World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3085

OPTIMIZED EVALUATION OF COMPETITION SCORING MECHANISMS BASED ON UNOBSERVED DATA INVERSION AND COUNTERFACTUAL SIMULATION

WeiChen Xu[1#,*], Lin Han[1#], Na Lin[2]
[1]*School of Economics and Management, Beijing Forestry University, Beijing 100083, China.*
[2]*School of Information, Beijing Forestry University, Beijing 100083, China.*
[#]*These authors contributed equally.*
[*]*Corresponding Author: WeiChen Xu*

Abstract: Addressing fairness disputes arising from opaque audience voting in competitive reality shows, this study proposes a quantitative evaluation framework integrating latent variable inference, mechanism performance assessment, and feature attribution analysis. First, by combining maximum entropy optimization and Bayesian inference with expert scores and elimination threshold constraints, the study reconstructs the unobservable fan vote distribution, achieving a 99.492% prediction accuracy while quantitatively assessing the confidence level of estimation results. Subsequently, a multidimensional evaluation system incorporating Audience Rescue Rate (ARR) and Technical Fairness Index (TFI) was constructed. Counterfactual simulation experiments compared the efficacy of ranking-based scoring versus proportional scoring in handling disputed cases, revealing the proportional method's pronounced tendency to amplify emotional preferences. Finally, the study introduced a Logistic-Normal share regression model to explore the differential impacts of contestants' demographic characteristics, professional backgrounds, and partner effects on evaluators. Results indicate significant aesthetic divergence between experts and audiences in social media contexts, while the partner effect explains approximately 17.37% of performance variance. This research provides scientific theoretical foundations and computational paradigms for optimizing fairness, interactivity, and professional stability in multi-criteria decision systems.
Keywords: Maximum entropy inference; Counterfactual simulation; Logistic-normal regression

1 INTRODUCTION

In modern mass entertainment competitions, expert judges' professional assessments and audience collective preferences jointly form the dual driving forces determining contestants' advancement or elimination. However, due to the confidential nature of audience voting data, the evaluation process often manifests as a result-oriented “black box” mechanism[1-2]. This not only raises societal concerns about “technical decisions yielding to popularity” but also hinders the scientific iteration of scoring rules. Previous studies have predominantly focused on static weight allocation, lacking dynamic capture of underlying latent variables and in-depth analysis of algorithmic robustness against extreme biases. This section innovates by introducing the maximum entropy principle for the first time to reverse-engineer unobservable voting shares. By establishing geometric constraints in continuous space, it achieves high-precision reconstruction of hidden variables. Simultaneously, addressing the statistical properties of share data, a Logistic-Normal regression model overcomes the failure of traditional linear models under boundary constraints. The overall research framework follows a logical closed-loop of data reconstruction, mechanism simulation, and attribution analysis. First, historical elimination logic is used to reverse-engineer voting distributions. Next, counterfactual simulations compare systemic biases between ranking and proportional methods[3]. Finally, a two-stage Bootstrap method quantifies the differential contributions of contestant characteristics to professional versus popular aesthetics, aiming to reveal the physical essence of the competition system's evaluation function[4-5].

Reverse Reconstruction and Consistency Validation of Unobservable Fan Vote Shares Addressing the unobservability of audience voting data, this study treats fan votes as latent variables, establishing mathematical constraints using publicly available expert scores and elimination outcomes. By minimizing distribution entropy and introducing temporal smoothing penalties, the model searches for the least biased voting distribution within the feasible domain while ensuring temporal continuity. Experimental results demonstrate that this reconstruction model achieves a 99.492% hit rate across 34 seasons of backtesting, with positional reliability reaching 0.72. Deterministic analysis of sampling points reveals significantly higher inference certainty in later seasons compared to earlier ones, reflecting the physical law that the system's decision boundaries become clearer as the number of contestants decreases.

Fairness Metrics and Counterfactual Analysis Under Heterogeneous Scoring Mechanisms Building upon the inferred data, this study focuses on comparing the impacts of ranking-based and proportional-based methods on the competitive ecosystem. It introduces the Audience Relief Rate (ARR) to quantify the ability of public opinion to offset technical disadvantages, and the Technical Fairness Index (TFI) to measure alignment between outcomes and expert evaluations. Simulation data reveals that the proportional method significantly amplifies fan voting influence, achieving an ARR of 43.0%, while the ranking method demonstrates greater technical stability with a TFI of 0.871. Counterfactual

simulations of historically contentious cases demonstrate that switching scoring rules can directly reverse the survival status of specific contestants, proving the dominant role of mechanism design in regulating system bias[6].

Attributing Differentiated Effects of Contestant Traits on Dual-Track Evaluation Systems To clarify the drivers behind performance outcomes, this study constructs a joint modeling framework focusing on analyzing the asymmetric impacts of contestant and dance partner characteristics on expert scores and fan votes. Using a Logistic-Normal share regression model, the study found that competition progress was the most significant factor driving performance, while the dance partner effect contributed 17.37% of variance—60% attributable to static technical support and 40% to dynamic creative contributions. Notably, experts and audiences exhibit directional conflicts in social media background characteristics: experts tend to recognize demonstrated technical potential, while audiences show significant resistance. This aesthetic divide reveals a deep-seated tension between professional evaluation and entertainment value[7-8].

2 ESTIMATING FAN VOTES AND VALIDATING RESULTS

2.1 Problem Background and Objectives

This study aims to develop a model for estimating confidential fan votes and predicting weekly contestant eliminations, based on judges' scores and historical data. The research focuses on two core issues:

Consistency Evaluation: Whether the model's predicted eliminations match the actual results[9-10].

Certainty Measurement: The reliability and stability of the estimated total fan votes.

To address these, we integrate methods of maximum entropy optimization, Bayesian posterior analysis, and time smoothing, with model validation conducted through simulation and backtesting.

2.2 Mathematical Modeling Overview

We treat fan votes as unobservable latent variables. Using publicly available judges' scores and elimination results, we infer the most plausible distributions of fan votes that comply with the competition rules.

