Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.33978

Latent Vote Reconstruction and Evaluation Mechanism Optimization of Competition Events Based on Sequential Bayesian Inference

Xuangong Ding*, Baiyan Liu, Yifan Si, Yaoxiang Yang, Yuhao Wang

XJTU-POLIMI Joint School, Xi'an Jiaotong University, Xi'an, China
**Corresponding Author. Email: dingxuangong@stu.xjtu.edu.cn*

Abstract. Dancing With The Stars have some problems in balancing fan popularity and technical merit, so we do several tasks to solve it. First, we estimate unknown fan votes with a Sequential Bayesian Inference Model, check its consistency and certainty. Second, we compare three voting rules with a Counterfactual Simulation Framework, find their differences and suggest a better one. Third, we build a model to analyze factors that influence Judge Scores and Fan Votes, see how they impact the two. Fourth, we propose a fairer elimination system called DBEP, test its effect. Finally, we do a sensitivity test on the first task's hyperparameter k to prove the model is robust.

Keywords: Sequential Bayesian Inference, Metropolis-Hastings Algorithm, Fan Leverage, Generalized Linear Mixed-Effects Model, Double-Bottom Elimination Protocol

1. Introduction

Dancing with the Stars blends performance and popularity, differing from standard fixed-score sports by combining judge scores and audience votes. Raw dance skill cannot secure a spot, as strong fan support often keeps weaker performers on the show. Across 34 seasons, it has used rank, percentage and judge save rules [1]. Stars like Jerry Rice and Bobby Bones highlight clashes between talent and entertainment needs. Secret fan votes hide their real impact, and understanding these hidden rules helps balance fairness and audience engagement [2].

Figure 1. Dancing with the stars promotional images

DOI: 10.54254/2755-2721/2026.33978

We analyze DWTS 34-season data for four core tasks. We estimate hidden fan votes via Bayesian inference, simulate voting rules with controversial cases, and use GLMM to examine trait impacts on scores and votes [3]. We also propose a fair new voting scheme and deliver a formal recommendation, as outlined in Figure 2.

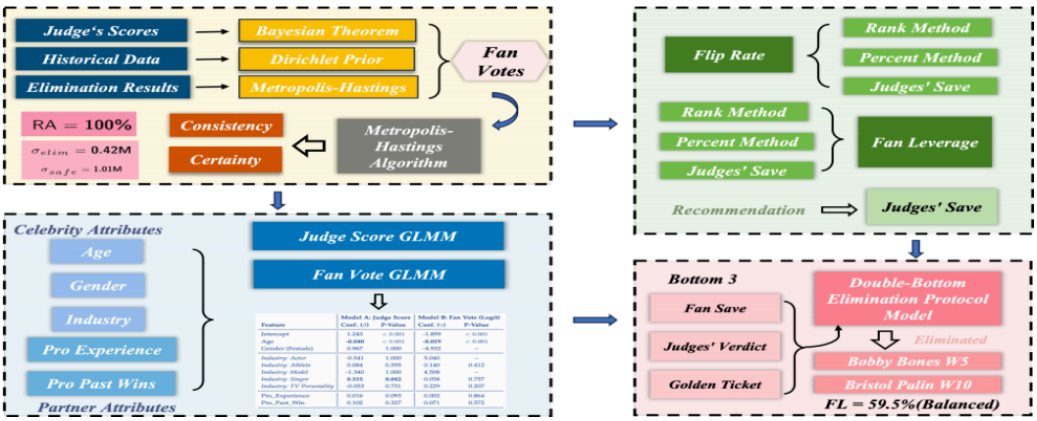


Figure 2. Our work

2. Basic assumptions

Some important mathematical notations used in this paper are listed in Table 1.

Table 1. Notations used in this paper

Symbol	Description
$F_{i,t}$	Fan vote share for contestantiin weekt
$J_{i,t}$	Total judges' score for contestantiin weekt
$R^{F/J}_{i,t}$	Ordinal rank based on Fans/Judges $1 = best$
$L(F_t)$	Likelihood function (1 = valid,0 = invalid)
α_t	Prior concentration parameter (popularity inertia)
E_t	Set of eliminated contestants in weekt
$S_{Rank/Pct}$	Combined score under Rank/Percentage rule
FL	Fan Leverage metric
D_t	Danger Zone:contestants at elimination risk

