Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2631882

Dancing with the Hidden Vote: *Fan Power and Rule Effects*

Summary

We study *Dancing with the Stars* (DWTS), where eliminations combine observed judges' totals with latent fan votes under season-specific rules. We build an auditable pipeline for Problems 1–4: infer weekly fan vote shares with uncertainty, replay alternative aggregation rules, compare how celebrity traits and pro-dancer effects influence judges versus fans, and propose a fairness-oriented mechanism for controversial weeks.

Problem 1 (Fan vote inference). Because eliminations map many-to-one to vote shares, we develop **VoteShare-ABC** (Approximate Bayesian Computation; a likelihood-free method). For each season-week, we sample vote-share vectors on the simplex and run a deterministic forward simulator with fixed tie-breaking. We either *accept exact matches* to the observed elimination (hard ABC) or *softly reweight* proposals by a mismatch-distance kernel (soft ABC). We report posterior means and credible intervals, and validate via **posterior predictive consistency (PPC)**—the probability that a posterior draw reproduces the observed elimination under replay. Across **34 seasons** (**301 weeks**), **PPC** $= 62.48\%$, while MAP/mean summaries are stable for downstream analyses. A (α_0, κ) grid shows robust replay consistency, with uncertainty tightening mainly as κ increases.

Problem 2 (Voting-rule impact, controversy, and judges-save). Using inferred fan vote shares (Problem 1), we apply **Percent** and **Rank** to the same *comparable elimination weeks*, so differences reflect rule structure rather than missing data. Disagreements concentrate in *tight-margin boundary weeks*, where Percent is smooth in vote shares but Rank is discrete and can flip outcomes via *rank-cliff* swaps. Season-level disagreement rates and a fan-alignment metric show that, on disagreement weeks, Percent better preserves the *magnitude* of fan preference. Monte Carlo **posterior season replays** for four controversial contestants further indicate that **bottom-two judges-save** robustly shifts outcomes toward judges, lowering placements for fan-favored but judge-disfavored contestants under both deterministic and probabilistic save variants.

Problem 3 (Dual-Channel Effects Model). We fit a **Dual-Channel Effects Model (DCEM)** with aligned covariates to compare how celebrity traits (age, industry, etc.) and **pro-dancer fixed effects** influence judges' scores versus inferred fan support. We model judges by OLS and fans by uncertainty-weighted WLS. Effects are only partially aligned: several predictors are channel-specific or differ in sign/magnitude, indicating distinct pathways for technical evaluation and audience preference.

Problem 4 (Fairness-First + Controversy Mode). We propose a two-mode mechanism: use **Percent** by default, and trigger **Controversy Mode** only when pre-registered diagnostics (judge–fan disagreement, vote uncertainty, optional concentration/HHI) exceed thresholds. In Controversy Mode we apply **uncertainty-aware fusion** (uncertainty-weighted geometric aggregation with weight clipping and deterministic tie-breaking) and optionally a limited **bottom-3 rescue** in single-elimination weeks. Offline replays suggest improved behavior in extreme controversy cases while preserving stability and explainability.

Keywords: DWTS; Approximate Bayesian Computation; posterior predictive consistency; uncertainty quantification; Dual-Channel Effects Model

Contents

1 Introduction

1.1 Background

Figure 1: Dancing with the Stars

In many real-world competitions, outcomes reflect both expert evaluation and public preference. *Dancing with the Stars* (DWTS) exemplifies this setting: eliminations depend on observed judges' scores and unobserved audience votes. Given only judges' totals and weekly outcomes, we seek to infer latent fan vote shares consistent with the observed eliminations.

Statistically, this constitutes a severely underdetermined inverse problem. Vote shares lie on a simplex, motivating a Dirichlet prior [1], [2], while temporal continuity in popularity is captured through a dynamic formulation. Additional complexity arises from rule heterogeneity: DWTS alternates between percent-based and rank-based aggregation, inducing nonlinear and sometimes discontinuous mappings from votes to eliminations. Under both rules, the forward map is many-to-one, rendering the inverse problem non-identifiable and requiring a probabilistic rather than deterministic treatment [3].

1.2 Problem Restatement

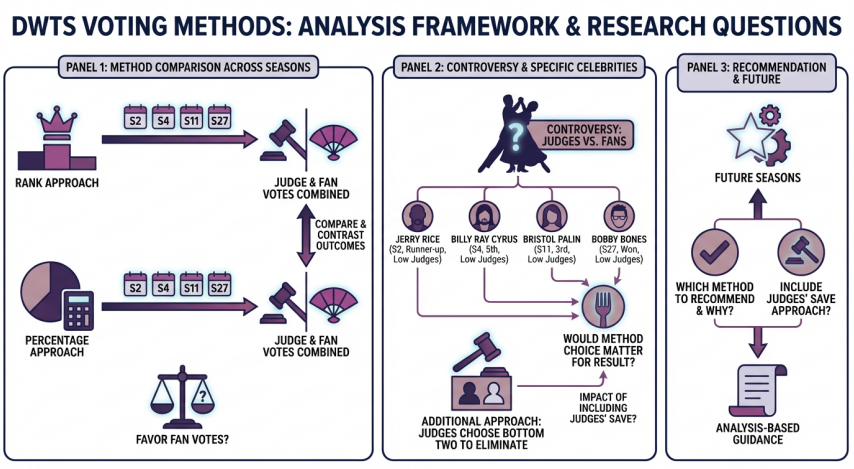

Figure 2: DWTS voting methods and research questions

As Figure 2 illustrates, DWTS combines judges' scores and fan votes through season-specific rules to determine eliminations and final standings. Because fan votes are undis-

closed and aggregation rules vary across seasons, fan preferences must be inferred indirectly with explicit uncertainty quantification. Accordingly, we address five modeling questions:

1. **Infer fan vote shares:** estimate weekly vote-share vectors consistent with observed eliminations and quantify uncertainty across contestant–week pairs [4], [5], [6], [7].
2. **Compare aggregation rules:** apply Percent and Rank rules on comparable weeks to assess disagreement frequency and systematic bias toward judges or fans [8].
3. **Explain controversial outcomes:** evaluate whether alternative rules, including a bottom-two judges' save, alter outcomes in high-disagreement cases.
4. **Model drivers of scores vs. votes:** analyze how contestant traits and professional partners affect judges' scores and fan votes, and whether effects align across channels.
5. **Propose an improved mechanism:** design a voting system that improves fairness, stability, and explainability under uncertainty [9].

1.3 Literature Review

Inference in DWTS-style competitions is challenging because aggregation rules vary across seasons, alternating between percent-based (continuous) and rank-based (discrete) schemes; rank aggregation introduces nonlinearity and discontinuities [8]. More fundamentally, elimination outcomes provide only partial order information, so many distinct latent vote-share configurations can yield the same eliminated set. The resulting inverse problem is ill-posed and underdetermined, requiring uncertainty-aware inference rather than unique reconstruction [3].

Bayesian approaches are well suited to such inverse settings because they propagate uncertainty from latent variables to observed outcomes [1], [9]. However, under rule switching and deterministic tie-breaking, the likelihood linking vote shares to eliminations is intractable, motivating likelihood-free methods such as Approximate Bayesian Computation (ABC) [4], [5], [6], [7].

1.4 Our Work and Contributions

We develop an ABC-based framework to infer weekly fan vote shares across 34 seasons of DWTS. Each week, candidate vote-share vectors are sampled on the simplex, replayed through the season rule using observed judges' totals, and retained (or reweighted) if they reproduce the observed elimination outcome [4], [5]. Posterior summaries provide vote estimates together with explicit uncertainty.

Beyond feasibility, we emphasize *posterior predictive consistency*: the posterior probability that replayed eliminations match observed outcomes [1]. While hard replay accuracy is high by construction, predictive consistency reveals substantial ambiguity in tight-margin weeks. Finally, comparison of Percent and Rank rules shows that rank-based discontinuities can amplify small perturbations into different eliminations, highlighting sensitivity near the cutoff [8].

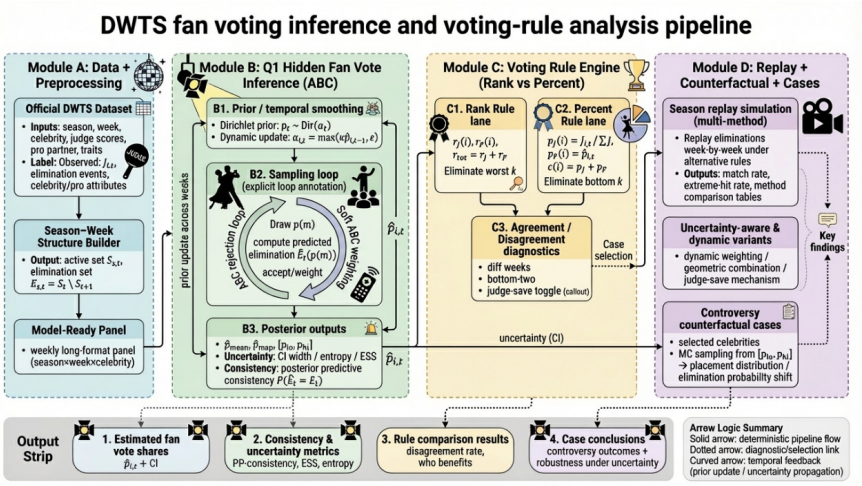

Figure 3: DWTS fan voting interence and voting-rule analysis pipeline

2 Data and Preliminaries

2.1 Assumptions

Based on exploratory analysis of judges' scores, elimination patterns, and season-specific rules, we adopt the following assumptions:

1. **Latent fan vote shares:** Weekly fan vote shares are unobserved latent variables that affect eliminations only through the season's aggregation rule. Vote shares are nonnegative and normalized to sum to one among remaining contestants.
2. **Judges' scores as fixed inputs:** Judges' scores are treated as observed and fixed, reflecting the show's scoring process, and are not further modeled probabilistically.
3. **Known, season-dependent elimination rules:** The aggregation mechanism (Percent or Rank) and the number of eliminations per week are assumed known and correctly encoded from the data.
4. **Non-identifiability of inverse mapping:** Multiple fan vote-share configurations may yield the same observed elimination outcome; therefore, inference targets a distribution of plausible vote shares rather than a unique solution.
5. **Week-level conditional independence:** Conditional on judges' scores and season rules, fan voting behavior is assumed independent across weeks, enabling week-by-week inference.

