Dancing with the Stars: Voting Method Comparison

MCM Problem C 2026 – Objective 2 Technical Documentation

Team Documentation

January 2026

Contents

1 Executive Summary

This document details our mathematical framework for comparing voting methods used in *Dancing with the Stars* (DWTS). We examined three distinct voting mechanisms:

1. **RANK Method** (Seasons 1–2): Sum of judge rank + fan vote rank
2. **PERCENT Method** (Seasons 3–27): Sum of judge percentage + fan vote percentage
3. **RANK+B2 Method** (Seasons 28–34): Judges choose bottom 2, fans save one

Key Findings

- Generated **counterfactual eliminations** for all 34 seasons under each method
- Computed **5 novel metrics**: ODS, JFAC, MoS, MSI, USP
- Discovered that **ODS = 0 for all rank-based seasons** (S1-2, S28-34) – a methodological concern
- Found **negative Judge-Fan Alignment** across all methods (JFAC ≈ -0.24 to -0.33)

Important Caveat: Our metrics measure **season characteristics**, not pure method effects. We cannot establish causation because different eras had different voting technologies, celebrity pools, and show formats.

2 Problem Formulation

2.1 The Counterfactual Problem

Given fan vote estimates from Objective 1, we ask: *What would have happened if DWTS had used a different voting method?*

This requires:

1. Applying each voting method's elimination rule to all seasons
2. Computing divergence metrics between method outcomes
3. Analyzing which method favors fan votes vs. judge scores

2.2 Voting Methods

2.2.1 Method 1: Rank-Based (RANK)

Combined score via rank addition (higher rank = worse):

$$S_i^{rank} = R_i^J + R_i^F \tag{1}$$

where $R_i^J = \text{rank}(-J_i)$ and $R_i^F = \text{rank}(-F_i)$.

Elimination rule: $e^{rank} = \arg\max_i S_i^{rank}$

2.2.2 Method 2: Percent-Based (PERCENT)

Combined score via percentage addition:

$$S_i^{pct} = P_i^J + P_i^F = \frac{J_i}{\sum_j J_j} + \frac{F_i}{\sum_j F_j} \tag{2}$$

Elimination rule: $e^{pct} = \arg\min_i S_i^{pct}$

2.2.3 Method 3: Judges' Bottom 2 (RANK+B2)

Two-step process:

1. Identify bottom 2 by combined rank: $B_2 = \arg\text{top2}_i(R_i^J + R_i^F)$
2. Judges choose elimination from B_2 (we assume they eliminate lower judge scorer)

Elimination rule: $e^{B2} = \arg\min_{i \in B_2} J_i$

3 Novel Metrics

3.1 Outcome Divergence Score (ODS)

The fraction of elimination weeks where methods produce different outcomes:

$$\text{ODS}_{season} = \frac{1}{W} \sum_{w=1}^{W} \mathbf{1}[e_w^{rank} \neq e_w^{pct}] \tag{3}$$

3.1.1 Empirical Results

Actual Method	Mean ODS	Range
RANK (S1–2)	0.000	[0.0, 0.0]
PERCENT (S3–27)	0.373	[0.0, 0.67]
RANK+B2 (S28–34)	0.000	[0.0, 0.0]

Table 1: ODS by actual voting method used

Methodological Concern: ODS = 0 for **all 9 rank-based seasons** (S1-2, S28-34). This is suspicious and suggests our counterfactual simulation may have issues with these eras, OR that these seasons had unusually high judge-fan alignment.

3.2 Judge-Fan Alignment Coefficient (JFAC)

Spearman rank correlation between judge rankings and fan vote rankings:

$$\text{JFAC}_w = \rho_s(R_w^J, R_w^F) \tag{4}$$

Range: $[-1, 1]$, where:

- $+1$ = perfect alignment (fans vote exactly as judges score)
- 0 = no correlation
- -1 = perfect anti-alignment (fans prefer low scorers)

3.2.1 Empirical Results

Method Era	Mean JFAC	95% CI
RANK (S1–2)	-0.309	$[-0.45, -0.17]$
PERCENT (S3–27)	-0.238	$[-0.31, -0.17]$
RANK+B2 (S28–34)	-0.331	$[-0.42, -0.24]$

Table 2: JFAC by voting method era

Key Insight: JFAC is **negative for all eras**, indicating that fans consistently vote differently from judges (slightly favoring lower-scoring contestants). This may reflect:

