Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2627255

Behind the Mirrorball: Vote Reconstruction

Summary

Eliminations in *Dancing with the Stars* depend on judges' scores and audience votes, yet the provided dataset omits raw fan-vote totals, creating an inverse problem in which latent popularity must be inferred from observed outcomes. We develop a rule-aware reconstruction–replay–redesign framework that (i) reconstructs within-week fan support, (ii) audits how aggregation rules alone can flip eliminations, (iii) attributes outcome drivers to celebrity versus professional-partner effects, and (iv) proposes an implementable voting rule that better balances merit and popularity.

Task 1: Latent Fan-Vote Reconstruction (Inverse Inference). We model fan support as a within-week vote share on the simplex and estimate it via constrained optimization that enforces elimination-order consistency (with slack for borderline flips), penalizes week-to-week volatility, and regularizes toward a smooth prior. For discontinuous rank-rule weeks, we characterize the feasible set using Monte Carlo rejection sampling and evaluate predictions in rank space. This yields vote-share reconstructions for Seasons 1–34 (421 season–celebrity records) with uncertainty quantified by feasible-sample dispersion.

Task 2: Counterfactual Rule Audit (Replay Under Identical Inputs). Using reconstructed vote shares, we replay each season under percentage-based versus rank-based aggregation to isolate pure rule effects. Across 265 elimination weeks, the two mechanisms disagree in 55.8% of weeks, showing that rule choice alone can frequently change who goes home. Disagreements concentrate in tight-cutoff (low-margin) weeks, accounting for most flips. We further evaluate a bottom-two Judges' Save counterfactual: across 226 analyzed weeks, it would change the outcome in 100 weeks (44.2%), indicating that safeguards matter most in high-sensitivity situations. Spearman alignment diagnostics reveal a structural tradeoff: percent-style combining is typically more judge-aligned, while rank-style combining is more fan-aligned.

Task 3: Explaining Outcome Drivers (Mixed-Effects Decomposition). To separate "skill" and "popularity" channels, we fit paired linear mixed-effects models to standardized judges' scores and logit-transformed inferred fan support, with fixed effects for celebrity attributes and random effects for professional partners and seasons. The results indicate a measurable Partner Halo Effect, reflected in professional-partner random effects that systematically shift inferred fan support beyond celebrity performance trends.

Task 4: A Dynamic "Circuit-Breaker" Voting Rule (Redesign). We keep a near-balanced split in most weeks ($\alpha_w = 0.55$); in the tightest $\sim$25% (critical) weeks, we increase judges' weight to 65%, with a merit-first tiebreak for extreme near ties. In simulation, the redesigned rule improves judge-consistency from 44.7% to 70.8% (+26.1 percentage points) while leaving the majority of weeks unchanged.

Keywords: latent vote reconstruction; counterfactual replay; mixed-effects modeling; dynamic circuit-breaker; fairness–engagement tradeoff

Contents

1 Introduction

1.1 Problem Background

Skill, Style, Spirit — Together

– A DWTS Motto

Figure 1: The DWTS Mirrorball Trophy

Dancing with the Stars (DWTS) is a long-running U.S. TV competition in which celebrity–professional pairs perform weekly routines. Outcomes are determined by two forces: *judges' scores* (technical quality) and *audience votes* (popularity and engagement). This blend of expert evaluation and public preference makes DWTS a natural testbed for studying fairness—the balance between technical merit and audience participation—in hybrid decision systems.

Across 30+ seasons, DWTS has repeatedly adjusted how scores and votes are aggregated. Early seasons relied on rank-based aggregation; later seasons adopted percentage-based aggregation; and some seasons introduced a bottom-two procedure that grants judges additional discretion. These changes often followed public controversy—especially when contestants with low judges' evaluations advanced far (or even won) due to strong audience support.

Crucially, the key driver behind these debates remains hidden: weekly vote totals are not public. Only judges' scores, eliminations, and final placements are observable. Using these signals, we build inverse models to infer weekly fan support and compare historical mechanisms via counterfactual replay, quantifying how alternative rules shift eliminations and final standings. Our analysis moves beyond anecdote to provide data-driven recommendations for a more transparent and better-balanced voting system.

1.2 Restatement of the Problem

Weekly fan vote totals are not publicly released. Using observed judges' scores, eliminations, and final placements, we are asked to:

1. **Reconstruct weekly fan-vote shares** for each contestant, enforce and validate consistency with observed eliminations, and quantify uncertainty.
2. **Compare aggregation rules** (rank-based vs. percentage-based) via counterfactual replay, including controversial weeks and the impact of a bottom-two *Judges' Save*.

3. **Explain outcome drivers** by relating celebrity/professional attributes to judges' scoring and inferred fan support.
4. **Propose and justify** an improved weekly vote–score aggregation rule that better meets a stated objective (e.g., fairness).

2 Assumptions and Justifications

We adopt the following assumptions to make the inverse inference tractable and the counterfactual comparisons interpretable.

Core Modeling Assumptions

A1: **Fan support is identifiable up to scale.** Outcomes depend on *relative* within-week support rather than absolute vote totals; therefore, a normalized vote share is sufficient for our analyses. We explicitly report uncertainty and test robustness across plausible reconstructions.

A2: **Within-season behavior is approximately stable.** The documented aggregation rule is applied consistently within a season, and audience preference evolves smoothly from week to week over the short voting window. Abrupt deviations (e.g., special-format episodes or ties) are treated as residual noise rather than unmodeled rule changes.

A3: **Outcomes are primarily driven by the official mechanism.** Eliminations are attributed to the published combination of judges' scores and fan voting. Unmodeled external shocks (e.g., withdrawals or production exceptions) are treated as noise so the analysis focuses on the structural consequences of rule design.

Analytical & Forward-Looking Assumptions

A4: **Attributes have structured, estimable effects.** Observable contestant and professional-partner characteristics can be represented as structured model terms with stable, interpretable effects on judges' scores and inferred fan support.

A5: **History is informative for future design.** Historical judging and voting dynamics are assumed broadly representative of near-future seasons, providing an evidence base for counterfactual comparisons and rule recommendations.

A6: **"Better" can be defined operationally.** Improvements can be evaluated by measurable criteria (e.g., reduced judge–fan disagreement and stronger alignment between skill signals and survival).

3 Notations

Some important mathematical notations used in this paper are listed in Table 1. All other symbols are defined at their first appearance in the text. We focus on symbols that are repeatedly used in Tasks 1–2.

Table 1: Key Notations (Tasks 1–2)

Symbol	Description
Indices and Sets	
i, w, s	Indices for contestant, week, and season.
A_w	Active contestants in week w (not eliminated before week w).
Observed / Inferred Quantities	
$S_{i,w}$	Judges' score of contestant i in week w.
$q_{i,w}$	Normalized judges' score share of contestant i in week w.
$\hat{p}_{i,w}$	Estimated fan vote share of contestant i in week w.
Mechanism Totals	
$C_{i,w}$	Combined score used by the percent mechanism in week w.
$R_{i,w}$	Combined rank-based total used by the rank mechanism in week w.
Optimization Parameters	
α, β	Regularization weights in the objective function.
$\xi_{e,j,w}$	Slack variables for elimination-order consistency constraints.

4 Model Preparation

4.1 Data Description

Data Source and Coverage. We use the official dataset provided by the MCM organizing committee (2026_MCM_Problem_C_Data.csv), covering Seasons 1–34 of *Dancing with the Stars*. It contains 421 season–celebrity records, where each row corresponds to one celebrity's participation in a specific season.

Unit of Analysis and Structure. The raw data are stored in a wide format (one row per season–celebrity). Week-level judges' scores are spread across columns in the form week<t>_judge<k>_score (t = 1,...,11; k = 1,...,4). For week-by-week inference and replay, we reshape the dataset into a long panel indexed by (season, week, couple), where a couple is identified by the (celebrity_name, ballroom_partner) pair within a season. We restrict all week-level computations to couples active in that week (post-elimination weeks are structurally missing).

Key Variables and Latent Target. Modeling-critical observed fields include: (i) identifiers and background covariates (celebrity_name, ballroom_partner, celebrity_industry, celebrity_homestate, celebrity_homecountry/region, celebrity_age_during_season, season); (ii) weekly performance signals (judges' scores); and (iii) season outcomes (placement and the textual results field, which we parse to recover elimination weeks for validation). Crucially, weekly fan votes are not included in the dataset; we therefore infer relative fan support (vote share) as a latent variable from judges' scores, observed eliminations, and the documented voting rule regime for each season.

4.2 Data Preprocessing and Cleaning

We preprocess the raw wide-format dataset into a clean, rule-aware, week-level panel for modeling. The pipeline consists of three steps: (i) standardizing and parsing core fields, (ii) reshaping the data from wide to long format for week-level analysis, and (iii) attaching each season's voting regime to ensure rule-consistent computation in both inverse inference (Task 1) and counterfactual replay (Task 2).

