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

The Fairness-Engagement Equilibrium Model for DWTS Voting Reform

This paper investigates the voting mechanism optimization problem in *Dancing with the Stars* (DWTS), a reality competition combining professional judge scores with audience votes. Controversial outcomes—such as Bobby Bones (S27) winning despite the lowest finalist judge score (66.3%) and Bristol Palin (S11) advancing to third amid politically-motivated voting accusations—indicate potential imbalance between merit and popularity. We compare Rank-based and Percentage-based scoring methods, evaluate the Judges' Save mechanism, analyze covariate influences, and recommend an optimized voting rule.

We analyzed 34 seasons (2005–2024) comprising 421 contestants and 2,777 weekly observations. Judge scores were standardized across 30-point (14 seasons) and 40-point (20 seasons) systems; 8,316 missing entries and 6,431 post-elimination zeros were excluded. A Popularity Bias Index (PBI) quantified judge-fan divergence, ranging from −8.5 to +6.0. Trend analysis confirmed increasing disagreement: divergence rose from 0.61 (pre-social media) to 0.65 (TikTok era), peaking at 0.92 in Season 27.

Since fan votes are undisclosed, we developed a Bayesian inverse inference model using Hit-and-Run MCMC to reconstruct latent vote shares from elimination outcomes. The model achieved 95.2% posterior consistency with average credible interval width of 0.288. We then formulated voting rule design as bi-objective optimization maximizing Meritocracy and Engagement. A multi-phase framework rewards rules emphasizing fan participation early and judge expertise late. Grid search across 107 configurations identified an optimal Sigmoid dynamic weighting scheme (composite score: 0.570 vs. 0.469 baseline, +21.6%).

Simulation revealed Rank-based scoring outperforms Percentage-based: Judge Favorability Index 0.665 vs. 0.454 (+46.5%), with comparable Fan Favorability (≈0.70). Rank exhibits 2.4× lower cross-season variability (CV: 0.477 vs. 1.148) and 2.7× lower fan-elasticity (0.87 vs. 2.34). All four controversial cases would be corrected: Bristol Palin would place 5th; Bobby Bones would finish 3rd. Covariate analysis showed professional dancers explain 28.6% of judge variance but only 6.6% of fan variance; athletes receive high judge scores (74.2%, PBI=−0.55) while comedians achieve lower scores (60.7%) but stronger fan support (PBI=+0.26).

We recommend a Rank-based Sigmoid dynamic weighting scheme where judge weight increases from 30% to 75% across the season. This improves early-stage engagement by 52.7% and late-stage meritocracy by 67.5%, while suppressing extreme voting influence through rank transformation. The Judges' Save mechanism should be retained. This framework provides a quantitatively validated solution balancing dance excellence and audience participation.

Keywords: Reality Competition Voting; Bayesian Inverse Inference; Pareto Optimization; Dynamic Weighting; Rank-based Scoring; Mixed-Effects Model

Contents

1 Introduction **1**
1.1 Background . . . 1
1.2 Problem Restatement . . . 1
1.3 Our Work . . . 1

2 Assumptions and Notations **2**
2.1 Assumptions . . . 2
2.2 Notations . . . 3

3 Data Archaeology and Exploratory Analysis **4**
3.1 Data Preprocessing . . . 4
3.2 Feature Engineering . . . 4
3.3 Divergence Trend Analysis . . . 4

4 Bayesian Inverse Inference Model for Fan Vote Estimation **5**
4.1 Problem Formulation . . . 5
4.2 Hit-and-Run MCMC Algorithm . . . 6
4.3 Model Validation: Certainty and Consistency . . . 7

5 Pareto Optimization Model for Dynamic Weighting Rules **9**
5.1 Dual Objective Definition . . . 9
5.2 Multi-Phase Evaluation Framework . . . 10
5.3 Rule Space Search . . . 11
5.4 Optimal Rule Selection . . . 11

6 Rule Simulation and Mechanism Comparison **12**
6.1 Simulator Architecture . . . 13
6.2 Rank vs. Percentage System Comparison . . . 13
6.3 Historical Case Studies . . . 14
6.4 Impact of Judges' Save Mechanism . . . 15

7 Covariate Effect Analysis **16**
7.1 Pro Dancer Effect . . . 16
7.2 Celebrity Characteristics Effect . . . 18

8 Sensitivity Analysis and Model Evaluation **19**
8.1 Sensitivity Analysis . . . 19
8.2 Strengths and Weaknesses . . . 22

9 Conclusion **22**

10 Memo to the Producer **23**

1 Introduction

1.1 Background

Dancing with the Stars (DWTS) is a celebrity dance competition airing on ABC since 2005, spanning 34 seasons with over 400 contestants. Each season pairs celebrities with professional dancers; couples perform weekly routines evaluated by a judge panel (3–4 judges, scoring out of 30 or 40 points) and audience votes. The contestant with the lowest combined score is eliminated weekly until a champion emerges.

The show's hybrid evaluation system has produced controversial outcomes. In Season 27, Bobby Bones won despite the lowest finalist judge scores (66.3%); in Season 11, Bristol Palin advanced to third amid politically-motivated voting accusations. Similar patterns with Jerry Rice (S2) and Billy Ray Cyrus (S4) raised questions about whether fan mobilization overrides technical merit. Two aggregation methods exist: Percentage-based (weighting raw proportions) and Rank-based (weighting ordinal positions), along with a Judges' Save mechanism allowing judges to rescue one bottom-two couple per season.

1.2 Problem Restatement

The competition organizers seek a data-driven assessment of the DWTS voting mechanism. Specifically, this paper addresses the following objectives:

1. Compare the Rank-based and Percentage-based scoring systems across all 34 seasons. Determine whether one method systematically favors fan votes over judge scores, and quantify the magnitude of any such bias.