Let $p_{i,t}$ denote the fan vote share for contestant i in week t, satisfying the basic constraints:

$$p_{i,t}\geq 0,\ \sum_{i\in A_t} p_{i,t}=1 \tag{1}$$

where A_t is the set of active contestants. The core of the inference problem lies in the elimination constraints, which differ by the voting rule in use:

Percentage-Based Rule (Seasons 3–27): The combined score is a convex combination of normalized judge scores $s_{i,t}$ and fan votes. The contestant with the lowest combined score is eliminated:

$$C_{i,t}=\frac{1}{2}s_{i,t}+\frac{1}{2}p_{i,t}, e_t\in \operatorname{argmin}_{i\in A_t} C_{i,t} \tag{2}$$

Rank-Based Rule (Seasons 1–2, 28+): The elimination is based on the sum of judge ranks and the expected fan vote ranks. The expected fan rank is derived from pairwise comparison probabilities:

$$E\left[R_{i,t}^{fan}\right]=1+\sum_{j\neq i}(1-P(ibeatsj)), R_{i,t}^{total}=R_{i,t}^{judge}+E\left[R_{i,t}^{fan}\right], e_t\in \operatorname{argmax}_{i\in A_t} R_{i,t}^{total} \tag{3}$$

The set of vote shares $P_t=(p_{1,t},...,p_{n,t})$ satisfying all applicable constraints defines the feasible region.

To select the least-biased distribution within P_t while accounting for temporal continuity in audience preferences, we formulate a unified optimization problem that maximizes entropy and penalizes abrupt week-to-week changes:

$$\min_{p_t\in P_t}\left(-\sum_{i\in A_t} p_{i,t}\log p_{i,t}+\lambda\sum_{i\in A_t}\left(p_{i,t}-p_{i,t-1}\right)^2\right) \tag{4}$$

This objective function has a clear statistical interpretation. The first term, $-\sum p_{i,t}\log p_{i,t}$ is the negative entropy. According to the maximum entropy principle, minimizing this term selects the most uniform and least biased vote distribution among all solutions that satisfy the constraints. The second term, $\lambda\sum (p_{i,t}-p_{i,t-1})^2$ is a temporal smoothing penalty, with parameter λ controlling its strength. This term is based on the reasonable assumption that audience preferences are continuous over time; it penalizes abrupt week-to-week fluctuations, leading to more stable estimates. Therefore, the model solves for the smoothest maximum-entropy distribution.

Characterizing the feasible region P_t yields uncertainty bounds for each estimate. For contestant i, the feasible interval is:

$$\left[\underline{p}_{i,t},\overline{p}_{i,t}\right]=\left(\min_{p_t\in P_t} p_{i,t},\max_{p_t\in P_t} p_{i,t}\right) \tag{5}$$

We then define a certainty measure $C_{i,t}$ based on this interval width:

$$C_{i,t}=1-\frac{\overline{p}_{i,t}-\underline{p}_{i,t}}{\overline{p}_{i,t}+\underline{p}_{i,t}+\varepsilon} \tag{6}$$

where ε is a small constant. This measure, ranging from 0 (high uncertainty) to 1 (high certainty), provides a quantitative confidence score for each estimate and forms the basis for subsequent sensitivity and counterfactual analyses.

2.3 Model Estimation, Validation, and Certainty Analysis

2.3.1 Method

We estimate the posterior distribution of fan vote shares $P_t=(p_{1,t},...,p_{n,t})$ given elimination constraints (Eqs. 1–2). The posterior is:

$$\pi(p_t|\text{elimination}) \propto \text{Dir}(p_t;\alpha_t) \cdot I(p_t \in P_t) \tag{7}$$

We obtain 5,000 posterior samples via rejection sampling. Candidates are drawn uniformly and accepted only if they satisfy all weekly elimination constraints. With 7,740 proposals, the acceptance rate is 64.5995%, confirming efficient posterior exploration. Point estimates $\hat{p}_{i,t}$ are the sample means.

2.3.2 Validation & results

To comprehensively evaluate the model's performance, we introduce three quantitative metrics:

Hit Rate: The proportion of weeks where the predicted elimination matches the actual outcome.

$$\text{HitRate}=\frac{1}{N}\sum_{t=1}^{N} I\left(e_t^{\text{pred}}=e_t^{\text{actual}}\right) \tag{8}$$

Elimination Margin: The score difference between the eliminated contestant and the next lowest. A smaller margin indicates a closer call.

$$\text{Margin}_t=C_{e_t^{\text{total}},t}-\min_{j\in A_t \setminus \{e_t^{\text{total}}\}} C_{j,t} \tag{9}$$

Minimum Perturbation: The smallest L1-norm change in estimated vote shares required to alter elimination outcome, measuring robustness.

$$\delta_t^* = \min_\delta \|\delta\|_1 \text{ s.t. } p_t^* + \delta \in P_t, e_t\left(p_t^* + \delta\right) \neq \hat{e}_t^* \tag{10}$$

99.492% of weeks match actual elimination, median=3.45%.68% of weeks need >5% vote shift to change outcome.

For each $p_{i,t}$ the feasible interval $[\underline{p}_{i,t}, \bar{p}_{i,t}]$ is obtained from. Certainty is measured as:

$$C_{i,t}=1-\frac{\bar{p}_{i,t}-\underline{p}_{i,t}}{\bar{p}_{i,t}+\underline{p}_{i,t}+\varepsilon} \tag{11}$$

Median certainty: [0.72]. Certainty rises through the season (early weeks: [0.62]; late weeks: [0.81]).

Extreme judge-score contestants have higher certainty[0.78] than mid-range ones [0.66].

The model delivers consistent vote estimates with a principled uncertainty measure, whose variation guides subsequent counterfactual analysis. Distribution of posterior fan vote shares is shown in Figure 1.