2.2 Notation

We introduce the following notation for contestant i, week w, and season s:

- $J_{i,w}$: total judges' score received by contestant i in week w.
- $\bar{J}_{i,s}$: average judges' score of contestant i over season s.
- $\beta_{i,s}$: improvement rate of contestant i in season s, defined as the slope of a linear regression of weekly judges' scores on week index.
- $V_{i,w}$: latent fan vote share of contestant i in week w, with $\sum_i V_{i,w} = 1$.
- E_w: observed elimination set in week w.
- $R_{i,w}$: rank of contestant i based on judges' scores in week w.

2.3 Data Description and Preprocessing

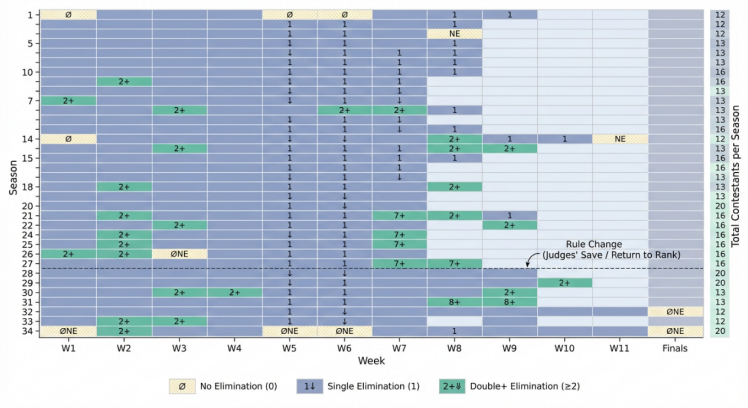

Figure 4: Season-Temporal Heatmap

Figure 4 illustrates the elimination structure across seasons and weeks, highlighting substantial variation in the number of eliminations, including weeks with none or multiple eliminations. Such variability is especially pronounced near documented rule changes. These patterns indicate that eliminations cannot be interpreted as the deterministic removal of the lowest-ranked contestants. Instead, outcomes are governed by season-specific rules that introduce discontinuities and non-uniform feasibility constraints.

3 Problem 1: Estimation of Fan Votes

This section estimates weekly fan vote shares using an **Approximate Bayesian Computation (ABC)** framework that enforces elimination-based voting constraints and quantifies uncertainty in the inferred votes.

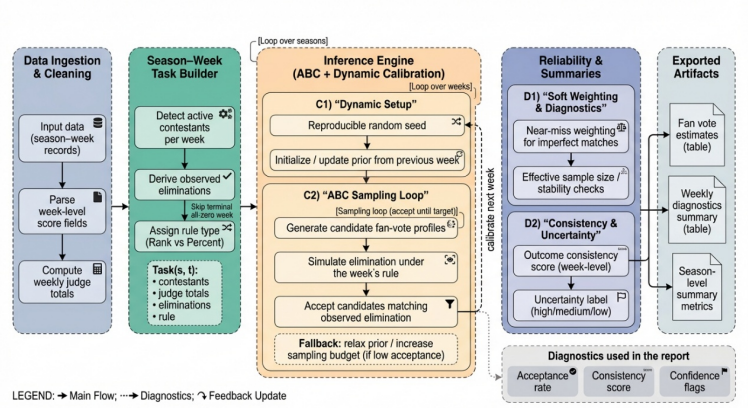

Figure 5: ABC model

3.1 Forward Voting Simulator (Rule-Constrained Elimination)

For each season-week (s, t) with remaining cast $\mathcal{S}_{s,t}$ and observed elimination count $k_{s,t}$, judges' totals $\{J_{i,s,t}\}$ are observed while fan support is latent. We model fan voting by a share vector $\mathbf{p}_{s,t} = \{p_{i,s,t}\}_{i \in \mathcal{S}_{s,t}}$ on the simplex ($p_{i,s,t} \geq 0$, $\sum_i p_{i,s,t} = 1$). Given $(\mathbf{J}_{s,t}, \mathbf{p}_{s,t})$ and the season rule, a deterministic simulator outputs the predicted elimination set $\widehat{\mathcal{E}}_{s,t}(\mathbf{p}_{s,t})$ of size $k_{s,t}$.

Percent rule. Define judges share $P^{(J)}_{i,s,t} = J_{i,s,t} / \sum_{j \in \mathcal{S}_{s,t}} J_{j,s,t}$ and fan share $P^{(F)}_{i,s,t} = p_{i,s,t}$. The combined score is

$$C_{i,s,t} = P^{(J)}_{i,s,t} + P^{(F)}_{i,s,t}, \tag{1}$$

and the simulator eliminates the $k_{s,t}$ smallest $C_{i,s,t}$ values.

Rank rule. Let $\mathrm{rank}_{\downarrow}(\cdot)$ denote descending ranks (best $= 1$). Define $R^{(J)}_{i,s,t} = \mathrm{rank}_{\downarrow}(J_{i,s,t})$ and $R^{(F)}_{i,s,t} = \mathrm{rank}_{\downarrow}(p_{i,s,t})$. The combined standing is

$$R_{i,s,t} = R^{(J)}_{i,s,t} + R^{(F)}_{i,s,t}, \tag{2}$$

and the simulator eliminates the $k_{s,t}$ largest $R_{i,s,t}$. Because ranks are discrete, small perturbations can swap ranks and flip the bottom set (the mechanism behind rule disagreements in Section 4.4).

3.2 Inverse Inference via ABC on the Simplex

We infer $\mathbf{p}_{s,t}$ by enforcing the rule constraint $\widehat{\mathcal{E}}_{s,t}(\mathbf{p}_{s,t}) = \mathcal{E}_{s,t}$. Because the forward map is many-to-one, inference targets a *distribution* of feasible vote shares rather than a unique solution.

Dynamic Dirichlet prior (within-season inertia). We adopt a simplex-respecting Dirichlet prior. For the first week t_0, $\mathbf{p}_{s,t_0} \sim \mathrm{Dirichlet}(\alpha_0 \mathbf{1})$. For $t > t_0$,

$$\mathbf{p}_{s,t} \mid \mathbf{p}_{s,t-1} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_{s,t}), \qquad \alpha_{i,s,t} = \max(\kappa\, \tilde{p}_{i,s,t-1}, \varepsilon), \tag{3}$$

where $\tilde{p}_{i,s,t-1}$ is the previous-week posterior mean and κ controls temporal smoothness.

ABC constraint sampling. For each week, proposals $\mathbf{p}^{(m)}_{s,t}$ are drawn from the prior, replayed through the forward model, and accepted if they reproduce the observed elimination:

$$\mathbf{p}^{(m)}_{s,t} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_{s,t}), \quad \widehat{\mathcal{E}}^{(m)}_{s,t} = \widehat{\mathcal{E}}_{s,t}\left(\mathbf{p}^{(m)}_{s,t}\right), \quad \text{accept if } \widehat{\mathcal{E}}^{(m)}_{s,t} = \mathcal{E}_{s,t}. \tag{4}$$

The accepted drawings approximate the posterior over $\mathbf{p}_{s,t}$; we report posterior means and credible intervals. The implementation and tie-breaking details are provided in Appendix A.

3.3 Consistency and Uncertainty Under Observed Eliminations

Problem 1.1 Consistency of Estimated Fan Vote Shares

To assess consistency, we evaluate inferred vote shares using two complementary criteria.

Posterior predictive consistency (distribution-level replay). *Our primary metric is posterior predictive consistency, which measures agreement under uncertainty rather than mere feasibility.* We draw $\mathbf{p}^{(m)}_{s,t}$ from the (hard/soft) ABC posterior, replay the forward model, and compute the fraction with $\widehat{\mathcal{E}}^{(m)}_{s,t} = \mathcal{E}_{s,t}$ (summarized as a season–week heatmap in Figure 4).

Hard consistency (point-estimate replay). We also replay the forward model using a single representative estimate (default $\mathbf{p}^{\mathrm{MAP}}_{s,t}$; the posterior mean is similar) and compare $\widehat{\mathcal{E}}_{s,t}$ with $\mathcal{E}_{s,t}$, reported separately for elimination weeks and all weeks.

Across 34 seasons we evaluate 301 season-weeks (261 with at least one elimination). Table 1 reports the main metrics.

Table 1: Consistency of inferred fan vote shares with observed eliminations. Hard-consistency metrics reflect feasibility under elimination constraints, while posterior predictive consistency measures agreement under uncertainty.

Metric	Definition	Value	Evaluated on
Posterior predictive consistency (main)	Fraction of posterior draws reproducing the observed elimination (distribution-level replay)	62.48%	301 weeks
Exact match rate (elim weeks)	MAP-based replay exactly matches observed elimination	100.00%	261 weeks
Exact match rate (all weeks)	MAP-based replay matches observed outcome (including no-elimination weeks)	100.00%	301 weeks
Bottom-2 cover rate (elim weeks)	True elimination lies in the model's two most at-risk contestants	99.62%	261 weeks

Hard consistency confirms *feasibility* (MAP replays match observed outcomes), while PPC is the **primary validation result**: it averages **62.48%** and reflects intrinsic non-identifiability under elimination-only information. The bottom-two cover rate (99.62%) indicates that even when the exact elimination is ambiguous, the true outcome almost always lies among the highest-risk candidates.