1. Standardization and week-level panel construction. We first convert raw records into analysis-ready variables. Specifically, we parse the textual "results" field to extract

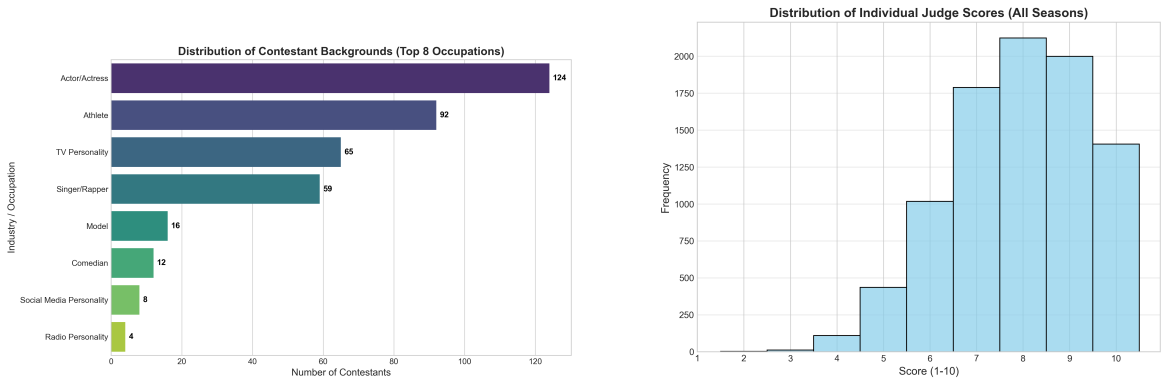


Figure 2: Exploratory distributions: contestant professions (left) and judges' score distribution (right).

elimination-week indices for validation, standardize numeric fields (e.g., placement and age) and categorical fields (e.g., "celebrity_industry"), and reshape the wide table into a long-format panel indexed by (season, week, couple). Structural missingness after elimination (blank weekly scores) is handled by restricting all week-level computations to couples who are *active* in that week, ensuring that totals, ranks, and shares are computed on the correct competitive field A_w.

2. Framework alignment: encoding historical voting regimes. Because the aggregation rule varies across seasons, we explicitly encode each season's rule regime ("rule_type") and generate regime-consistent week-level quantities. In percentage-based seasons, we work in share space (e.g., within-week judges' score shares and inferred vote shares); in rank-based seasons, we operate in ordinal rank space (e.g., within-week judge ranks). This regime annotation is essential for correctly defining model constraints in Task 1 and for running counterfactual replays in Task 2.

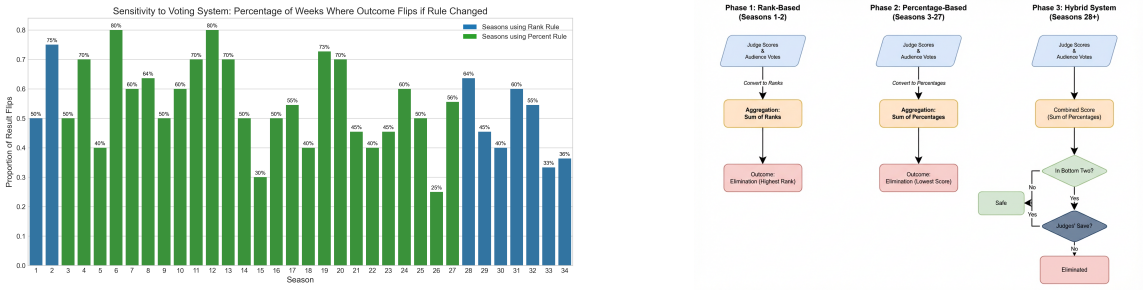


Figure 3: Comparative view of rank-based vs. percentage-based systems (left) and historical evolution of voting mechanisms (right).

Figure 3 motivates our rule-aware preprocessing: since seasons follow different aggregation mechanisms, all subsequent computations must be conditioned on the correct regime to avoid mixing incomparable decision rules in validation and counterfactual analysis.

3. Validation metrics for model evaluation. To evaluate whether reconstructed fan-vote shares can explain observed eliminations, we define two complementary metrics that reflect strict correctness and near-miss robustness:

- **Strict Hit:** the predicted eliminated set matches the observed eliminated set exactly.
- **Relaxed Hit:** the true eliminated contestant(s) fall within a small bottom-group "danger zone" predicted by the model (e.g., Bottom-K with $K = 3$), reducing sensitivity to near-ties.

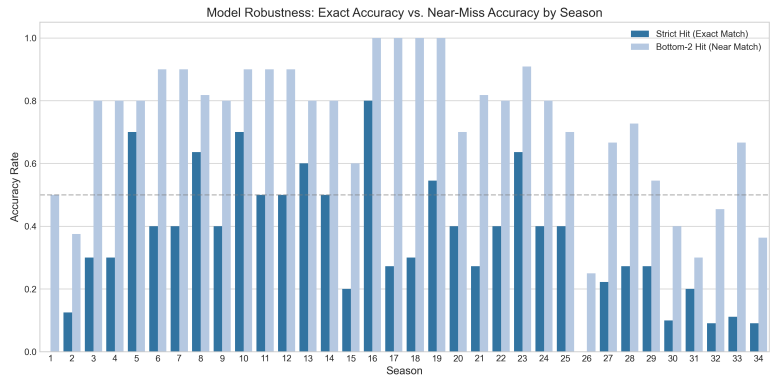


Figure 4: Illustration of strict vs. relaxed hit metrics for model evaluation.

Figure 4 illustrates how the two metrics capture different notions of success: *Strict Hit* rewards exact reproduction of eliminations, while *Relaxed Hit* emphasizes stability in borderline weeks where multiple contestants cluster near the cutoff.

Table 2: Example Rows from the Cleaned, Rule-Aware Long-Format Panel

Season	Week	Contestant	Judge	#J	Share	Rule	Status
5	1	Sabrina Bryan	26.0	3	9.85%	Pct	Safe
5	1	Helio Castroneves	25.0	3	9.47%	Pct	Safe
5	1	Jane Seymour	24.0	3	9.09%	Pct	Safe
5	1	Josie Maran	16.0	3	6.06%	Pct	Elim
		⋮					
29	1	Skai Jackson	21.0	3	N/A	Rank	Safe
29	1	Justina Machado	21.0	3	N/A	Rank	Safe

Note: **Share** is the judges' score share within week w and is defined only for percent-rule seasons. For rank-rule seasons, it is not applicable and shown as **N/A**. **Pct** = percentage-based rule; **Elim** = eliminated.

Table 2 shows the final modeling panel indexed by (season, week, couple). This rule-aware structure ensures that all subsequent computations (shares, ranks, elimination constraints, and counterfactual eliminations) are applied consistently with the historical mechanism in each season.

4.3 Exploratory Data Analysis (EDA)

Before constructing the inverse-inference model, we conduct exploratory analysis to characterize (i) the informativeness of judges' scores, (ii) the relationship between contestant attributes and survival, (iii) early-week volatility, and (iv) the evolution of voting regimes across seasons. These observations motivate our core modeling choices: treating fan support as a decisive latent driver in close eliminations, using robust validation criteria, and stratifying all computations by rule regime.

(1) Judges' score concentration and limited separability. Weekly judges' totals are often clustered within a narrow upper range, producing a small effective spread and frequent near-ties. Consequently, rankings based on judges' scores alone can be unstable and weakly separable. *Implication:* When score separability is limited, eliminations become highly sensitive to the unobserved fan-vote component, supporting our treatment of fan support as a decisive latent factor rather than a minor adjustment.

(2) Attribute imbalance and judge–fan mismatch. The contestant pool is dominated by public-facing professions (e.g., entertainers), and some groups (e.g., athletes) often begin from a higher scoring baseline. However, survival is not fully explained by judges' scores: some categories receive lower average scores yet remain longer, suggesting a mismatch between technical evaluation and audience support. *Implication:* Contestant and partner attributes are meaningful covariates, but survival cannot be explained by judges' scores alone. This motivates reconstructing fan support as a latent variable and analyzing its dynamics separately from judges' scoring.

(3) Early-week volatility as a survival "barrier." Eliminations are noisier and less stable in the early weeks (roughly Weeks 1–4). Contestants who survive this initial stage tend to exhibit more stable trajectories, consistent with an early high-variance phase (smaller observed performance differences and still-forming fan bases). *Implication:* Our reconstruction and validation emphasize robustness in early weeks, rather than enforcing overly strict consistency when the observed signal is weak.

(4) Rule regimes as a time-varying mechanism. The score–vote aggregation mechanism is not constant over the 34-season history. Distinct regimes (rank-based, percentage-based, and later hybrid variants) change how judges' scores and fan votes translate into eliminations. *Implication:* A single monolithic model would be misspecified. We therefore stratify computations by rule regime and ensure that all counterfactual comparisons respect the historically documented mechanism.