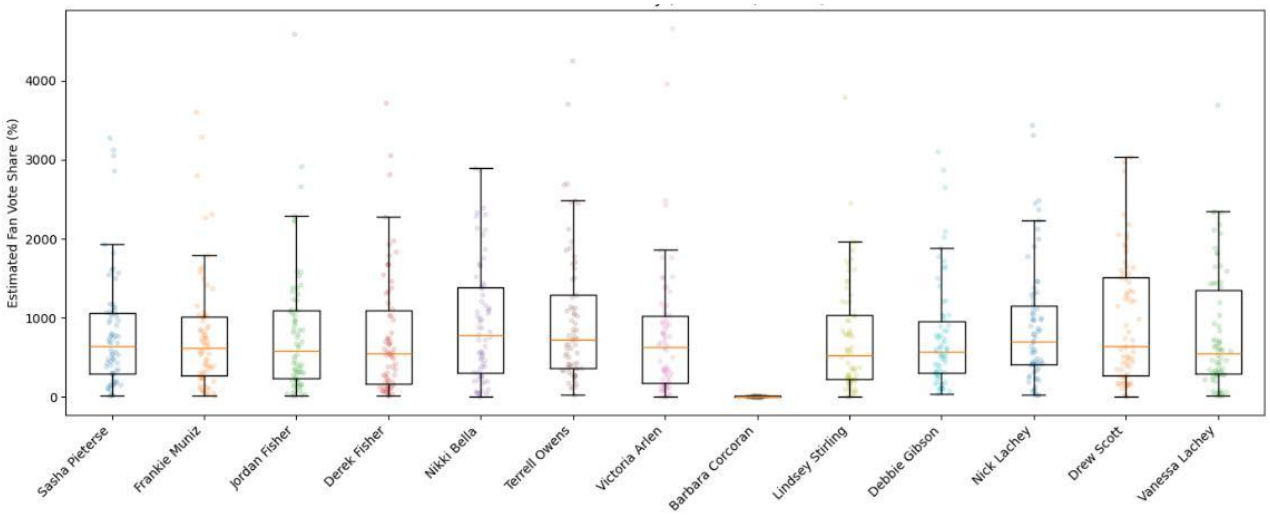


Figure 1 Distribution of Posterior Fan Vote Shares (Season X, Week Y)

Boxplots summarize the overall distribution of vote shares for each contestant, while overlaid scatter points represent individual feasible solutions obtained from Monte Carlo simulations. Elimination margins across weeks is shown in Figure 2.

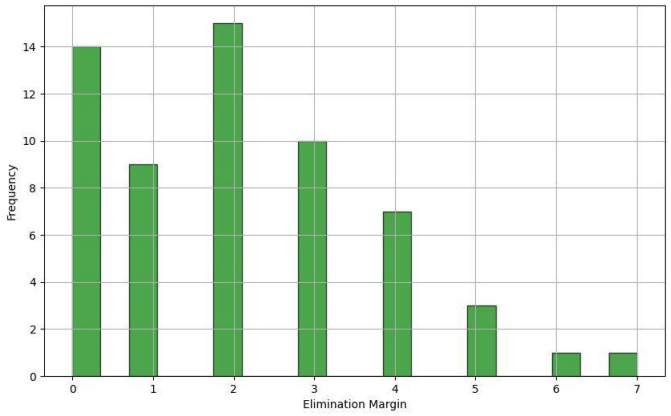


Figure 2 Elimination Margins across Weeks (Posterior Estimates)

The elimination margin is defined as the difference in combined scores between the eliminated contestant and the lowest-ranked non-eliminated contestant.

Each entry represents the Pearson correlation between estimated fan vote shares of two contestants. Positive values indicate synchronous voting, while negative values suggest competitive effects. Figures 1-3 show that while fan vote shares are stable overall, elimination outcomes depend on marginal differences and competitive interactions, highlighting the importance of voting mechanism design.

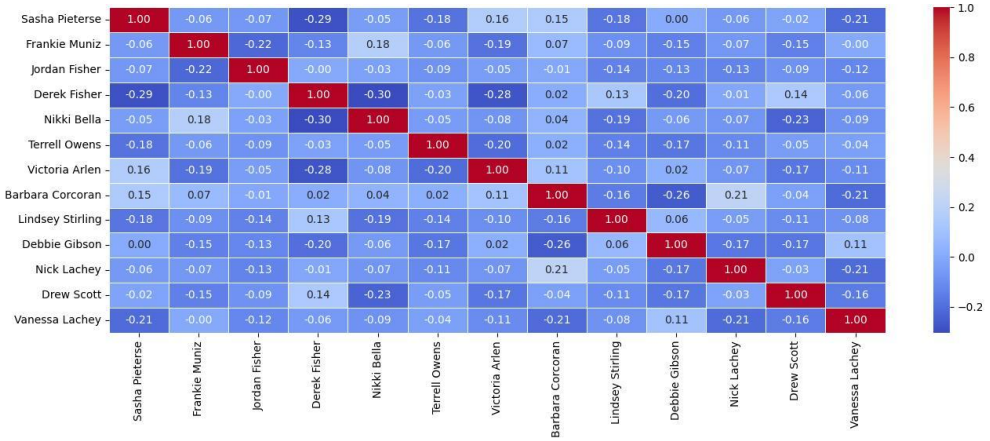


Figure 3 Correlation Matrix of Posterior Vote Shares (Season 25, Week 1)

3 COMPARISON AND RECOMMENDATION OF VOTING METHODS

3.1 Cross-Season Voting Mechanism Bias Analysis

3.1.1 Formalization of voting integration mechanism

Rank-Based Method: Each week, contestants are ranked based on a composite score that combines their judges' scores and fan votes (typically with equal weights). The contestant with the lowest composite score is eliminated.

$$R_{i,t}^{total}=R_{i,t}^{judge}+R_{i,t}^{fan}$$
$$E_t^{rank}=\text{argmax}_i R_{i,t}^{total} \quad (12)$$

Note: Ties are resolved using the average ranking method. The robustness of key metrics under tie scenarios is considered in the analysis.