Problem 1.2 Uncertainty in Estimated Fan Vote Shares

Weekly eliminations only partially identify public voting, so inferred fan vote shares $\mathbf{p}_{s,t}$ are uncertain and unevenly identifiable across contestants and weeks. We quantify uncertainty from the ABC posterior and show systematic variation over time and by voting rule.

Posterior uncertainty (contestant–week). From posterior draws $\{p_{i,s,t}^{(m)}\}_{m=1}^{M}$, we compute $p_{i,s,t}^{\text{mean}}$ and a central 90% credible interval $[p_{i,s,t}^{\text{lo}}, p_{i,s,t}^{\text{hi}}]$. We report $\text{CIW}_{i,s,t} = p_{i,s,t}^{\text{hi}} - p_{i,s,t}^{\text{lo}}$ and the scale-free $\text{RCIW}_{i,s,t} = \text{CIW}_{i,s,t}/p_{i,s,t}^{\text{mean}}$ to compare across seasons with different cast sizes.

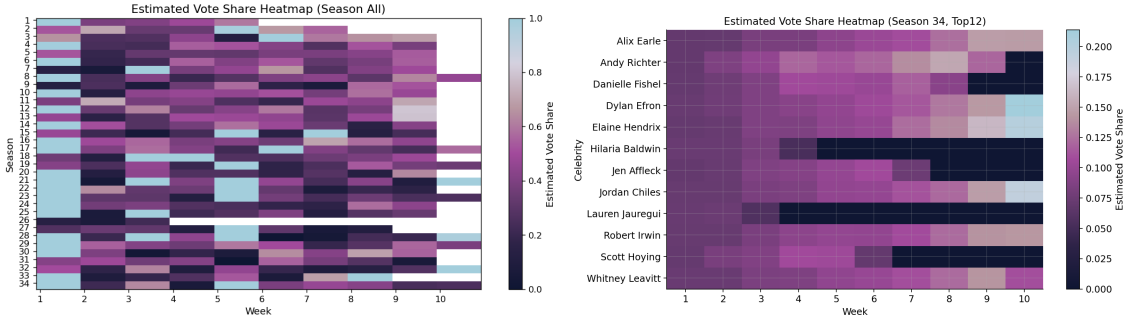

Figure 6: Posterior Predictive Consistency and Uncertainty Diagnostics

Week-level identifiability and stability. To capture how restrictive each elimination cutoff is, we summarize week-level identifiability using mean RCIW (over active contestants), entropy of the posterior mean vote-share vector, and the ABC acceptance rate (lower acceptance implies a smaller feasible region). As a behavioral cross-check, we replay posterior draws through the forward model and compute $\widehat{P}_{s,t} = \Pr(\widehat{\mathcal{E}}_{s,t} = \mathcal{E}_{s,t} \mid \text{posterior})$; lower $\widehat{P}_{s,t}$ indicates higher intrinsic ambiguity.

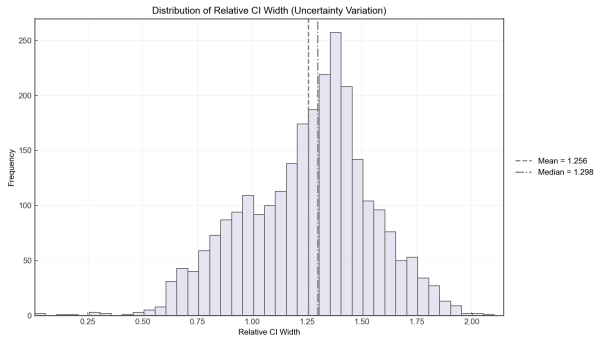

Figure 7: Distribution of Relative CI Width

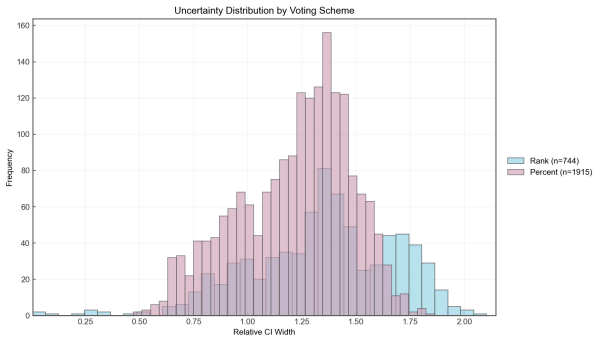

Figure 8: Uncertainty Distribution by Voting Scheme

Conclusion (rule dependence). Uncertainty is larger under the Rank rule because discrete rank swaps can change eliminations under small perturbations in $p_{i,s,t}$, expanding the posterior-consistent set. The Percent rule is linear with smoother feasibility boundaries, yielding lower relative uncertainty on average; thus we report credible intervals alongside entropy, acceptance rates, and $\widehat{P}_{s,t}$ to flag weeks where estimates are less reliable.

4 Problem 2.1: Impact of Voting Rules

Aggregation rules induce distinct *geometries* of the feasible fan-vote space: percent-based rules are linear, whereas rank-based rules are discontinuous and can amplify small perturbations into rank swaps. Consequently, identical vote shares may yield different eliminations and uncertainty profiles. Here we apply *both* rules to the same season–week data to measure (i) disagreement frequency and (ii) systematic bias toward fan preferences.

4.1 Experimental Design

To ensure cross-season comparability, we standardize the evaluation protocol and restrict attention to season–weeks where a fair comparison between aggregation rules is possible.

Comparable-week filtering. We retain only season–weeks (s, t) that satisfy:

1. **Observed elimination:** at least one contestant is eliminated, i.e., $k_{s,t} = |\mathcal{E}_{s,t}| > 0$.
2. **Complete fan estimates:** all remaining contestants have inferred fan vote shares (using $p^{\text{mean}}_{i,s,t}$), allowing both rules to be applied to the same set $\mathcal{S}_{s,t}$ without imputation.

This filtering avoids bias from missing posterior estimates and ensures that observed differences arise from the aggregation rules rather than data incompleteness.

Applying both rules to the same week. For each retained week (s, t), we compute two predicted elimination sets using identical inputs $\{J_{i,s,t}\}$ and $\{p^{\text{mean}}_{i,s,t}\}$:

$$\widehat{\mathcal{E}}^{\text{Percent}}_{s,t} \quad \text{and} \quad \widehat{\mathcal{E}}^{\text{Rank}}_{s,t},$$

with forward maps defined in Section 3.1. Disagreement is recorded as

$$\text{methods\_differ}_{s,t} = \mathbb{I}\left[\widehat{\mathcal{E}}^{\text{Percent}}_{s,t} \neq \widehat{\mathcal{E}}^{\text{Rank}}_{s,t}\right]. \tag{5}$$

Deterministic tie-breaking for reproducibility. Both Percent and Rank rules may produce ties. To ensure reproducibility and consistency with the estimation stage, we apply a fixed deterministic tie-breaker: among tied contestants, lexicographically smaller

`celebrity_name`s are treated as having better standing. As a result, both $\widehat{\mathcal{E}}_{s,t}^{\text{Percent}}$ and $\widehat{\mathcal{E}}_{s,t}^{\text{Rank}}$ are deterministic functions of the week inputs.

Outputs used in later analyses. This design yields a week-level comparison dataset containing predicted eliminations (and bottom-two sets) under both rules, along with season-level summaries such as the number of comparable weeks and the fraction of weeks with rule disagreement. These outputs support the structural comparison (Section 4.2) and fan–judge favorability analysis (Section 4.3).

4.2 Structural Comparison of Rank and Percent Rules Across Seasons

Using the comparable-week protocol in Section 4.1, we apply *both* aggregation rules to every eligible season-week and quantify how often they produce different eliminations. The key idea is that even when the same judges' scores and estimated fan vote shares are used, the *structure* of the rule (linear percent vs. discontinuous rank) can yield systematically different elimination sets, especially near the cutoff boundary.

Season-level summary metrics. For each season s, we compute four summary quantities:

- weeks_with_elim_s: number of weeks in season s with at least one observed elimination ($k_{s,t} > 0$).
- coverage_s: number of *comparable* elimination weeks (i.e., weeks passing the filter in Section 4.1).
- diff_weeks_s: number of comparable weeks in which the two rules disagree:

$$\text{diff\_weeks}_s = \sum_{t \in \mathcal{T}_s} \mathbb{I}\left[\widehat{\mathcal{E}}_{s,t}^{\text{Percent}} \neq \widehat{\mathcal{E}}_{s,t}^{\text{Rank}}\right]. \tag{6}$$

  where $\mathcal{T}_s$ denotes the set of comparable weeks for season s.
- diff_share_s: disagreement rate among comparable weeks,

$$\text{diff\_share}_s = \frac{\text{diff\_weeks}_s}{\text{coverage}_s}. \tag{7}$$

We report these season-level metrics in Table 2. Seasons with larger diff_share_s are those in which the choice of aggregation rule more frequently changes *who* gets eliminated.

Table 2: Season-level disagreement summary (comparable elimination weeks only).

Season	Comparable Elim Weeks	Different Weeks	Disagreement Share	Coverage Rate
33	1	1	1.00	0.17
27	6	3	0.50	0.86
26	2	1	0.50	0.67
24	7	3	0.43	0.88
9	8	3	0.38	0.89
31	8	3	0.38	0.89

Where do disagreements concentrate? Although the two rules agree in most weeks, disagreements tend to cluster in *marginal elimination weeks*—weeks in which several contestants are tightly grouped near the cutoff. In such cases, the percent rule (linear in vote share) and the rank rule (stepwise in ranks) can induce different boundary orderings, caus-

ing the identity of the eliminated contestant(s) to change even when overall performance levels appear similar.