- Celebrity popularity independent of dance ability
- Underdog support from audiences
- Regional/demographic voting patterns

3.3 Fan Vote Leverage Index (FVLI)

The sensitivity of survival probability to fan vote changes:

$$\text{FVLI}_i = \frac{\partial P(\text{survive})}{\partial F_i} \tag{5}$$

For the percent method (analytical):

$$\text{FVLI}_i^{pct} = \frac{1}{\sum_j F_j}\left(1 - \frac{F_i}{\sum_j F_j}\right) \tag{6}$$

For the rank method, FVLI depends on vote gaps between contestants (computed empirically via Monte Carlo).

3.3.1 Scaled FVLI Results

Method	Mean FVLI (scaled)	Interpretation
RANK	0.833	Moderate fan influence
PERCENT	0.874	High fan influence
RANK+B2	0.899	Highest fan influence (in bottom 2)

Table 3: FVLI comparison across methods

3.4 Margin of Safety (MoS)

How close was the non-eliminated contestant to being eliminated?

$$\text{MoS}_w = |S_{e_w} - S_{runner\_up}| \tag{7}$$

Lower MoS indicates more competitive eliminations.

3.5 Method Sensitivity Index (MSI)

The minimum vote change required to flip an elimination outcome:

$$\text{MSI}_w = \min_{\delta} \{|\delta| : e(F + \delta) \neq e(F)\} \tag{8}$$

Higher MSI indicates more robust/stable outcomes.

3.6 Underdog Survival Probability (USP)

Probability that the lowest judge scorer survives elimination:

$$\text{USP} = \frac{1}{W} \sum_{w=1}^{W} \mathbf{1}[\text{lowest judge scorer}_w \text{ survives}] \tag{9}$$

Higher USP indicates the method is more forgiving to poor performers (fans can "save" them).

4 Counterfactual Analysis

4.1 Methodology

For each elimination week across all 34 seasons:

1. Apply RANK elimination rule: Find $e^{rank} = \arg\max_i(R_i^J + R_i^F)$
2. Apply PERCENT elimination rule: Find $e^{pct} = \arg\min_i(P_i^J + P_i^F)$
3. Apply RANK+B2 elimination rule: Identify bottom 2 by rank, eliminate lower scorer
4. Record whether methods agree or disagree

4.2 Monte Carlo Uncertainty

Since fan vote estimates have uncertainty bounds, we run Monte Carlo simulations:

Algorithm 1 Counterfactual with Uncertainty

Require: Fan vote estimates $\hat{F}_i$, bounds $[F_i^{min}, F_i^{max}]$
Ensure: Disagreement probability $P(\text{disagree})$
1: **for** $k = 1$ to 500 MC samples **do**
2: Sample $F_i \sim \text{Uniform}(F_i^{min}, F_i^{max})$ for all i
3: Compute e^{rank} and e^{pct} for sampled votes
4: $\text{disagree}_k \leftarrow \mathbf{1}[e^{rank} \neq e^{pct}]$
5: **end for**
6: $P(\text{disagree}) \leftarrow \frac{1}{500} \sum_k \text{disagree}_k$

4.3 Results Summary

Metric	Point Estimate	MC Mean	MC 95% CI
Overall Disagreement Rate	15.3%	18.7%	[12%, 25%]
Mean JFAC when methods agree	0.42	–	–
Mean JFAC when methods disagree	−0.15	–	–

Table 4: Counterfactual analysis summary statistics

5 Controversial Contestants Case Studies

We analyzed four controversial contestants who benefited from strong fan voting:

5.1 Jerry Rice (Season 2)

- **Actual Placement:** 2nd place
- **Method Used:** RANK
- **Counterfactual:** Under PERCENT simulation, would have been eliminated earlier in some weeks
- **Caveat:** Only 2 seasons used pure RANK, limiting comparison power

5.2 Bristol Palin (Season 11)

- **Actual Placement:** 3rd place
- **Method Used:** PERCENT

- **Counterfactual:** Under RANK+B2 simulation, survived fewer weeks
- **Context:** Political fan mobilization during 2010 election cycle

5.3 Bobby Bones (Season 27)

- **Actual Placement:** Winner
- **Method Used:** PERCENT
- **Pattern:** Consistently high fan rankings despite lower judge scores
- **Context:** Large radio audience provided voting base

5.4 Billy Ray Cyrus (Season 4)

- **Actual Placement:** 5th place
- **Method Used:** PERCENT
- **Pattern:** Country music fanbase provided support but couldn't sustain advancement

Limitation: These are counterfactual **simulations**, not causal analyses. Fan voting behavior would likely change under different rules, and we cannot isolate method effects from era/technology/culture changes.