5 Task 1: Fan-Vote Reconstruction and Elimination Consistency Modeling

5.1 Mechanism and Variables

Because DWTS does not release weekly vote totals, we treat fan support as a latent *within-week vote share*. For week w, let A_w be the set of active contestants. We define

$$p_{i,w} \geq 0 \ \ (i \in A_w), \qquad \sum_{i \in A_w} p_{i,w} = 1. \tag{1}$$

Let $S_{i,w}$ be the observed total judges' score. For within-week comparability, we normalize it into a score share:

$$q_{i,w} = \frac{S_{i,w}}{\sum_{j \in A_w} S_{j,w}}, \qquad 0 \leq q_{i,w} \leq 1, \quad \sum_{i \in A_w} q_{i,w} = 1. \tag{2}$$

Percent-rule seasons

In percent-rule seasons, judges and fans are combined in share space. We model the composite score as

$$C_{i,w} = q_{i,w} + p_{i,w}. \tag{3}$$

Let $E_w \subseteq A_w$ be the eliminated set in week w and $k_w = |E_w|$. The rule implies

$$E_w = \text{Bottom-}k_w(C_{\cdot,w}), \tag{4}$$

where Bottom-$k(\cdot)$ returns the set of the k smallest values.

Rank-rule seasons

In rank-rule seasons, elimination is based on ordinal ranks. Let $r^J_{i,w}$ and $r^F_{i,w}$ be ranks of judges' scores and fan shares (rank 1 is best):

$$r^J_{i,w} = \text{rank}_{\downarrow}(S_{i,w})\,, \qquad r^F_{i,w} = \text{rank}_{\downarrow}(p_{i,w})\,. \tag{5}$$

Paired-comparison models (e.g., Bradley–Terry) provide a classical statistical foundation for modeling ordinal preferences [3]. A probabilistic alternative is to model full rankings via permutation models such as Plackett–Luce [4]. We define the combined rank

$$R_{i,w} = r^J_{i,w} + r^F_{i,w}. \tag{6}$$

Eliminated contestants are those with the largest combined ranks:

$$E_w = \text{Worst-}k_w(R_{\cdot,w})\,. \tag{7}$$

Tie handling (deterministic). To ensure reproducibility, any ties in ranking are broken deterministically using a fixed contestant ordering (e.g., lexicographic contestant ID). This convention is applied consistently whenever $\text{rank}_{\downarrow}(\cdot)$, Bottom-$k(\cdot)$, or Worst-$k(\cdot)$ is computed.

Note. Our goal is to reconstruct latent vote *shares* consistent with observed eliminations under the documented rule, not raw vote counts.

5.2 Linear Prior

The inverse problem is generally non-unique. We therefore anchor estimation with a linear prior that captures a plausible "popularity direction." Given a feature vector $\mathbf{x}_{i,w}$, we compute

$$z_{i,w} = \boldsymbol{\theta}^\top \mathbf{x}_{i,w}, \qquad \tilde{p}_{i,w} = \frac{\exp(z_{i,w})}{\sum_{j\in A_w}\exp(z_{j,w})}. \tag{8}$$

Features may include score share $q_{i,w}$, historical trend terms, and week effects, and exclude future information.

5.3 Inconsistency and Smoothing Optimization

We estimate $\{p_{i,w}\}$ by balancing three goals: (i) match observed eliminations, (ii) smooth evolution across adjacent weeks, and (iii) remain close to the prior $\tilde{p}_{i,w}$.

Elimination inconsistency for percent-rule weeks

For percent-rule weeks, the observed elimination implies eliminated contestants should not exceed survivors in composite score. We allow small deviations via slacks $\xi_{e,j,w} \geq 0$:

$$C_{e,w} \leq C_{j,w} + \xi_{e,j,w}, \qquad \xi_{e,j,w} \geq 0, \quad e \in E_w,\ j \in A_w \setminus E_w. \tag{9}$$

We define the weekly inconsistency as the average required slack:

$$inc_w = \frac{1}{k_w\,(|A_w| - k_w)} \sum_{e\in E_w} \sum_{j\in A_w\setminus E_w} \xi_{e,j,w}, \qquad (1 \leq k_w \leq |A_w| - 1), \tag{10}$$

and set $inc_w = 0$ for non-elimination weeks.

Objective function (percent-rule weeks)

For percent-rule weeks, we solve:

$$\min_{\{p_{i,w}\},\{\xi_{e,j,w}\}} \sum_w inc_w + \beta \sum_w \sum_{i \in A_w \cap A_{w-1}} (p_{i,w} - p_{i,w-1})^2 + \alpha \sum_w \sum_{i \in A_w} (p_{i,w} - \tilde{p}_{i,w})^2, \quad (11)$$

subject to

$$p_{i,w} \geq 0, \qquad \sum_{i \in A_w} p_{i,w} = 1, \qquad \forall w,\ \forall i \in A_w, \quad (12)$$

$$C_{i,w} = q_{i,w} + p_{i,w}, \qquad \forall w,\ \forall i \in A_w, \quad (13)$$

$$C_{e,w} \leq C_{j,w} + \xi_{e,j,w}, \qquad \xi_{e,j,w} \geq 0, \qquad \forall w,\ \forall e \in E_w,\ \forall j \in A_w \setminus E_w. \quad (14)$$

Here, α controls fidelity to the prior and β controls temporal smoothness. For rank-rule weeks, the elimination rule is ordinal and discontinuous in share space, so we estimate $\hat{\mathbf{p}}_w$ via feasible-set rejection sampling in Sec. 5.5 (accepted-sample mean or MAP), and then evaluate consistency and predictions in rank space using $R_{i,w}$.

5.4 Outputs and Evaluation Metrics

After obtaining $\hat{p}_{i,w}$, we compute model-implied eliminations under the season rule.

- **Percent-rule.** $\hat{C}_{i,w} = q_{i,w} + \hat{p}_{i,w}$ and $\hat{E}_w = \text{Bottom-}k_w(\hat{C}_{\cdot,w})$.
- **Rank-rule.** $\hat{r}^F_{i,w} = \text{rank}_{\downarrow}(\hat{p}_{i,w})$, $\hat{R}_{i,w} = r^J_{i,w} + \hat{r}^F_{i,w}$, and $\hat{E}_w = \text{Worst-}k_w(\hat{R}_{\cdot,w})$.

We report:

- **Strict Hit:** $\text{Strict}_w = \mathbf{1}\{\hat{E}_w = E_w\}$.
- **Relaxed Hit:** the true eliminated contestant(s) fall within a Bottom-/Worst-K "danger zone" (e.g., $K = 3$).
- **Boundary margin (percent-rule):**

$$\hat{m}_w = \hat{C}_{\text{best safe},w} - \hat{C}_{\text{worst elim},w}. \quad (15)$$

Smaller $\hat{m}_w$ indicates a more borderline week.

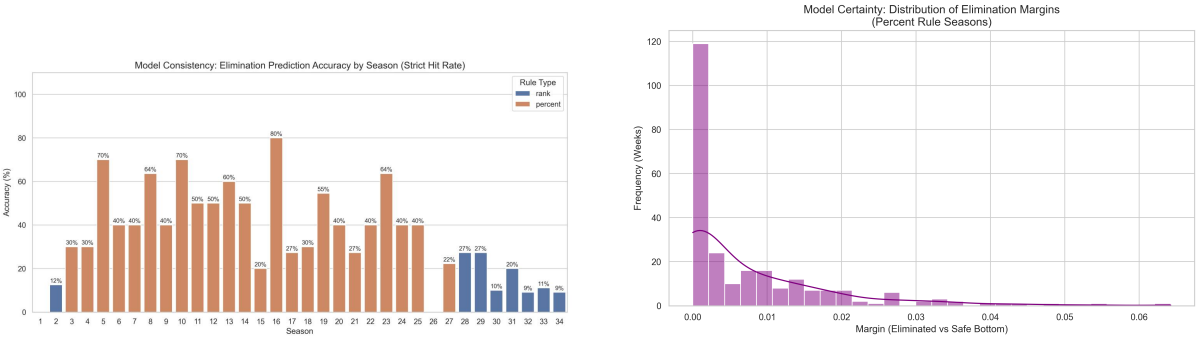


Figure 5: Task 1 performance summary: strict-hit rate by season (left) and distribution of elimination margins (right).

Across seasons, strict-hit rates differ by rule regime: percent-rule aggregation preserves continuous information, while rank-rule aggregation compresses gaps into ordinal ranks and is more sensitive to ties. Elimination margins are frequently near zero, implying many weeks are inherently borderline and thus sensitive to latent fan support. This motivates explicitly modeling fan support as a primary latent driver in borderline weeks rather than a minor perturbation.