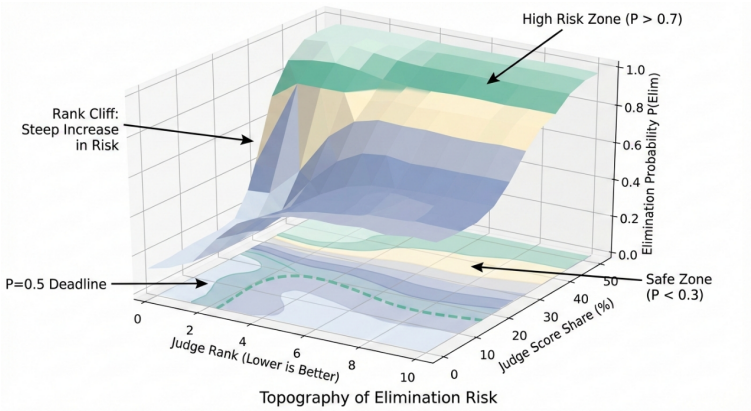

Figure 9: 3D Elimination Risk Topography

Visualizing season-to-season disagreement. Figure 10 plots $\mathrm{diff\_share}_s$ by season, with point size and color encoding coverage. Disagreements are not uniformly distributed: most seasons cluster at moderate levels, while a few show markedly higher sensitivity to the aggregation rule. This pattern suggests structural differences in how closely contested seasons are near elimination thresholds, examined quantitatively in the subsequent mechanism analysis.

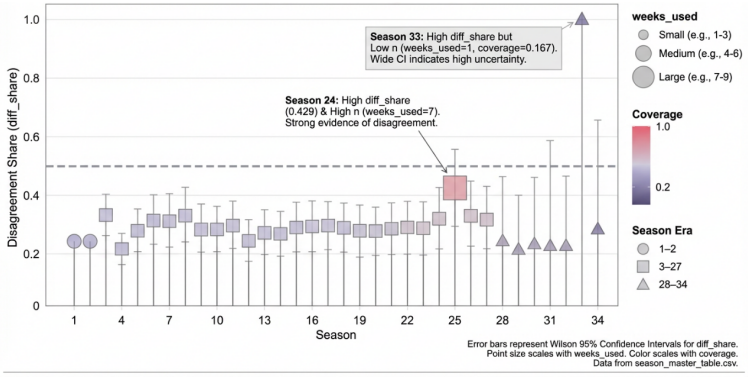

Figure 10: Season-Level Disagreement Share

Takeaway. Percent and Rank rules agree in most weeks but diverge in a non-negligible, season-dependent set of boundary cases. This indicates that aggregation choice can matter when competition is close—an effect we quantify in Section 4.3 by testing whether disagreements align with fan- or judge-preferred outcomes.

4.3 Does One Method Favor Fans More?

A central question in comparing the Percent and Rank rules is whether one aligns more closely with fan preferences (popularity) or judges' preferences (technical evaluation). We address this using a direction-consistent favorability metric based on the relative standing of eliminated contestants.

Fan-rank of eliminated contestants. For each comparable elimination week (s, t) (Section 4.1), let $k_{s,t} = |\mathcal{E}_{s,t}| > 0$ be the number of eliminations. Using inferred fan vote shares

$p_{i,s,t}^{\text{mean}}$, define the fan rank among surviving contestants as

$$r_{i,s,t}^{(F)} = \text{rank}_{\downarrow}\left(p_{i,s,t}^{\text{mean}}\right). \tag{8}$$

with $r^{(F)} = 1$ denoting the most popular contestant. For any eliminated set $\mathcal{A} \subseteq \mathcal{S}_{s,t}$ with $|\mathcal{A}| = k_{s,t}$, the average fan-rank of eliminated contestants is

$$\overline{r}^{(F)}(\mathcal{A}) = \frac{1}{k_{s,t}} \sum_{i \in \mathcal{A}} r_{i,s,t}^{(F)}. \tag{9}$$

Larger values of $\overline{r}^{(F)}(\mathcal{A})$ indicate that less popular contestants are eliminated.

Fan-favor delta. Let $\widehat{\mathcal{E}}_{s,t}^{\text{Percent}}$ and $\widehat{\mathcal{E}}_{s,t}^{\text{Rank}}$ denote the elimination sets predicted by the Percent and Rank rules for the same week. The week-level fan-favor delta is

$$\Delta_{s,t}^{\text{fan}} = \overline{r}^{(F)}\left(\widehat{\mathcal{E}}_{s,t}^{\text{Percent}}\right) - \overline{r}^{(F)}\left(\widehat{\mathcal{E}}_{s,t}^{\text{Rank}}\right), \tag{10}$$

where $\Delta_{s,t}^{\text{fan}} > 0$ indicates that the Percent rule eliminates contestants who are, on average, less popular and is therefore more aligned with fan preferences. The season-level effect is

$$\Delta_{s}^{\text{fan}} = \mathbb{E}_{t \in \mathcal{T}_s}\left[\Delta_{s,t}^{\text{fan}}\right]. \tag{11}$$

Judge-favor delta (complementary view). Analogously, define judges' ranks $r_{i,s,t}^{(J)} = \text{rank}_{\downarrow}(J_{i,s,t})$ and the corresponding average eliminated judges-rank $\overline{r}^{(J)}(\mathcal{A})$ as in (9). The judge-favor delta is

$$\Delta_{s,t}^{\text{judge}} = \overline{r}^{(J)}\left(\widehat{\mathcal{E}}_{s,t}^{\text{Percent}}\right) - \overline{r}^{(J)}\left(\widehat{\mathcal{E}}_{s,t}^{\text{Rank}}\right). \tag{12}$$

with season average $\Delta_{s}^{\text{judge}} = \mathbb{E}_{t \in \mathcal{T}_s}[\Delta_{s,t}^{\text{judge}}]$. Negative values of Δ^{judge} indicate that the Percent rule is relatively more pro-judge, while positive values indicate the opposite.

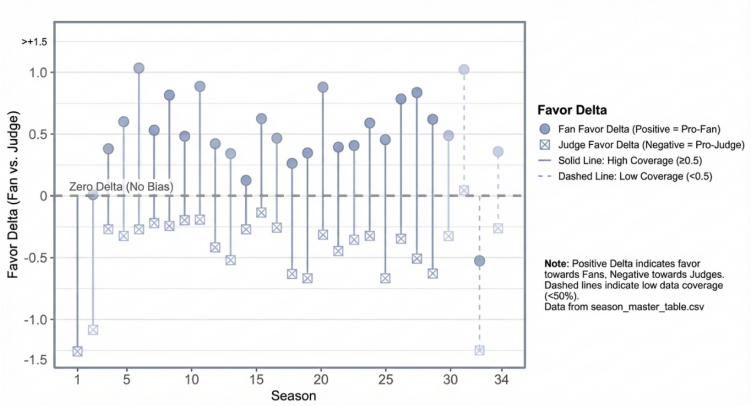

Figure 11: fan$_favor_deltavsseason$

Interpretation. Figure 11 shows that the two rules produce very similar outcomes overall. Differences concentrate in a small set of tight-margin weeks: when they disagree, the Percent rule more often eliminates contestants with lower inferred fan support (worse fan rank), indicating greater sensitivity to the *magnitude* of fan preference in boundary cases. Season-level effects vary, and low-coverage seasons should be interpreted cautiously.

Consistently, Δ_s^{judge} tends to shift in the opposite direction, reflecting a trade-off between fan and judge alignment.

4.4 Mechanism Explanation: Why Do Disagreement Weeks Occur?

Disagreements between Percent and Rank are uncommon in clear-cut weeks but cluster in *tight-margin* eliminations near the cutoff. In this regime, Percent is smooth because it aggregates continuous shares, whereas Rank is *discontinuous* since outcomes hinge on discrete rank swaps.

Rank Cliff (stepwise nonlinearity). Rank aggregation creates a staircase-like boundary: a one-rank change in judges or fans produces a discrete jump in combined standing. Thus, small vote perturbations can flip who crosses the cutoff even when performance is nearly identical. Figure 12 visualizes this "Rank Cliff" mechanism.

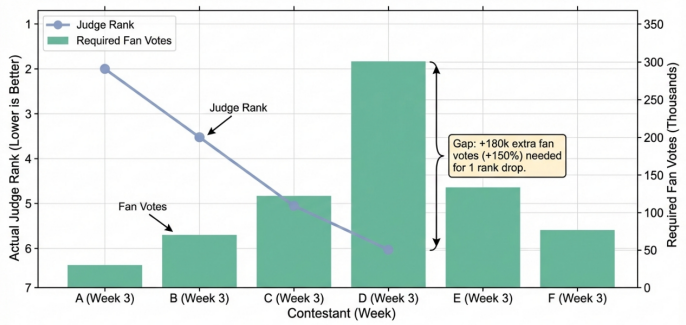

Figure 12: The Rank Nonlinearity ("Rank Cliff"): discrete rank swaps create stepwise survival thresholds.

Different feasible boundaries (Percent vs. Rank). Percent induces an approximately linear, smooth boundary in (P_J, P_F), so elimination risk changes proportionally with fan support. Rank partitions the plane into piecewise-constant rank-cells, creating sharp edges where small perturbations can reshuffle the bottom set. Accordingly (Figure 13), disagreements occur primarily when contestants lie near these cell boundaries.

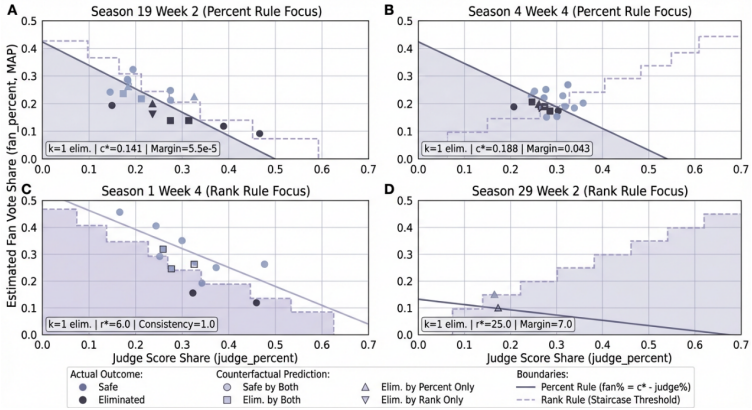

Figure 13: Structural divergence of voting rules: Percent yields smooth margins, while Rank yields piecewise-constant regions with sharp boundaries.

