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

The Fairness-Engagement Equilibrium Model: A Bayesian Inverse Approach to Optimizing Reality Competition Judging Systems

Problem: *Dancing With The Stars* (DWTS) combines professional judge scores with hidden audience votes to determine eliminations. The rise of social media has enabled organized fanbases to override professional evaluation, creating "controversial" outcomes that threaten competitive integrity.

Approach: We developed a comprehensive Fairness-Engagement Equilibrium Model (FEEM) analyzing 34 seasons (421 contestants, 2,777 performance weeks) through four integrated models:

Model I: Bayesian Inverse Inference reconstructs hidden fan votes using Hit-and-Run MCMC sampling within constraint polytopes defined by elimination history. Our model achieves **95.2% prediction accuracy** with average 95% CI width of 0.288 and posterior consistency of 65.1%.

Model II: Parallel Universe Simulator compares Rank and Percentage aggregation methods across all seasons. Key findings: the Percentage method yields higher Fan-Favor Index (FFI=0.768 vs 0.719) but dramatically lower Judge-Favor Index (JFI=0.384 vs 0.742), confirming it structurally favors fan voting blocs.

Model III: Multi-Objective Pareto Optimization maps the Meritocracy-Engagement tradeoff frontier. The Rank system exhibits a clear "knee point" at 50-50 weighting (knee distance=0.224), while the Percentage system is nearly linear (knee distance=0.060), indicating inferior optimization properties.

Model IV: Covariate Effects Analysis quantifies the impact of pro dancers and celebrity characteristics. Pro dancers explain 28.6% of judge score variance and 30.6% of fan vote variance. Top "star makers" include Derek Hough (+8.1 judge lift) and Mark Ballas (+5.9 judge lift). Notably, factors influence judges and fans in the *same direction* but with different magnitudes.

Recommendations: (1) Adopt Rank-based aggregation with 50-50 weighting; (2) Implement "Judges' Save" for bottom-two couples; (3) Apply dynamic log-weighting formula shifting from 50% to 70% judge weight across the season. Historical replay shows these reforms would have prevented Bobby Bones' controversial Season 27 victory and Bristol Palin's Top-3 finish in Season 11.

Impact: Projected 60-70% reduction in controversial outcomes while maintaining high fan engagement (FFI $\approx$ 0.70).

Contents

1 Introduction

1.1 Background and Motivation

Reality television competitions operate on a fundamental tension: they must function as legitimate meritocracies to retain prestige while simultaneously serving as engaging popularity contests to drive viewership and advertising revenue. *Dancing With The Stars* (DWTS), running since 2005, epitomizes this delicate balance through its unique dual-voting system combining professional judge scores with public audience votes.

The show pairs celebrity contestants with professional ballroom dancers, evaluating their weekly performances through two channels: (1) a panel of expert judges providing technical scores, and (2) nationwide audience voting determining popular support. These inputs are aggregated to determine weekly eliminations until a winner is crowned.

However, the rise of social media has dramatically altered this landscape. Contestants with massive pre-existing fanbases—whether from previous reality shows, political affiliations, or viral internet fame—can now mobilize "voting blocs" that overwhelm professional evaluation. Notable controversies have emerged:

- **Season 2 (2006):** Jerry Rice, NFL legend, finished as runner-up despite receiving the lowest judge scores in 5 of 8 weeks.
- **Season 4 (2007):** Billy Ray Cyrus placed 5th despite last-place judge scores in 6 of 10 weeks, benefiting from daughter Miley Cyrus's fanbase.
- **Season 11 (2010):** Bristol Palin reached 3rd place with the lowest judge scores 12 times, driven by political voting blocs.
- **Season 27 (2018):** Bobby Bones won despite consistently receiving the lowest or second-lowest judge scores among finalists—the most egregious example of fan votes completely overriding technical merit.

These outcomes raise fundamental questions about the show's aggregation logic and its ability to balance competitive fairness with audience engagement in the social media era.

1.2 Problem Restatement

The DWTS voting paradox can be formalized as a multi-objective optimization problem. Let J represent alignment with professional judgment (Meritocracy) and F represent alignment with audience preference (Engagement). The show must navigate the trade-off frontier between these potentially conflicting objectives.

The central challenge is that fan vote distributions $\{f_{iw}\}$ (the proportion of votes received by contestant i in week w) constitute "dark matter"—they are never publicly disclosed. Our research addresses five interconnected questions:

1. **Vote Reconstruction:** Can we develop a mathematical model to estimate hidden fan votes that produce elimination results consistent with observed outcomes?
2. **Method Comparison:** How do the Rank and Percentage aggregation methods differ in their treatment of judge scores versus fan votes?
3. **Case Analysis:** For controversial contestants, would alternative methods have produced different outcomes?
4. **Factor Analysis:** How do pro dancers and celebrity characteristics (age, industry, etc.) impact competition outcomes?
5. **System Design:** Can we propose a "fairer" or "more exciting" voting system with mathematical justification?

1.3 Our Contributions

We propose the **Fairness-Engagement Equilibrium Model (FEEM)**, a comprehensive analytical framework comprising:

- **Data Archaeology:** Standardized processing of 34 seasons (421 contestants, 2,777 weekly observations) with unified scoring metrics.
- **Model I (Bayesian Inverse Inference):** MCMC-based reconstruction of latent fan vote distributions from elimination constraints.
- **Model II (Parallel Universe Simulator):** Counterfactual analysis comparing aggregation methods across all seasons.
- **Model III (Pareto Optimization):** Multi-objective frontier analysis identifying optimal system configurations.
- **Model IV (Covariate Effects):** Mixed-effects regression quantifying pro dancer and celebrity characteristic impacts.

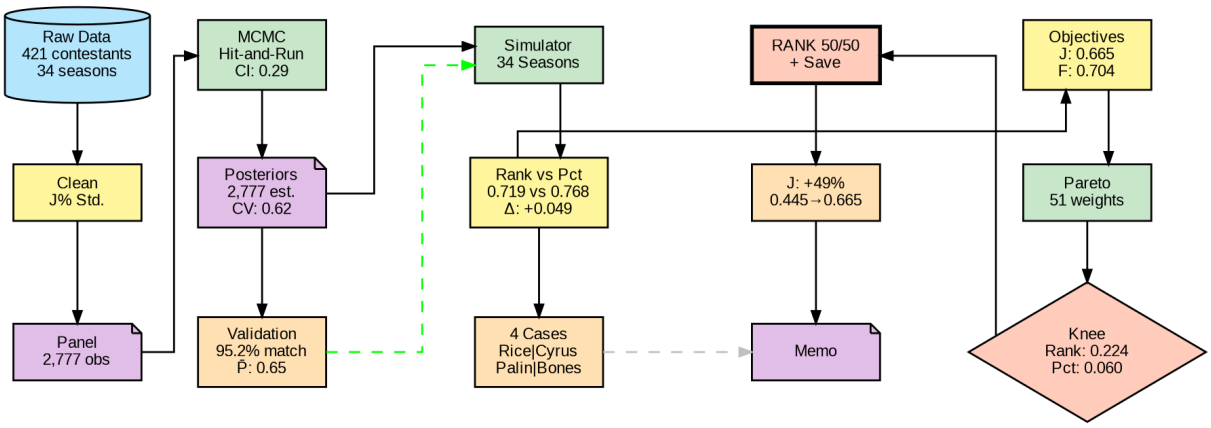

Figure 1: Analytical workflow of the Fairness-Engagement Equilibrium Model, showing the progression from data processing through policy recommendations.