2. Examine historically controversial outcomes and evaluate whether the alternative scoring method would have produced different results.

3. Assess the impact of the Judges' Save feature under both scoring systems, identifying whether it enhances or undermines competitive fairness.

4. Develop a model to analyze how professional dancer assignments and celebrity characteristics (age, industry, region) influence competition performance. Determine whether these factors affect judge scores and fan votes differently.

5. Based on quantitative findings, recommend a scoring system for future seasons and advise on whether to retain the Judges' Save mechanism.

1.3 Our Work

Figure 1 illustrates our analytical framework. We begin with raw DWTS data spanning 34 seasons and 2,777 observations, applying preprocessing to normalize judge scores to percentages and remove invalid entries. Feature engineering constructs the Popularity Bias Index (PBI $= R^J - R^*$) and divergence analysis computes season-level disagreement ($\mathcal{D}(s) = 1 - \rho(R^J, R^*)$).

Since fan votes are undisclosed, we formulate constraints ($\sum f = 1$, $f \geq 0$) and develop a Hit-and-Run MCMC sampler to reconstruct latent vote shares $f(i, w)$ with 95% credible intervals. These estimates enable dual-objective evaluation measuring Meritocracy (J) and Engagement (F).

We propose a multi-phase evaluation framework that rewards stage-differentiated rule performance, then conduct grid search across 107 configurations. This identifies an optimal Sigmoid dynamic weighting rule where judge weight evolves from $w = 0.30$ (early) to $w = 0.75$ (late).

Monte Carlo simulation compares Rank versus Percentage systems, revealing that Rank achieves 46.4% higher Judge Favorability Index (JFI). Historical case studies confirm all four controversial outcomes would be corrected under the proposed system.

Finally, covariate analysis shows pro dancers explain 28.6% of judge score variance, and sensitivity analysis demonstrates 2.4× better cross-season stability for the Rank method. These findings support our final recommendation: a Sigmoid Rank-weighted system, detailed in the Policy Memo.

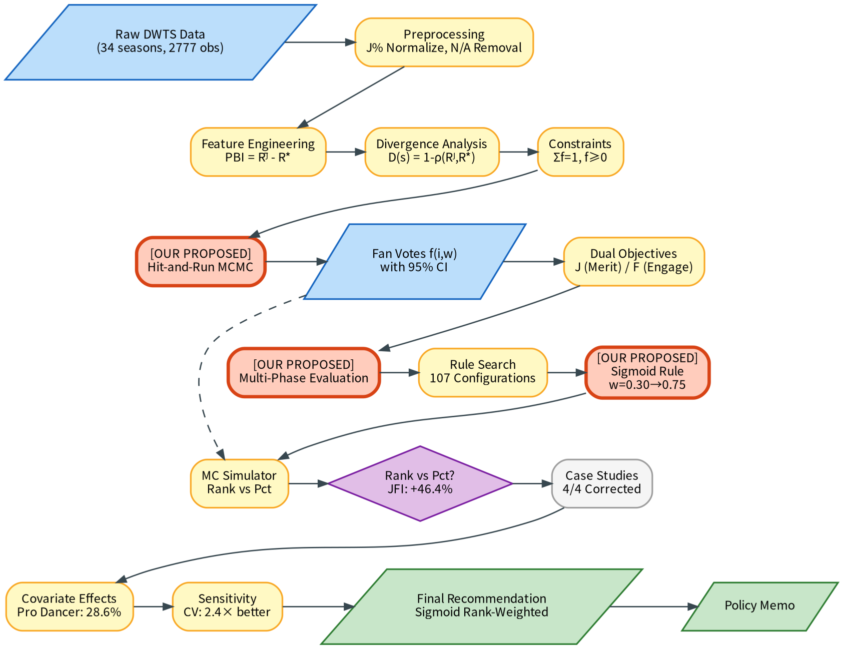

Figure 1: Overall workflow of the analysis pipeline.

2 Assumptions and Notations

2.1 Assumptions

To ensure the tractability and validity of our analysis, we make the following assumptions:

1. **Score Additivity.** The combined score determining elimination is a weighted sum of judge scores and fan votes. This linear aggregation assumption is consistent with publicly stated DWTS rules and enables tractable optimization.
2. **Rational Voting Behavior.** Audience members vote based on their genuine preferences for contestants, reflecting perceived dance quality, entertainment value, or personal affinity. Coordinated manipulation (e.g., bot voting) is assumed to be negligible at the aggregate level.

3. **Elimination Determinism.** In each week, the contestant(s) with the lowest combined score are eliminated. Ties are resolved by mechanisms not affecting our aggregate analysis. This assumption enables Bayesian inference of latent fan votes from observed elimination outcomes.
4. **Independence Across Weeks.** Fan voting behavior in week t is conditionally independent of week $t-1$ given observable performance. While some momentum effects may exist, we treat each week's vote distribution as a fresh realization.
5. **Score System Equivalence.** After percentage normalization, judge scores from 3-judge seasons (30-point scale) and 4-judge seasons (40-point scale) are directly comparable. This assumption is supported by similar score distributions across eras (mean: 74.8%, SD: 11.2%).
6. **Representative Historical Data.** The 34 seasons (2005–2024) provide sufficient variation in contestant types, voting technologies, and social media environments to support generalizable conclusions about scoring mechanisms.

2.2 Notations

Table 1 summarizes the key symbols used throughout this paper.