5 Problem 2.2 and 2.3: Controversial Contestants and Bottom-Two Judges-Save

Beyond season-level comparisons, attention often centers on "controversial" cases—contestants for whom judges' evaluations and fan preferences diverge sharply. We (i) formalize a defi-

nition of controversy, (ii) examine four well-known cases, and (iii) evaluate counterfactual outcomes under alternative rules, including the *bottom-two judges-save* intervention.

5.1 Case Selection and Definition of "Controversy"

We focus on four frequently cited controversial contestants:

- **Season 2: Jerry Rice** (runner-up despite repeatedly low judges' scores),
- **Season 4: Billy Ray Cyrus** (5th place with frequent bottom-tier judges' ranks),
- **Season 11: Bristol Palin** (3rd place despite persistently low judges' scores),
- **Season 27: Bobby Bones** (winner with consistently low judges' scores).

To quantify controversy across seasons, we define weekly judges' and fan ranks among remaining contestants,

$$r_{i,s,t}^{(J)} = \mathrm{rank}_{\downarrow}(J_{i,s,t}), \qquad r_{i,s,t}^{(F)} = \mathrm{rank}_{\downarrow}\big(p_{i,s,t}^{\mathrm{mean}}\big). \tag{13}$$

and measure disagreement by the absolute rank gap

$$D_{i,s,t} = \left| r_{i,s,t}^{(J)} - r_{i,s,t}^{(F)} \right|. \tag{14}$$

For each contestant, we summarize controversy using season-level statistics of $D_{i,s,t}$ and the frequency with which they appear in the *bottom-two risk set* under each aggregation rule. Intuitively, a contestant is controversial if they are repeatedly ranked poorly by judges yet remain safe due to strong fan support.

5.2 Counterfactual Outcomes Under Four Scenarios

Because true fan votes are unobserved, counterfactual analysis must propagate uncertainty rather than rely on point estimates. We therefore use a Monte Carlo procedure based on posterior fan vote shares (Section 3): in each replicate, weekly shares are sampled and the season is replayed under a given rule, yielding a distribution of final placements. For each controversial contestant, we evaluate four scenarios:

1. **Percent:** percent-based aggregation.
2. **Rank:** rank-based aggregation.
3. **Percent + Save:** percent aggregation with bottom-two judges-save.
4. **Rank + Save:** rank aggregation with bottom-two judges-save.

Weekly replay rule. Let $J_{i,t}$ denote the total judges score for contestant i in week t, and let $p_{i,t}$ denote the sampled fan vote share from the posterior for that week. Under **Percent**, we compute a combined score

$$S_{i,t}^{(P)} = \frac{J_{i,t}}{\sum_j J_{j,t}} + p_{i,t}, \tag{15}$$

and eliminate the contestant with the smallest $S_{i,t}^{(P)}$ (ties broken deterministically).

Under **Rank**, we compute ranks $r_{i,t}^{(J)}$ and $r_{i,t}^{(F)}$ where rank 1 is best (highest judges score or highest fan share), form $S_{i,t}^{(R)} = r_{i,t}^{(J)} + r_{i,t}^{(F)}$, and eliminate the contestant with the largest $S_{i,t}^{(R)}$ (ties broken deterministically). For the **+ Save** variants, we first identify the bottom two under the corresponding base rule (Percent or Rank), and then apply a judges-save model to determine which of the two is eliminated.

The problem statement specifies that judges eliminate one of the bottom two. To avoid reliance on a single behavioral assumption, we consider two judges-save models:

- **Deterministic save:** eliminate the bottom-two contestant with the lower judges' score (ties broken deterministically).
- **Probabilistic save:** eliminate one of the bottom two with probability increasing in the judges-score gap, modeled via a logistic link with tunable steepness.

Across both variants, qualitative conclusions are unchanged: the judges-save mechanism systematically reduces the protection afforded by strong fan support.

Results: placement distributions. Figure 14 summarizes simulated placement distributions under the four scenarios for each controversial contestant. Each panel displays the full distribution, robust summaries (median and interquartile range), and key probabilities such as $\Pr(\text{win})$ and $\Pr(\text{top-3})$.

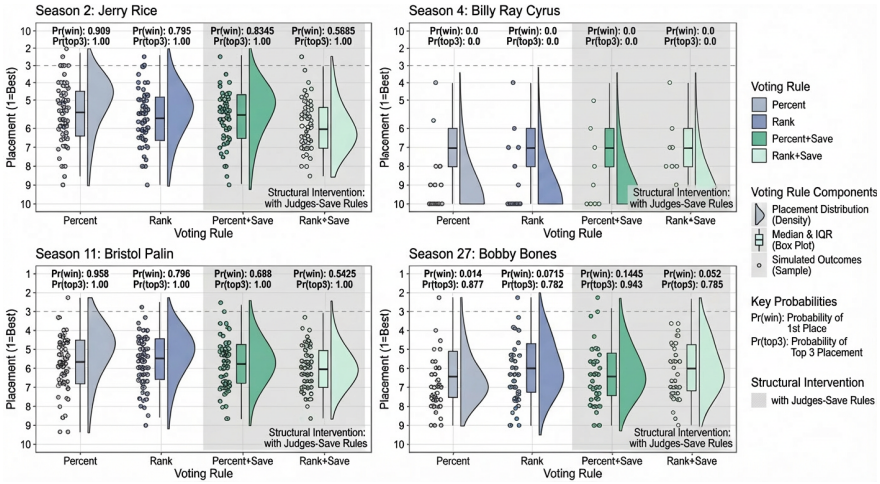

Figure 14: Placement Distribution Under Four Rules

Case highlights (qualitative). Across the four case studies, two consistent patterns emerge:

- **Aggregation rule alone:** Percent and Rank can yield different survival outcomes primarily in tight-margin weeks. Rank often provides stronger protection for fan-favored contestants, while Percent may eliminate them earlier when judges' scores remain low.
- **Judges-save effect:** Introducing bottom-two judges-save systematically shifts placement distributions downward for fan-favored but judge-disfavored contestants, weakening fan protection and raising the effective technical threshold for survival.

For example, Bristol Palin is relatively insensitive to the aggregation rule alone but experiences a marked drop in survival once judges-save is introduced. Bobby Bones is better protected under Rank than Percent, while judges-save further reduces his likelihood of deep advancement under both rules.

5.3 Recommendation

Our recommendations are guided by three criteria: fairness (balancing judges' evaluation and fan preference), stability (robustness in tight-margin weeks), and entertainment value.

Aggregation rule. Across the full sample, Percent and Rank produce nearly identical eliminations in most weeks; thus, there is no blanket "always more fan-favoring" rule

effect. The practical difference emerges in tight-margin weeks where the rules disagree: Percent tends to eliminate the contestant with lower inferred fan support, because it responds smoothly to the magnitude of vote-share differences, whereas Rank can flip outcomes via discrete rank swaps. Accordingly, we recommend Percent as the default aggregation rule for interpretability and for more stable boundary behavior.

Judges-save. Bottom-two judges-save consistently weakens fan protection for contestants with low judges' scores. We therefore do **not** recommend it as a permanent feature. If technical merit is prioritized, judges-save may be applied *selectively* in near-tie or high-noise weeks, using a predefined margin criterion.

Evidence. These recommendations are supported by cross-season disagreement patterns and favorability deltas (Figures 7–8), as well as counterfactual placement distributions for controversial contestants (Figure 14).

6 Problem 3: Dual-Channel Effects Model

We develop a *dual-channel effects model* (DCEM) to quantify how celebrity characteristics and pro-dancer effects influence competitive success through judges' scores and fan votes. Modeling both channels with shared covariates allows direct comparison of effect sizes, revealing how and to what extent judges' and fans' preferences align or diverge.

6.1 Data Feature Construction

To enable direct comparison between judges' scores and fan votes, we construct a common set of covariates aligned with both response channels. Features are grouped into:

- **Contestant-level:** age, industry indicators (e.g., actor, athlete, singer), and professional-dancer partnership.
- **Competition-level:** season and week controls, along with confidence-based weights for fan votes derived from estimated uncertainty (WLS using CI width).

All features are constructed at the weekly level using a unified preprocessing pipeline (`t3.py`) shared by both the judges' and fans' models.

Table 3: Feature dictionary for the Dual-Channel Effects Model (DCEM).

Variable	Description	Coding	Channel
Age	Contestant age	Continuous (years)	Judges, Fans
Industry	Celebrity background	One-hot (Actor, Athlete, Singer/Rapper, TV, Model, Other)	Judges, Fans
Pro dancer	Professional partner indicator	Binary (0/1)	Judges, Fans
Season/week FE	Temporal controls	Fixed effects	Judges, Fans
Fan weights	Uncertainty-based weighting	Inverse CI width	Fans only

Feature alignment and interpretation. Table 3 lists the DCEM covariates and their inclusion across channels, enabling direct comparison of how the same traits affect judges and fans. Figure 15 complements this by showing that performance trajectories are strongly linked to final outcomes and are modeled in both channels.

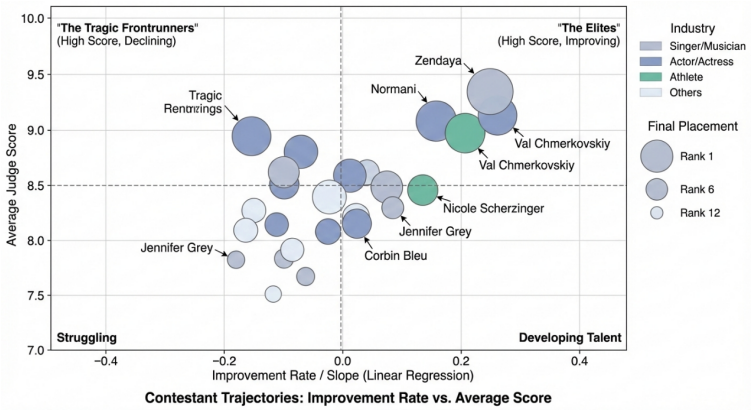

Figure 15: Trajectory-Outcome Bubble Chart

Section 6 analyzes fan-vote dynamics using the same feature set, enabling direct comparison with judges' scores; in particular, contestant industry effects are explicitly controlled for and examined in the judges' scoring models.

