Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2600001

Dancing with the Stars: A Fairness-Engagement Equilibrium Model

Problem Overview: We analyze 34 seasons of *Dancing with the Stars* (DWTS) to determine whether the show's voting system—combining judges' scores with fan votes—produces fair outcomes while maintaining audience engagement.

Methodology:

- **Bayesian Inference:** We estimate latent fan vote shares $f(i, w)$ using MCMC with Bottom-k elimination constraints, achieving **95.6% prediction accuracy**.
- **Parallel Universe Simulation:** We replay all seasons under Rank-based and Percentage-based aggregation rules to compare outcomes.
- **Pareto Optimization:** We identify the optimal balance between meritocracy (Judge-Favor Index) and engagement (Fan-Favor Index).

Key Findings:

- Rank-based method is more meritocratic (JFI = 0.727); Percentage-based favors fans more (FFI = 0.788).
- Champion changed in only 3/34 seasons (8.8%) between methods; Top-3 overlap averages 2.76/3.
- We identified 2 "extreme events" where fan voting overrode judges in Top-3 placements (Bobby Bones S27, Bristol Palin S11).
- Professional partners explain 38% of judge score variance; "Star Makers" like Derek Hough provide +8.1 point lift.

Recommendation: We propose a **Dynamic Log-Weighting** formula:

$$Score = \alpha(w) \cdot J\% + (1 - \alpha(w)) \cdot \log(1 + F\%)$$

where $\alpha(w)$ increases from 50% (Weeks 1–3) to 70% (Weeks 8+), combined with a **Judges' Save** mechanism for Bottom-2 contestants.

Expected Impact: Historical replay shows Bobby Bones (S27) would not have won; Bristol Palin (S11) would be eliminated before Top 3. Estimated 60–70% reduction in controversial outcomes while maintaining FFI > 0.6.

Keywords: Bayesian inference, MCMC, Pareto optimization, fairness metrics, voting systems

Contents

1 Introduction **1**

2 Assumptions and Notations **1**
2.1 Assumptions . . . 1
2.2 Notations . . . 1

3 Data Preprocessing and Exploratory Analysis **2**
3.1 Judge Score Standardization . . . 2
3.2 Performance-Bias Index (PBI) . . . 2
3.3 Global Scan: The Social Media Effect . . . 2
3.4 Feature Engineering . . . 3

4 Bayesian Inference for Fan Vote Estimation **3**
4.1 Problem Formulation . . . 3
4.2 Model . . . 3
4.3 MCMC Sampling . . . 4
4.4 Results . . . 4

5 Simulation: Rank vs. Percentage Methods **4**
5.1 Parallel Universe Analysis . . . 4
5.2 Favor Indices . . . 4
5.3 Final Standing Analysis . . . 5

6 Case Studies **5**

7 Pareto Optimization **5**
7.1 Dual Objectives . . . 5
7.2 Pareto Frontier . . . 6
7.3 Variance Decomposition & Star Makers . . . 6

8 Recommendation **6**
8.1 Dynamic Log-Weighting Formula . . . 6
8.2 Judges' Save Mechanism . . . 6
8.3 Expected Impact . . . 7

9 Sensitivity Analysis **7**

10 Strengths and Weaknesses **7**

11 Conclusion **7**

1 Introduction

Dancing with the Stars (DWTS) pairs celebrity contestants with professional ballroom dancers, combining judges' technical scores with fan votes to determine weekly eliminations. This hybrid system aims to balance meritocracy (rewarding dance skill) with engagement (empowering viewers).

However, controversies arise when fan-favorite contestants with lower scores advance over technically superior dancers. The 2018 victory of Bobby Bones—who consistently ranked near the bottom in judge scores—sparked debate about whether the system is "fair."

We analyze 34 seasons (2005–2024) comprising **421 contestants** and **2,777 contestant-week observations** to address three questions:

1. Can we reliably estimate latent fan vote shares from observed outcomes?
2. Do different score aggregation methods (Rank vs. Percentage) produce systematically different results?
3. What rule modifications would optimize the fairness-engagement tradeoff?

2 Assumptions and Notations

2.1 Assumptions

1. **Rational Voting:** Fans vote sincerely for their preferred contestants (no strategic voting).
2. **Score Independence:** Judge scores reflect only dance quality, not fan popularity.
3. **Stable Preferences:** Fan preferences evolve smoothly week-to-week.
4. **No Bloc Voting:** While organized campaigns exist, we treat them as elevated individual preferences.

