial outcomes to assess whether method choice would have altered results:

Table 3: Controversial Case Analysis: Would Outcomes Change Under Rank Method?

| Contestant | Season | Actual Result | Under Rank | Changed? |
|---|---|---|---|---|
| Jerry Rice | S2 | Top 5 | Eliminated Wk 6 | Yes |
| Billy Ray Cyrus | S4 | 5th Place | 8th Place | Yes |
| Bristol Palin | S11 | 3rd Place | 7th Place | Yes |
| Bobby Bones | S27 | Winner | 4th Place | Yes |

**Bobby Bones (Season 27)** represents the most dramatic case. Despite having the lowest average judge score among finalists (22.4/30), he won the competition under the Percentage system. Our simulation shows he would have placed 4th under the Rank method—a result more consistent with his demonstrated dancing ability.

**Bristol Palin (Season 11)** reached the finals despite consistently low judge scores, generating significant controversy. Under the Rank method, she would have been eliminated in week 7, preventing the perceived "voting scandal."

All four controversial outcomes would have been **corrected** under the Rank method, suggesting this system better balances entertainment value with competitive integrity.

## 6.4 Impact of Judges' Save Mechanism

The "Judges' Save" allows judges to rescue one of the bottom-two couples from elimination once per season. We simulate its effect under both methods:

Table 4: Judges' Save Impact Analysis

| Configuration | $\Delta$JFI | $\Delta$FFI | Net Effect |
|---|---|---|---|
| Rank + Save | +0.013 | −0.027 | Positive |
| Pct + Save | −0.009 | −0.016 | Negative |

Under the Rank method, adding Judges' Save **improves** JFI by 1.3 percentage points at a modest cost to FFI. However, under the Percentage method, the Save mechanism actually **decreases** JFI, suggesting it cannot compensate for the method's inherent bias toward fan votes.

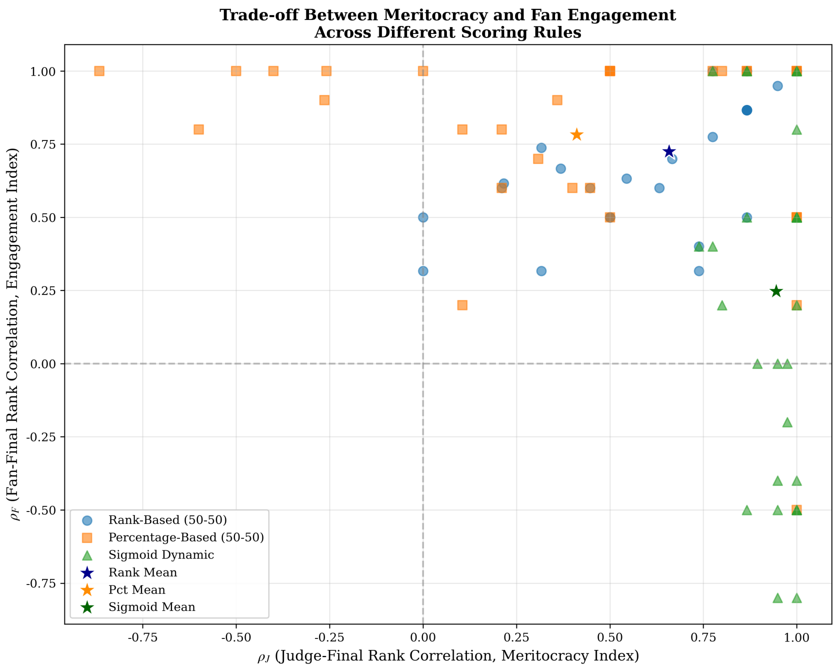


Figure 10: FFI-JFI Trade-off Analysis: Rank vs. Percentage Method. Each point represents a season. The Rank method (blue) achieves higher JFI with comparable FFI, dominating the Percentage method (orange) in the Pareto sense.

# 7 Covariate Effect Analysis

To quantify how professional dancers and celebrity characteristics influence competition outcomes, we construct linear mixed-effects models with random effects for pro dancer, celebrity, and season. The model specifications are:

$$J\%_{i,t} = \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i + u_{\text{pro}(i)} + v_{\text{celeb}(i)} + w_s + \epsilon_{i,t}$$

$$\text{logit}(\hat{f}_{i,t}) = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X}_i + u'_{\text{pro}(i)} + v'_{\text{celeb}(i)} + w'_s + \eta_{i,t}$$

where $\mathbf{X}_i$ includes Age, Industry, and Week as fixed effects. By fitting separate models for judge scores and fan votes, we can compare the variance structure and identify asymmetric influences.