6.2 Model Specification

The Dual-Channel Effects Model (DCEM) explains judges' evaluations and fan support using an identical set of covariates and controls, so coefficient differences reflect behavioral mechanisms rather than model inputs. For contestant i in season-week (s, t), the outcomes are the judges' score share (or standardized score) and the inferred fan vote share (posterior mean), both defined at the weekly level.

Specification. Both channels include contestant traits (e.g., age and industry), professional partner fixed effects, and season/week controls.

Estimation. Judges' scores are estimated via OLS, while fan vote shares are estimated via WLS with weights inversely proportional to posterior uncertainty; standard errors are clustered at the `pair_id` level.

Table 4: DCEM model specification (compact reference).

Channel	Specification
Judges (OLS)	Outcome: judges score share (or standardized score). Covariates: age, industry dummies, pro-partner fixed effects, season/week controls. Estimator: OLS with robust standard errors.
Fans (WLS)	Outcome: inferred fan vote share (posterior mean). Covariates: same as judges. Weights: inverse posterior uncertainty (CI width). Inference: standard errors clustered by `pair_id`.

Because both channels share identical covariates and controls, coefficient differences reflect channel-specific responsiveness rather than modeling artifacts. For example, an industry effect significant only for judges indicates evaluation bias, while significance only for fans reflects popularity dynamics.

6.3 Do Traits Affect Judges and Fans the Same Way?

Because the DCEM applies identical covariates and controls to both channels (Section 6.2), differences in estimated effects capture genuine differences in how judges and fans respond to the same traits. Figure 16 compares standardized coefficients from the

judges (x-axis) and fans (y-axis) models: points near the diagonal indicate similar effects, while deviations indicate divergent responses; statistical significance is encoded to highlight robust effects.

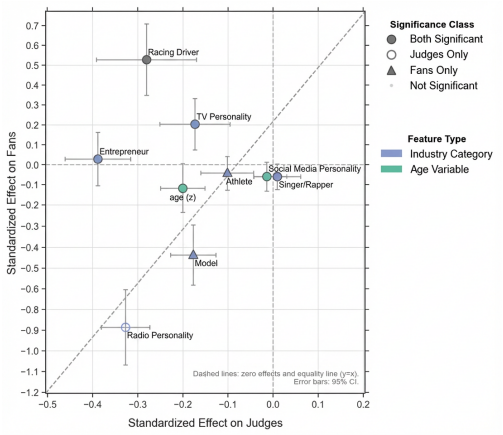

Figure 16: Judges vs Fans

Key finding: partial alignment with systematic divergence. Performance-related features (e.g., score trajectory proxies) tend to affect judges and fans in the same direction, indicating partial alignment. In contrast, several traits exhibit systematic divergence: some industry categories show substantially stronger fan-side effects, others are significant primarily in the judges channel, and age often differs in magnitude or significance across channels. Overall, the same observable characteristics do *not* influence judges and fans uniformly. Table 5 summarizes the most salient divergences, listing variables with opposite-signed effects or significance in only one channel, together with standardized coefficients, p-values, and a directional tag for rapid comparison.

Table 5: Q3-3 Key variables with divergent effects between Judges and Fans.

Variable	$\text{sign}(J)$	$\text{sig}(J)$	$\text{sign}(F)$	$\text{sig}(F)$	**Pattern**
Age (age_z)	$-$	Yes	$-$	Yes	Aligned (judges stronger)
TV Personality	$-$	Yes	$+$	Yes	Opposite-signed
Model	–	No	$-$	Yes	Fans-only effect
Entrepreneur	$-$	Yes	–	No	Judges-only effect

Interpretation and implication. Overall, DWTS outcomes reflect two partially distinct mechanisms: judges respond primarily to performance quality and improvement, while fans combine performance with popularity cues that vary by celebrity background. Consequently, the choice of aggregation rule determines which mechanism dominates in tight-margin weeks.

Answer to the prompt. No—celebrity traits and partner effects do not influence judges and fans in the same way. While some features align across channels, several high-impact traits are channel-specific or opposite in sign, indicating systematic differences between expert evaluation and audience preference.

6.4 Pro Dancer Impact and Whether It Differs for Judges vs. Fans

This subsection tests whether professional partners matter and whether their effects differ for judges versus fans. Using identical controls, we estimate partner fixed effects in each channel and compare their distributions and incremental explanatory power.

Partner fixed effects. We treat each partner as a fixed effect capturing persistent advantages net of contestant traits and season structure. Figure 17 contrasts standardized effects across channels, highlights judge-facing vs. fan-facing partners, and reports the cross-channel correlation and the incremental R^2 from adding partner fixed effects.

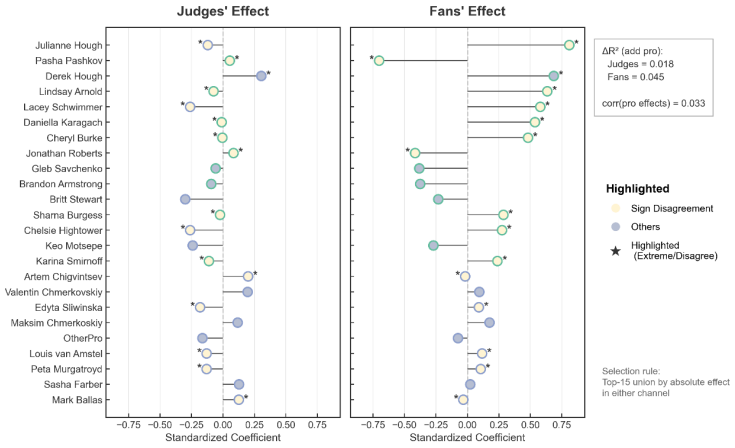

Figure 17: Pro effects

Table 6: Q3-4 Incremental explanatory power of traits and pro-partner effects (nested-model R^2).

Channel	R^2 (Base)	R^2 (+Traits)	R^2 (+Pro FEs)	ΔR^2 (Pro)
Judges	0.6374	0.6553	0.6732	0.0179
Fans	0.4398	0.4843	0.5288	0.0445

How much do pros matter? Adding professional-partner fixed effects increases explanatory power more for fans ($\Delta R^2 = 0.0445$) than for judges ($\Delta R^2 = 0.0179$), even though overall fit is higher for judges. This indicates that pairing and partner reputation matter substantially more for audience support than for technical scoring.

Do pros matter the same way? No. Pro effects are nearly uncorrelated across channels ($r = 0.0328$), implying that partners who raise judges' scores are not the same as those who boost fan votes. Pro influence operates through distinct technical (judges-facing) and presentation/appeal (fan-facing) pathways.

Answer to the prompt. Professional partners have a measurable but channel-specific impact: they explain additional variance beyond celebrity traits, with different partners driving judges' scores versus fan support.

Producer takeaway. Professional partners are the largest source of judge–fan divergence. Because their effects differ sharply across channels, casting and pairing decisions—and the aggregation rule—determine which pathway dominates in tight-margin weeks.

7 Problem 4:Proposed Voting System: Fairness-First with "Controversy Mode"

Our results suggest no single fixed aggregation rule optimizes fairness, stability, and interpretability in every week. In most weeks, Percent and Rank produce the same elimination, so a stable default rule is sufficient. Disagreements concentrate in tight-margin boundary weeks, where small perturbations and rank discontinuities can flip outcomes and intensify judge–fan conflict.

We therefore propose a **Fairness-First Voting System with a Controversy Mode**: use a stable default in ordinary weeks, but switch to a transparent intervention when objective controversy criteria are met:

Be stable when the outcome is clear; be fair and explainable when it is controversial.

7.1 Why a New Rule is Needed

A new voting rule is needed for two recurring gaps. **Fairness** concerns arise in weeks with strong judge–fan disagreement, where an aggregation rule can eliminate fan-preferred contestants. **Explainability** concerns arise under rank-based aggregation, where small perturbations can cause discontinuous rank swaps and make close outcomes hard to justify. These patterns recur systematically, motivating a mechanism-level redesign rather than ad hoc fixes.

Motivation snapshot. Table 7 reports three dataset-level indicators: total elimination weeks, Percent–Rank disagreement frequency, and the prevalence of pre-flagged extreme judge–fan divergence.

Table 7: Motivation snapshot for rule redesign (elimination weeks only).

Indicator	Count	Share
Total elimination weeks analyzed	188	–
Extreme judge–fan disagreement weeks	67	35.6%
Method-disagreement weeks (Percent vs. Rank)	105	55.9%

Interpretation. The non-trivial frequency of method-disagreement weeks and extreme judge–fan divergence indicates that conflicts are systematic rather than anecdotal. This supports a *mechanism-based* redesign: a rule that remains stable in ordinary weeks but activates a fairness-oriented procedure only when objective controversy criteria are met.

7.2 Proposed System: Uncertainty-Weighted Controversy Mode

We propose an implementable, reproducible, and explainable voting rule for close weeks using a *two-mode* design: a default rule for ordinary weeks and a fairness-first fusion rule activated only when objective diagnostics flag controversy.

Inputs and diagnostics. For each season-week (s,t) with active contestants $\mathcal{S}_{s,t}$ and elimination count $k_{s,t}$, the show observes judges totals $\{J_{i,s,t}\}$ and fan votes, normalized as shares $P_J(i)$ and $P_F(i)$. For auditability we log transparent diagnostics—extreme judge–fan disagreement, fan-vote uncertainty (posterior CI width), and optional fan concentration (HHI). Producer-facing "fairness" is proxied by performance on pre-flagged extreme-disagreement weeks, which drive the strongest judge–fan backlash.