6 Judges' Bottom 2 Rule Analysis

6.1 Rule Description

Starting in Season 28, DWTS adopted a hybrid rule:

1. Judges and fans combine to identify bottom 2 (by rank sum)
2. Judges then choose which of the bottom 2 to eliminate

This gives judges a "veto" on the final elimination decision.

6.2 Simulation Results

We simulated applying the B2 rule to all 34 seasons:

- **Outcome Difference Rate:** In approximately 20% of weeks, B2 produces a different elimination than pure PERCENT
- **Pattern:** Differences are larger when the score gap in the bottom 2 is large
- **Mechanical Effect:** B2 tends to eliminate the lower judge scorer by construction

6.3 Historical Context

The B2 rule was adopted partly in response to controversial outcomes:

- Bristol Palin (S11) advanced despite lower judge scores
- Bobby Bones (S27) won despite consistent criticism from judges

The rule represents a compromise: fans have influence, but judges get final say in close calls.

7 Methodological Limitations

7.1 Confounding Factors

We cannot causally attribute outcome differences to voting methods because:

1. **Temporal Confounding:** Different eras had different voting technologies (phone, SMS, online, app)
2. **Celebrity Variation:** The popularity and skill mix of contestants varies by season
3. **Format Evolution:** Show format, episode length, and elimination timing changed over 34 seasons
4. **Sample Size:** Only 2 seasons (S1-2) used pure RANK method

7.2 The ODS = 0 Problem

Our finding that ODS = 0 for all rank-based seasons is suspicious. Possible explanations:

1. **Data Issue:** Our fan vote estimates for S1-2 and S28-34 may be overly constrained
2. **True Pattern:** These eras genuinely had higher judge-fan alignment
3. **Method Artifact:** The rank method may compress differences in ways that mask divergence

7.3 Recommended Interpretation

Readers should interpret our metrics as **descriptive statistics of different eras**, not as causal effects of voting methods. The question "Which method is better?" cannot be answered purely from observational data.

8 Synthesis and Recommendations

8.1 Method Comparison Summary

Criterion	RANK	PERCENT	RANK+B2
Fan Influence (FVLI)	0.833	0.874	0.899
Judge-Fan Alignment (JFAC)	−0.309	−0.238	−0.331
Underdog Survival (USP)	Higher	Moderate	Depends on judges
Stability (MSI)	High	Moderate	High

Table 5: Method comparison across key metrics

8.2 Scenario-Based Recommendations

Based on our analysis (with caveats acknowledged):

1. **If prioritizing skill (judge agreement):** RANK+B2 gives judges final say in close calls
2. **If prioritizing fan engagement:** PERCENT method has highest fan leverage for non-bottom-2 contestants
3. **If prioritizing controversy avoidance:** RANK+B2 reduces extreme upset scenarios
4. **If prioritizing simplicity:** RANK method is easiest to explain and most robust to small changes

8.3 Final Caveat

Our analysis computes **what would have happened under different rules given observed data**, but it cannot predict how fan behavior would change under different incentive structures. A proper causal analysis would require randomized experiments, which are not available in this context.

9 Data Files Generated

The following files were created by our Objective 2 analysis:

File	Description
`counterfactual_history.csv`	Point estimates for all weeks
`counterfactual_history_with_uncertainty.csv`	MC uncertainty bounds
`ods_by_season.csv`	ODS scores by season
`fvli_analysis.csv`	Fan vote leverage metrics
`breakeven_analysis.csv`	Break-even fan share analysis
`msi_analysis.csv`	Method sensitivity index
`judges_b2_simulation.csv`	B2 rule simulation results
`controversial_counterfactual_summary.csv`	Case study summaries

Table 6: Output files from Objective 2 analysis

10 References

1. Spearman, C. (1904). "The proof and measurement of association between two things." *American Journal of Psychology*, 15(1), 72-101.

2. DWTS Official Scoring Rules (various seasons), ABC Entertainment.

3. Objective 1 Documentation: Fan Vote Estimation Model.