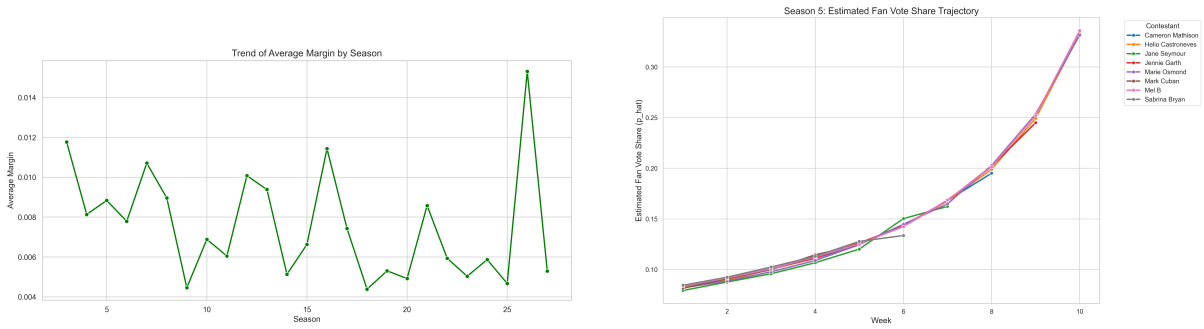


Figure 6: Temporal dynamics: average elimination margin over weeks (left) and average fan-vote share trend (right).

Early weeks are typically noisier and more variable; later weeks stabilize as the field narrows. Vote-share trajectories also evolve over a season, reinforcing the need for week-to-week smoothing.

5.5 Uncertainty Quantification via Monte Carlo Rejection Sampling

Multiple vote-share vectors can be consistent with the same observed elimination. To characterize this non-uniqueness, we use Monte Carlo rejection sampling to approximate the conditional distribution of feasible fan-vote shares [5].

Method. For each elimination week w, we draw N candidate vectors $\mathbf{p}_w$ from a symmetric Dirichlet proposal on the simplex and accept a sample if it reproduces the observed elimination under the season-specific rule. Accepted samples form an empirical conditional distribution of plausible vote shares.

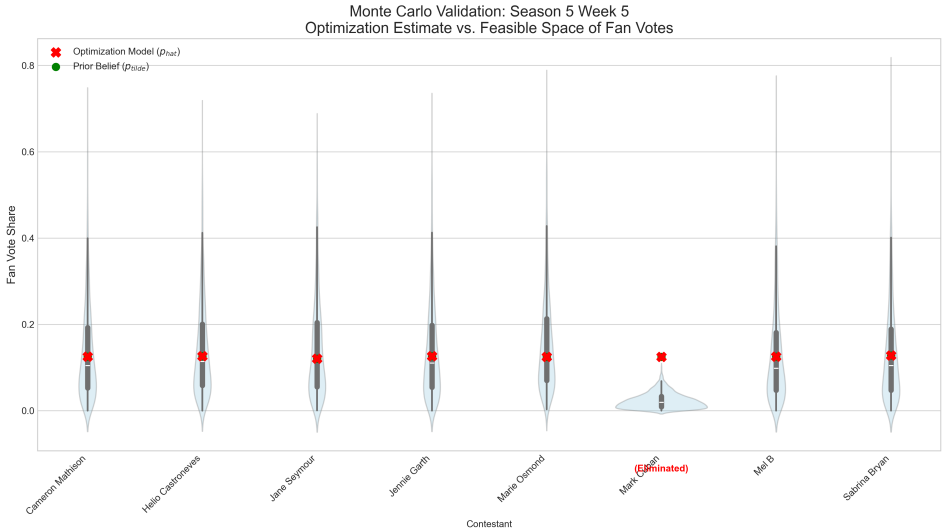


Figure 7: **Feasible fan-vote share distribution under outcome constraints (MC rejection sampling, S05–W06).** Blue density: accepted samples; red "X": optimized estimate $\hat{p}$; green marker: prior anchor.

The optimized estimate typically lies in a high-density region of the accepted-sample distribution, suggesting it is a representative feasible solution rather than an extreme corner case.

Certainty metric. We summarize dispersion using the standard deviation (Std Dev) and a 95% credible interval computed from accepted samples. Here, **CI95 denotes**

the 95% credible-interval width (i.e., the difference between the 97.5th and 2.5th percentiles) for each contestant's vote share. Smaller Std Dev (and narrower CI95) indicate tighter constraints and higher certainty.

Table 3: **Example Monte Carlo summary (text output, S05–W06).**

Reported are $\hat{p}$, Std Dev (certainty), and the 95% credible-interval width for each contestant.

Contestant	Judge Score	$\hat{p}$	Std Dev	95% CI (width)	Status
Cameron Mathison	25.0	0.1432	0.1219	0.4554	Safe
Helio Castroneves	28.0	0.1425	0.1212	0.4518	Safe
Jane Seymour	22.0	0.1503	0.1219	0.4576	Safe
Jennie Garth	27.0	0.1433	0.1219	0.4552	Safe
Marie Osmond	23.0	0.1447	0.1214	0.4536	Safe
Mel B	30.0	0.1424	0.1218	0.4519	Safe
Sabrina Bryan	25.0	0.1336	0.0183	0.0678	Eliminated

This uncertainty analysis verifies that the optimized solution is statistically typical within the feasible set, supporting downstream conclusions in borderline weeks and controversy analyses. Although the reconstructed fan-vote share is not unique, the implied elimination outcome remains stable across feasible solutions, indicating that our conclusions are robust to this non-uniqueness.

6 Task 2: Counterfactual Analysis of Voting Mechanisms and Rule Recommendations

6.1 Objective and Counterfactual Setup

This task isolates the *rule effect*: how the aggregation mechanism changes eliminations and season outcomes when the underlying inputs are held fixed. Using reconstructed vote shares from Task 1, we conduct a *counterfactual replay* for each season: for every elimination week w, we apply both aggregation mechanisms to the same inputs $\{S_{i,w}, \hat{p}_{i,w}\}_{i \in A_w}$, where $S_{i,w}$ is the judges' score for couple i in week w, and $\hat{p}_{i,w}$ is the estimated fan-vote share. A_w denotes the set of active couples that week. Because eliminations are path-dependent, a different elimination early in the season changes the active set in later weeks and can cascade into different matchups and final placements. Therefore, we simulate each season week-by-week using the observed weekly elimination count k_w. Weeks with no elimination ($k_w = 0$) are excluded from disagreement statistics to avoid artificially inflating agreement.

We focus on three goals:

1. **Quantify disagreement** between rank- and percent-based aggregation across seasons.
2. **Examine controversial weeks** where judges and fans diverge, and assess whether rule choice could alter outcomes.
3. **Evaluate a Judges' Save intervention**, where judges choose which of the bottom two contestants to eliminate, and determine when this intervention matters most.

6.2 Two Aggregation Mechanisms

6.2.1 Percent-based Aggregation

Define the within-week judges' share as:

$$q_{i,w} = \frac{S_{i,w}}{\sum_{j \in A_w} S_{j,w}}. \tag{16}$$

We combine this with the fan-vote share to compute the percent-based combined score:

$$C_{i,w}^{\text{pct}} = q_{i,w} + \hat{p}_{i,w}, \tag{17}$$

and predict the eliminated set as the bottom k_w contestants by C^{pct}:

$$\hat{E}_{\text{pct}}(w) = \text{Bottom-}k_w \left(C_{\cdot,w}^{\text{pct}} \right). \tag{18}$$

6.2.2 Rank-based Aggregation

Let $r_{i,w}^J$ and $r_{i,w}^F$ be the ranks of judges' scores and fan-vote shares, respectively (rank 1 is best):

$$r_{i,w}^J = \text{rank}_{\downarrow}(S_{i,w}), \quad r_{i,w}^F = \text{rank}_{\downarrow}(\hat{p}_{i,w}). \tag{19}$$

The combined rank is:

$$R_{i,w}^{\text{rank}} = r_{i,w}^J + r_{i,w}^F, \tag{20}$$

and the predicted eliminated set is:

$$\hat{E}_{\text{rank}}(w) = \text{Worst-}k_w \left(R_{\cdot,w}^{\text{rank}} \right). \tag{21}$$

To ensure reproducibility in case of ties, we apply a deterministic tie-break (e.g., lexicographic contestant ID).

6.3 Frequency and Patterns of Voting-Mechanism Divergence

We define the weekly disagreement indicator:

$$D_w = \mathbf{1}\{\hat{E}_{\text{pct}}(w) \neq \hat{E}_{\text{rank}}(w)\}, \tag{22}$$

and the season-level disagreement rate:

$$\text{DisagreeRate}_s = \frac{1}{|W_s|} \sum_{w \in W_s} D_w, \tag{23}$$

where W_s contains only weeks with $k_w \geq 1$.

The key finding is that **rule choice alone can flip eliminations even under identical inputs**. Since elimination is path-dependent, these flips propagate, influencing later weeks and final outcomes.