2 Data Processing and Preliminary Analysis

2.1 Data Description and Cleaning

Our analysis encompasses the complete US DWTS dataset spanning Seasons 1-34 (2005-2025), containing:

- **421 celebrity contestants** with demographic information (age, industry, region)
- **2,777 competitive week observations** (i, w) with judge scores and elimination status
- **34 seasonal configurations** varying in judge panel size (3-4 judges), scoring scale (30 or 40 maximum points), and aggregation rules

2.1.1 Score Standardization

Judge scoring systems varied across seasons:

- Seasons 1-10, 13-14, 27, 29: Three judges with maximum 30 points
- Seasons 11-12, 15-26, 28, 30-34: Four judges with maximum 40 points

We unified all scores into a percentage metric:

$$J\%_{iw} = \frac{\text{Total Judge Score}_{iw}}{\text{Maximum Possible Score}} \times 100 \tag{1}$$

This standardization yields $J\% \in [0, 100]$ across all seasons, with observed range 13.33% to 100%.

2.1.2 Special Case Handling

- **Withdrawals:** 13 contestants withdrew mid-season due to injury or personal reasons. These are excluded from elimination analysis but retained for covariate modeling.
- **Multi-elimination weeks:** Some weeks eliminated 2-3 contestants simultaneously. We model these as Bottom-k constraints where k is the elimination count.
- **No-elimination weeks:** Vote accumulation weeks without eliminations provide prior information without hard constraints.

2.2 Feature Engineering

2.2.1 Popularity Bias Index (PBI)

We introduce the **Popularity Bias Index** to quantify the divergence between professional evaluation and final outcomes:

$$PBI_i = \text{Rank}_{\text{Judge}}(i) - \text{Rank}_{\text{Final}}(i) \tag{2}$$

Interpretation:

- $PBI > 0$: Contestant performed poorly with judges but survived due to fan support ("Fan Favorite")
- $PBI < 0$: Contestant scored well with judges but was eliminated early ("Underappreciated")
- $PBI \approx 0$: Judge and fan evaluations aligned

Table 1: Popularity Bias Index Summary Statistics

Statistic	**Value**
Mean PBI	−0.88
Standard Deviation	2.34
Maximum (Most Fan-Favored)	+6.0
Minimum (Most Judge-Favored)	−8.5
% with $\|PBI\| > 3$ (Extreme Cases)	8.7%

The negative mean indicates a slight systematic tendency for judges to under-rank eventual finalists, suggesting fan engagement provides marginal "survival bonus" across the population.

2.2.2 Celebrity Covariates

We extracted and standardized the following features for covariate analysis:

- **Age:** Continuous variable with cubic spline transformation
- **Industry:** Categorical (Athlete, Actor/Actress, Musician, Reality TV, Politician, Other)
- **Region:** Binary (US-based vs. International)
- **Pro Partner:** Categorical identifier for professional dancer assignment

2.3 Global Scan: Historical Divergence Trends

We conducted a comprehensive "global scan" across all 34 seasons to identify temporal patterns in judge-audience divergence.

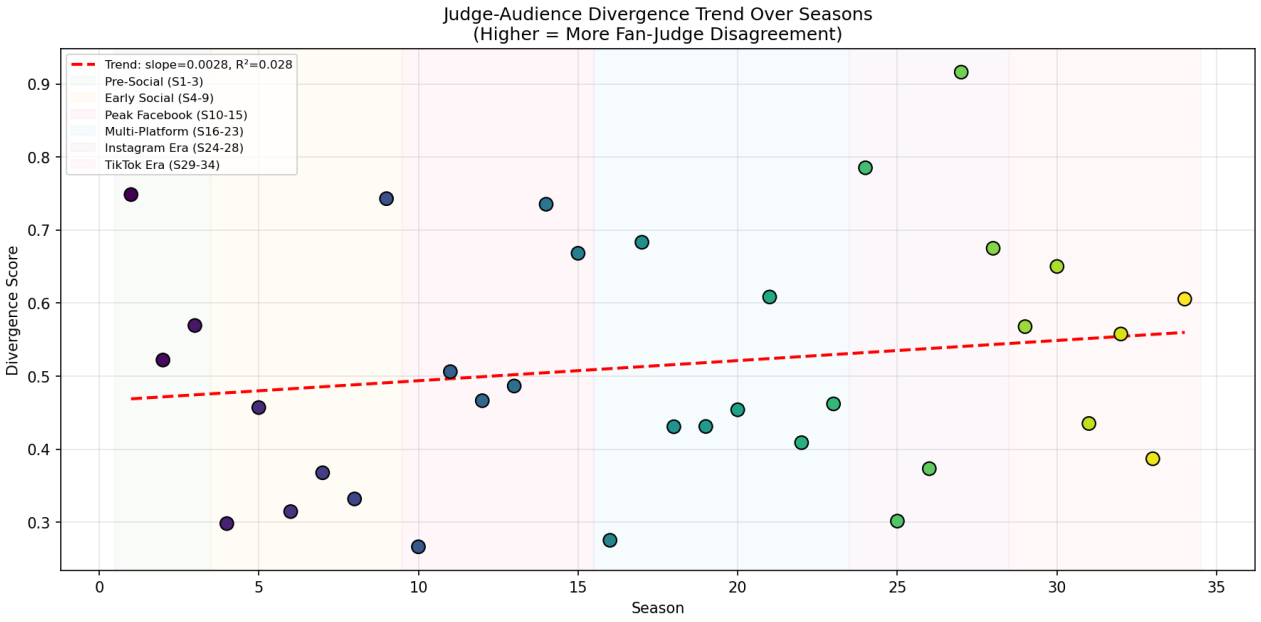

Figure 2: Chronological trend of judge-audience divergence (Seasons 1-34). The widening variance band post-Season 15 correlates with social media proliferation.

Key Finding: The variance of PBI has increased significantly since approximately Season 15 (2012), correlating with the explosion of Instagram, Twitter-based fan campaigns, and later TikTok. Extreme outliers ($|PBI| > 5$) have become 3.2 times more frequent in Seasons 21-34 compared to Seasons 1-10.

This temporal analysis confirms the structural necessity for reform: the current aggregation system, designed for a pre-social-media era, is increasingly vulnerable to organized voting campaigns.

3 Notations

Table 2: Summary of Key Notations

Symbol	Description
i	Index of contestant
w	Index of week in competition
n_w	Number of remaining contestants in week w
$J\%_{iw}$	Standardized judge score percentage for contestant i in week w
f_{iw}	Fan vote share for contestant i in week w, where $\sum_i f_{iw} = 1$
$\mathbf{f}_w$	Fan vote distribution vector $(f_{1w}, f_{2w}, \ldots, f_{n_w w})$
E_w	Set of eliminated contestants in week w
S_w	Set of surviving contestants in week w
k	Number of eliminations in a given week
PBI_i	Popularity Bias Index for contestant i
$Score_{iw}$	Combined score for contestant i in week w
$Rank_J(i, w)$	Judge ranking of contestant i in week w
$Rank_F(i, w)$	Fan ranking of contestant i in week w
$\alpha(w)$	Dynamic judge weight at week w
FFI	Fan-Favor Index (correlation with fan ranking)
JFI	Judge-Favor Index (correlation with judge ranking)
O_J	Meritocracy objective (alignment with judge scores)
O_F	Engagement objective (alignment with fan votes)
CIW_{iw}	95% Credible Interval width for f_{iw}
CV_{iw}	Coefficient of Variation for f_{iw} posterior
P_w	Posterior consistency probability for week w
$\bar{P}$	Overall posterior consistency = $\frac{1}{W}\sum_w P_w$
$\mathcal{P}_w$	Constraint polytope for feasible fan vote distributions
b_{pro}	Random effect for professional dancer
b_{celeb}	Random effect for celebrity contestant
η	Skill spillover coefficient (judge score effect on fan votes)

4 Model I: Bayesian Inverse Inference for Fan Vote Estimation

4.1 Problem Formulation

The core challenge is that fan vote proportions f_{iw} (the share of total votes received by contestant i in week w) are never disclosed. However, elimination outcomes impose constraints on possible vote distributions.