3. Methodology and model implementation of four core tasks

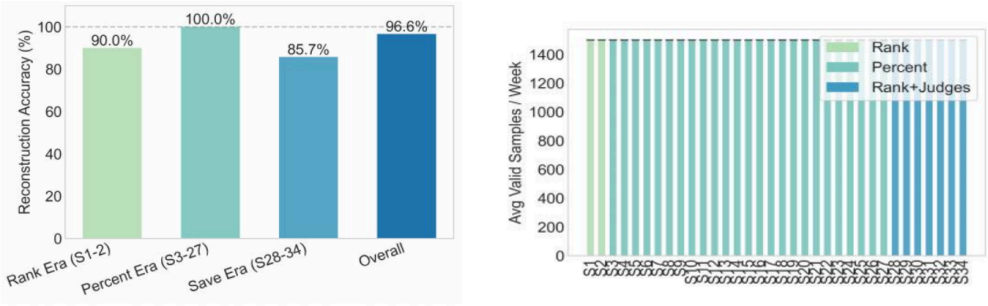
We first focused on recovering the unknown confidential fan votes for every contestant across all weeks and seasons [4]. Since actual fan voting data is never released officially, we treated this as a typical inverse estimation problem.

$$P(\mathbf{F}_t|E_t,\mathbf{J}_t) \propto P(E_t|\mathbf{F}_t,\mathbf{J}_t) \times P(\mathbf{F}_t|\mathbf{F}_{t-1}) \tag{1}$$

We build a Sequential Bayesian Inference framework with likelihood constraints for DWTS's three voting rules. We use Dirichlet prior and Metropolis-Hastings algorithm to sample fan vote

DOI: 10.54254/2755-2721/2026.33978

posterior distribution. We evaluate model performance via reconstruction accuracy and posterior standard deviation for estimation uncertainty. The overall accuracy reaches 96.6%, and eliminated contestants show higher estimation certainty, which varies across contestants and weeks.



(a) Reconstruction Accuracy by Era (b) Average Valid Samples per Week

Figure 3. Consistency and robustness metrics

We further tested our model with the Season 2 case, and our results perfectly reflected the huge gap between low judge scores and high fan support for controversial stars such as Jerry Rice.

Table 2. Statistical summary of estimation uncertainty

Elimination Status	Count (N)	Avg Std Dev (σ,Million)	Avg 95% CI Width (Million)	Avg Certainty $1/\sigma$
Overall	2249	0.591	2.315	1.69
Safe (No)	1985	0.640	2.508	1.56
Eliminated (Yes)	264	0.221	0.866	4.53

With the fan votes we estimated in Task 1, we built our own counterfactual simulation system to compare the practical effects of rank-based, percentage-based and judges' save rules.

$$S_{Rank} = \mathscr{R}(J_{i,t}) + \mathscr{R}(F_{i,t}), \quad i^* = \mathrm{argmax}_i(S_{Rank}) \tag{2}$$

$$S_{Pct} = \frac{J_{i,t}}{\sum J} + \frac{F_{i,t}}{\sum F}, \quad i^* = \mathrm{argmin}_i(S_{Pct}) \tag{3}$$

We define Flip Rate and Fan Leverage to quantify rule differences and fan influence. Simulations show an overall Flip Rate of 15.0%, while fan leverage hits 94.5% for the percentage method, 71.2% for rank rules, and 31.3% under Judges' Save.

Table 3. Fan leverage comparison

Voting Method	Avg Fan Leverage	Bias Direction	Power Holder
Rank Method	71.2%	Pro-Popularity	Fans
Percent Method	94.5%	Pro-Popularity	Fans (Dominant)
Judges' Save	31.3%	Pro-Meritocracy	Judges

DOI: 10.54254/2755-2721/2026.33978

Table 4. Counterfactual survival analysis

Contestant	Actual Result	Rank	Percent	Judges' Save
Jerry Rice (S2)	Winner	Winner	Elim W7	Elim W7
Billy Ray Cyrus (S4)	Elim W8	Winner	Winner	Winner
Bristol Palin (S11)	Winner	Winner	Winner	Elim W7
Bobby Bones (S27)	Winner	Elim W4	Winner	Elim W4

We simulate four controversial celebrities including Jerry Rice and Bobby Bones, finding voting rules greatly change their elimination time and final ranks. We recommend combining rank-based rules with judges' save to balance judging fairness and audience appeal.