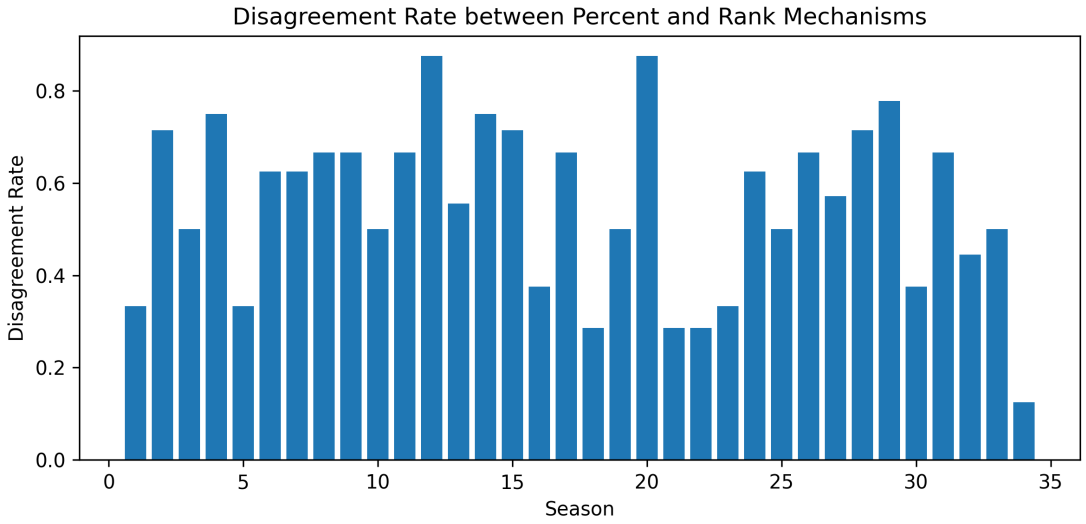


Figure 8: Season-Level Disagreement Rate Between Rank and Percent Aggregation

6.4 Bias Analysis: Alignment with Judges versus Fans

To quantify whether a mechanism favors judges or fans, we compare each mechanism's induced weekly ordering to (i) the judges' ordering and (ii) the fan-share ordering using Spearman rank correlation [1]. We use Spearman's ρ as our primary alignment metric and note Kendall's τ as a common alternative for measuring rank agreement [2].

Across weeks, **percent aggregation is typically more judge-aligned**, while **rank aggregation is more fan-aligned** [7, 8].

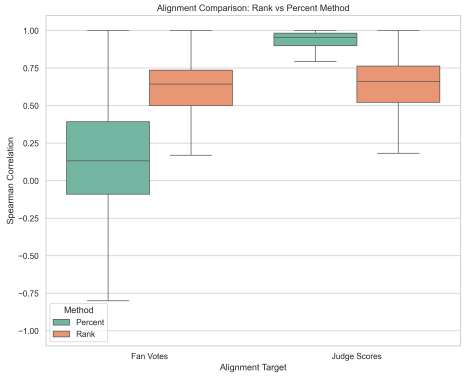


Figure 9: Alignment Comparison (Spearman): Rank vs Percent Against Judges and Fans

6.5 Drivers of Divergence: A Margin-Based Inquiry

To determine whether disagreements are structural or mainly due to borderline flips, we define a percent-space elimination margin:

$$m_w = C^{\text{pct}}_{\text{best safe}} - C^{\text{pct}}_{\text{worst eliminated}}. \tag{24}$$

where $C^{\text{pct}}_{\text{best safe}}$ is the smallest percent score among survivors and $C^{\text{pct}}_{\text{worst eliminated}}$ is the largest percent score among eliminated couples.

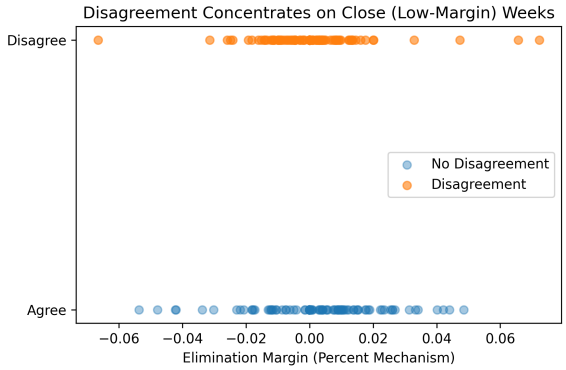


Figure 10: Disagreement Occurs Primarily in Low-Margin (Borderline) Weeks

Empirically, **disagreements concentrate around small** $|m_w|$, indicating that the two mechanisms mainly differ in high-sensitivity weeks where the cutoff is tight.

6.6 Counterfactual Impact on Outcomes and Judges' Save

We evaluate the Judges' Save intervention: judges choose which of the bottom two contestants to eliminate. In our counterfactual implementation, the bottom two are determined by the week's mechanism, and judges eliminate the lower-scoring couple.

Across 226 analyzed elimination weeks, the Judges' Save intervention would have changed the outcome in 100 weeks (44.2%).

Table 4: Impact of a Judges' Save Mechanism on Elimination Outcome

Total Weeks Analyzed	Result Changes	Implication
226	100 weeks (44.2%)	Rank-sum dynamics often allow popularity to override technical merit; Judges' Save mitigates extremes.

6.6.1 Case Studies: Jerry Rice (Season 2) and Bobby Bones (Season 27)

Table 5: Case study summary: Jerry Rice (Season 2)

Week	Rank Elim	Pct Elim	Bottom 2	Judges Save	Actual
3	False	True	False	False	Safe
7	False	True	False	False	Safe
8	True	True	True	False	Eliminated

Case 1: Jerry Rice (Season 2)

Table 6: Case study summary: Bobby Bones (Season 27)

Week	Rank Elim	Pct Elim	Bottom 2	Judges Save	Actual
7	False	True	False	False	Safe
9	True	True	True	False	Safe

Case 2: Bobby Bones (Season 27)

6.6.2 Conclusion

Discrepancies are most salient in bottom-two or near-cutoff situations. The Judges Save mechanism acts as a stabilizer in these high-sensitivity weeks by enabling judges to override extreme vote-dominant reversals while preserving meaningful audience influence. Overall, **percent aggregation amplifies fan influence**, while **rank aggregation suppresses extreme fan-preference swings** by compressing vote-share differences into ranks.

6.7 Rule Recommendations for Future Seasons

1. **If fairness and professional evaluation are prioritized,** percent aggregation is preferable, as it is more judge-aligned and less prone to large vote-driven reversals in clear-cut weeks.

2. **If audience influence and engagement are prioritized,** rank aggregation is preferable, as it is more fan-aligned and increases the effective influence of fan votes.
3. **On Judges' Save,** adopt it *conditionally* (e.g., only after mid-season) or as a *limited safeguard* (e.g., once per season) to balance fairness with engagement.

The aggregation rule matters most in borderline weeks, where $m_w \approx 0$. Producers can manage this by focusing on these critical weeks where rule sensitivity and narrative shifts are most pronounced.

7 Task 3: Analysis of Professional Dancer and Celebrity Features on Performance

7.1 Data and Variables

Our goal is to separate two drivers of success in *Dancing with the Stars*: **technical performance** (judges) and **popularity** (fans), and to quantify how *celebrity attributes* and *professional partners* contribute to each channel. The unit of analysis is contestant i in week w of season s.

Response 1: Judges (technical proficiency). Let $S_{i,w}$ be the observed total judges' score for contestant i in week w. To reduce cross-week scoring baseline shifts, we standardize within the active set A_w:

$$y^J_{i,w} = \frac{S_{i,w} - \text{mean}_{j \in A_w}(S_{j,w})}{\text{std}_{j \in A_w}(S_{j,w})}. \tag{25}$$

Response 2: Fans (popularity). Using the reconstructed vote share $\hat{p}_{i,w}$ from Task 1, we apply a logit transform:

$$y^F_{i,w} = \log\left(\frac{\hat{p}_{i,w} + \epsilon}{1 - \hat{p}_{i,w} + \epsilon}\right), \tag{26}$$

where ϵ is a small constant (we use $\epsilon = 10^{-6}$) to avoid numerical issues when $\hat{p}_{i,w}$ is extremely close to 0 or 1.

Covariates. We include two types of explanatory variables:

- **Celebrity features X_i:** standardized age; industry indicators (one-hot encoding with a baseline category); and other available static descriptors (e.g., region/country if present).
- **Controls $W_{s,w}$:** week index (or stage proxy) and field size $|A_w|$ to capture season progression and increasing competition intensity.