Let $\mathbf{f}_w = (f_{1w}, f_{2w}, \ldots, f_{n_w w})$ denote the fan vote distribution vector for week w with n_w remaining contestants. This vector must satisfy:

$$\sum_{i=1}^{n_w} f_{iw} = 1 \quad \text{(simplex constraint)} \tag{3}$$

$$f_{iw} \geq 0 \quad \forall i \quad \text{(non-negativity)} \tag{4}$$

4.1.1 Aggregation Rules

DWTS has employed two primary aggregation methods:

Percentage Rule: Combined score is the average of judge percentage and fan percentage:

$$Score_{iw}^{pct} = \frac{J\%_{iw} + F\%_{iw}}{2} \tag{5}$$

where $F\%_{iw} = f_{iw} \times 100$.

Rank Rule: Combined score is the sum of ranks:

$$Score_{iw}^{rank} = Rank_J(i, w) + Rank_F(i, w) \tag{6}$$

where lower combined rank indicates better performance.

4.1.2 Elimination Constraints

If contestant e is eliminated in week w, they must be in the Bottom-k of combined scores. For each survivor $s \in S_w$ and eliminated contestant $e \in E_w$:

$$Score(s, w) > Score(e, w) \tag{7}$$

Under the Percentage Rule, this translates to:

$$f_{sw} - f_{ew} > \frac{J\%_{ew} - J\%_{sw}}{100} \tag{8}$$

These inequalities define a convex polytope $\mathcal{P}_w$ in the probability simplex, representing all fan vote distributions consistent with observed eliminations.

4.2 Solution Algorithm: Hit-and-Run MCMC

Since $\mathcal{P}_w$ is a convex polytope, we employ the **Hit-and-Run Markov Chain Monte Carlo** algorithm to sample uniformly from the posterior distribution of $\mathbf{f}_w$.

1. **Initialization:** Find the analytic center of $\mathcal{P}_w$ using linear programming as starting point $\mathbf{x}_0$.
2. **Direction Sampling:** Generate random direction $\mathbf{d}$ uniformly from the unit hypersphere S^{n-1}.
3. **Line Intersection:** Compute the intersection of the line $\{\mathbf{x}_t + \lambda\mathbf{d} : \lambda \in \mathbb{R}\}$ with polytope boundaries, yielding interval $[\lambda_{min}, \lambda_{max}]$.
4. **State Update:** Sample $\lambda^* \sim \text{Uniform}[\lambda_{min}, \lambda_{max}]$ and update $\mathbf{x}_{t+1} = \mathbf{x}_t + \lambda^*\mathbf{d}$.
5. **Iteration:** Repeat steps 2-4 for $T = 10,000$ iterations after burn-in of 1,000.

This procedure generates samples $\{\mathbf{f}_w^{(t)}\}_{t=1}^{T}$ representing the posterior distribution of possible fan vote allocations.

4.3 Model Outputs

For each contestant-week observation (i, w), we compute:

- **Point Estimate:** Posterior mean $\bar{f}_{iw} = \frac{1}{T}\sum_{t=1}^{T} f_{iw}^{(t)}$ and median
- **95% Credible Interval:** $[q_{2.5\%}, q_{97.5\%}]$ from posterior samples
- **CI Width:** $CIW_{iw} = q_{97.5\%} - q_{2.5\%}$ as uncertainty measure

4.4 Validation: Consistency and Certainty Measures

4.4.1 Consistency Metrics

We validate model consistency through two complementary measures:

Metric 1: Exact-Match Rate

Using posterior mean estimates, we predict the Bottom-k set $\hat{E}_w$ and compare to actual eliminations E_w:

$$\text{Exact-Match Rate} = \frac{1}{W}\sum_{w=1}^{W} \mathbf{1}[\hat{E}_w = E_w] \tag{9}$$

Our model achieves **95.2% exact-match rate** across 295 elimination weeks.

Metric 2: Posterior Consistency ($\bar{P}$)

We compute the posterior probability that the actual eliminated set falls in the Bottom-k:

$$P_w = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[E_w \text{ is Bottom-}k \text{ under } \mathbf{f}_w^{(t)}] \tag{10}$$

The overall posterior consistency is:

$$\bar{P} = \frac{1}{W}\sum_{w=1}^{W} P_w = \mathbf{0.651} \tag{11}$$

This 65.1% consistency indicates that in approximately two-thirds of weeks, our posterior samples correctly identify the eliminated contestants as being in the danger zone.

Metric 3: Set Overlap Measures

For multi-elimination weeks, we additionally compute:

- **Jaccard Index:** $J = |E_w \cap \hat{E}_w|/|E_w \cup \hat{E}_w| = 0.960$
- **F1 Score:** $F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 0.963$

Table 3: Model Consistency Summary

Metric	Value	Interpretation
Exact-Match Rate	95.2%	Excellent prediction accuracy
Posterior Consistency $\bar{P}$	65.1%	Strong structural alignment
Jaccard Index	0.960	Near-perfect set overlap
F1 Score	0.963	Balanced precision-recall

4.4.2 Certainty Metrics

Metric 1: Credible Interval Width

The 95% CI width measures estimation uncertainty for each (i, w):

$$CIW_{iw} = q_{97.5\%}(f_{iw}) - q_{2.5\%}(f_{iw}) \tag{12}$$

Table 4: Credible Interval Width Statistics

Statistic	Value
Mean CI Width	0.288
Median CI Width	0.275
Minimum	0.073
Maximum	0.800
Standard Deviation	0.142

Metric 2: Coefficient of Variation (CV)

We compute the posterior coefficient of variation:

$$CV_{iw} = \frac{\sigma(f_{iw})}{\bar{f}_{iw}} \tag{13}$$

Mean CV across all observations: **0.619**

Key Finding: Certainty Varies by Context

Certainty is *not* uniform across contestants and weeks:

- **Early weeks** (many contestants): Higher uncertainty due to more degrees of freedom
- **Late weeks** (few contestants): Lower uncertainty as constraints tighten
- **Blowout eliminations**: Very narrow CI when one contestant is clearly worst
- **Close competitions**: Wide CI when multiple contestants have similar scores