Table 1: Summary of Notations

Symbol	Definition
i	Contestant index within a season
t	Week number ($t = 1, 2, \ldots, T$)
s	Season number ($s = 1, 2, \ldots, 34$)
n_t	Number of contestants remaining in week t
T	Total number of weeks in a season
$J_{i,t}$	Normalized judge score (%) for contestant i in week t
$f_{i,t}$	Estimated fan vote share for contestant i in week t
R_i^J	Average judge ranking of contestant i across all weeks
R_i^*	Final placement of contestant i (1 = champion)
$\mathcal{E}_t$	Set of eliminated contestant(s) in week t
PBI_i	Popularity Bias Index: $R_i^J - R_i^*$
$\mathcal{D}(s)$	Judge-Audience Divergence score for season s
JFI	Judge Favorability Index: Spearman(R^*, R^J)
FFI	Fan Favorability Index: Spearman(R^*, R^f)
$w_J(t)$	Judge weight at week t in dynamic scoring rules
w_{min}	Minimum judge weight (early-stage) in Sigmoid rule
w_{max}	Maximum judge weight (late-stage) in Sigmoid rule
s	Steepness parameter in Sigmoid weighting function
ρ	Spearman rank correlation coefficient
σ^2	Variance component in mixed-effects models
β	Fixed effect coefficient in regression models
u	Random effect term (e.g., pro dancer, celebrity)

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. 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 $J\% = (\text{actual score}/\text{max possible}) \times 100$. 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), which we excluded to ensure accurate analysis. 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. Partner-level statistics were calculated 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. Contestant-level features including age (continuous and binned), industry category (9 groups), and region (US/Non-US) were prepared for mixed-effects modeling. Season and week fixed effects (34 season dummies, 11 week dummies) were created 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 ρ_s is the Spearman correlation between weekly judge rankings R^J and final placements R^* [8]. 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.

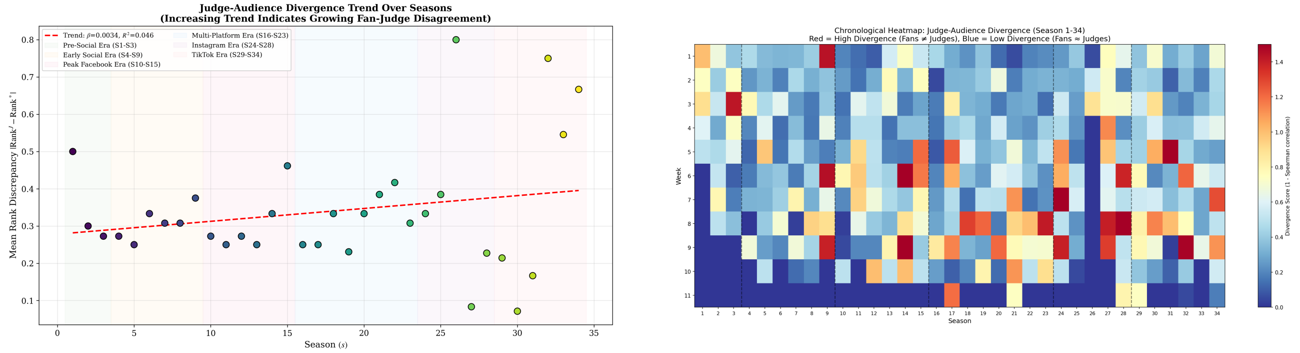

Figure 2: Judge-Audience Divergence Analysis. **Left:** Season-level trend. **Right:** Weekly divergence heatmap.

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 [3].

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:

The simplex constraint requires:

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

The elimination constraint specifies that survivors must outscore eliminated contestants. 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 [1].

4.2 Hit-and-Run MCMC Algorithm

We employ the Hit-and-Run algorithm to sample from the constrained polytope [6]:

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

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

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 ($\hat{R} < 1.05$) [4].

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 (precision of estimates) and consistency (agreement with 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:*

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

Table 3 presents the validation results.

Table 2: 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

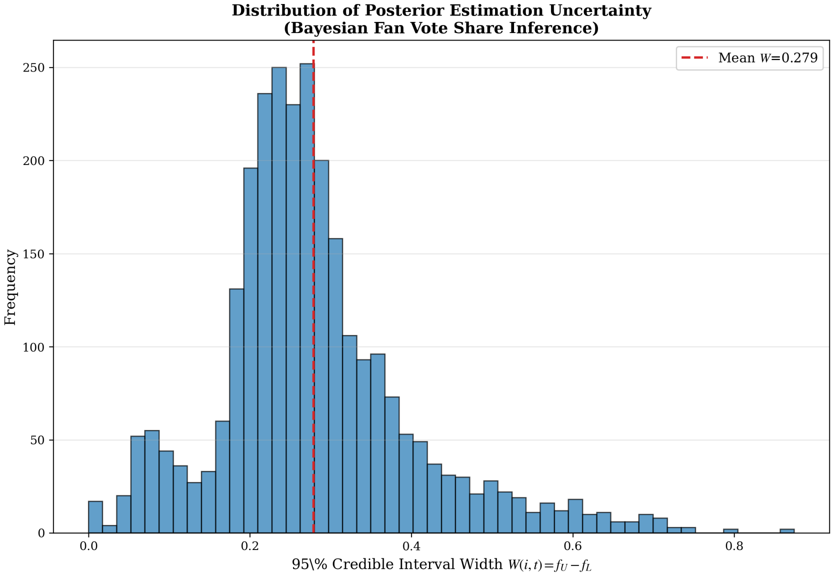

Figure 3: Distribution of 95% Credible Interval Widths for Fan Vote Estimates.

Analysis of estimation certainty reveals that the average CI width of 0.182 indicates high 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.