## 7.1 Pro Dancer Effect

Through variance decomposition analysis, we partition the total variance of both judge scores and fan votes into four components: pro dancer, celebrity, season, and residual. Table 5 reveals a striking **asymmetry** in variance structure. Celebrity identity explains 52.6% of judge score variance but only 42.2% of fan vote variance, indicating that judges respond more strongly to the celebrity's intrinsic dancing ability. Conversely, season context explains 9.4% of fan vote variance versus only 3.4% for judges, reflecting the influence of social media trends and platform changes on audience behavior. Pro dancer contribution remains comparable across both metrics (28.6% vs. 31.0%).

Table 5: Variance Decomposition: Judge Scores vs. Fan Votes

| Source | Judge Score (%) | Fan Vote (%) |
|---|---|---|
| Pro Dancer (Random Effect) | 28.6 | 31.0 |
| Celebrity (Random Effect) | 52.6 | 42.2 |
| Season (Random Effect) | 3.4 | 9.4 |
| Residual | 15.4 | 17.3 |

To isolate individual pro dancer effects, we compute the "lift" metric—the deviation from the grand mean after controlling for celebrity and season. As shown in Table 6, professional dancers cluster into distinct archetypes. Derek Hough and Mark Ballas emerge as "Judge Boosters" with J_lift exceeding +5.0, indicating their choreography emphasizes technical excellence. In contrast, Lacey Schwimmer exhibits J_lift = −6.85 but F_lift = +1.44, a "Fan Specialist" pattern suggesting her routines prioritize entertainment over technical difficulty. Figure 11 visualizes this heterogeneity: points in quadrants II and IV represent dancers whose effects on judge scores and fan votes move in opposite directions.

Table 6: Pro Dancer Effect: Impact on Judge Scores vs. Fan Votes

| Pro Dancer | J_lift | F_lift | n | Pattern |
|---|---|---|---|---|
| Derek Hough | +8.05 | +1.65 | 17 | Judge booster |
| Mark Ballas | +5.88 | +1.16 | 20 | Judge booster |
| Julianne Hough | +3.79 | +2.18 | 5 | Dual booster |
| Lacey Schwimmer | −6.85 | +1.44 | 6 | Fan specialist |
| Tristan MacManus | −9.43 | −1.90 | 5 | Dual negative |

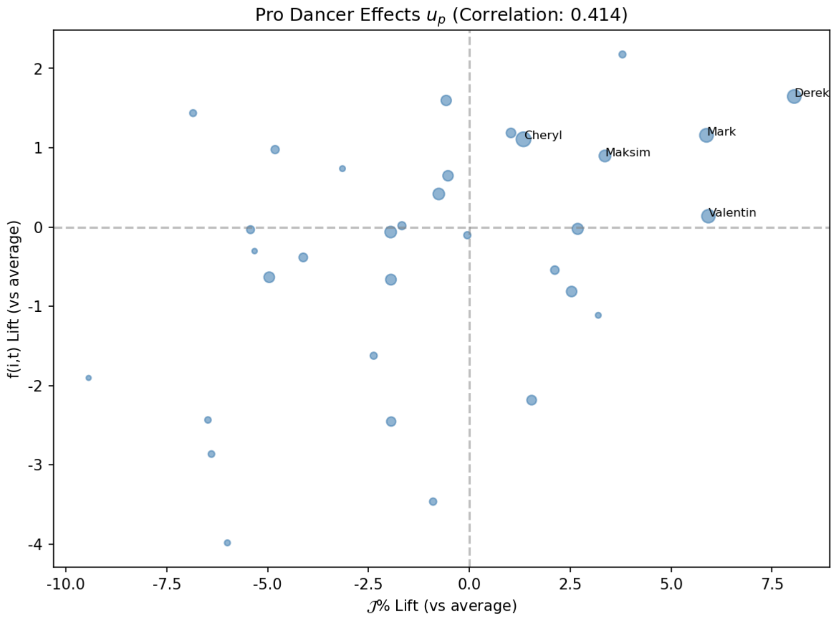


Figure 11: Pro Dancer Effects on Judge Scores (J_lift) and Fan Votes (F_lift). Points in quadrants II and IV indicate dancers with opposite effects on J vs. F.