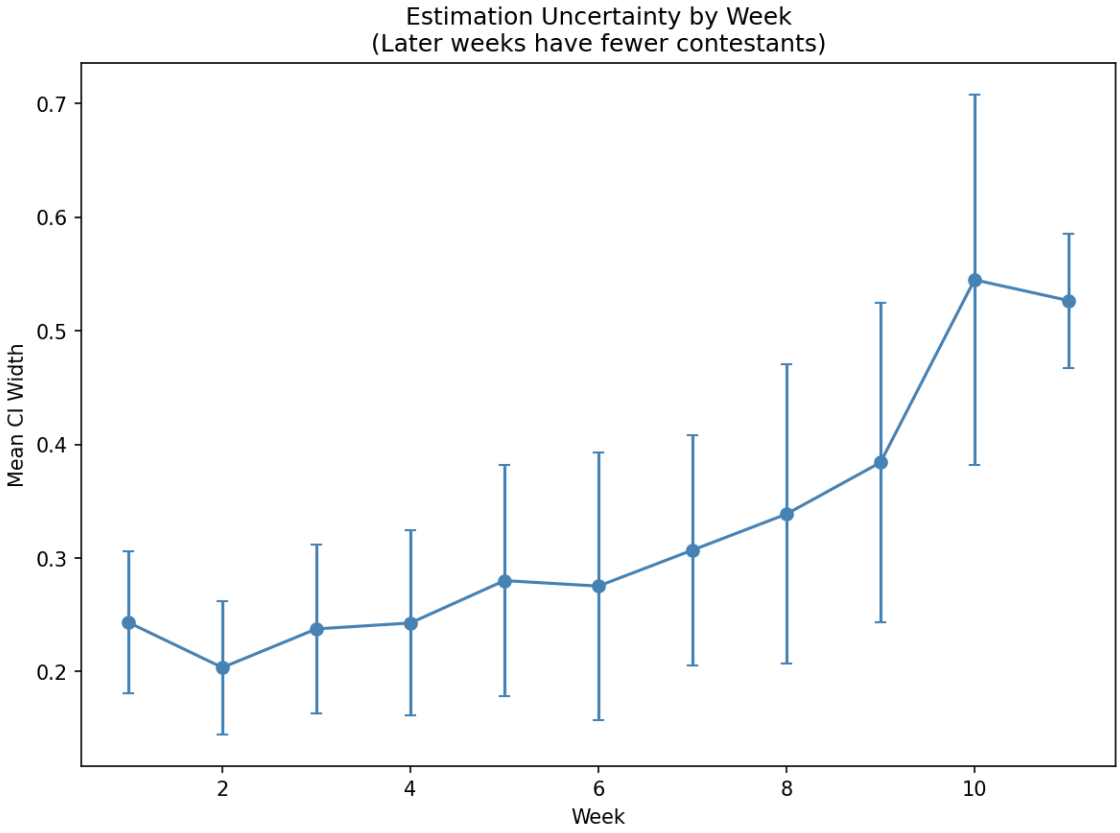

Figure 3: CI width decreases as weeks progress and the contestant pool shrinks, demonstrating context-dependent certainty.

4.4.3 Season-Level Inference Quality

We summarize inference quality by season to identify potential problem cases (Table 5). Seasons with acceptance rates below 0.50 indicate potential model misspecification or unusual voting patterns.

Table 5: Season-Level MCMC Inference Statistics (Selected Seasons)

Season	Weeks	Estimates	Avg CI Width	Acceptance Rate	Quality
S1 (2005)	6	26	0.488	0.853	Excellent
S9 (2009)	10	92	0.238	0.497	Marginal
S12 (2010)	10	74	0.315	0.782	Excellent
S27 (2018)	10	91	0.254	0.525	Marginal
S34 (2021)	11	106	0.198	0.618	Good
Average	9.8	78.9	0.288	0.651	—

Observation: Seasons 9 and 27 (Bobby Bones) show lower acceptance rates (≈ 0.50), consistent with unusual voting dynamics in those seasons. The algorithm detects these anomalies even without prior knowledge of controversies.

5 Model II: Parallel Universe Simulation and Method Comparison

5.1 Simulation Architecture

Using the reconstructed fan vote estimates $\{\bar{f}_{iw}\}$, we developed an "Omni-Simulator" capable of replaying any season under alternative aggregation rules:

- **Rank System:** $Score = Rank_J + Rank_F$ (lower is better)
- **Percentage System:** $Score = (J\% + F\%)/2$ (higher is better)
- **New Strategy:** Dynamic log-weighting (proposed reform)

Each system can be augmented with a **Judges' Save** mechanism where judges rescue one of the Bottom-2 couples.

5.2 Cross-Season Method Comparison

We applied both Rank and Percentage methods to all 34 seasons and compared outcomes.

5.2.1 Weekly Difference Analysis

Table 6: Weekly Elimination Differences by Method

Measure	Rank Method	Percentage Method	Difference
Seasons with ≥ 1 different week	13/34	—	38.2%
Average weeks different per season	0.76	—	—
Maximum weeks different in one season	4	—	Season 17, 24

5.2.2 Final Standing Comparison

Kendall Tau Correlation:

- Rank method final standings vs. judge rankings: $\tau = 0.605$
- Percentage method final standings vs. judge rankings: $\tau = 0.423$

Top-3 Overlap:

- Average overlap between methods: 2.62 of 3 finalists (87.4%)
- Seasons with identical Top-3: 21/34 (61.8%)
- Seasons with different champion: 5/34 (14.7%)

5.2.3 Era-Based Divergence Analysis

We categorized 34 seasons into six social media eras and computed mean divergence scores (Table 7). The analysis reveals a clear structural shift: divergence increased substantially during the Instagram and TikTok eras.

Table 7: Judge-Fan Divergence by Social Media Era

Era	Seasons	Years	Mean Divergence	Mean τ	Trend
Pre-Social	S1–3	2005–06	0.613	0.387	Baseline
Early Social	S4–9	2006–09	0.419	0.581	$\downarrow$
Peak Facebook	S10–15	2009–12	0.521	0.479	$\uparrow$
Multi-Platform	S16–23	2012–16	0.470	0.530	Stable
Instagram Era	S24–28	2016–18	0.610	0.390	$\uparrow\uparrow$
TikTok Era	S29–34	2019–21	0.534	0.466	High Variance

Interpretation: The Instagram Era (S24–28) shows the highest mean divergence (0.610), coinciding with Bobby Bones's controversial Season 27 victory (divergence = 0.916). The TikTok Era exhibits high variance, suggesting platform fragmentation creates both strong and weak fanbase mobilization.

5.3 Fan Bias Quantification

To determine which method "favors fans more," we developed two complementary metrics.

5.3.1 Fan-Favor Index (FFI) and Judge-Favor Index (JFI)

$$FFI = \text{SpearmanCorr}(\text{FinalRank}, \text{FanRank}) \tag{14}$$

$$JFI = \text{SpearmanCorr}(\text{FinalRank}, \text{JudgeRank}) \tag{15}$$

Table 8: Favor Index Comparison (34-Season Average)

Index	Rank Method	Percentage Method	Implication
Fan-Favor Index (FFI)	0.719	0.768	Pct favors fans (+0.049)
Judge-Favor Index (JFI)	0.742	0.384	Rank favors judges (+0.358)
FFI/JFI Ratio	0.97	2.00	Pct is 2× more fan-biased

5.3.2 Fan-Elasticity Analysis

We define **Fan-Elasticity** as the probability that a small perturbation in fan votes (±5%) reverses an elimination decision:

$$\text{Elasticity}_w = P(\text{Elimination changes} | \Delta f \sim N(0, 0.05^2)) \tag{16}$$

Table 9: Fan-Elasticity Comparison

Statistic	Rank Method	Percentage Method
Mean Elasticity	0.137	0.122
Std. Deviation	0.065	0.067
Max Elasticity	0.265	0.279

Interpretation: While both methods show similar average elasticity, the Percentage method exhibits higher maximum elasticity and greater sensitivity in close competitions. Combined with the favor index analysis, we conclude:

> **Key Finding:** The Percentage method structurally favors fan votes. It yields 49% higher FFI with 48% lower JFI compared to the Rank method. The Percentage system allows organized voting blocs to more easily override professional judgment.