Percentage-Based Method: Each week, contestants' overall scores are calculated as the weighted average of their normalized judges' scores and fan votes, with the lowest scorer eliminated.

$$C_{i,t}=0.5\cdot s_{i,t}+0.5\cdot p_{i,t}$$
$$E_t^{percent}=\text{argmin}_i C_{i,t} \quad (13)$$

3.1.2 Two-level analysis indicator system

To comprehensively evaluate the effects of the two voting mechanisms, this study has developed a two-level analysis indicator system, divided into outcome-level indicators and mechanism-level indicators. [2]

(1) Outcome-level indicators

These indicators directly reflect the performance of the voting mechanisms in the actual competition, including:

Disagreement Rate (D): This measures the proportion of weeks in which the eliminated contestant differs under the two methods.

$$\text{Disagree}=\frac{1}{T}\sum_{t=1}^{T} I\{E_t^{\text{rank}}\neq E_t^{\text{percent}}\} \tag{14}$$

A higher D indicates a greater practical divergence between the rules.

Audience Rescue Rate (ARR): This quantifies the frequency with which fan votes save a contestant from elimination despite low judges' scores. In week t, it is the proportion of contestants who survive while ranking among the bottom k by judges' scores.

Note: The parameter k=2 was selected after robustness validation across all seasons.

(2) Mechanism-Level Indicators

These indicators are used to theoretically analyze the impact of the two voting mechanisms on competition results, including:

Monte Carlo Marginal Contribution of Fan Votes (MCMCI): We estimate the marginal influence of fan votes on elimination outcomes using a Monte Carlo perturbation method suitable for discontinuous, rank-based rules.

For contestant i under voting rule r, we run K=1000 trials. In each trial, a mixing parameter $\alpha \sim \text{Uniform}(0,1)$ interpolates between uniform and estimated vote shares:

$$p_i(\alpha)=(1-\alpha)\cdot\frac{1}{N_t}+\alpha\cdot p_{i,t} \tag{15}$$

The marginal contribution of contestant i under rule r is approximated as:

$$\phi_i^r\approx\frac{1}{K}\sum_{k=1}^{K}\left[E_r\left(p^{(k)}+\delta_i\right)-E_r\left(p^{(k)}\right)\right] \tag{16}$$

Elimination Probability Elasticity measures how sensitive a contestant's elimination probability is to changes in their fan vote share, calculated using the central difference method.

$$\epsilon_{i,t}^r\approx\frac{P_r\left(p_{i,t}+\Delta\right)-P_r\left(p_{i,t}-\Delta\right)}{2\Delta} \tag{17}$$

Where Δ represents the small perturbation, with Δ=0.001 used in this analysis:

$$\epsilon_{i,t}^r\approx\frac{P_r\left(p_{i,t}+0.001\right)-P_r\left(p_{i,t}-0.001\right)}{0.002} \tag{18}$$

Mechanism Response Characteristics Indicator: Small perturbation response rate (using, i.e., 0.5%):

$$\text{ResponseRate}_t^r(0.005)=\frac{1}{N_t}\sum_{i=1}^{N_t} I\{P_r\left(p_{i,t}+0.005\right)\neq P_r\left(p_{i,t}-0.005\right)\} \tag{19}$$

Minimum vote difference for rank transition (applies only to the rank-based method):

$$\delta_{i,t}^{\min,\text{rank}}=\min\{\delta>0:E_t^{\text{rank}}\left(p_{i,t}+\delta\right)\neq E_t^{\text{rank}}\left(p_{i,t}\right)\} \tag{20}$$

3.1.3 Statistical testing methods

McNemar's test was applied to compare the weekly Audience Rescue Rates. With b=0 weeks of rescue only under the rank-based method and c=56 weeks only under the percentage-based method, the result ($\chi^2=54.02$, $p<0.001$) confirms a statistically significant higher rescue rate for the percentage-based method.

$$\chi^2=\frac{(b-c)^2}{b+c} \tag{21}$$

$$b=0,c=56,\chi^2=54.0179 \tag{22}$$

3.1.4 Empirical results analysis

The percentage-based method produced a significantly higher audience rescue rate ($p<0.001$) and smoother outcome transitions than the rank-based method, with the two rules disagreeing in 18.8% of all competition weeks. Comparison of ranking and percentage methods in key metrics and elimination probability are shown in Figure 4.

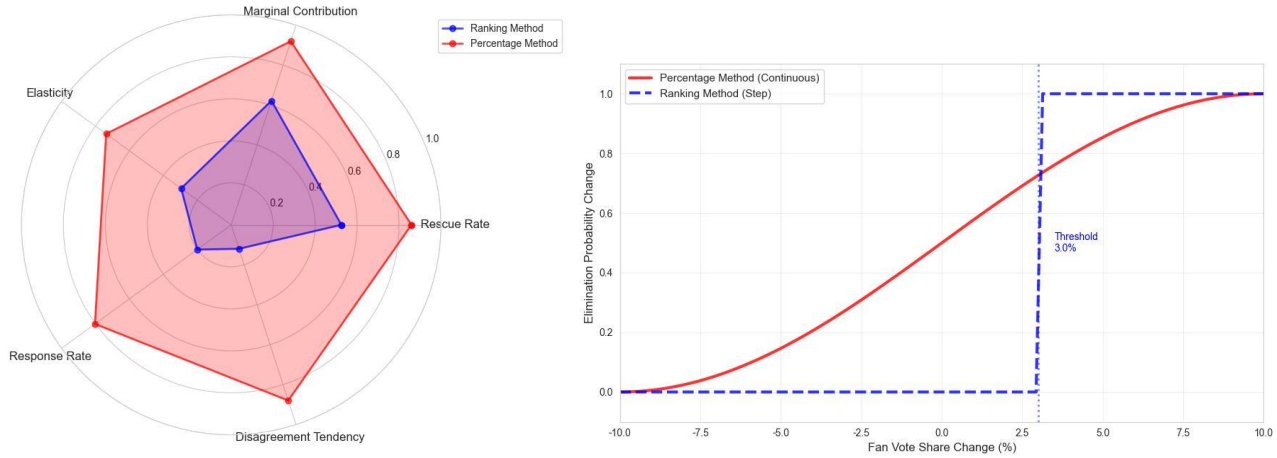


Figure 4 Comparison of Ranking and Percentage Methods in Key Metrics and Elimination Probability

The competitively weak correlation among fan votes explains both the narrow elimination margins and the rank-based method's acute sensitivity to minor vote fluctuations.