7.2 Model Specification

We use linear mixed-effects models to disentangle **fixed effects** (celebrity traits and controls) from **random effects** (professional partner and season heterogeneity). We fit two parallel models:

Judges-score model.

$$y^J_{i,w} = \alpha_J + \beta_J^\top X_i + \gamma_J^\top W_{s,w} + u_{d(i)} + u_s + \varepsilon^J_{i,w}. \tag{27}$$

Fan-vote model.

$$y^F_{i,w} = \alpha_F + \beta_F^\top X_i + \gamma_F^\top W_{s,w} + v_{d(i)} + v_s + \varepsilon^F_{i,w}. \tag{28}$$

Here $d(i)$ denotes the professional dancer paired with contestant i. We assume random intercepts $u_d \sim \mathcal{N}(0, \sigma_u^2)$, $v_d \sim \mathcal{N}(0, \sigma_v^2)$, and season effects $u_s \sim \mathcal{N}(0, \sigma_{u_s}^2)$, $v_s \sim \mathcal{N}(0, \sigma_{v_s}^2)$, capturing baseline shifts across seasons (e.g., scoring strictness or format differences).

Notation note. Subscripts J and F denote the judges-score and fan-vote models, respectively. X_i collects celebrity features and $W_{s,w}$ are week/season controls; $u_{d(i)}, v_{d(i)}$ and u_s, v_s are random intercepts for professional partners and seasons, and $\varepsilon^J_{i,w}, \varepsilon^F_{i,w}$ are residuals.

7.3 Statistical Analysis Framework

7.3.1 Fixed Effects and Judge–Fan Mismatch

We interpret celebrity effects via the estimated coefficients β_J and β_F with 95% confidence intervals. To directly test whether a trait impacts judges and fans differently, we examine the *differential effect*

$$\Delta = \beta_F - \beta_J, \tag{29}$$

where positive components of Δ indicate traits that are relatively more beneficial for fan support than for judges' scores, and negative components indicate the opposite.

7.3.2 Quantifying Professional Partner Importance (ICC)

To measure how much the professional partner contributes beyond celebrity traits and week controls, we compute ICC:

$$ICC_J = \frac{\mathrm{Var}(u_d)}{\mathrm{Var}(u_d) + \mathrm{Var}(\varepsilon^J)}, \qquad ICC_F = \frac{\mathrm{Var}(v_d)}{\mathrm{Var}(v_d) + \mathrm{Var}(\varepsilon^F)}. \tag{30}$$

A larger ICC implies that *who the contestant is paired with* explains a larger share of residual variation.

7.3.3 Dancer Typology: Technical Specialists vs. Fan Favorites

We compare dancer effects across the two channels by computing the Spearman rank correlation between $\{u_d\}$ and $\{v_d\}$. A strong positive correlation suggests "dual-impact" partners; weak/negative correlation suggests specialization (high u_d but low v_d as technical specialists; high v_d but low u_d as fan favorites).

7.4 Visual Interpretation of Results

(A) Fixed and differential effects. Figure 11 (left) compares fixed-effect estimates between the judges-score and fan-vote models. Figure 11 (right) visualizes the coefficient gap $\Delta = \beta_F - \beta_J$, highlighting attributes that systematically favor popularity over technical scoring (or vice versa).

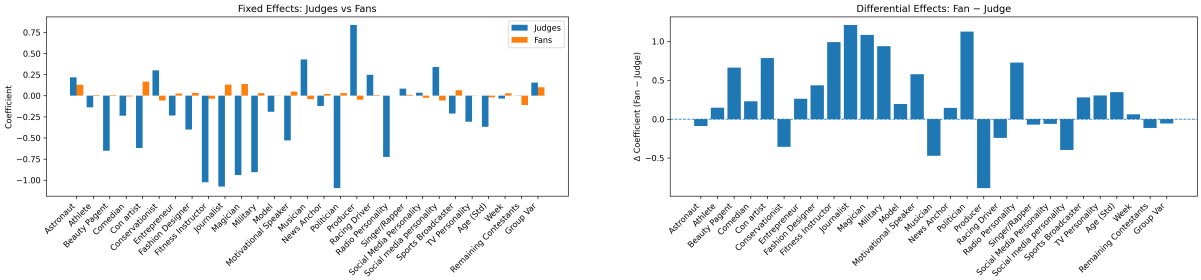


Figure 11: Comparison of effects: fixed effects (judges vs. fans) (left) and differential effects (fan – judge) (right).

(B) Professional partner effects and variance partitioning. Figure 12 (left) maps professional dancers in a two-dimensional effect space (judge-impact u_d vs. fan-impact v_d), suggesting a typology of "dual-impact" vs. specialized partners. Figure 12 (right) summarizes partner importance using ICC for the two models.

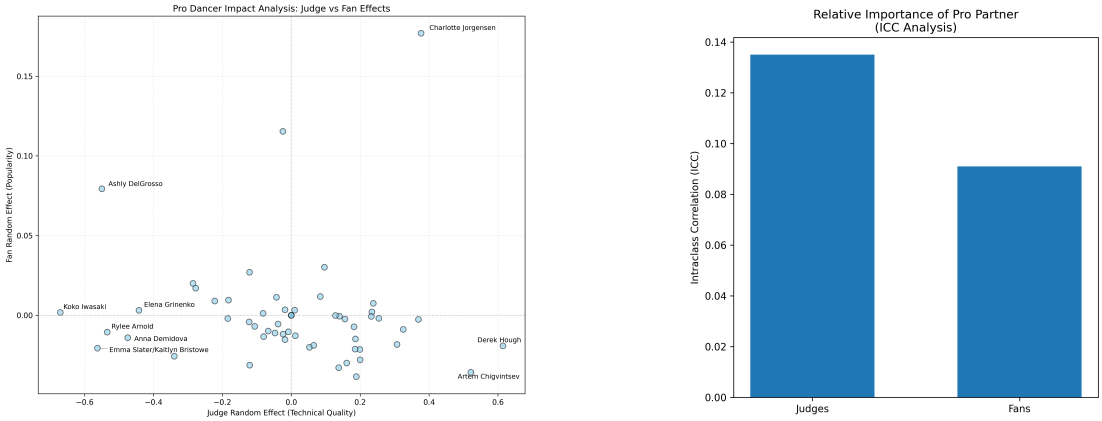


Figure 12: Partner analysis: partner effect landscape (judge-impact vs. fan-impact) (left) and ICC comparison (right).

8 Task 4: A New Weekly Vote–Score Aggregation Rule

8.1 Design Objectives

Task 4 asks for a new weekly aggregation system that is "fairer" (or otherwise better, e.g., more exciting) while remaining implementable using only weekly judges' scores and fan votes. Guided by Task 2, we propose a rule that explicitly targets the weeks where rule choice matters most: *borderline* eliminations with small cutoff margins. Our design pursues three operational objectives:

1. **Protect merit in borderline weeks (fairness).** When the cutoff is intrinsically tight, small vote noise should not overturn clearly stronger technical performance.
2. **Preserve meaningful fan influence in typical weeks (engagement).** When the cutoff is not tight, fan votes should remain consequential to sustain participation and excitement.
3. **Increase stability without removing drama.** Reduce outcome sensitivity to nearties while keeping most weeks driven by the audience–judge balance.

8.2 Proposed Rule: Dynamic Weighting with a Minimal Merit Safeguard

Let A_w be the active set in week w. Let $S_{i,w}$ be the raw judges' total for contestant i, and define the within-week judges' share

$$q_{i,w} = \frac{S_{i,w}}{\sum_{j \in A_w} S_{j,w}}, \qquad i \in A_w. \tag{31}$$

Let $\hat{p}_{i,w}$ be the reconstructed fan vote share from Task 1. We define the new weekly composite score

$$C_{i,w}^{new} = \alpha_w \, q_{i,w} + (1 - \alpha_w) \, \hat{p}_{i,w}, \qquad i \in A_w, \tag{32}$$

where $\alpha_w \in (0, 1)$ is a *week-specific* judge weight. Unlike a fixed 50–50 split, α_w increases only when the week is borderline.

8.3 Borderline Detection and the Dynamic Weight α_w

Task 2 showed that disagreements between aggregation mechanisms concentrate around small elimination margins. We therefore use a percent-space margin as a borderline diagnostic:

$$m_w = C_{\text{best safe},w}^{pct} - C_{\text{worst elim},w}^{pct}, \tag{33}$$

where $C_{i,w}^{pct} = q_{i,w} + \hat{p}_{i,w}$, $C_{\text{best safe},w}^{pct}$ is the smallest percent score among survivors, and $C_{\text{worst elim},w}^{pct}$ is the largest percent score among eliminated contestants. Smaller m_w implies a tighter cutoff and higher sensitivity to fan-vote fluctuations.

We set α_w via a piecewise rule:

$$\alpha_w = \begin{cases} 0.65, & m_w \le \tau \quad \text{(Borderline week: prioritize merit)}, \\ 0.55, & m_w > \tau \quad \text{(Regular week: preserve engagement)}, \end{cases} \tag{34}$$

where τ is chosen from historical margins. In our implementation, we use $\tau = 0.015$ (approximately the lower-quartile of historical margins), which flags the most fragile weeks while keeping most weeks in the engagement-friendly regime.