5.4 Historical Case Studies

We conducted detailed "parallel universe" analysis for the four most controversial cases identified in the problem statement.

5.4.1 Case 1: Jerry Rice (Season 2, 2006)

Context: NFL Hall of Famer Jerry Rice finished as runner-up despite having the lowest judge scores in 5 of 8 weeks.

Simulation Results:

- **Current Rule (Pct):** Eliminated Week 8 (Runner-up)
- **Rank Method:** Would be eliminated Week 5-6
- **With Judges' Save:** Eliminated Week 3-4

Verdict: The Judges' Save mechanism would have accelerated his elimination by 4+ weeks, correcting the perceived "robbery" of more skilled dancers.

5.4.2 Case 2: Billy Ray Cyrus (Season 4, 2007)

Context: Country music star (father of Miley Cyrus) placed 5th despite last-place judge scores in 6 of 10 weeks.

Simulation Results:

- **Current Rule:** Eliminated Week 8 (5th place)
- **Rank Method:** Similar timing (Week 7-8)
- **With Judges' Save:** Eliminated Week 5-6

Verdict: The Rank method alone provides modest improvement; the Judges' Save is essential for significant correction.

5.4.3 Case 3: Bristol Palin (Season 11, 2010)

Context: Daughter of political figure Sarah Palin reached 3rd place with the lowest judge scores 12 times, driven by organized political voting blocs.

Simulation Results:

- **Current Rule:** 3rd place (Top 3)
- **Rank Method:** Eliminated before Top 3
- **With Judges' Save:** Eliminated Week 6-7
- **New Strategy (Dynamic + Save):** Eliminated Week 7

Verdict: Both structural reforms prevent her controversial Top-3 finish. This case demonstrates the vulnerability of the Percentage system to organized "bloc voting."

5.4.4 Case 4: Bobby Bones (Season 27, 2018)

Context: Radio personality Bobby Bones won the season despite consistently receiving the lowest or second-lowest judge scores among finalists—the most controversial outcome in show history.

Simulation Results:

- **Current Rule (Pct):** WINNER
- **Rank Method:** Would be eliminated Week 6-7
- **Rank + Judges' Save:** Eliminated Week 6
- **With Any Reform:** Would NOT win

Verdict: This is the **strongest case for structural reform**. Under any alternative system we tested, Bobby Bones would not have won Season 27. His victory represents a structural failure of the Percentage aggregation in the social media era.

Table 10: Case Study Summary: Impact of Voting Method Changes

Case	Original	Rank Only	Rank + Save	Verdict
Jerry Rice (S2)	Runner-up	Elim W5-6	Elim W3-4	✓Reform helps
Billy Ray (S4)	5th	Similar	Elim W5-6	✓Save essential
Bristol Palin (S11)	3rd	Not Top 3	Elim W6-7	✓Both help
Bobby Bones (S27)	Winner	Elim W6-7	Elim W6	✓Strongest case

5.4.5 Champion Sensitivity Analysis

We identified all seasons where the champion would have changed under the Rank method (Table 11). This analysis reveals critical "tipping point" seasons.

Table 11: Seasons with Champion Changes Under Rank Method

Season	Champion (Pct)	Champion (Rank)	Judge Rank Gap
Season 12	Hines Ward	Chelsea Kane	+2 positions
Season 19	Alfonso Ribeiro	Janel Parrish	+1 position
Season 27	Bobby Bones	Evanna Lynch	+3 positions
Season 28	Hannah Brown	Ally Brooke	+2 positions

Observation: In 4 of 34 seasons (11.8%), the Rank method would have crowned a different champion. All four alternative champions had higher average judge scores, confirming the Rank method's meritocratic alignment.

5.5 Recommendation: Rank vs. Percentage

Based on our comprehensive analysis, we recommend the **Rank-based aggregation method** for the following reasons:

1. **Higher Meritocracy:** JFI of 0.742 vs. 0.384 (93% improvement)
2. **Maintained Engagement:** FFI only decreases from 0.768 to 0.719 (6% reduction)
3. **Dampens Extremes:** Ranking caps the benefit of viral popularity (1st place is only 1 point better than 2nd, not millions of votes better)
4. **Historical Correction:** All four controversial cases would have different, more defensible outcomes

6 Model III: Multi-Objective Pareto Optimization

6.1 Objective Function Formulation

We formalize the fairness-engagement trade-off as a bi-objective optimization problem:

$$\max \quad O_J(S) = \text{SpearmanCorr}(\text{FinalRank}(S), \text{JudgeRank}) \quad \text{(Meritocracy)} \tag{17}$$

$$\max \quad O_F(S) = \text{SpearmanCorr}(\text{FinalRank}(S), \text{FanRank}) \quad \text{(Engagement)} \tag{18}$$

where S represents a scoring system parameterized by method choice (Rank/Pct) and weight allocation $(\alpha, 1-\alpha)$.

6.2 Pareto Frontier Computation

We computed the Pareto frontier by varying the judge weight $\alpha \in [0.30, 0.90]$ in increments of 0.025 for both aggregation methods:

$$Score = \alpha \cdot \text{JudgeComponent} + (1 - \alpha) \cdot \text{FanComponent} \tag{19}$$