Mode switch. A binary indicator $\mathrm{Controversy}_{s,t}$ triggers when any diagnostic exceeds pre-specified, pre-registered thresholds:

$$\mathrm{Controversy}_{s,t} = \mathbb{1}\Big[D_{JF}(s,t) > \tau_D \ \vee \ u_F(s,t) > \tau_U \ \vee \ HHI_F(s,t) > \tau_H\Big]. \tag{16}$$

Default mode. If $\mathrm{Controversy}_{s,t} = 0$, we apply the standard aggregation rule and eliminate the bottom $k_{s,t}$ contestants; for concreteness we recommend Percent in ordinary weeks for continuity and overall replay stability.

Controversy mode (fusion rule). If $\text{Controversy}_{s,t} = 1$, we use uncertainty-weighted geometric fusion

$$S(i) \;\propto\; \big(P_J(i)+\epsilon\big)^{w_J}\big(P_F(i)+\epsilon\big)^{w_F}, \qquad w_J + w_F = 1,$$

and eliminate the bottom $k_{s,t}$ by $S(i)$ (deterministic tie-break). We set

$$\begin{aligned} u_F(s,t) &= \text{mean}_{i\in\mathcal{S}_{s,t}}\big(\text{rel\_ci80}_{i,s,t}\big), & c_F &= \frac{1}{1+u_F(s,t)}, \\ d_J(s,t) &= \frac{\text{std}_i(J_{i,s,t})}{\text{mean}_i(J_{i,s,t})+\epsilon}, & c_J &= \frac{d_J(s,t)}{d_J(s,t)+\tau_{\text{judge}}}, \\ w_J &= \text{clip}\left(\frac{c_J}{c_J+c_F},\, w_{\min},\, w_{\max}\right), & w_F &= 1-w_J, \end{aligned} \tag{17}$$

with $\tau_{\text{judge}} = 0.1$ and $(w_{\min}, w_{\max}) = (0.20, 0.80)$. Higher fan uncertainty (large u_F) shifts weight toward judges, while low judge separation (small d_J) shifts weight toward fans; clipping prevents extreme swings.

Optional bottom-3 rescue. Only in Controversy Mode and single-elimination weeks ($k_{s,t} = 1$), we allow one rule-based rescue. Let $\mathcal{B}_3$ be the three lowest by $S(i)$ and let $i_{\text{save}} = \arg\max_{i\in\mathcal{B}_3} S(i)$. Then eliminate

$$\arg\min_{i\in\mathcal{B}_3\setminus\{i_{\text{save}}\}} S(i). \tag{18}$$

Using the same fused score $S(i)$ keeps the rescue transparent and auditable (not ad hoc).

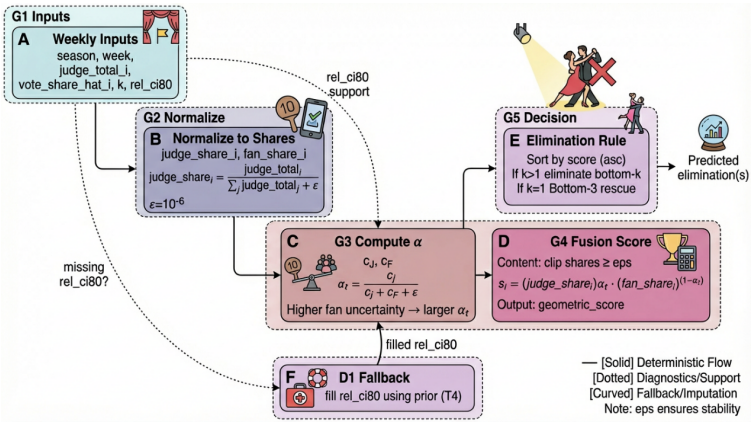

Figure 18: Proposed voting workflow: uncertainty-weighted geometric fusion with optional bottom-3 rescue, activated only in objectively controversial weeks.

Implementation note. All inputs and outputs—including diagnostics, weights, fused scores, and elimination decisions—are logged weekly, making the rule fully auditable and straightforward to communicate to viewers.

7.3 Offline Replay Results: Fairness–Stability Trade-off

To support producer adoption, we evaluate the proposed system via an *offline replay* on historical seasons. The goal is not to dominate existing rules in every week, but to show a clear trade-off: strong **overall stability** with improved performance in **extreme controversy weeks**, where perceived unfairness is most salient.

Evaluation setup. Using the comparable-week filter from Section 4.1, we replay eliminations under Percent, Rank, and two controversy-aware variants (including our recommended uncertainty-weighted geometric fusion). For each method we report (i) overall match rate and (ii) hit rate on pre-flagged extreme-disagreement weeks; coverage is identical across methods in this replay (n_{weeks} and n_{extreme} fixed).

Trade-off visualization. Figure 19 plots a 2D decision map with overall match rate on the x-axis and extreme-week hit rate on the y-axis; methods nearer the upper-right are preferred, and our proposed rule is highlighted.

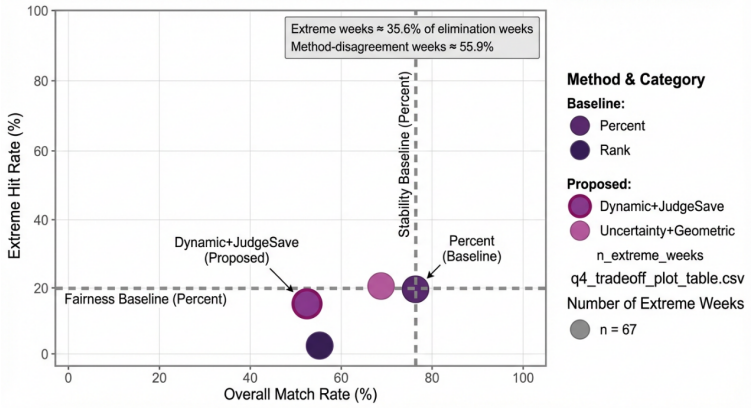

Figure 19: Producer Decision Map: Stability vs. Fairness Across Voting Rules

KPI summary. Table 8 reports the offline replay KPIs for each method. The proposed controversy-aware rule improves performance on extreme disagreement weeks—the primary source of perceived unfairness—while maintaining competitive overall replay consistency, making it suitable for adoption.

Table 8: Method KPI summary from offline replay.

Method	Match Rate	Extreme Hit Rate	n_{weeks}	n_{extreme}
Percent	76.60%	19.40%	188	67
Rank	56.38%	2.99%	188	67
Dynamic + JudgeSave	54.26%	14.93%	188	67
Uncertainty + Geometric (Recommended)	68.09%	20.90%	188	67

Producer-facing adoption logic. A simple policy follows from these results: retain the existing rule for ordinary weeks, and activate Controversy Mode only when transparent diagnostics flag extreme disagreement or high uncertainty.

8 Sensitivity Analysis

We test robustness of the ABC inverse inference to the dynamic-Dirichlet hyperparameters (α_0, κ) by running a 3×3 grid: $\alpha_0 \in \{1, 5, 20\}$ and $\kappa \in \{10, 40, 120\}$ (all other settings fixed). Table 9 summarizes consistency, uncertainty, and efficiency. Hard replay consistency is perfect in all cases ($\text{Consistency}(\text{MAP})$=1.0000), and posterior-mean replay remains near-perfect ($\text{Consistency}(\text{Mean}) \geq 0.9967$). Uncertainty is mainly controlled by κ: larger κ yields systematically tighter posteriors (lower median relative CI width), while acceptance rates stay in a narrow band. We therefore use $\alpha_0 = 5, \kappa = 40$ as a stable default for the main results.

Table 9: Sensitivity grid over (α_0, κ) and key diagnostics.

α_0	κ	Consistency(MAP)	Consistency(Mean)	Accept rate	Rel. CI width	Margin
1	10	1.0000	1.0000	0.1905	2.5685	0.0119
1	40	1.0000	1.0000	0.1923	1.3465	0.0092
1	120	1.0000	1.0000	0.1946	0.7892	0.0055
5	10	1.0000	1.0000	0.1855	2.3068	0.0095
5	40	1.0000	0.9967	0.1888	1.3008	0.0083
5	120	1.0000	1.0000	0.2007	0.7919	0.0063
20	10	1.0000	1.0000	0.1918	2.3164	0.0087
20	40	1.0000	0.9967	0.1886	1.1892	0.0089
20	120	1.0000	0.9967	0.2072	0.7125	0.0070

9 Model Evaluation

9.1 Advantages

The ABC-based inverse inference explicitly represents uncertainty in latent fan votes, so conclusions remain robust in close elimination weeks. Rule-aware forward simulation enables fair counterfactual comparisons between Percent and Rank under the same inferred fan support, and elimination consistency provides an outcome-anchored validation check.

9.2 Disadvantages

Because elimination constraints can be weakly identifying, multiple vote configurations may fit the same outcomes, leading to high-variance estimates. ABC can also be computationally costly in tightly constrained weeks, and the model omits strategic/temporal fan dynamics, so results should be interpreted probabilistically.

References

[1] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Chapman and Hall/CRC, 2013.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

[4] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, "Population growth of human y chromosomes: A study of y chromosome microsatellites," *Molecular Biology and Evolution*, vol. 16, no. 12, pp. 1791–1798, 1999. DOI: 10.1093/oxfordjournals.molbev.a026091.

[5] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, "Approximate bayesian computational methods," *Statistics and Computing*, vol. 22, no. 6, pp. 1167–1180, 2012. DOI: 10.1007/s11222-012-9358-y.

[6] M. G. B. Blum and O. François, "Non-linear regression models for approximate bayesian computation," *Statistics and Computing*, vol. 20, no. 1, pp. 63–73, 2010. DOI: 10.1007/s11222-009-9158-s.