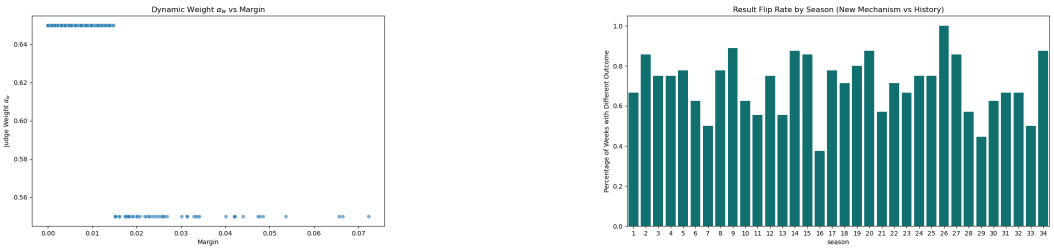


Figure 13: Diagnostics for the proposed rule: the judge weight increases in borderline weeks (left), while the season-level flip rate summarizes how often the new rule changes weekly eliminations relative to history (right).

8.4 A Minimal Safeguard for Extremely Tight Boundaries

Dynamic weighting reduces sensitivity in borderline weeks, but we also include a *minimal, transparent* safeguard for rare cases where the elimination boundary remains nearly indistinguishable even after reweighting.

Let $\hat{E}_w^{new}$ denote the eliminated set under $C_{\cdot,w}^{new}$ (bottom k_w), and let $e^\star$ be the strongest eliminated contestant and $j^\star$ be the weakest survivor under the new score:

$$e^\star = \arg\max_{e \in \hat{E}_w^{new}} C_{e,w}^{new}, \qquad j^\star = \arg\min_{j \in A_w \setminus \hat{E}_w^{new}} C_{j,w}^{new}. \tag{35}$$

Define the new-score boundary gap

$$g_w = C_{j^\star,w}^{new} - C_{e^\star,w}^{new}. \tag{36}$$

If $g_w < \delta$ (an extreme near-tie), we apply a merit-first tiebreak *only* to the two contestants closest to the cutoff (the pair $\{e^\star, j^\star\}$):

> *Eliminate the contestant with the lower judges' score share $q_{i,w}$ (equivalently, lower raw judges' total $S_{i,w}$).*

We use $\delta = 0.02$, so the safeguard activates only when the boundary is nearly indistinguishable, making the intervention rare and defensible.

8.5 Evidence of Improvement: Overall Fairness and a Controversy Case

(A) Overall fairness improvement (season-level). Figure 14 summarizes an aggregate fairness diagnostic: the fraction of elimination outcomes that are consistent with judges' ordering increases substantially under the new rule. This indicates that the proposed mechanism reduces cases where small vote shocks overturn clearly weaker technical performance, especially in borderline weeks (where α_w is increased by design).

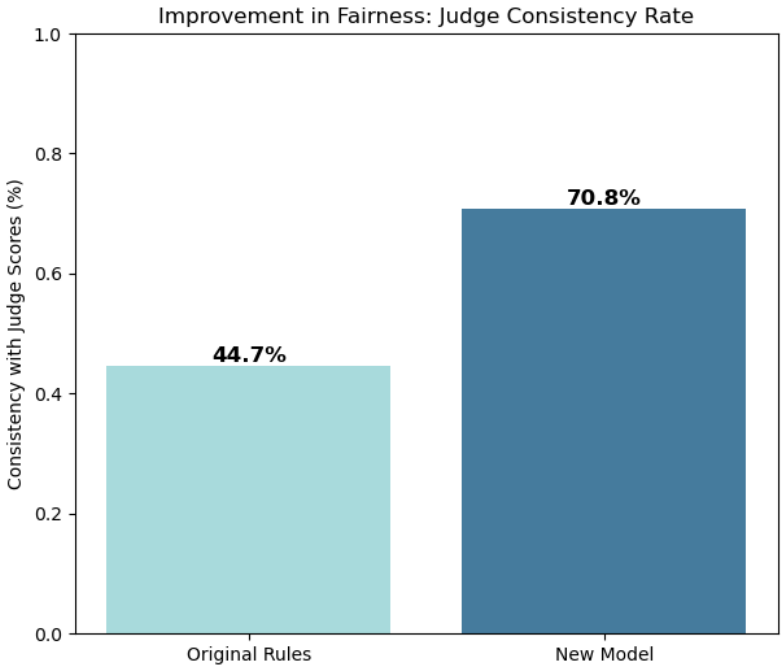


Figure 14: Fairness improvement under the proposed mechanism: judge-consistency rate (original rules vs. new model).

(B) A controversy correction example (Sean Spicer, Season 28). To illustrate how the mechanism behaves in a high-controversy setting, Figure 15 tracks a specific contestant's weekly trajectory under the new composite score. The gray curve/zone represents the survival threshold (i.e., the cutoff separating safe contestants from those at elimination risk), while the red curve shows the contestant's C^{new} across weeks. Under our mechanism, the contestant falls below the threshold and would be eliminated around Week 4, whereas historically they remained in the competition until Week 9 (blue dashed line). This case study demonstrates the intended behavior of the rule: it corrects extreme vote-dominant survivals when technical merit is persistently weak, without requiring extra information beyond weekly scores and votes.

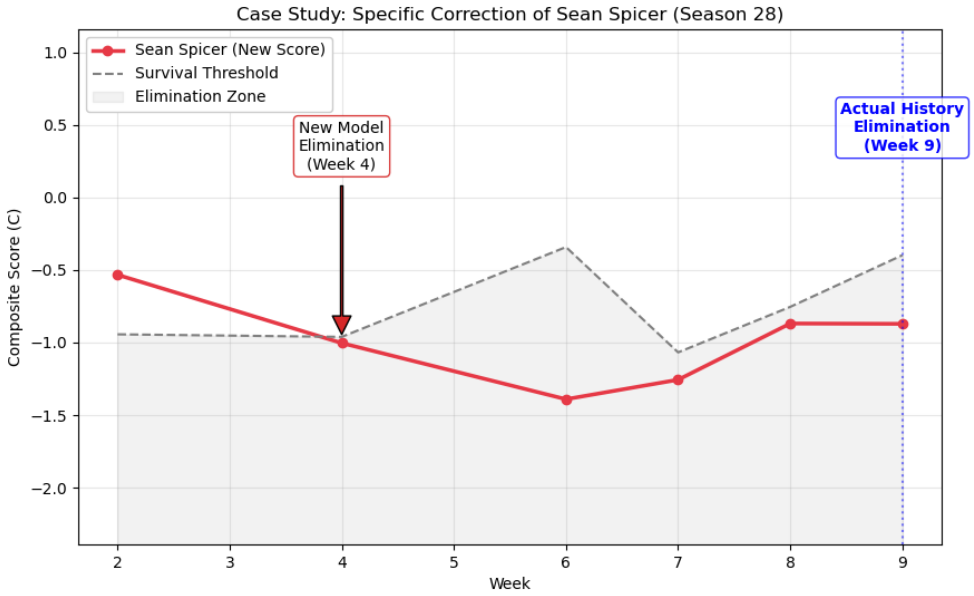


Figure 15: Case study (Season 28): the proposed rule eliminates a persistent low-merit contestant earlier (Week 4) compared with the historical elimination (Week 9).

8.6 Why Producers Should Adopt This Rule

(1) It targets the failure mode identified in Task 2. Task 2 found that mechanism disagreements and "controversy" concentrate in low-margin weeks. Our rule intervenes *only* when $m_w \leq \tau$, precisely where outcomes are most sensitive to small vote fluctuations, thereby improving fairness when merit is most vulnerable.

(2) It preserves fan influence for most of the season. Because most weeks satisfy $m_w > \tau$, the system uses $\alpha_w = 0.55$ in typical weeks, ensuring that fan votes remain meaningful and the show retains audience-driven excitement.

(3) The safeguard is minimal, transparent, and easy to justify publicly. Unlike an opaque judges' override, the safeguard triggers only under a measurable condition ($g_w < \delta$). When activated, it resolves only an extreme near-tie by deferring to the most objective signal available that week (judges' scoring), improving perceived legitimacy without routinely overriding the audience.

8.7 Implementation and Evaluation Criteria

The proposed rule uses the same weekly inputs as our counterfactual replay, $\{S_{i,w}, \hat{p}_{i,w}\}_{i \in A_w}$, and the observed elimination count k_w. Producers can monitor four simple evaluation metrics:

- **Flip Rate:** fraction of weeks where eliminations differ from history (see Figure 13, right).
- **Merit Protection Index:** overlap between $\hat{E}_w^{new}$ and the bottom-k_w by judges' score.
- **Fan Influence Retention:** average $(1 - \alpha_w)$ across weeks (should remain high because most weeks are regular).
- **Safeguard Activation Rate:** fraction of weeks with $g_w < \delta$ (should be small by design).