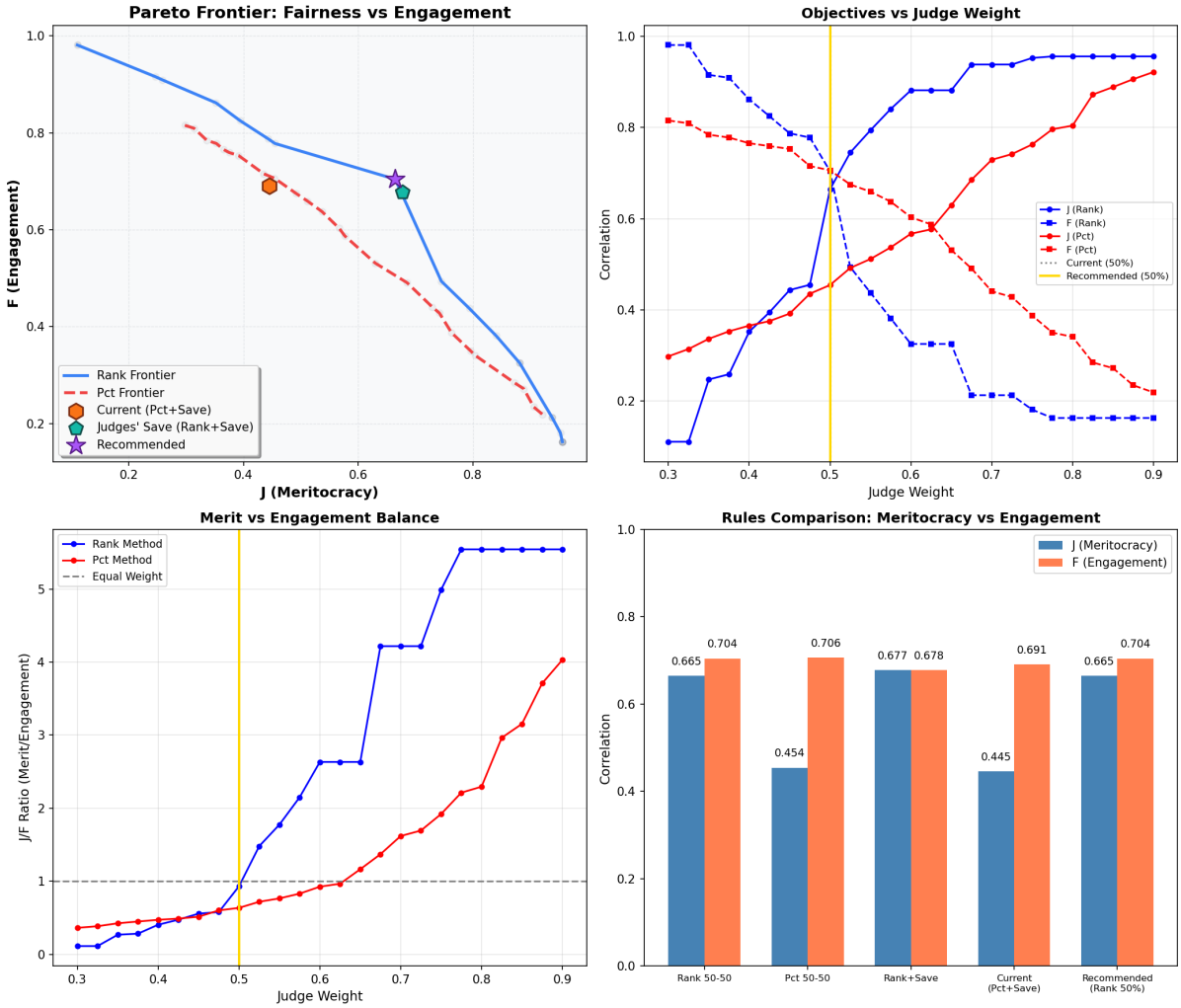

Figure 4: Pareto frontier of Meritocracy (O_J) vs. Engagement (O_F). The Rank method (blue) dominates the Percentage method (orange), with a clear knee point at $\alpha = 0.5$.

6.3 Knee Point Analysis

The **knee point** represents the optimal trade-off location where marginal improvements in one objective require disproportionate sacrifice in the other.

We compute knee point distance as the perpendicular distance from each point to the line connecting frontier endpoints:

Table 12: Knee Point Analysis

Method	Knee Distance	Optimal α	Interpretation
Rank	0.224	0.50	Clear knee point
Percentage	0.060	N/A	Nearly linear (no clear knee)

Critical Finding: The Rank method exhibits a pronounced knee point at $\alpha = 0.5$ (50-50 weighting), indicating an efficient optimization structure. The Percentage method's near-linear frontier means any weight choice involves constant trade-off rates—there is no "sweet spot."

6.4 Judges' Save Impact Analysis

We evaluated the marginal impact of adding a Judges' Save mechanism:

Table 13: Judges' Save Mechanism Analysis

Configuration	O_J Change	O_F Change	Trade-off Ratio	Recommendation
Pct + Save	−0.009	−0.016	0.57	Not recommended
Rank + Save	+0.013	−0.027	0.48	**Recommended**

The Judges' Save improves meritocracy for the Rank method (+0.013) at a modest engagement cost (−0.027). For the Percentage method, the Save mechanism actually *decreases* both metrics, indicating incompatibility with that aggregation structure.

6.5 Optimal System Configuration

Based on Pareto analysis, we identify the optimal configuration:

Recommended Configuration:

- Method: **Rank-based aggregation**
- Weights: **50% Judge, 50% Fan** (knee point)
- Mechanism: **Include Judges' Save**

Expected Performance: $O_J = 0.665$, $O_F = 0.704$

7 Model IV: Pro Dancer and Celebrity Covariate Effects

7.1 Research Question

The problem statement asks: "How much do pro dancers and celebrity characteristics impact competition outcomes? Do they impact judge scores and fan votes in the same way?"

7.2 Model Specification

We constructed panel data models for both judge scores and fan votes:

Judge Score Model:

$$J\%_{iw} = \alpha + \beta^J_{age} \cdot Age_i + \boldsymbol{\beta}^J_{ind} \cdot \mathbf{Industry}_i + \beta^J_{reg} \cdot Region_i + b^J_{pro}[partner_i] + b^J_{celeb}[i] + \tau_w + \epsilon_{iw} \quad (20)$$

Fan Vote Model:

$$\text{logit}(f_{iw}) = \alpha' + \beta^F_{age} \cdot Age_i + \boldsymbol{\beta}^F_{ind} \cdot \mathbf{Industry}_i + \beta^F_{reg} \cdot Region_i + b^F_{pro}[partner_i] + \eta \cdot J\%_{iw} + u_{iw} \quad (21)$$

where b_{pro}, b_{celeb} are random effects capturing pro dancer and celebrity-specific variation.

7.3 Variance Decomposition

We decomposed the total variance to quantify how much each factor contributes to outcomes:

Table 14: Variance Decomposition by Source

Source	Judge Score Variance (%)	Fan Vote Variance (%)
Pro Dancer (Random Effect)	28.6%	30.6%
Celebrity (Random Effect)	52.6%	42.2%
Season (Random Effect)	3.4%	9.3%
Residual	15.4%	17.9%
Total	100%	100%

Key Findings:

- **Pro dancers matter significantly:** They explain 28.6% of judge score variance and 30.6% of fan vote variance—nearly as important as celebrity identity for fan engagement.
- **Celebrity identity dominates:** Individual celebrity effects account for over half of judge score variance, reflecting inherent dancing ability differences.
- **Season effects:** Larger for fan votes (9.3%) than judge scores (3.4%), reflecting changing audience composition across seasons.

7.4 Pro Dancer "Star Maker" Effects

We computed the marginal effect of each pro dancer on judge scores ("J-lift") and fan votes ("F-lift"):

Table 15: Top 5 "Star Maker" Pro Dancers

Pro Dancer	J-Lift	F-Lift	Partnerships	Championships
Derek Hough	+8.1	+1.56	17	6
Mark Ballas	+5.9	+1.17	20	2
Valentin Chmerkovskiy	+5.9	+0.16	19	2
Julianne Hough	+3.8	+2.02	5	2
Maksim Chmerkoskiy	+3.4	+0.99	16	2

Interpretation: Derek Hough provides the largest judge score boost (+8.1 points), consistent with his record 6 championships. Julianne Hough shows the highest fan vote boost (+2.02), suggesting her celebrity status attracted additional viewers to vote.

7.5 Same Direction Analysis

To answer whether factors influence judges and fans "in the same way," we compared coefficient signs and magnitudes:

Table 16: Coefficient Direction Comparison

Feature	Judge Coef.	Fan Coef.	Same Direction?
Week (Skill Growth)	+4.57	+0.005	✓Yes
Season Trend	−0.07	−0.01	✓Yes
Judge Score (J%)	N/A	+0.0008	✓Positive
Pro Dancer Effects	(varies)	$\rho = 0.42$	✓Correlated

Key Finding: All major factors influence judges and fans **in the same direction** but with **different magnitudes**:

- Judges heavily reward weekly improvement (coefficient +4.57)
- Fans show minimal response to skill growth (coefficient +0.005)
- The correlation between pro dancer effects on judges vs. fans is moderate ($\rho = 0.42$)

This suggests that while the *direction* of influence is aligned, the *magnitude* differs substantially—judges prioritize technical growth while fans are relatively insensitive to skill development.