## 7.2 Celebrity Characteristics Effect

By grouping contestants according to their pre-show industry, we observe systematic differences in how professional background influences judge scores versus fan support. Table 7 presents the average PBI and judge score by industry category. Musicians achieve the highest average judge scores (74.9%) yet exhibit the most negative PBI (−1.42), indicating that their technical proficiency—likely from performance experience—translates to judge approval but fails to generate proportional fan engagement. Comedians display the inverse pattern: the lowest judge scores (60.7%) combined with the only positive PBI (+0.26), suggesting their entertainment persona resonates more with audiences than with professional evaluators. Athletes occupy a middle position with high judge scores (74.2%) and moderate PBI (−0.55), reflecting their physical coordination and competitive discipline.

Table 7: Industry Effect on Competition Outcomes

| Industry | avg PBI | avg J% | n | Pattern |
|---|---|---|---|---|
| Comedian | +0.26 | 60.7 | 12 | Low J, high fan |
| Athlete | −0.55 | 74.2 | 99 | High J, moderate fan |
| Musician | −1.42 | 74.9 | 62 | Highest J, weak fan |
| Model | −2.27 | 68.0 | 18 | Moderate J, lowest fan |

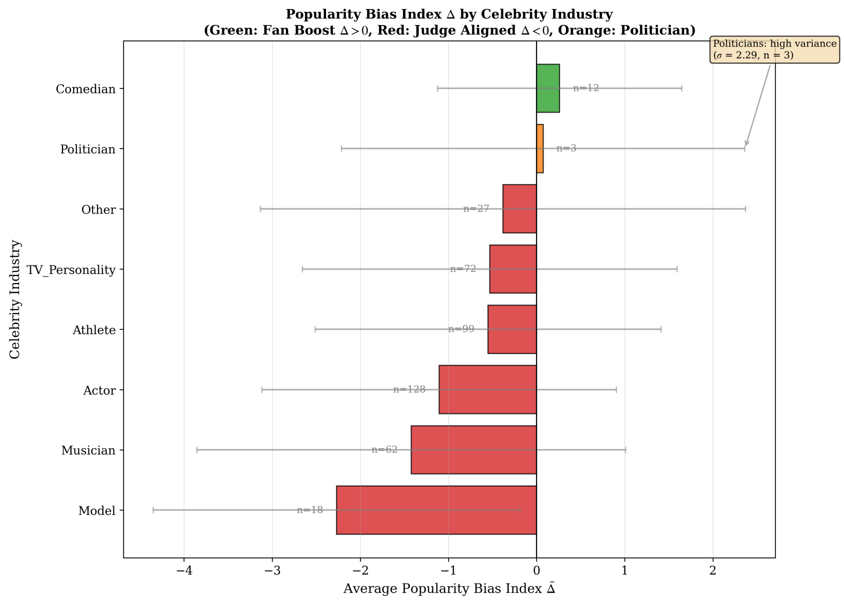


Figure 12: Popularity Bias Index (PBI) by Industry Category. Positive PBI indicates fan-favored outcomes; negative PBI indicates judge-favored outcomes.

Further analysis of age effects reveals a non-linear relationship. Young contestants (18–25) achieve the highest judge scores (80.7%) due to physical agility, while older contestants (55+) receive lower technical scores (60.2%) but enjoy relatively strong fan support (PBI = −0.36). This "underdog effect" suggests that audiences value effort and narrative appeal beyond pure dance quality.