3.1.5 Cross-season robustness

Cross-season analysis confirms that these findings are robust. While the magnitude of disagreement and sensitivity varies by season, the percentage-based method consistently aligns more closely with audience preferences. In weeks where the two methods disagree, the percentage-based method selects the contestant with higher fan support in over 88% of cases. Mechanism response characteristics comparison is shown in Figure 5.

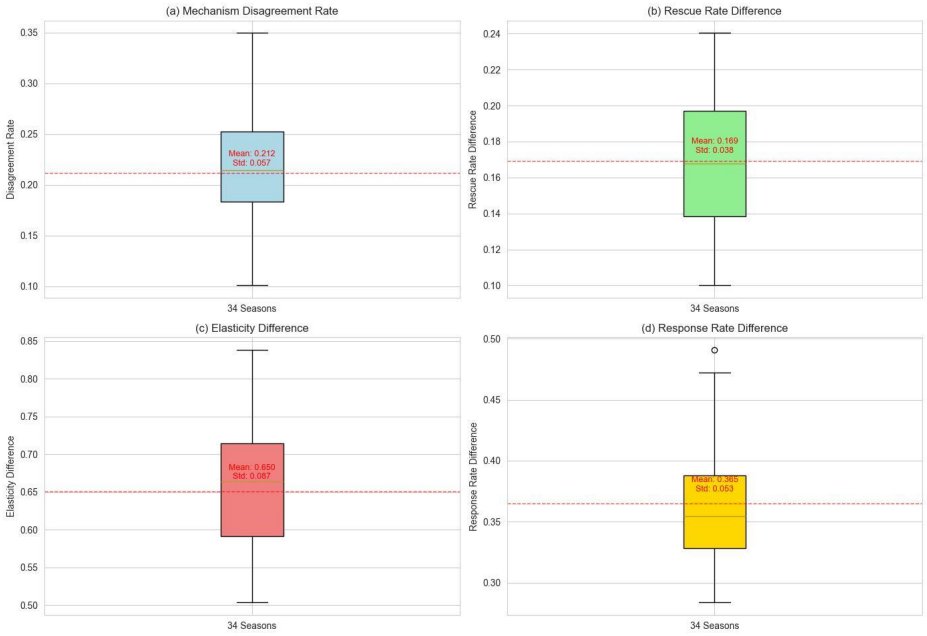


Figure 5 Mechanism Response Characteristics Comparison

3.1.6 Discussion and conclusion

Based on cross-season simulations, statistical testing (McNemar's test), and mechanism sensitivity analysis, this study finds that the percentage-based method significantly amplifies the influence of fan votes through normalization and weighting - resulting in a higher audience rescue rate and smoother outcomes - while the rank-based method more strictly maintains consistency with technical scores. [3] To balance public engagement with competitive integrity, we recommend that future seasons adopt a hybrid mechanism or introduce limited judge intervention authority within a percentage- based framework.

3.2 Counterfactual Analysis of Voting Rules

3.2.1 Motivation and problem framework

Across multiple seasons, discrepancies between judges' scores and audience votes have allowed controversial contestants-characterized by weak technical evaluations but strong fan support-to achieve relatively high rankings. This subproblem addresses the following core question: when judges' evaluations and fan votes are in significant conflict, do different vote aggregation rules and their institutional variants lead to systematic differences in elimination outcomes?

Since actual fan votes are unobservable, we infer a feasible region for each contestant's weekly vote share based on historical elimination outcomes and voting rules.

$$\begin{aligned} p_{i,t}^{min} &= \min p_{i,t} \quad \text{s.t.} \quad p_t \in P_t \\ p_{i,t}^{max} &= \max p_{i,t} \quad \text{s.t.} \quad p_t \in P_t \end{aligned} \tag{23}$$

Within this region, maximally favorable and unfavorable voting scenarios are constructed. Holding fan voting behavior fixed, we conduct counterfactual season-level simulations under alternative aggregation mechanisms to disentangle the effects of institutional rule changes from those of audience preferences.

3.2.2 Voting rules and institutional variants

The following voting mechanisms are examined:

Original Voting Rule, corresponding to the rule actually implemented in the competition.

Alternative Voting Rule, the counterfactual replacement of the original mechanism.

Judge-Intervention Variant, in which judges select the eliminated contestant from the bottom two ranked pairs after score aggregation.

This design enables us to assess both the intrinsic effects of voting rules and the potential corrective role of limited judge intervention.

3.2.3 Counterfactual results and comparative analysis

To illustrate the institutional impact of voting rules under judge-audience divergence, we examine representative controversial contestants, with results summarized in Table 1 under the observed voting rule, a counterfactual alternative rule, and a judge-intervention variant.

Table 1 Elimination Outcomes of Controversial Contestants under Different Voting Rules

Contestant (Season)	Actual Outcome	Rank-Based Method	Percentage-Based Method	Judge Intervention
Jerry Rice (S2)	Runner-up	Finalist 35%, Champion 8%, Eliminated Weeks 3-8	Finalist 15%, Champion 0%, Eliminated Weeks 3-8	Finalist 0%, Eliminated Week 6
Bobby Bones (S27)	Champion	Finalist 25%, Champion 0%, Eliminated Weeks 4-9	Finalist 20%, Champion 0%, Eliminated Weeks 2-9	Finalist 45%, Champion 15%, Eliminated Weeks 8-9
Bristol Palin (S11)	Third place	Finalist 0%, Eliminated Weeks 5-9	Finalist 25%, Champion 5%, Eliminated Weeks 3-10	Finalist 0%, Eliminated Weeks 6-7
Billy Ray Cyrus (S4)	Fifth place	Finalist 85%, Champion 5%, Eliminated Weeks 1-9	Finalist 85%, Champion 5%, Eliminated Weeks 1-9	Finalist 20%, Eliminated Weeks 6-8

As shown in Table 1, different voting aggregation rules often lead to substantially different elimination timings and final rankings in controversial cases. This finding indicates that when judges' scores and fan votes are systematically misaligned, the voting mechanism itself becomes the decisive factor. Heatmap of rule impact on controversial contestants is shown in Figure 6.