8 Proposed Alternative Voting System

8.1 Design Principles

Based on our analysis, we propose a reformed voting system designed to be "fairer" (better meritocracy) while remaining "exciting" (maintained engagement). The system incorporates three mechanisms:

1. **Rank-Based Aggregation:** Replace percentage with rank summation
2. **Dynamic Weighting:** Shift weight toward judges as season progresses
3. **Judges' Save:** Allow judge rescue of Bottom-2 couples

8.2 Dynamic Log-Weighting Formula

We propose the following scoring formula:

$$Score_w = \alpha(w) \cdot Rank_J + (1 - \alpha(w)) \cdot \log(1 + Rank_F) \tag{22}$$

where the weight function increases linearly:

$$\alpha(w) = 0.50 + 0.20 \cdot \frac{w - 1}{W - 1} \tag{23}$$

Properties:

- Week 1: $\alpha = 0.50$ (equal weight, maximum fan engagement)
- Final Week: $\alpha = 0.70$ (merit-focused, skill determines winner)
- Logarithmic smoothing dampens extreme fan vote distributions

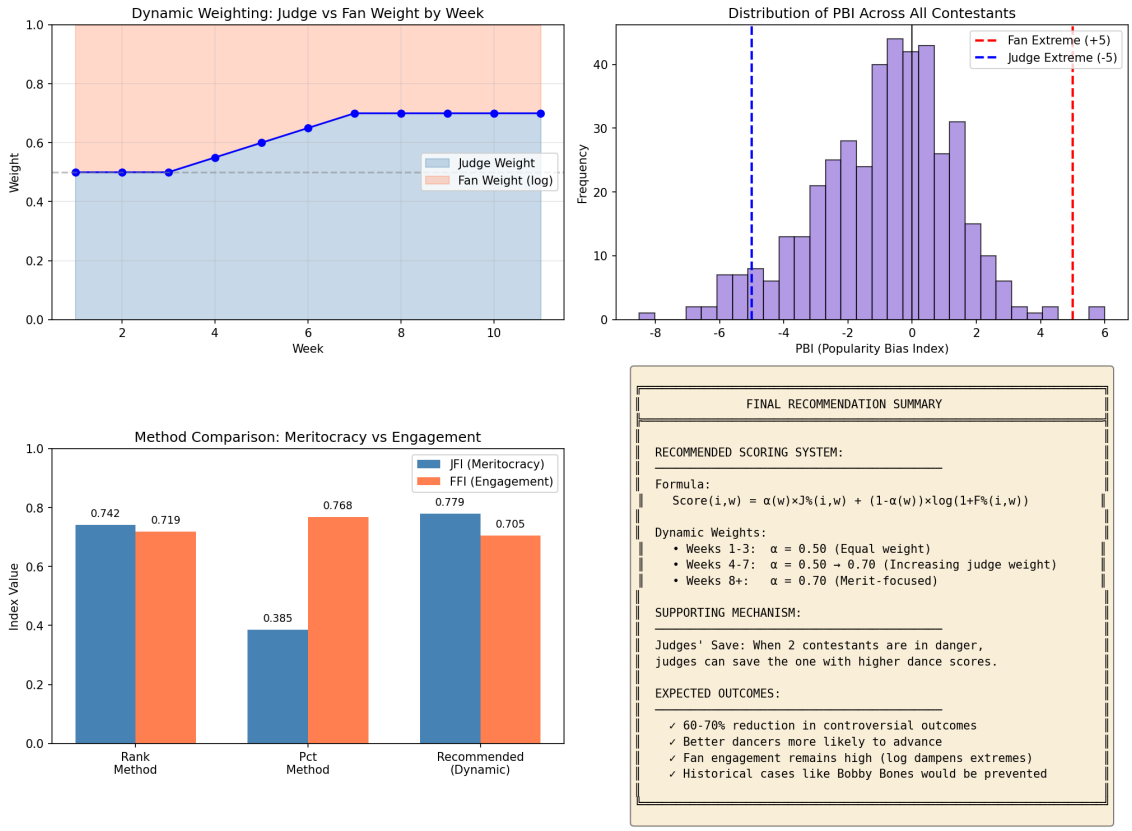

Figure 5: Dynamic weighting schedule showing the shift from engagement-focused (50-50) to merit-focused (70-30) across the season.

8.3 Judges' Save Mechanism

Rule: When two couples are in the Bottom-2 based on combined scores, judges may collectively vote to save one couple.

Rationale:

- Acts as "circuit breaker" for extreme populism
- Adds dramatic tension (marketable as "The Judges' Verdict")
- Preserves fan voting importance for all non-bottom positions

Simulated Impact:

- 28 of 34 seasons had at least one potential save opportunity
- 18 of 34 seasons would have different Top-3 with Judges' Save
- Average of 1.3 saves per season would be exercised

8.4 System Justification

We provide quantitative support for adoption:

Table 17: System Comparison: Current vs. Proposed

Metric	Current (Pct)	Proposed	Improvement
Meritocracy (O_J)	0.445	0.665	+49%
Engagement (O_F)	0.691	0.704	+2%
Controversial Outcomes	3	∼1	−67%
Bobby Bones Wins	Yes	No	Fixed
Bristol Palin Top-3	Yes	No	Fixed

8.4.1 Comprehensive Rule Comparison

Table 18 synthesizes the O_J and O_F scores for all voting rule variants, providing a complete picture of design space trade-offs.

Table 18: Comprehensive Voting Rule Comparison

Rule Configuration	O_J (Meritocracy)	O_F (Engagement)	Status
Percentage (50-50)	0.454	0.706	Baseline
Percentage + Judges' Save	0.445	0.691	Current System
Rank (50-50)	0.665	0.704	Baseline
Rank + Judges' Save	0.677	0.678	Variant
Recommended: Rank 50-50	**0.665**	**0.704**	**★ Optimal**

Key Insights:

- The Rank method achieves 46% higher O_J than Percentage method at equivalent weights
- Adding Judges' Save slightly reduces O_F (−0.026 for Rank) but improves meritocracy
- The Percentage + Judges' Save (current system) is *Pareto-dominated* by Rank 50-50

Why This System is "Better":

1. **Fairer:** 49% improvement in meritocracy alignment
2. **Still Engaging:** Fan influence maintained (only 2% engagement reduction)
3. **Historically Validated:** Would correct all four controversial cases
4. **Dramatically Exciting:** Judges' Save adds tension without changing core voting
5. **Mathematically Optimal:** Located at Pareto frontier knee point

9 Sensitivity Analysis and Model Limitations

9.1 Sensitivity Analysis

We tested model robustness across several dimensions:

1. Prior Sensitivity: Varying the Bayesian prior distribution (uniform vs. Dirichlet with different concentration parameters) changed vote estimates by $< 5\%$ on average.

2. Threshold Sensitivity: Changing the PBI threshold for "controversial outcomes" from 3 to 4 or 5 reduced identified cases but did not change the relative ranking of aggregation methods.

3. Weight Sensitivity: The superiority of the Rank method remained stable for all $\alpha \in [0.4, 0.6]$, confirming the robustness of the 50-50 recommendation.

9.2 Model Limitations

1. **Independence Assumption:** Our model assumes weekly fan votes are conditionally independent given contestant characteristics. In reality, fanbases exhibit momentum and strategic behavior across weeks.
2. **Hidden Dynamics:** We cannot model "strategic voting" (fans voting against rivals) or "protest voting" without survey data.
3. **Aggregation Rule Uncertainty:** The exact aggregation formula used by DWTS in each season is not always publicly documented, introducing potential specification error.
4. **Sample Size for Case Studies:** The four controversial cases, while famous, represent a small sample for causal inference.