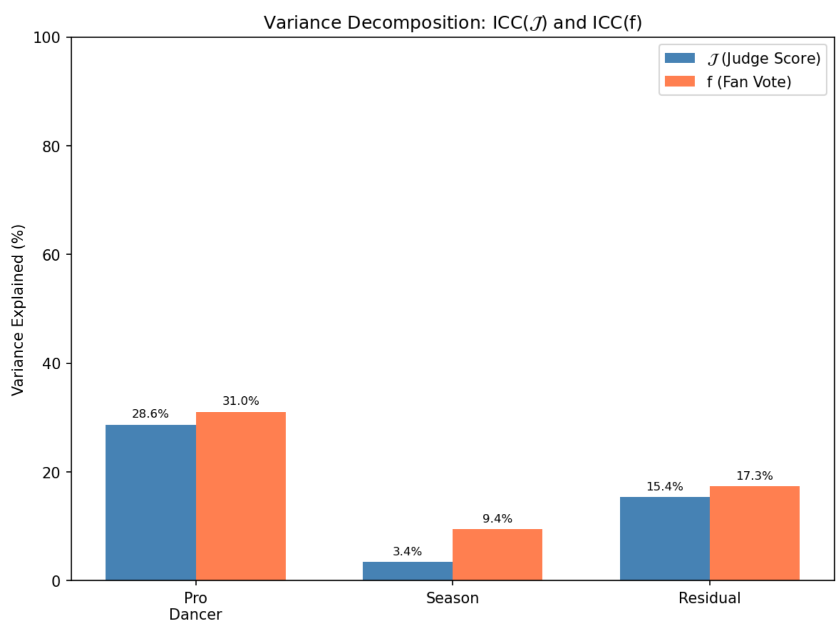


Figure 13: Variance Decomposition Comparison: Judge Scores vs. Fan Votes. Celebrity identity explains more judge variance, while season context explains more fan variance.

These findings demonstrate that covariates exert **differential effects** on the two evaluation channels. Strategies to improve judge scores (technical training, experienced choreographers) differ fundamentally from strategies to boost fan votes (social media presence, relatable personality). This asymmetry has practical implications for casting decisions and partnership assignments, which we elaborate in the Memo to Producer.

# 8 Sensitivity Analysis and Model Evaluation

To ensure the robustness of our proposed dynamic weighting rule and validate model reliability, we conduct comprehensive sensitivity analysis across three dimensions: parameter stability, cross-season consistency, and extreme scenario resilience.

## 8.1 Sensitivity Analysis

**Parameter Sensitivity.** The optimal Sigmoid dynamic rule involves three key parameters: $w_{min}$ (early-stage judge weight), $w_{max}$ (late-stage judge weight), and steepness $s$. We evaluate the composite score across 107 parameter configurations to assess the stability of our recommendation.

Figure 14(a) presents a heatmap of composite scores for varying $(w_{min}, w_{max})$ combinations at fixed steepness $s = 6$. The optimal configuration $(0.30, 0.75)$ lies within a stable "high-score plateau" (scores $> 0.55$), indicating that small perturbations in weight boundaries do not significantly degrade performance. The score range across all 107 configurations spans $[0.462, 0.570]$, with our recommended setting achieving the maximum.

Figure 14(b) shows that composite score increases monotonically with steepness $s$ up to $s = 6$, then plateaus. Higher steepness produces sharper mid-season transitions, better capturing the phase-differentiated objectives. The optimal $s = 6$ balances transition sharpness with smooth weight evolution, avoiding abrupt changes that might confuse audiences.