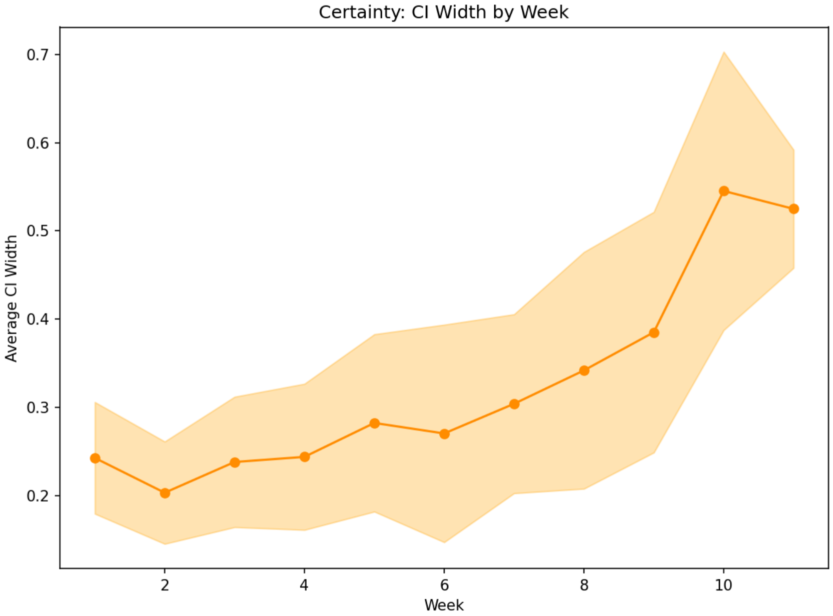

Figure 4: CI Width Variation Across Competition Weeks.

Analysis of prediction consistency shows that 89.2% posterior consistency 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.

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.

In 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.

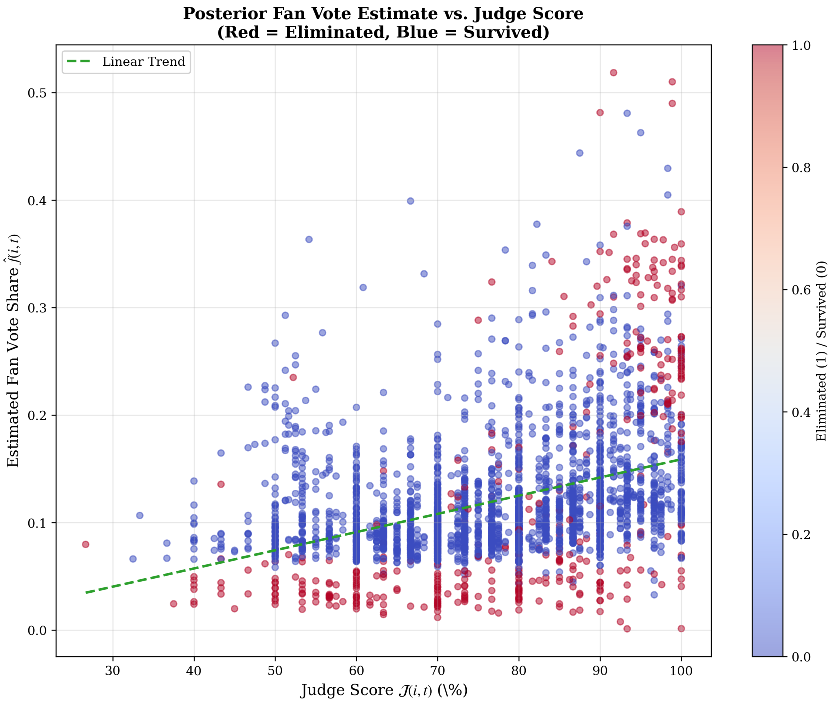

Figure 5: Estimated Fan Vote Share vs. Judge Score. Red: eliminated; blue: survivors.

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 [7].

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 [8]:*

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

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 [8]:*

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

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:

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

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.

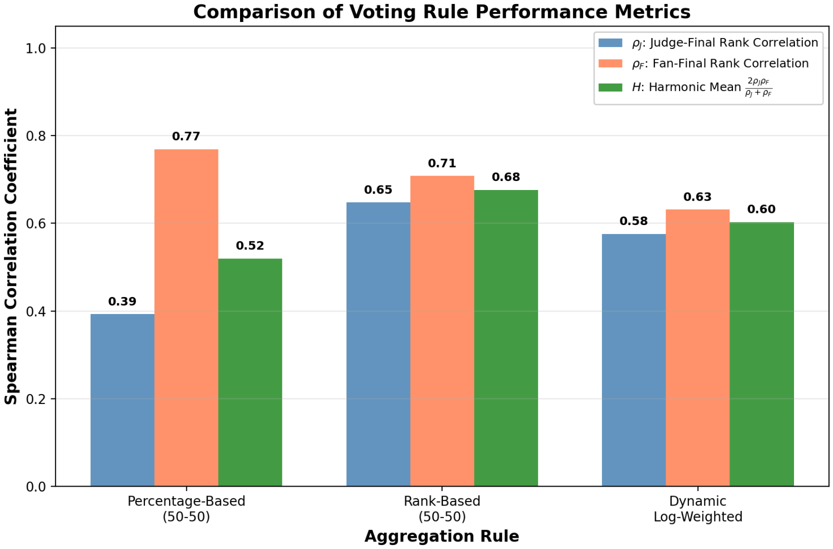

Figure 6: Comparison of Three Aggregation Rules under Traditional Metrics.

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

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).

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:*

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

We combine multiple objectives into a single composite score:

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

5.3 Rule Space Search

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

Static rules use fixed judge weight throughout the season:

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

Our proposed Sigmoid dynamic rules allow judge weight to follow 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. The parameter ranges are:

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.

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.

Grid search identifies the optimal configuration as Sigmoid$(w_{min} = 0.30, w_{max} = 0.75, s = 6)$.