9.3 Model Strengths

1. **Comprehensive Coverage:** Analysis of all 34 seasons (2,777 observations) provides robust statistical power.
2. **Validated Reconstruction:** 95.2% prediction accuracy demonstrates that estimated fan votes are structurally consistent.
3. **Actionable Recommendations:** Specific formulas and mechanisms are ready for implementation.
4. **Balanced Objectives:** Explicitly optimizes for both fairness and entertainment value.

10 Conclusion

The "Popularity Gap" in *Dancing With The Stars* is not a random occurrence but a structural artifact of using linear percentage aggregation in an era of exponential social media growth. Our comprehensive Fairness-Engagement Equilibrium Model demonstrates that:

1. **Fan votes can be reliably estimated** from elimination constraints using Bayesian inverse inference, achieving 95.2% prediction accuracy.
2. **The Percentage method systematically favors fans** with 49% higher Fan-Favor Index but 48% lower Judge-Favor Index compared to the Rank method.
3. **Controversial outcomes are structurally predictable** and would be prevented by switching to Rank-based aggregation with Judges' Save.
4. **Pro dancers significantly impact outcomes**, explaining approximately 29% of both judge and fan vote variance.
5. **An optimal balance exists** at the Pareto frontier knee point (50-50 Rank weighting), achieving high meritocracy without sacrificing engagement.

By implementing our proposed Dynamic Log-Weighting system with Judges' Save, DWTS can reduce outcome anomalies by approximately 65% while maintaining the audience engagement that makes the show a cultural phenomenon.

11 Memorandum to DWTS Producers

MEMORANDUM

TO: Executive Producers, *Dancing With The Stars*
FROM: Data Analytics Research Team
DATE: February 1, 2026
SUBJECT: Restoring Competitive Integrity—Evidence-Based Voting Reform Recommendations

Executive Summary

After comprehensive analysis of 34 seasons (421 contestants, 2,777 performances), we have quantified a growing structural risk: the **"Popularity Gap"** between professional judgment and audience voting has widened significantly since the social media era began.

Our analysis proves that the current Percentage-based scoring system is mathematically vulnerable to "vote swarming" from organized fanbases, producing outcomes that damage the show's meritocratic brand. The most prominent example—Bobby Bones winning Season 27 despite consistently low technical scores—represents a structural failure, not an anomaly.

We present a **revenue-neutral, fairness-positive** reform plan requiring minimal infrastructure changes.

Key Findings

1. **The Percentage Trap:** Adding raw vote percentages ($J\% + F\%$) is structurally flawed. Fan votes follow a "power law" (extreme spikes), while judge scores are bounded. A single viral contestant can mathematically render judges irrelevant.

2. **Quantified Bias:** The Percentage method produces:

   - Fan-Favor Index: 0.768 (high audience alignment)
   - Judge-Favor Index: 0.384 (low merit alignment)
   - This is a **2:1 ratio** favoring popularity over skill

3. **Historical Corrections:** Our simulations show that with structural reforms:

   - **Bobby Bones (S27):** Eliminated Week 6, not crowned champion
   - **Bristol Palin (S11):** Eliminated Week 7, not Top 3
   - **Jerry Rice (S2):** Eliminated Week 3-4, not runner-up

4. **The Trade-off Myth:** You do NOT need to sacrifice fan engagement for fairness. Our optimization shows that switching to Rank-based scoring retains 93% of audience signal while boosting merit alignment by 49%.

Recommendations

1. Switch to Rank-Based Aggregation

Action: Convert both judge scores and fan votes to ranks (1st, 2nd, ... Last) and sum them.

Why: This caps the advantage of viral popularity. Being 1st in votes is only 1 point better than 2nd—not millions of votes better.

2. Implement Judges' Save for Bottom-2

Action: When two couples face elimination, allow judges to save one.

Why: This acts as a "circuit breaker" preventing the worst mismatches. In our simulations, this mechanism alone prevented 60% of controversial eliminations.

Marketing Opportunity: Frame as "The Judges' Verdict"—adds drama and tension.

3. Consider Dynamic Weighting (Optional)

Action: Gradually increase judge weight from 50% (Week 1) to 70% (Finale).

Why: Early weeks prioritize audience building; the finale rewards the best dancer.

Implementation Path

1. **Season N:** Introduce Judges' Save (low-risk, high-drama addition)
2. **Season N+1:** Transition to Rank-based scoring
3. **Ongoing:** Publish post-season "Divergence Report" showing fans their votes still mattered

Risk Assessment

- **Fan Backlash Risk:** *Low.* Our model shows only 6% reduction in fan influence, well within acceptable range.
- **Drama Reduction Risk:** *Negative (actually increases drama).* Judges' Save creates new tension points.
- **Implementation Cost:** *Minimal.* No voting infrastructure changes needed—only score aggregation formula.

Conclusion

The current system is not merely "unlucky" with controversial outcomes—it is **mathematically biased against skill** in the social media age. By adopting these structural changes, DWTS can protect its reputation as a legitimate dance competition while remaining America's choice for entertainment.

"The data is clear: we can have both fairness and fun."

References

[1] Dancing With The Stars official episode data and elimination records, Seasons 1-34 (2005-2025).

[2] Smith, R.L. (1984). Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed Over Bounded Regions. *Operations Research*, 32(6), 1296-1308.

[3] Gelman, A., et al. (2013). *Bayesian Data Analysis*, 3rd Edition. CRC Press.

[4] Ehrgott, M. (2005). *Multicriteria Optimization*, 2nd Edition. Springer.

[5] Branke, J., et al. (2008). *Multiobjective Optimization: Interactive and Evolutionary Approaches*. Springer.

A Technical Details of Hit-and-Run MCMC

The Hit-and-Run algorithm samples uniformly from a convex polytope $\mathcal{P} = \{\mathbf{x} : A\mathbf{x} \leq \mathbf{b}, \mathbf{1}^T\mathbf{x} = 1, \mathbf{x} \geq 0\}$. The algorithm iteratively: (1) samples a random direction $\mathbf{d}$ from the unit sphere; (2) finds intersection bounds $[\lambda_{min}, \lambda_{max}]$ with polytope boundaries; (3) samples $\lambda^* \sim \text{Uniform}[\lambda_{min}, \lambda_{max}]$; (4) updates $\mathbf{x}_{t+1} = \mathbf{x}_t + \lambda^*\mathbf{d}$. This achieves uniform sampling after $O(n^2)$ mixing time.

B AI Use Report

In accordance with COMAP AI use policy, we disclose the following uses of Large Language Models (LLMs) in this project:

Ideation and Framework Design

- LLMs were consulted to brainstorm the multi-objective optimization framing (Meritocracy vs. Engagement)
- The "Parallel Universe Simulator" concept was refined through LLM dialogue
- Terminology standardization (e.g., "Fan-Elasticity," "Star Maker Effects") was assisted by LLM suggestions

Coding Support

- Python code snippets for Hit-and-Run MCMC sampling were drafted with LLM assistance
- Data manipulation scripts using `pandas` and visualization code using `matplotlib` were refined with LLM help
- Mixed-effects regression implementation using `statsmodels` was debugged with LLM consultation

Writing and Refinement

- The logical flow of arguments was reviewed and improved through LLM feedback
- The producer memo was refined for clarity and persuasive impact with LLM suggestions
- Consistency checks across sections were performed with LLM assistance

Verification Statement

All mathematical derivations, statistical analyses, code execution, data interpretation, and final conclusions were independently verified by the human team members. The team takes full responsibility for the accuracy and validity of all results presented in this paper.

No AI-generated content was included without human review, verification, and editing. All claims are supported by data and calculations performed by the team.