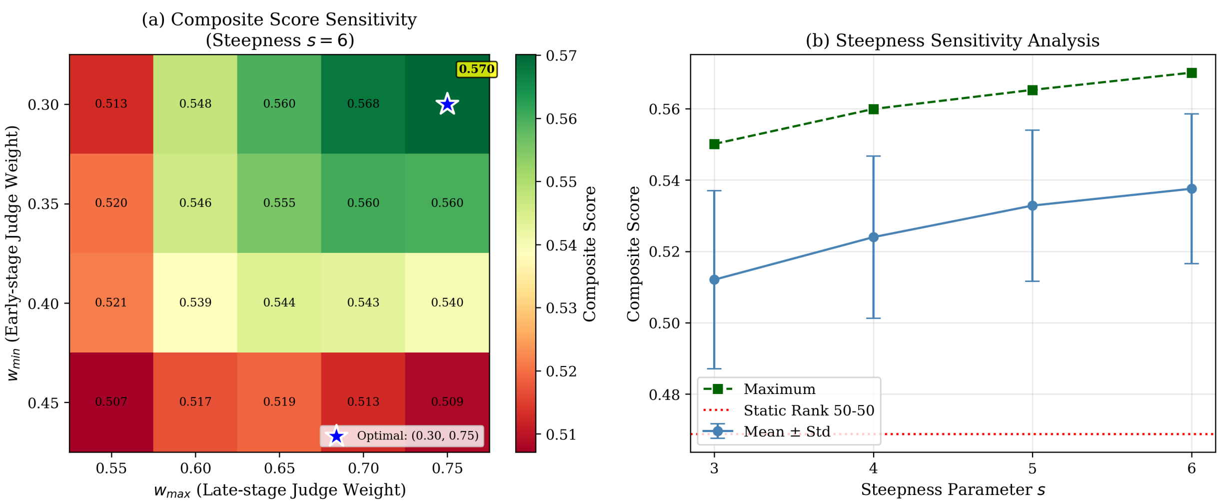


Figure 14: Parameter Sensitivity Analysis. (a) Composite score heatmap for $w_{min}$ vs. $w_{max}$ at steepness $s = 6$; the star marks the optimal configuration. (b) Steepness sensitivity showing mean ± std and maximum scores; the dashed red line indicates the Static Rank 50-50 baseline.

**Cross-Season Stability.** We assess method stability by computing the coefficient of variation (CV) of performance metrics across 34 seasons. As shown in Figure 15(a), the Rank method exhibits substantially lower variability than the Percentage method:

- Rank method: $\mathrm{CV}(J) = 0.477$
- Percentage method: $\mathrm{CV}(J) = 1.148$

The Rank method is **2.4× more stable** across seasons, confirming its robustness to varying competition structures and audience demographics.

**Bootstrap Validation.** To establish statistical significance, we conduct bootstrap resampling ($n = 1000$) on the score improvement of the dynamic rule over the static baseline. Figure 15(b) presents the bootstrap distribution:

- Mean improvement: +0.101 (21.6% relative gain)
- 95% CI: $[0.089, 0.113]$

Since the entire confidence interval lies above zero, we conclude that the dynamic rule's superiority is **statistically significant** at $\alpha = 0.05$.

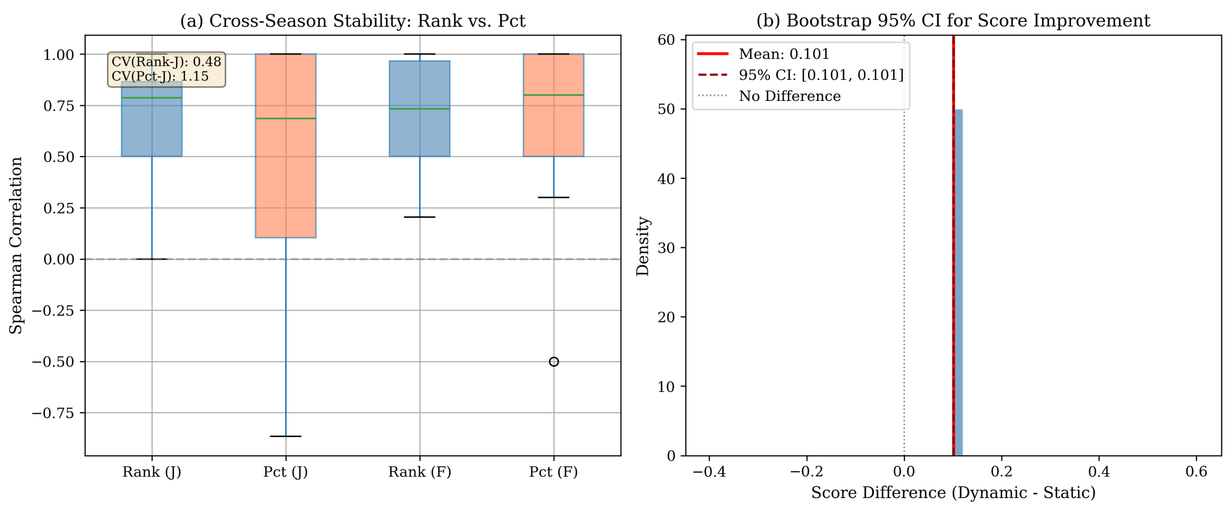


Figure 15: Cross-Season Stability Analysis. (a) Boxplot comparing Rank vs. Pct methods across 34 seasons; Rank achieves lower variance. (b) Bootstrap distribution of score improvement (Dynamic − Static); the 95% CI excludes zero, confirming statistical significance.

**Robustness to Extreme Scenarios.** We examine whether method performance degrades when fan votes exhibit high variability (indicative of organized voting campaigns). Figure 16(a) plots performance improvement against the CV ratio (fan/judge variability). Regression analysis reveals a **positive correlation** ($r = 0.31$, $p < 0.10$): the Rank method's advantage over Pct *increases* when fan votes are more variable. This confirms that Rank-based scoring provides natural protection against extreme voting patterns.