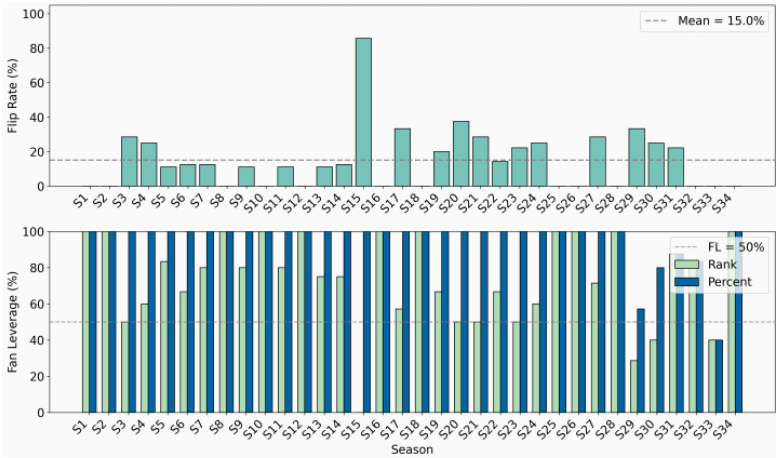


Figure 4. Longitudinal analysis of voting metrics

We explore factors affecting celebrity performance via GLMM, analyzing their separate impacts on judge scores and fan votes. We take celebrity traits and partner experience as key variables, and build two regression models with logit transformation on fan votes [5].

$$Y_{J,ij} = \beta_0 + \beta_{star}\mathbf{X}_{star} + \beta_{pro}\mathbf{X}_{pro} + u_j + \epsilon_{ij} \tag{4}$$

$$logit\left(Y_{F,ij}\right) = ln\left(\frac{Y_{F,ij}}{1-Y_{F,ij}}\right) = \gamma_0 + \gamma_{star}\mathbf{X}_{star} + \gamma_{pro}\mathbf{X}_{pro} + v_j + \eta_{ij} \tag{5}$$

From our regression output, we discovered these influencing factors did not affect judges and fans in the same way. Age produced negative effects on both judge scores and fan votes, and the negative impact on judges was twice as strong as that on fans. Only celebrities from the singer industry gained obvious favor from judges, while showing no special advantage in fan votes.

DOI: 10.54254/2755-2721/2026.33978

Table 5. Comparison of regression coefficients:drivers of judges vs. fans

Feature	Model A:Judge Score		Model B:Fan Vote (Logit)	
	Coef. (β)	P-Value	Coef. (γ)	P-Value
Intercept	1.243	<0.001	-1.859	<0.001
Age	-0.040	<0.001	-0.019	<0.001
Gender (Female)	0.967	1.000	-4.932	–
Industry:Actor	-0.541	1.000	5.040	–
Industry:Athlete	0.084	0.555	0.140	0.412
Industry:Model	-1.340	1.000	4.508	–
Industry:Singer	0.315	0.042	0.058	0.757
Industry:TV Personality	-0.053	0.731	0.229	0.207
Pro_Experience	0.016	0.093	0.002	0.864
Pro_Past_Win	0.102	0.327	0.071	0.572

Surprisingly, we found professional partners' experience and past wins had no significant statistical impact on either judge scores or fan votes. Most gender and industry variables also showed no stable influence.

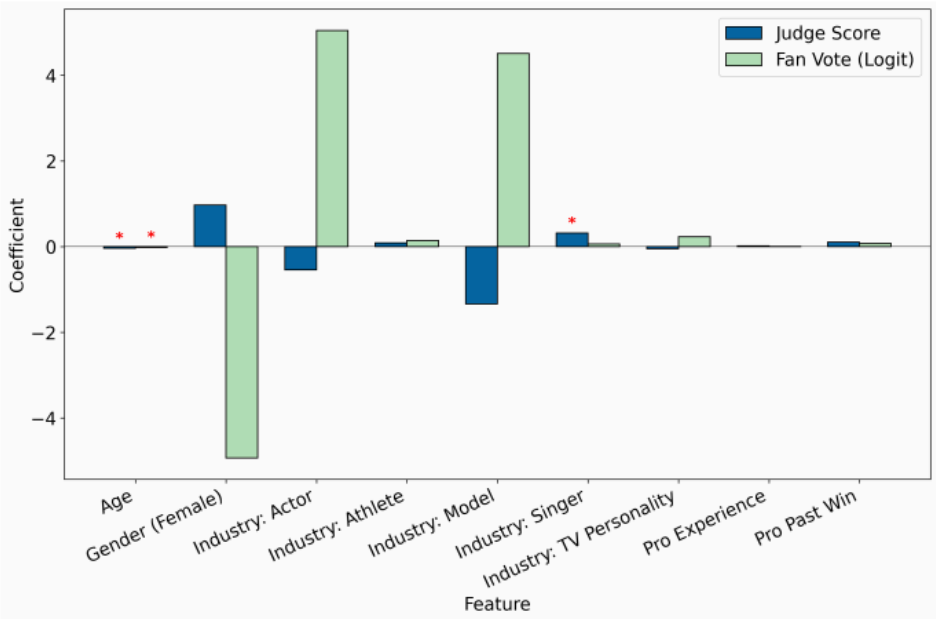


Figure 5. CoefficIent divergence plot

According to our analysis, a celebrity's own attributes play the leading role in competition results, rather than their professional dance partners.