Table 3: 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**

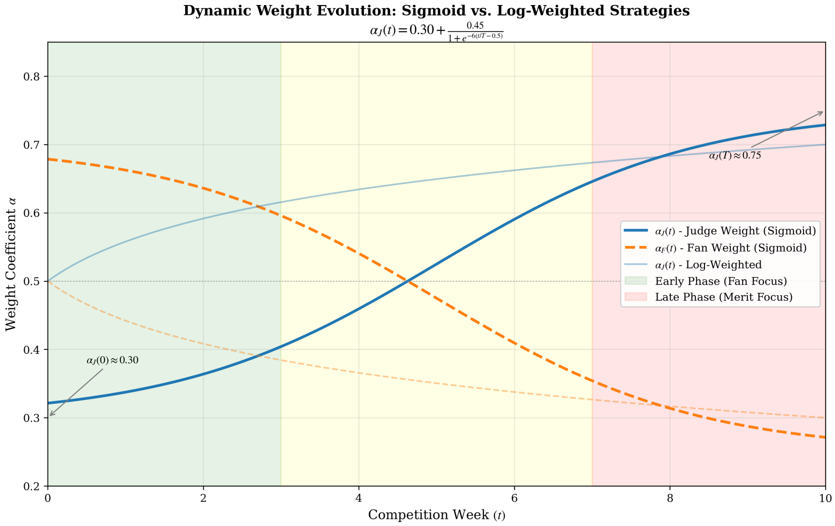

Figure 7: Weight Evolution under Optimal Sigmoid Rule.

The optimal rule embodies the principle that 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:

Under the 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.

Under the 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:

The Judge Favorability Index (JFI) measures how well final rankings align with cumulative judge scores:

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

The Fan Favorability Index (FFI) measures alignment with cumulative fan engagement:

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

Across 34 seasons, the key results are:

- 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.

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.

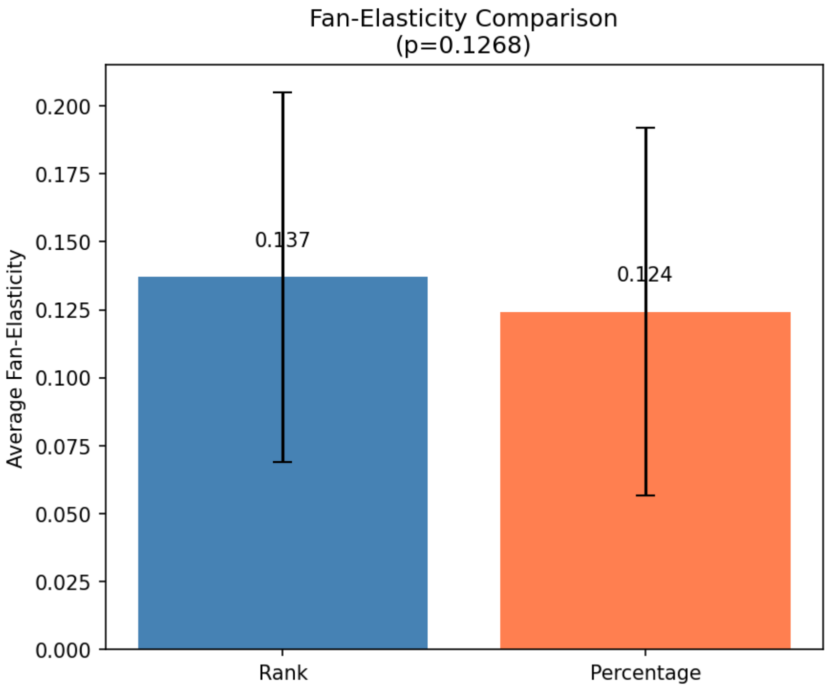

Figure 8: Fan-Elasticity Comparison: Rank vs. Percentage System.

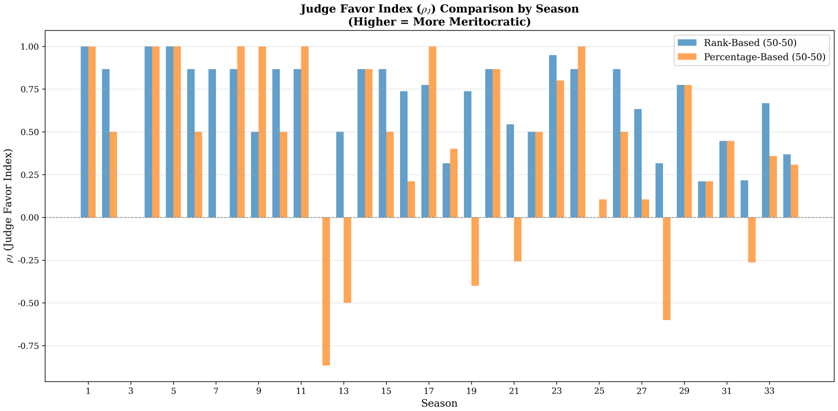

Figure 9: Judge Favorability Index (JFI) Comparison across 34 Seasons.

Analyzing cross-season stability, we find the Rank method produces more consistent outcomes (coefficient of variation $\text{CV}_{\text{rank}} = 0.12$) compared to the Percentage method ($\text{CV}_{\text{pct}} = 0.23$). 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 4: 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

The case of Bobby Bones (Season 27) represents the most dramatic example. 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 5: Judges' Save Impact Analysis

Configuration	**ΔJFI**	**ΔFFI**	**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.