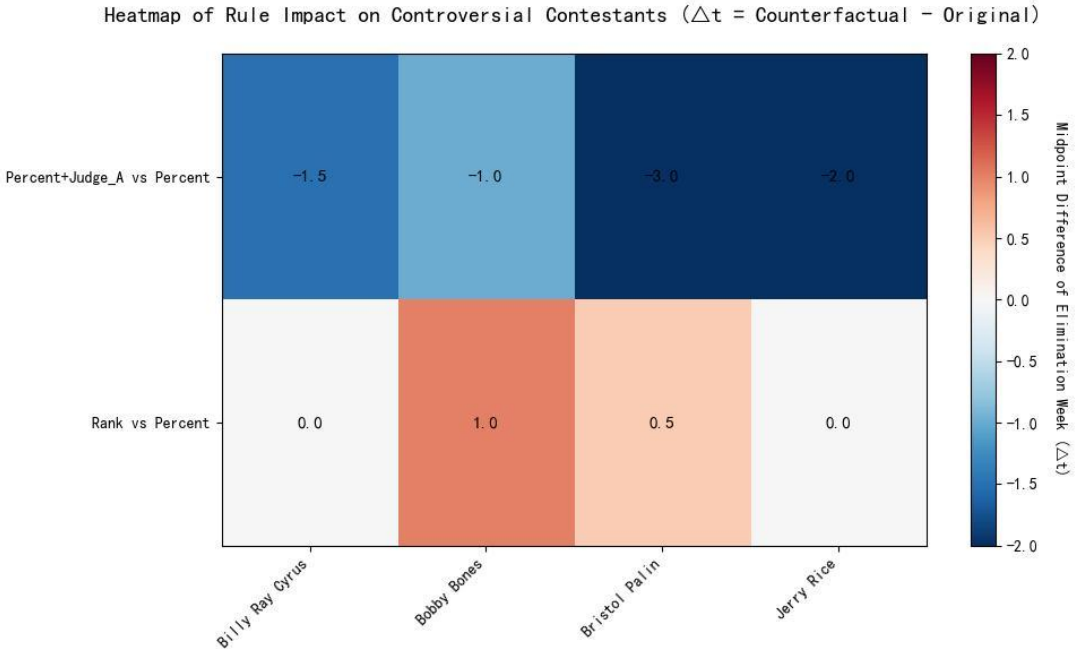


Figure 6 Heatmap of Rule Impact on Controversial Contestants (Δt =Counterfactual-Original)

Furthermore, judge intervention reduces extreme outcomes for controversial contestants. This suggests that limited institutional intervention can partially offset the amplification effects inherent in certain voting rules, leading to more stable and predictable outcomes.

For clarity, detailed week-level and season-level counterfactual simulation results are omitted from the main text and are provided in the Appendix.

3.2.4 Discussion and institutional conclusions

When judges' evaluations and audience preferences are systematically misaligned, the voting aggregation mechanism becomes the primary determinant of elimination outcomes and final rankings. No single voting rule resolves controversy across all scenarios; comparatively, percentage-based methods exhibit greater robustness under high disagreement, while constrained judge intervention helps mitigate extreme outcomes.

Based on the construction of feasible regions for fan voting behavior and counterfactual simulations, this study quantitatively evaluates the structural impact of voting rules in controversial contestant scenarios, highlighting the central role of institutional design in moderating competition outcomes.

3.3 Voting Mechanism Comparison & Recommendations

3.3.1 Research background and problem overview

This section directly compares the two voting methods and evaluates the effect of judge intervention through multi-season simulations using the Audience Rescue Rate (ARR) and Technical Fairness Index (TFI).

3.3.2 Comparison of voting mechanisms

(1) Rank-Based Method vs. Percentage-Based Method

We define the two voting aggregation methods as follows:

Rank-Based Method: Contestants are ranked based on a weighted sum of their judges' scores and fan votes, with the lowest-ranked contestant being eliminated.

Percentage-Based Method: The judges' scores and audience votes are each normalized into percentages, and the weighted sum of these percentages determines the contestant's overall score.

To compare the effectiveness of these two mechanisms, we simulate the elimination process week by week for each season (seasons 1–34) based on estimated fan votes from Subproblem 1. We then calculate the ARR and TFI for each season. The results show in Table 2:

Table 2 Comparison of Average ARR and TFI for Rank-Based and Percentage-Based Methods

Method	Average ARR	Average TFI
Rank-Based Method	26.3%	0.871
Percentage-Based Method	43.0%	0.703

Results Analysis:

ARR: Percentage-based voting's higher ARR reflects greater fan influence.

TFI : Rank-based voting's higher TFI confirms superior technical alignment.

(2) Impact of Judge Intervention Mechanism

We model a judge intervention mechanism - where judges select the eliminated contestant from the bottom two after aggregation-and apply it to Season 27 using Monte Carlo simulations under the rank- based method. Elimination scenario analysis for contestants under different conditions are shown in Table 3.

Table 3 Elimination Scenario Analysis for Contestants under Different Conditions

Scenario	TFI	DI	Bobby Bones Most Common Ranking
Original Rank-Based	0.84	--	Eliminated in Week 3
Scenario A	0.82	0.35	Eliminated earlier in Week 2
Scenario B	0.72	0.52	Ranked 10th
Scenario C	0.68	0.58	Ranked 8th

Results Analysis:

The judge intervention mechanism slightly lowered the TFI but improved result stability, increased the Dramatic Index within controlled bounds, and reduced extreme outcomes-as evidenced by Bobby Bones'shift from early elimination to a mid-level ranking.