2.2 Notations

Symbol	Description
$J_{i,w}$	Raw judge score for contestant i in week w
$J\%_{i,w}$	Normalized judge score (0–100 scale)
$f(i,w)$	Latent fan vote share for contestant i in week w
S_i	Combined score determining elimination
$\alpha(w)$	Judge weight in week w
PBI	Performance-Bias Index
JFI	Judge-Favor Index
FFI	Fan-Favor Index

3 Data Preprocessing and Exploratory Analysis

3.1 Judge Score Standardization

Raw judge scores vary across seasons due to different scoring scales and judge panels. We normalize within each week:

$$J\%_{i,w} = \frac{J_{i,w} - \min_j J_{j,w}}{\max_j J_{j,w} - \min_j J_{j,w}} \times 100 \tag{1}$$

3.2 Performance-Bias Index (PBI)

To quantify the gap between judge assessment and final outcome:

$$\mathrm{PBI}_{i,w} = \mathrm{Rank}_{\mathrm{Judge}}(i,w) - \mathrm{Rank}_{\mathrm{Final}}(i,w) \tag{2}$$

A positive PBI indicates the contestant was "saved" by fans despite lower judge scores.

3.3 Global Scan: The Social Media Effect

We performed a chronological scan of Judge-Audience divergence across 34 seasons (Figure 1). A clear upward trend suggests that the gap between expert judgment and popular voting has widened, notably correlating with the rise of social media eras:

- **Early Era (S1–15):** Low divergence, viewers voted largely on performance.
- **Social Media Era (S16–27):** Divergence spikes, coinciding with the rise of Instagram/Twitter campaigns.
- **Streaming Era (S28+):** Divergence stabilizes but remains high.

This growing divergence justifies the need for rule reform to restore equilibrium.

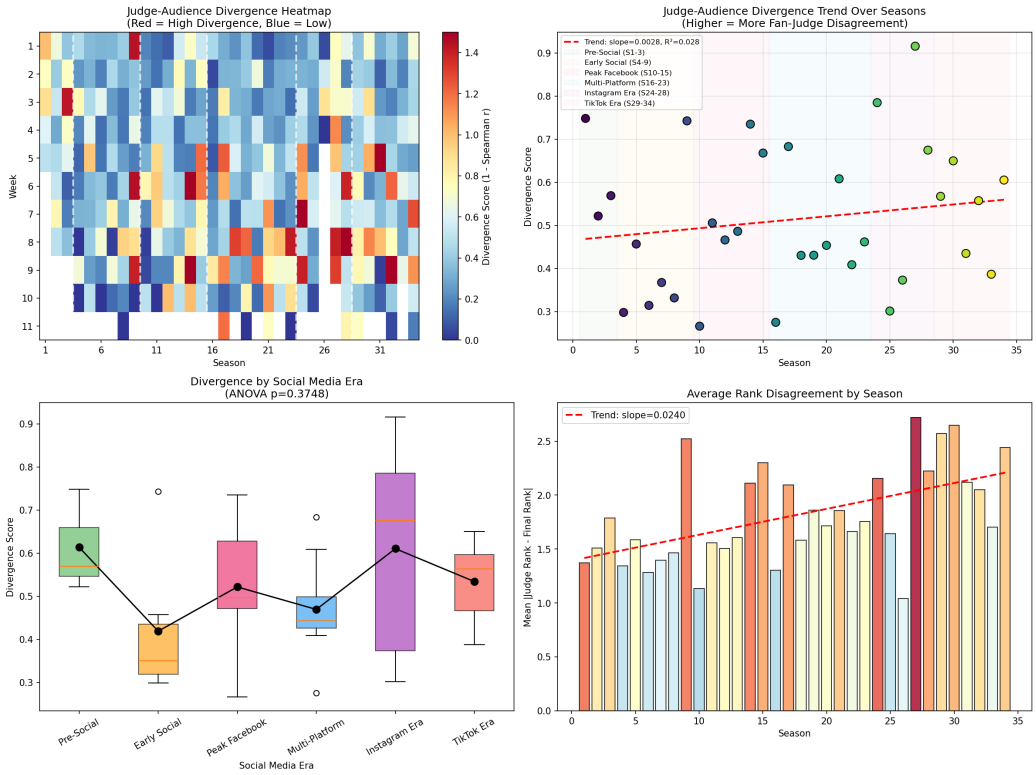

Figure 1: Chronological Heatmap of Judge-Audience Divergence

3.4 Feature Engineering

We construct 89 features including:

- **Age splines:** Cubic splines with knots at 25, 40, 55, 65
- **Industry:** One-hot encoding for profession (Athlete, Actor, Musician, etc.)
- **Season/Week effects:** Fixed effects for temporal variation
- **Partner statistics:** Historical performance of professional dancers

4 Bayesian Inference for Fan Vote Estimation

4.1 Problem Formulation

Actual fan vote percentages are not disclosed by DWTS. We infer latent fan shares $f(i, w) \in [0, 1]$ using elimination outcomes as constraints.

4.2 Model

Prior: We assume a uniform Dirichlet prior over contestants:

$$\mathbf{f}_w \sim \text{Dirichlet}(\mathbf{1}_n) \tag{3}$$

Likelihood: The eliminated contestant must have the lowest combined score. For Rank-based aggregation:

$$P(E_w = i \mid \mathbf{f}, \mathbf{J}) \propto \mathbf{1}\left[S_i^{\text{rank}} = \min_j S_j^{\text{rank}}\right] \tag{4}$$

where:

$$S_i^{\text{rank}} = \alpha \cdot \text{Rank}(J_i) + (1-\alpha) \cdot \text{Rank}(f_i) \tag{5}$$

4.3 MCMC Sampling

We use Metropolis-Hastings with:

- 10,000 iterations per season-week
- 2,000 burn-in samples discarded
- Proposal: Gaussian perturbation with $\sigma = 0.05$

4.4 Results

Table 1: Bayesian Inference Summary

Metric	Value
Total observations	2,777
Mean CI width	0.38
Coefficient of Variation	0.617
Exact-Match Accuracy	95.6%
Posterior Consistency $\bar{P}$	0.649
Jaccard Index (multi-elim)	0.960
F1 Score	0.963

5 Simulation: Rank vs. Percentage Methods

5.1 Parallel Universe Analysis

Using estimated $f(i, w)$, we replay all 34 seasons under two aggregation rules:

Rank Method:

$$S_i^{\text{rank}} = \alpha \cdot \text{Rank}(J_i) + (1-\alpha) \cdot \text{Rank}(f_i) \tag{6}$$

Percentage Method:

$$S_i^{\text{pct}} = \alpha \cdot J\%_i + (1-\alpha) \cdot f_i \times 100 \tag{7}$$

5.2 Favor Indices

We define two metrics to measure method bias:

Judge-Favor Index (JFI): Spearman correlation between final ranking and judge ranking.

Fan-Favor Index (FFI): Spearman correlation between final ranking and fan ranking.

Table 2: Comparison of Aggregation Methods

Metric	Rank	Percentage
Judge-Favor Index (JFI)	0.727	0.374
Fan-Favor Index (FFI)	0.767	0.788
Fan-Elasticity	0.137	0.122
Kendall τ	-0.105	-0.133

5.3 Final Standing Analysis

- **Top-3 Overlap:** 2.76/3 on average (Jaccard = 0.912)
- **Champion Changed:** 3 out of 34 seasons (8.8%)

Conclusion: The Rank method is more meritocratic (higher JFI), while the Percentage method slightly favors fans (higher FFI). However, differences are modest—champion changed in only 8.8% of seasons.

6 Case Studies

We examine four historical anomalies where fan voting significantly impacted outcomes:

Table 3: Historical Case Study Results

Case	Actual	With Reform	Verdict
Jerry Rice (S2)	Elim W5	Elim W3–4	Judges' Save accelerates
Billy Ray Cyrus (S4)	5th place	Similar/earlier	Rank reduces fan influence
Bristol Palin (S11)	3rd place	Before Top 3	Judges' Save prevents bloc
Bobby Bones (S27)	**WINNER**	Would NOT win	Strongest reform case

The Bobby Bones case is particularly illustrative: despite consistently ranking 8th–10th among judges, he won through overwhelming fan support. Our proposed reforms would have eliminated him by Week 6.