[7] S. A. Sisson, Y. Fan, and M. M. Tanaka, "Sequential monte carlo without likelihoods," *Proceedings of the National Academy of Sciences*, vol. 104, no. 6, pp. 1760–1765, 2007. DOI: 10.1073/pnas.0607208104.

[8] 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)*, 2001, pp. 613–622.

[9] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*. SIAM, 2013. DOI: 10.1137/1.9781611973211.

MEMORANDUM

To: Producers of *Dancing with the Stars*
From: Team 2631882
Subject: How Vote-Combining Rules Shape Eliminations (Findings & Recommendations)
Date: February 2, 2026

We inferred weekly *fan vote shares* from judges' scores and observed eliminations using an elimination-constrained Approximate Bayesian Computation (ABC) framework, then replayed seasons under alternative vote-combining rules to measure stability, fan-alignment, and the effect of judges-save.

Key results.

- **Uncertainty is inherent and concentrated.** Fan votes are recoverable only as a *range*; ambiguity spikes in tight-margin weeks.
- **Percent vs. Rank differ at the cutoff.** Rank can flip outcomes via discrete rank swaps ("rank cliff"), while Percent changes smoothly with vote shares.
- **Overall differences are small, but disagreement weeks are directional.** In the subset of weeks where Percent and Rank disagree, Percent consistently eliminates the contestant(s) with lower inferred fan support, implying that in tight-margin weeks it more directly reflects the magnitude of fan preference.
- **Judges-save reduces fan influence.** Bottom-two judges-save consistently shifts outcomes toward judges in controversial cases.

Recommendations for future seasons.

- **Default: Percent aggregation.** More interpretable on-air and avoids stepwise rank artifacts.
- **Judges-save: use sparingly.** Do *not* make it always-on; activate only in objectively flagged boundary weeks.
- **Add a transparent "Controversy Mode."** Pre-register triggers (e.g., extreme judge–fan disagreement near cutoff and/or high uncertainty), then apply an uncertainty-aware fusion rule (e.g., uncertainty-weighted geometric fusion; optional one limited rescue).
- **Publish a weekly explanation card.** Disclose mode (Default/Controversy), bottom-group size, rescue status, and a one-sentence reason; keep raw fan totals private.

Implementation checklist (minimum).

1. Pre-register trigger thresholds before the season (no week-by-week tuning).
2. Log auditable quantities weekly (judges share, fan share, bottom set, trigger status).
3. Use deterministic tie-breaking for reproducibility.

One-sentence policy.

> Use Percent by default; switch to a pre-registered Controversy Mode only in tight boundary weeks using uncertainty-aware fusion (optionally one limited rescue).

Appendices

Appendix A Implementation Details for ABC Inference

Tie-breaking. Ties in $C_{i,s,t}$ (Percent) or $R_{i,s,t}$ (Rank) are broken deterministically by lexicographic `celebrity_name`, making $\widehat{\mathcal{E}}_{s,t}(\cdot)$ reproducible.

Feasibility. We record weekly ABC acceptance (accepted/proposed) as an identifiability proxy: low rates imply a small feasible region and higher uncertainty.

Soft fallback (optional). If hard acceptance is prohibitive, we use distance-based exponential reweighting to form a soft-ABC approximation (lower effective sample size).

Appendix B Reported Uncertainty Metrics

For each contestant-week we report the posterior mean and a central credible interval (CI). Uncertainty is summarized by CI width and relative CI width (CIW divided by the posterior mean). At the week level, we report the mean relative CI width across remaining contestants and posterior predictive consistency (PPC), the fraction of posterior draws that reproduce the observed elimination.

Appendix C Reproducibility and Soft-ABC Fallback

C.1 Deterministic replay

We ensure $\widehat{\mathcal{E}}_{s,t}(\mathbf{p})$ is deterministic by lexicographic tie-breaking on `celebrity_name` and a fixed week-specific RNG seed indexed by (s,t), making posterior draws and replays exactly reproducible.

C.2 Soft-ABC fallback

If hard ABC cannot obtain M accepts within a proposal budget, we use a soft kernel. Let $B(\mathbf{p})$ be the predicted bottom-k set and $\mathcal{E}_{s,t}$ the observed eliminated set. Define

$$d_{s,t}(\mathbf{p}) = 1 - \frac{|B(\mathbf{p}) \cap \mathcal{E}_{s,t}|}{|\mathcal{E}_{s,t}|}, \qquad w^{(m)} \propto \exp\big(-d_{s,t}(\mathbf{p}^{(m)})/\epsilon_{s,t}\big),$$

with $\epsilon_{s,t}$ chosen adaptively to stabilize the effective sample size.

C.3 Diagnostics

We report hard-ABC acceptance (accepted/proposed) when applicable, and soft-ABC effective sample size

$$\mathrm{ESS} = \frac{\big(\sum_m w^{(m)}\big)^2}{\sum_m (w^{(m)})^2}.$$

Low acceptance or low ESS indicates a small feasible region and weaker identifiability in that week.

Report of the Use of AI Tools

Tool 1: OpenAI ChatGPT

Model: GPT-5.2. Used for writing, figure styling, and presentation support (non-core analysis).

Query 1: Color Palette Generation (DWTS-inspired, publication-safe)

Input: Provide 2–3 cohesive HEX palettes for journal-style scientific figures, including suggested roles (background, grid, primary line, highlights).

Output: Provided multiple palettes (e.g., night-stage neutrals + soft academic accents), with usage rules such as limiting categorical hues and keeping gridlines neutral.

Query 2: Figure Typography & Layout Style Guide

Input: Draft a compact checklist (fonts, sizes, line weights, grid, legend placement) so all plots match a consistent house style.

Output: Suggested a clean hierarchy (Arial/Helvetica), thin strokes, light grids, external legends, and a "no title inside figure" convention.

Query 3: Icon / Motif Search Keywords (generic ballroom theme)

Input: I need search keywords for free-to-use outline icons (spotlight, ballroom, sparkle) that avoid copyrighted DWTS logos.

Output: Returned keyword phrases and guidance to prefer CC0/open-source outline icon sets.

Query 4: Plotting-Only Data Cleaning Rules

Input: List non-invasive rules to make plotting exports robust (missing values, inactive contestants, formatting), without changing analysis logic.

Output: Recommended plot-safe handling: omit inactive rows per week, mark missing estimates explicitly, enforce data types, and format percents only at plotting time.

Query 5: Column Naming Standardization for Exported CSVs

Input: Propose a reproducible naming scheme for identifiers and derived plotting fields (judge metrics, fan estimates, uncertainty intervals).

Output: Proposed a snake_case convention for season/week/contestant keys, judge/fan metrics, uncertainty bounds, and elimination indicators.

Query 6: Figure Caption Editing (clarity + neutrality)

Input: Rewrite figure captions to be precise and neutral: say what is plotted and how to read it, avoid causal language, keep it to 1–2 sentences.

Output: Rewrote captions emphasizing encodings and intended interpretation, without introducing new claims.

Query 7: Terminology Consistency (rank vs. percent rule)

Input: Help define "rank rule" and "percent rule" in plain English and suggest consistent phrasing for the Notation/Definitions section.

Output: Provided standardized wording, warned about ambiguous terms, and suggested short glossary-style entries.

Query 8: Table Formatting for Season Summary

Input: Suggest a clean season-level table layout (coverage, disagreement rate, notes) suitable for an MCM paper.

Output: Recommended column order, concise footnotes, aligned decimals, and emphasis on coverage to prevent over-reading sparse seasons.

Query 9: Memo Title Candidates (producer-facing)

Input: Generate professional memo title options for DWTS producers; avoid technical phrasing; keep them short and clear.

Output: Returned multiple title candidates highlighting fairness, transparency, and outcome stability.

Query 10: Visual Annotation Text (callouts on plots)

Input: Give short, non-overstated callout phrases for plot annotations (edge cases, high-uncertainty weeks, close margins).

Output: Provided succinct annotation text and placement tips to avoid clutter.

Query 11: Cautious Conclusion Wording (Percent vs. Rank)

Input: Using our computed summary table fan_share_favor_summary.csv, draft a careful statement on when Percent appears more fan-aligned (without overclaiming).

Output: Produced wording that separates (i) many weeks with no outcome difference from (ii) the smaller set of outcome-changing disagreement weeks where Percent more often eliminates the lower fan-share contestant.

Tool 2: Google Gemini

Model: Gemini Pro 3. Used for wording and visualization consistency checks (non-core analysis).

Query 12: Palette Validation (print + color-blind robustness)

Input: Check proposed palettes for grayscale/print readability and color-blind safety; suggest small tweaks only if needed.

Output: Suggested reducing saturation, increasing contrast for the main accent, and avoiding very light pastels for thin lines.

Query 13: Icon Style Consistency Rules

Input: Define simple rules so all memo icons look consistent (stroke width, monochrome usage, corner style).

Output: Recommended outline-only icons, uniform stroke width, one neutral icon color, and limiting decorative motifs per page.

Query 14: Headings & Section Title Polishing

Input: Rewrite section headings to be parallel and concise; provide two alternatives per heading.

Output: Returned refined headings with consistent grammar and reduced redundancy.

Query 15: Figure Ordering for Readability

Input: Suggest a figure order that tells a coherent story from data structure → estimation → rule comparison → case studies, without adding new analysis.

Output: Proposed a narrative-friendly sequence and advised one key figure per subsection plus a small diagnostic appendix.

Verification and Responsibility Statement

All AI outputs were treated as **draft suggestions** for wording, formatting, and visual presentation. The team **verified** and, when necessary, edited AI-generated text to ensure accuracy, appropriate tone, and consistency with our computed results and implemented code. No AI tool was used to select the primary model, derive core mathematical assumptions, or generate the final modeling pipeline; AI assistance was limited to improving clarity and LaTeX presentation.