8.8 Recommended Parameters and Edge Cases

We recommend selecting τ from the empirical distribution of historical margins (e.g., 25th percentile) and choosing δ as a strict near-tie threshold. Weeks with $k_w = 0$ are excluded. The same logic extends to multi-elimination weeks by applying the boundary-gap test at the cutoff between the last survivor and the strongest eliminated. For rank-era seasons, the rule can be applied in share space using $q_{i,w}$ and $\hat{p}_{i,w}$ as defined above, enabling a unified weekly computation across regimes.

Conclusion. This dynamic-weight + minimal-safeguard mechanism is a targeted, data-informed improvement over fixed-weight rules: it protects fairness when outcomes are most fragile (borderline weeks), keeps fan influence meaningful in typical weeks, and introduces a transparent intervention only when the result is statistically indistinguishable.

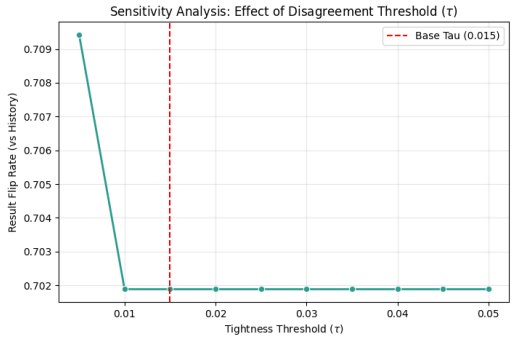
9 Sensitivity Analysis

We conduct a one-at-a-time sensitivity analysis on key hyperparameters in our proposed rule to assess robustness.

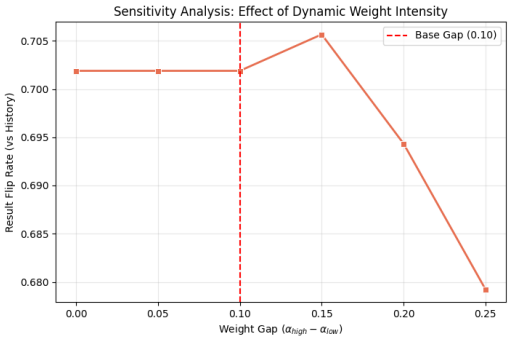
Sensitivity to the tightness threshold τ. The critical-week (tightness) threshold τ determines when the protection mode is activated (i.e., when the elimination margin is sufficiently small). Fig 16a reports the *result flip rate* (relative to historical outcomes) as τ varies. The curve flattens once τ exceeds approximately 0.01, indicating that the flip rate is **robust** to the exact choice of τ within a reasonable range. This plateau suggests that beyond a moderate cutoff, the set of weeks classified as "tight" remains essentially stable, and hence induced outcomes do not change appreciably as τ increases.

Sensitivity to dynamic-weight intensity. We further test sensitivity to the intensity of the dynamic weighting mechanism, measured by the weight gap $\Delta\alpha = \alpha_{\text{high}} - \alpha_{\text{low}}$. As shown in Fig 16b, the flip rate changes only marginally when $\Delta\alpha$ is small (e.g., 0–0.10), while a stronger re-weighting regime (around 0.20–0.25) yields a noticeable reduction in flip rate. Overall, these results indicate that the model is robust to moderate parameter perturbations; we therefore adopt baseline settings that deliver stability gains without requiring extreme re-weighting.

As discussed by Hansen [6], methods like L-curve can be employed for analyzing the trade-off between model stability and sensitivity, which is relevant to understanding the robustness of our re-weighting mechanism.



Effect of tightness threshold τ.



Effect of dynamic-weight gap $\Delta\alpha$.

Figure 16: One-at-a-time sensitivity analysis of key hyperparameters. Dashed vertical lines indicate the baseline settings used in our main experiments.

10 Model Evaluation and Limitations

Strengths:

- **Principled and interpretable rule design.** The dynamic weighting mechanism is not a black box; it adjusts influence based on a clear, data-driven principle (tightness of the elimination margin), making the rule both flexible and transparent.
- **Targeted fairness safeguards.** The tight-week protection and the "save"-style intervention explicitly target high-sensitivity cutoffs, mitigating controversial outcomes driven by near-ties and small perturbations.
- **Empirical robustness.** The model exhibits stable behavior across a reasonable range of hyperparameters, as demonstrated by the sensitivity analysis in Fig 16.
- **Uncertainty-aware framework.** The approach complements point estimates with uncertainty quantification (e.g., Monte Carlo feasibility sampling), enhancing practical credibility for decision support.

Weaknesses and limitations:

- **Computational overhead.** The reconstruction and replay pipeline requires repeated optimization and simulation, which can be computationally intensive during exhaustive hyperparameter tuning and robustness checks.
- **Dependence on historical dynamics.** External validity relies on the assumption that past judging and voting patterns are representative; substantial shifts in show format or audience behavior may require recalibration.
- **Potential for unobserved confounding.** The model only partially captures factors such as judge-specific subjectivity, coordinated fan voting, or external media shocks, which could introduce confounding effects.

- **Limited causal attribution of rule changes.** A causal evaluation of real-world rule changes could be pursued using difference-in-differences designs [9].

11 Memorandum

To: Executive Producers, *Dancing with the Stars*
From: MCM Analysis Team 2627255
Subject: A Dynamic "Circuit-Breaker" Rule to Protect Merit and Preserve Popularity
Date: February 1, 2026

Dear Executive Producers,

We are pleased to share our findings and a practical recommendation based on the provided MCM dataset (Seasons 1–34). Because raw fan-vote totals are not included, we reconstructed within-week fan support (vote shares) in a way that is consistent with observed eliminations, and then used these reconstructions to compare how different vote–score combination rules would have behaved under the same inputs.

What we found. Our analysis shows that the way judges scores and audience votes are combined can meaningfully affect who goes home, and differences between reasonable rules are most noticeable when the result is extremely close (a bottom-two kind of week). We also see a clear contrast between the two common approaches: percent-style combining uses the actual vote and score shares, so outcomes can be more sensitive to week-to-week changes in audience support, while rank-style combining turns those shares into ranks, which compresses large gaps and can help limit extreme, vote-driven reversals.

Recommendation. We recommend keeping the current format for most weeks, but adding a simple **Circuit-Breaker** safeguard that activates only in the tightest weeks. In ordinary weeks, the show runs exactly as it does now. When a week is clearly "too close to call," the safeguard temporarily puts slightly more emphasis on judges' scores to prevent a highly volatile, hard-to-explain elimination.

How it would work in practice. The production team can pre-define a "close-week" trigger based on the gap between the bottom two contestants. If the gap is not small, nothing changes. If the gap is very small, the safeguard turns on for that week only by increasing judges' influence in the combined result. This is easy to communicate on-air as a fairness protection used only in razor-thin situations.

Bottom line. A Dynamic Circuit-Breaker is a minimally intrusive adjustment: it preserves the entertainment value of audience participation, but reduces the risk that near-cutoff volatility produces outcomes that feel out of step with technical performance.

Thank you for the opportunity to participate in this project. We hope our work provides clear, actionable guidance for future seasons.

Sincerely yours,
MCM Analysis Team 2627255

References

[1] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi:10.2307/1412159.

[2] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938. doi:10.1093/biomet/30.1-2.81.

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952. doi:10.1093/biomet/39.3-4.324.

[4] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975. doi:10.2307/2346567.

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi:10.1063/1.1699114.

[6] P. C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, 1992. doi:10.1137/1034115.

[7] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*, pages 613–622, 2001. doi:10.1145/371920.372165.

[8] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM*, 55(5):23, 2008. doi:10.1145/1411509.1411513.

[9] A. Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021. doi:10.1016/j.jeconom.2021.03.014.

Report on Use of AI

Tool Used: OpenAI ChatGPT (GPT-5).

Statement of Use. We used the AI assistant only for minor writing and formatting support (language polishing and LaTeX formatting suggestions). All model formulation, computations, experiments, and conclusions were completed and verified by the team.

Usage Log.

1. **OpenAI ChatGPT (February 1, 2026 version, GPT-5)**

   Query: Polish English wording for the Introduction/Problem Background to improve clarity and concision without changing meaning.
   Output: Provided alternative phrasing, grammar fixes, and smoother transitions between paragraphs.
   How we used it: We adopted selected edits to improve readability; no technical content was added or altered.

2. **OpenAI ChatGPT (February 1, 2026 version, GPT-5)**

   Query: Generate a clean LaTeX template for figure/table captions and consistent cross-references (Fig./Table/Eq.).
   Output: Suggested caption styles, labeling conventions, and a consistent referencing format in LaTeX.
   How we used it: We applied these formatting conventions to improve presentation consistency.

3. **OpenAI ChatGPT (February 1, 2026 version, GPT-5)**

   Query: Provide report-friendly one-sentence descriptions for common plots (bar chart, boxplot, scatter) to use in figure notes.
   Output: Drafted short, neutral descriptions explaining what each plot displays and how to read axes/legends.
   How we used it: We used the phrasing as figure notes; it did not affect any analysis or results.