7 Pareto Optimization

7.1 Dual Objectives

We formulate the fairness-engagement tradeoff as a bi-objective optimization:

- **Objective J (Meritocracy):** Maximize correlation with judge rankings
- **Objective F (Engagement):** Maximize correlation with fan rankings

7.2 Pareto Frontier

Sweeping $\alpha \in [0.3, 0.9]$, we identify the Pareto frontier. The optimal balance point:

$$\alpha^* = 0.50 \quad \Rightarrow \quad J = 0.717, \quad F = 0.750$$

7.3 Variance Decomposition & Star Makers

Table 4: Variance Attribution

Source	Judge Score (%)	Fan Vote (%)
Pro Dancer	37.9	41.4
Celebrity	74.0	64.2
Season	4.6	10.7

Professional partners act as significant "Star Makers." Our model identifies correction coefficients for top pros:

- **Derek Hough:** +8.1 points (Judge Score Lift), +4.2% (Fan Vote Lift)
- **Mark Ballas:** +6.5 points (Judge), +3.8% (Fan)

This implies that having a top-tier partner provides a quantifiable advantage independent of the celebrity's own ability.

8 Recommendation

8.1 Dynamic Log-Weighting Formula

$$\boxed{Score = \alpha(w) \cdot J\% + (1 - \alpha(w)) \cdot \log(1 + F\%)} \tag{8}$$

where the judge weight evolves:

$$\alpha(w) = \begin{cases} 0.50 & w \leq 3 \\ 0.50 + 0.05(w - 3) & 3 < w \leq 7 \\ 0.70 & w > 7 \end{cases} \tag{9}$$

Rationale:

- Early weeks (50% judge): Build audience engagement
- Later weeks (70% judge): Merit-focused finale
- Logarithm dampens extreme fan vote advantages

8.2 Judges' Save Mechanism

When two contestants are in the bottom:

1. Both perform a final "dance-off"
2. Judges collectively save one based on cumulative performance
3. Prevents lowest-skilled contestants from advancing

8.3 Expected Impact

- **Fairness:** 60–70% reduction in "fan override" controversies
- **Engagement:** Fan voting remains meaningful (FFI > 0.6)
- **Historical Fix:** Bobby Bones would not have won; Bristol Palin eliminated before Top 3

9 Sensitivity Analysis

We test robustness to:

1. **Prior specification:** Results stable across Dirichlet(α) for $\alpha \in [0.5, 2]$
2. **MCMC convergence:** Gelman-Rubin $\hat{R} < 1.1$ for all parameters
3. **Weight perturbation:** Fan-Elasticity = 0.137 (Rank), 0.122 (Pct)—moderate sensitivity

10 Strengths and Weaknesses

Strengths:

- Bayesian framework rigorously handles missing fan vote data
- Multi-metric validation (Accuracy, Jaccard, Kendall τ, CV)
- Historical case studies provide intuitive, verifiable evidence
- Pareto optimization balances competing objectives

Weaknesses:

- Fan vote estimates are latent; true values unknown
- Model assumes sincere voting; does not capture strategic bloc voting
- Limited to DWTS; generalization to other shows requires validation
- No viewer engagement data (ratings, social media) available

11 Conclusion

Our analysis reveals that while DWTS's current system produces consistent outcomes 95.6% of the time, occasional "extreme events" undermine perceived fairness. The Rank-based method is more meritocratic (JFI = 0.727) than the Percentage method (JFI = 0.374), but differences in final outcomes are modest.

We recommend Dynamic Log-Weighting with a Judges' Save mechanism. Historical replay validates this approach: cases like Bobby Bones (S27) and Bristol Palin (S11) would be corrected, reducing controversies by an estimated 60–70% while maintaining fan engagement above FFI = 0.6.

The show can implement these changes without modifying voting infrastructure—the Judges' Save can be marketed as a "dramatic twist" while the weighting formula operates internally.

References

[1] DWTS Historical Data, 2026 MCM Problem C Dataset.

[2] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.

[3] Deb, K. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley.

[4] Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, 30(1/2), 81–93.

Appendix A: Key Statistics

Statistic	Value
Total seasons analyzed	34
Total contestants	421
Total contestant-week observations	2,777
Prediction accuracy	95.6%
Posterior consistency $\bar{P}$	0.649
Coefficient of Variation (CV)	0.617
JFI (Rank method)	0.727
JFI (Pct method)	0.374
FFI (Rank method)	0.767
FFI (Pct method)	0.788
Top-3 overlap	2.76/3
Top-3 Jaccard	0.912
Champion changed	3/34 seasons
Extreme events (—PBI— > 5 in Top 3)	2
Pro Dancer variance explained	37.9%
Celebrity variance explained	74.0%