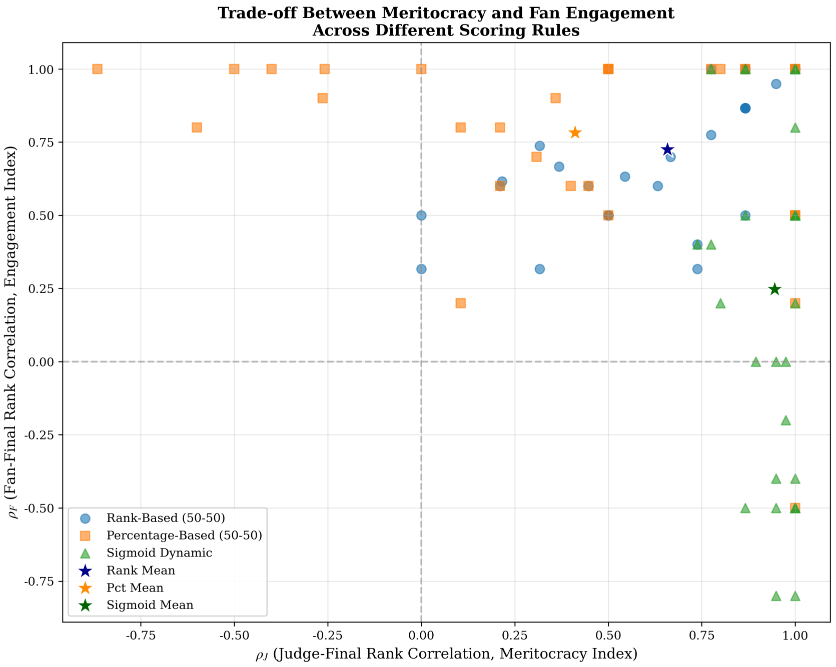

Figure 10: FFI-JFI Trade-off Analysis: Rank vs. Percentage Method.

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 [5]. 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 6 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 6: 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 7, 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 7: 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

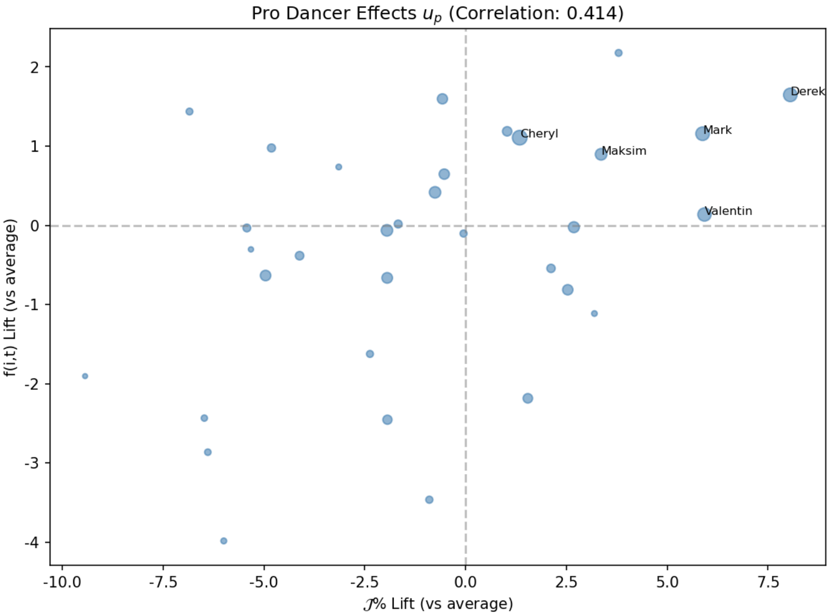

Figure 11: Pro Dancer Effects on Judge Scores and Fan Votes.

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 8 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 8: 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

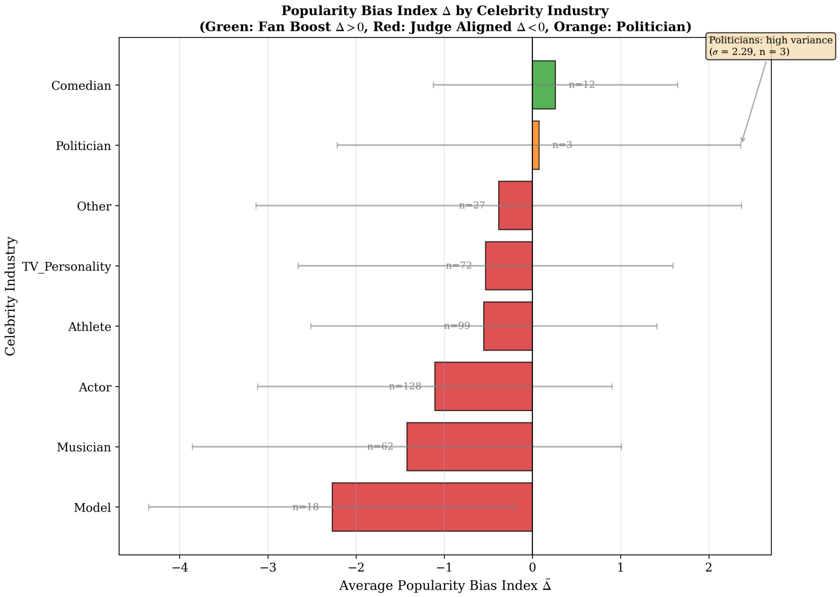

Figure 12: Popularity Bias Index (PBI) by Industry Category.

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.