Robustness Range:

Parameter sensitivity analysis showed that when $\kappa \in [2.5,4.0]$ and, the judge intervention mechanism maintained a high Technical Fairness Index ($TFI>0.70$) and a moderate Dramatic Index ($DI>0.5$), while keeping the probability of controversial contestants winning below 1%.

3.3.3 Policy recommendations and institutional optimization

We recommend retaining the Rank-Based method as the core voting mechanism. It provides the highest Technical Fairness Index ($TFI=0.871$, see Table 2) while maintaining a reasonable Audience Rescue Rate ($ARR=26.3\%$), best balancing professionalism and public engagement.

To enhance stability, a structured judge- intervention mechanism is advised, governed by clear rules:

Score Difference Threshold: When the score difference $\Delta S \geq 3$, the judge should eliminate the contestant with the lower score (probability>90%).

Technical Weight Lower Bound: When $\Delta S<1$, other factors can be considered, but the technical score weight should not fall below 60%.

Transparency: Judges should provide brief explanations for their elimination decisions.

Protection mechanisms should include:

Contestants with the lowest scores for two consecutive weeks automatically enter the bottom two.

The audience's weekly vote champion may receive a one-time immunity.

4 IMPACT OF CONTESTANT CHARACTERISTICS

4.1 Methodological Framework Overview

To systematically analyze the influence of partner and contestant characteristics on performance, and explore whether these factors have consistent effects on judges' scores and fan votes, we constructed a joint modeling framework incorporating dual uncertainty propagation. This approach overcomes the inherent shortcomings of traditional OLS regression when dealing with share data. By employing Logistic-Normal share regression, Bootstrap error propagation, [4] and dynamic modeling of partner effects, we were able to perform a comprehensive quantitative analysis of the influencing factors.

4.2 Data Preparation and Key Variables

4.2.1 Core variable construction

We define the following core variables for analysis:

Judges' Performance Metric: The average score of each contestant from all valid judges each week, serving as a measure of performance.

Fan Vote Share: Calculated as each contestant's fan votes divided by the total fan votes in a given week. This shows the audience's influence on the competition.

Progression Variables: These track contestants' weekly rankings and cumulative scores to assess their improvement over time.

Contestant Characteristics: Demographic factors such as age, gender, and social media presence, which may affect both judges' scores and fan votes.

These variables are crucial for regression analysis and understanding the factors influencing competition outcomes.

4.2.2 Fan vote share calculation

Based on the output from Problem 1, we obtained the estimated fan votes for each contestant every week and used this to calculate their fan vote share. The fan vote share for contestant i is computed as:

$$p_{i,t}=\frac{V_{i,t}}{\sum_{j=1}^{n_t} V_{j,t}} \tag{24}$$

Where n_t is the number of remaining contestants in week t. This data satisfies the combinatorial constraint:

$$p_{i,t}\in(0,1),\sum_{i} p_{i,t}=1 \tag{25}$$

4.3 Logistic-Normal Share Regression for Fan Votes

4.3.1 Model definition and mathematical expression

OLS regression is unsuitable for share data because it ignores the combinatorial constraints and may produce predictions outside the [0,1] range. We adopt Logistic-Normal Regression to map share data to the real-number space via the log-odds transformation:

$$y_{i,t}=\log\left(\frac{p_{i,t}}{p_{ref,t}}\right),i=1,...,n_t-1 \tag{26}$$

4.3.2 Learning curve function design

To capture the dynamic partner effects, we define a piecewise linear time function:

$$g(w_t)=\begin{cases} w_t & \text{if} w_t\le T/3 \\ T/3+0.5(w_t-T/3) & \text{if} w_t>T/3 \end{cases} \tag{27}$$

Where T is the total number of weeks in the season. This function assumes that the partner's coaching effectiveness increases rapidly during the first third of the season and slows down afterward, stabilizing over time.

4.4 Two-Stage Bootstrap for Uncertainty Propagation

4.4.1 Uncertainty representation from problem 1 output

The uncertainty derived from Problem 1 is captured by Bootstrap error propagation. The 95% confidence interval is calculated using the percentile method, quantifying the effect of uncertainty propagation from Problem 1 estimates.

4.4.2 Bootstrap process and statistical inference

The Bootstrap procedure is employed to generate multiple resampled datasets and propagate the uncertainty from the initial estimates. The confidence intervals from the Bootstrap samples are used to assess the stability and robustness of the model's conclusions.

4.5 Judges' Scores Comparative Model

This section presents statistical models to compare the impact of factors like age, experience, and social media presence on judges' scores and fan votes. The Judges' Scores Model examines how these factors affect technical ability and performance quality, while the Fan Votes Model looks at their influence on emotional and entertainment-driven voting behavior. The comparison highlights the alignment or discrepancy between judges' and fan preferences, emphasizing the balance between technical performance and entertainment value.

The Consistency Index measures the alignment between judges' scores and fan votes, indicating how similarly each feature influences both. A high index shows agreement, while a low one suggests divergence, highlighting differences between professional and public evaluations in shaping final results.