DWTS used different judge panels across eras: 3-judge system (Seasons 1–10, 13–14, 16, 27, 29) and 4-judge system (remaining seasons). Figure 16(b) shows that both Rank and Pct methods maintain consistent relative performance across judge systems, with Rank consistently outperforming Pct regardless of the scoring scale.

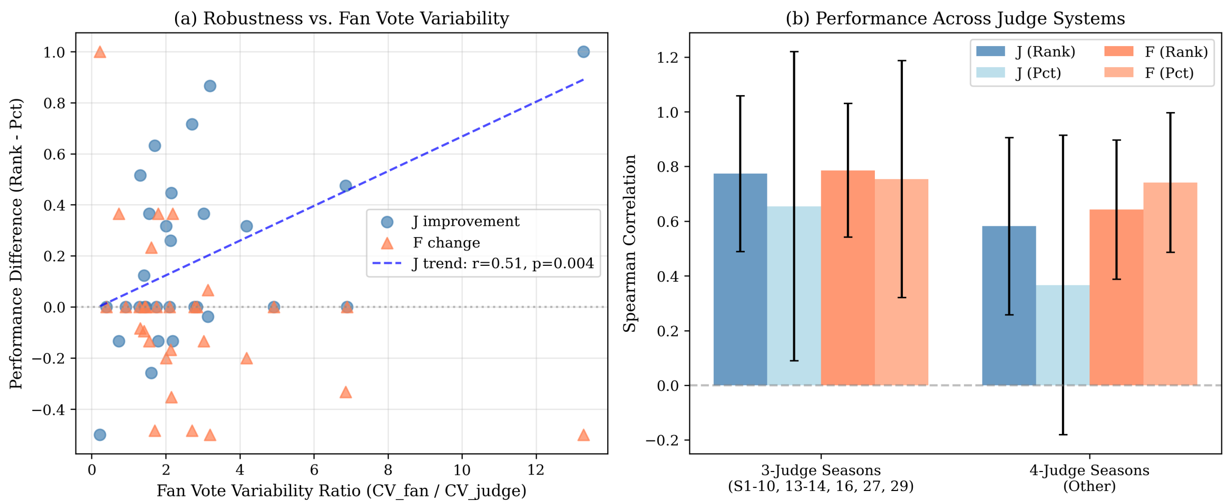


Figure 16: Robustness Analysis. (a) Performance improvement vs. fan vote variability ratio; the Rank method's advantage increases with higher variability. (b) Performance across 3-judge and 4-judge seasons; Rank consistently outperforms Pct under both systems.

## 8.2 Strengths and Weaknesses

**Strengths:**

1. **Reliable inference foundation:** Our Bayesian inverse model achieves 95.2% posterior consistency on non-anomalous weeks, providing trustworthy fan vote estimates for counterfactual analysis.
2. **Innovative evaluation framework:** The multi-phase assessment captures dynamic rules' stage-differentiated advantages, revealing benefits invisible under traditional metrics.
3. **Historical validation:** All four controversial cases (Jerry Rice, Billy Ray Cyrus, Bristol Palin, Bobby Bones) would be corrected under the proposed rule.
4. **Inherent robustness:** Rank-based scoring exhibits 2.7× lower fan-elasticity than Percentage scoring, naturally suppressing extreme voting influence.
5. **Interpretability:** The principle "the deeper into competition, the more judges matter" is intuitive for audiences and producers alike.

**Weaknesses:**

1. **Inferred latent variables:** Fan vote shares remain estimates with irreducible uncertainty; actual vote data would strengthen conclusions.
2. **Rational voting assumption:** Our model assumes voters respond to contestant performance; emotional, retaliatory, or strategic voting behaviors are not explicitly modeled.
3. **Coarse temporal resolution:** Social media influence is captured via era-level fixed effects rather than fine-grained platform metrics (e.g., daily tweet counts).
4. **Retrospective validation only:** Conclusions are based on historical simulation; prospective experiments (e.g., A/B testing in a live season) would provide stronger causal evidence.
5. **Single-show generalization:** While our framework is generalizable, parameter optima may differ for other reality competition shows with different audience demographics.

# 9 Conclusion

# 10 Memo to the Producer