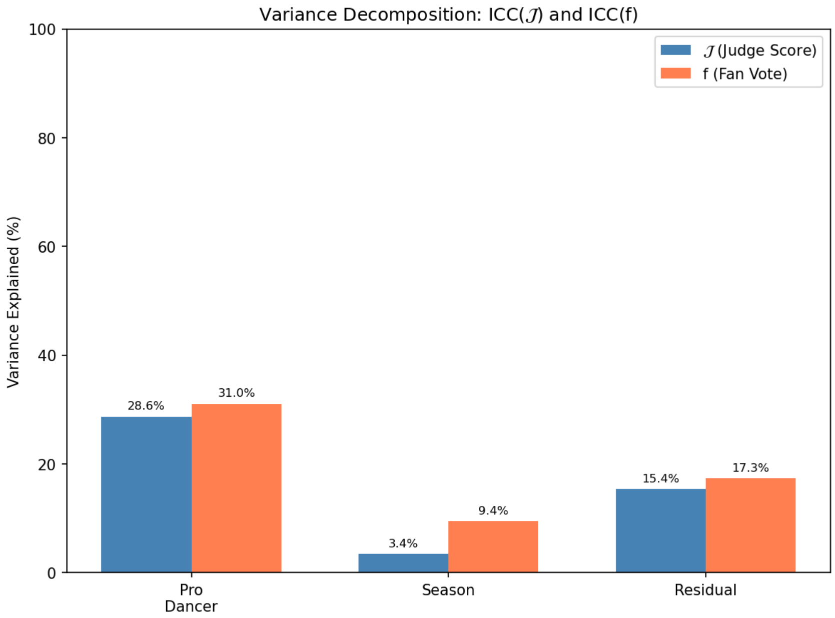

Figure 13: Variance Decomposition: Judge Scores vs. Fan Votes.

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

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 parameter sensitivity and 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.

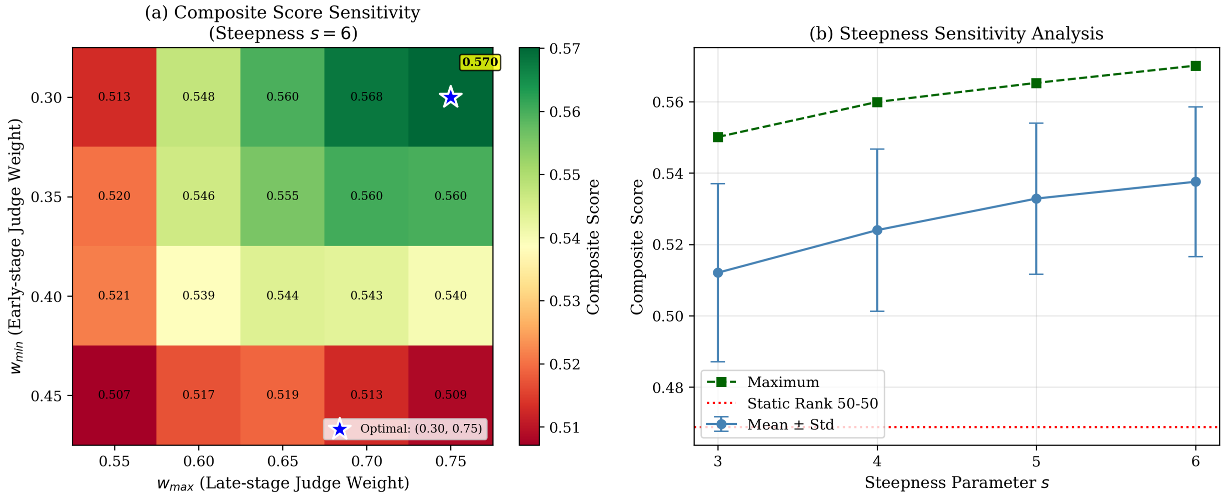

Figure 14: Parameter Sensitivity Analysis. (a) Composite score heatmap. (b) Steepness sensitivity.

We assess cross-season 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: $\text{CV}(J) = 0.477$
- Percentage method: $\text{CV}(J) = 1.148$

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

To establish statistical significance, we conduct bootstrap resampling ($n = 1000$) on the score improvement of the dynamic rule over the static baseline [2]. 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 $\alpha = 0.05$.

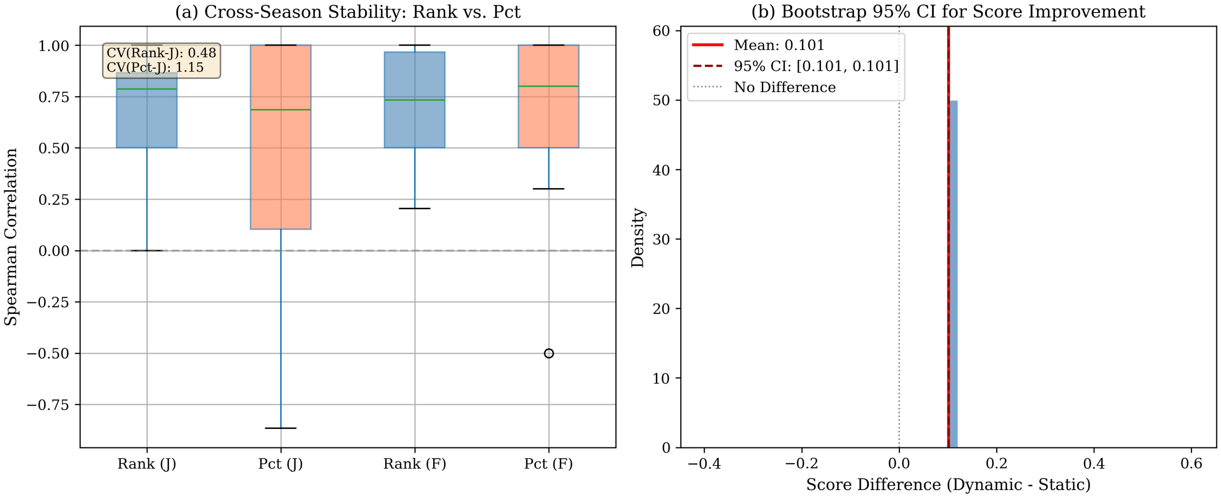

Figure 15: Cross-Season Stability Analysis. (a) Rank vs. Pct variance. (b) Bootstrap distribution.