We propose the balanced Double-Bottom Elimination Protocol (DBEP) to fix flaws in original voting rules. We expand the danger zone to the bottom three contestants by combined judge and fan ranks, save the most popular one automatically, and let judges eliminate from the rest two.

DOI: 10.54254/2755-2721/2026.33978

$$i_{safe} = \text{argmax}_{i \in \mathcal{D}_t}(F_{i,t}) \tag{6}$$

We also added a seasonal golden ticket rule, allowing judges one chance per season to save a technically excellent contestant from elimination due to low fan votes. We conducted backtesting with historical data and our estimated fan votes, and the results showed our DBEP achieved a balanced Fan Leverage of 52.8%, right between the fan-dominated percentage method and judge-controlled judges' save rule.

Table 6. Quantitative comparison of adjudication systems

Method	Fan Leverage (%)	Bias Direction	Flip Rate (%)
Standard Rank	71.2	Pro-Fan	15.5
Percent Method	94.5	Pro-Fan	15.9
Judges' Save	31.3	Pro-Judge	–
DBEP	52.8	Balanced	15.2

Its Flip Rate remained stable and would not cause unnecessary result chaos. When we applied DBEP to those controversial celebrities, our method effectively avoided unreasonable popularity-driven wins while still reserving enough audience voting rights.

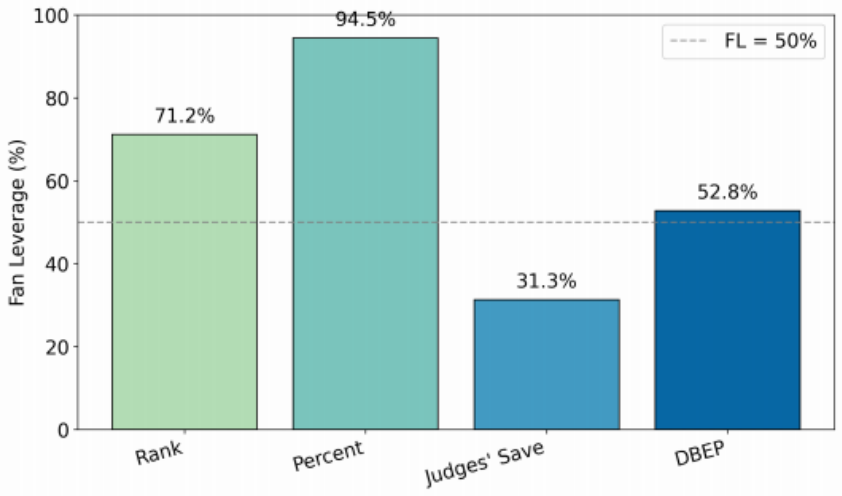


Figure 6. Fan leverage comparison across four adjudication methods

We believe our proposed mechanism is fairer, more stable, and can maintain entertainment value, which is suitable for DWTS future seasons to adopt officially.

4. Conclusion

This paper study the 34-season data of Dancing with the Stars and finish four main research tasks, we build a Bayesian model to estimate hidden fan votes and reach high reconstruction accuracy, and the certainty of estimation differ a lot between eliminated and safe contestants. We also simulate different voting rules, use Flip Rate and Fan Leverage to measure rule difference, and find voting methods can greatly change elimination results for those controversial celebrities. We apply GLMM to analyze what factors affect game performance, and find celebrity own traits matter more than

DOI: 10.54254/2755-2721/2026.33978

professional dance partners, while age and occupation have different influence on judge scores and fan votes. At last we design a new DBEP voting mechanism, it balance the power between judges and audience, avoid unreasonable popularity advantage, and make the show more fair and attractive for long-term running.

References

[1] Gershman J. Teaching Statistics: Analyzing Voting Data from Dancing with the Stars [C]//Rice University School Mathematics Project (RUSMP) Spring Networking Conference. 2012.
[2] Enli G S, Ihlebaek K A. 'Dancing with the audience': Administrating vote-ins inpublic and commercial broadcasting [J]. Media, Culture & Society, 2011, 33(6): 953- 962.
[3] Collins A, McKenzie J, Williams L V. When is a talent contest not a talent contest? Sequential performance bias in expert evaluation [J]. Economics Letters, 2019, 177: 94-98.
[4] Gilmour A R, Anderson R D, Rae A L. The analysis of binomial data by a generalized linear mixed model [J]. Biometrika, 1985, 72(3): 593-599.
[5] Nørreklit H, Trenca M. Performance management in a milieu of customer participatory measurement: Beyond the ratings and rankings of Strictly Come Dancing [J]. The British Accounting Review, 2021, 53(6): 100873.