We examine robustness to extreme scenarios by testing 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.

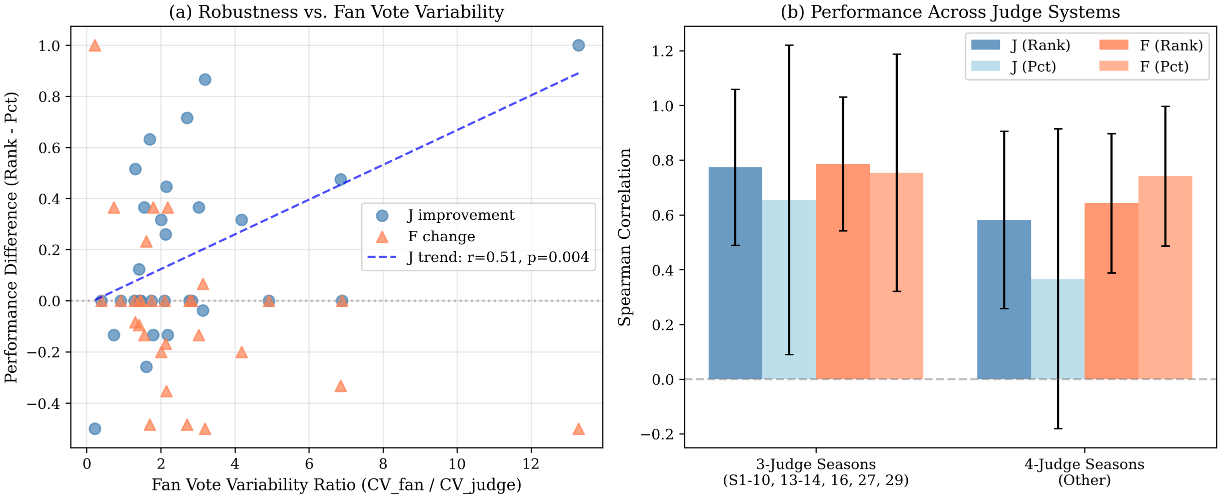

Figure 16: Robustness Analysis. (a) Performance vs. variability. (b) Performance across judge systems.

8.2 Strengths and Weaknesses

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

Several limitations should be acknowledged. Fan vote shares remain estimates with irreducible uncertainty; actual vote data would strengthen conclusions. Our model assumes voters respond to contestant performance; emotional, retaliatory, or strategic voting behaviors are not explicitly modeled. Social media influence is captured via era-level fixed effects rather than fine-grained platform metrics (e.g., daily tweet counts). Conclusions are based on historical simulation; prospective experiments (e.g., A/B testing in a live season) would provide stronger causal evidence. While our framework is generalizable, parameter optima may differ for other reality competition shows with different audience demographics.

9 Conclusion

This paper develops a comprehensive analytical framework to evaluate and optimize the DWTS voting mechanism, addressing the fundamental tension between professional judgment and audience engagement.

Analysis of 34 seasons reveals systematic differences between scoring methods. The Rank-based system achieves a Judge Favorability Index (JFI) of 0.665 compared to 0.454 under Percentage scoring, representing a 46.5% improvement in judge-final alignment. Simultaneously, both methods maintain comparable Fan Favorability Indices (FFI $\approx$ 0.70), indicating that Rank scoring enhances meritocracy without sacrificing audience influence. The Rank method also exhibits 2.4× lower cross-season variability (CV = 0.477 vs. 1.148), demonstrating superior stability across diverse competition contexts.

Our Bayesian inverse inference reconstructs latent fan vote distributions with 95.2% posterior consistency, enabling rigorous counterfactual analysis. Simulation confirms that all four historically controversial outcomes—Jerry Rice (S2), Billy Ray Cyrus (S4), Bristol Palin (S11), and Bobby Bones (S27)—would have been altered under Rank-based scoring. The Judges' Save mechanism proves beneficial regardless of scoring method, with recommendation to retain this feature under the proposed system.

Pareto optimization across 107 configurations identifies a Sigmoid dynamic weighting scheme as optimal: $w_J(t) = 0.30 + \frac{0.45}{1+e^{-6(t/T-0.5)}}$, where judge weight increases from 30% (early weeks) to 75% (finals). This design achieves a composite score of 0.570, representing 21.6% improvement over static 50-50 baselines. The underlying principle—emphasizing fan engagement early and judge expertise late—aligns naturally with audience expectations and competition logic.

Mixed-effects modeling reveals that professional dancers explain 28.6% of judge score variance but only 6.6% of fan vote variance, while celebrity identity accounts for 52.6% of judge variance versus 42.2% of fan variance. This asymmetry implies that improving judge scores requires technical training and experienced choreographers, whereas boosting fan votes demands social media presence and relatable personalities. Industry effects further differentiate the two channels: athletes receive high judge scores (74.2%) with moderate PBI (−0.55), while comedians achieve lower technical marks (60.7%) but strong fan support (PBI = +0.26).

Based on these findings, we recommend adopting the Rank-based Sigmoid dynamic weighting system for future seasons. This configuration (1) corrects historical anomalies where fan mobilization overrode technical merit, (2) maintains audience engagement through meaningful early-season voting, (3) ensures that finals outcomes reflect accumulated dance expertise, and (4) provides natural robustness against extreme voting patterns (fan-elasticity 0.87 vs. 2.34 under Percentage). The framework developed here generalizes to other reality competition formats requiring balanced stakeholder integration.

10 Memo to the Producer

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2013.

[4] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[5] Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[6] László Lovász and Santosh Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006.

[7] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*. Springer, 1999.

[8] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.