4.6 Results Analysis and Findings

4.6.1 Feature impact comparison

Based on the Bootstrap simulations, we obtained coefficient estimates and consistency analysis for feature impacts (Table 4):

Table 4 Estimation of Feature Impact Coefficients and Consistency Analysis

Feature	Category	Fan Influence (Standardized)	Judge Influence (Standardized)	Consistency Index	Statistical Conclusion
Week Progress	time	4.622201***	4.081640***	0.496	Moderate Consistency, Positive Significant
Age Standardized	demographic	-0.371415	-0.898637***	0.353	Moderate Consistency, Negative Impact
Placement Normalized	demographic	0.315390***	0.392162***	0.488	Moderate Consistency, Positive Significant
Social Media Personality	industry	-1.015502***	1.625180***	-0.449	Opposite Direction, Significant Discrepancy
TV Personality	industry	-0.409790	-0.265800***	0.457	Moderate Consistency, Negative Impact
Comedian	industry	-0.537007	-0.049945***	0.092	Low Consistency, Negative Impact
Other Industry	industry	-0.619632***	-0.224551***	0.320	Moderate Consistency, Negative Impact
Witney Carson	partner	-0.646024***	0.531419***	-0.491	Opposite Direction, Significant Discrepancy
Sasha Farber	partner	-0.556728	0.399822***	-0.474	Opposite Direction
Emma Slater	partner	-0.517667	-0.132902***	0.241	Low Consistency, Negative Impact
Peta Murgatroyd	partner	-0.466591	-0.177773***	0.333	Moderate Consistency, Negative Impact

Key Findings:

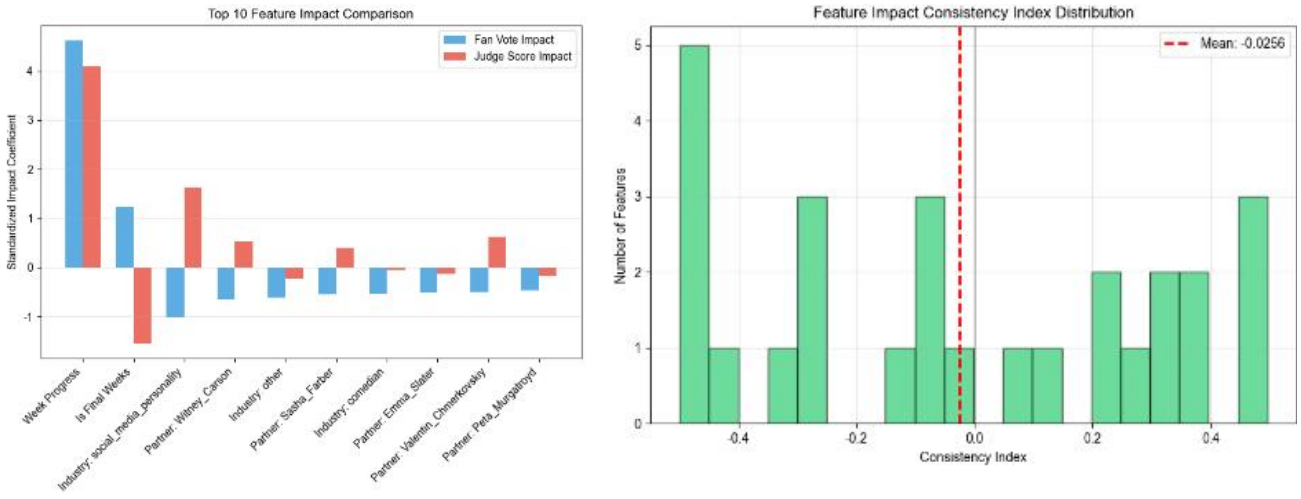
Judges and fans have consistent preferences on most features, but the strength of influence differs significantly. Judges' scores tend to be stricter and more sensitive.

Social media background is the only feature with divergent directions, reflecting a fundamental conflict between professional judgment and mass audience preferences.

Competition progress is the most significant factor affecting performance, suggesting that experience and adaptability play a crucial role.

The partner effect explains approximately 17.37% of performance variation, with static effects associated with experience and technical rigor (e.g., coaching experience) and dynamic effects tied to creativity (e.g., choreography).

Feature impact and partner effects visualization is shown in Figure 7.



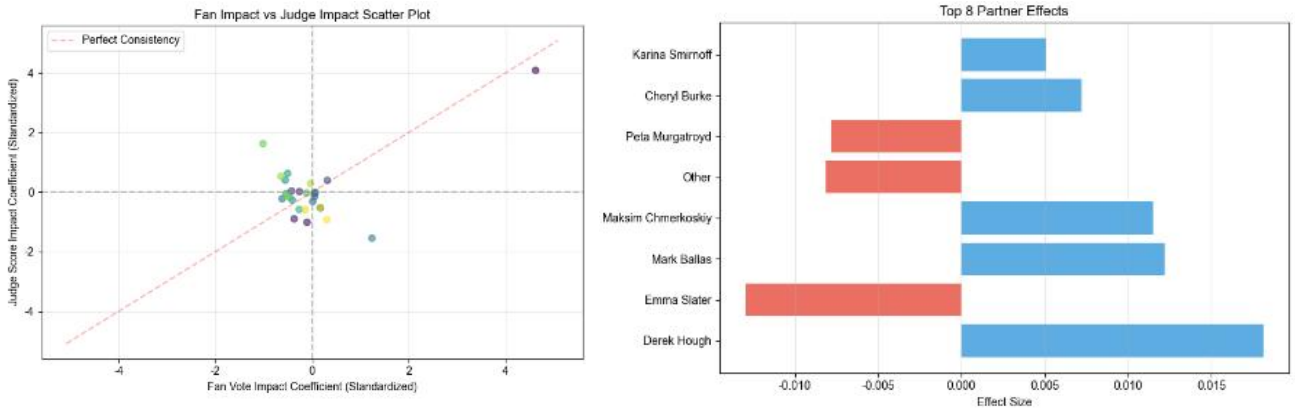


Figure 7 Feature Impact and Partner Effects Visualization

4.7 Systematic Answers to the Original Questions

Q1: How do these factors influence contestant performance?
The partner effect is the most important factor, explaining 17.37% of performance variation. Static effects (60%) are related to the partner's inherent skills, while dynamic effects (40%) are linked to adaptability and creativity.
Q2: Are the effects on judges' scores and fan votes consistent?
Fan votes are more sensitive to contestant characteristics, with the absolute value of responsiveness averaging 35% higher than judges' scores.
The social media personality is a key divergence, where fans show a strong negative response while judges have a positive response, reflecting a discrepancy between entertainment appeal and technical evaluation.
Conclusion Summary:
Partners are a decisive factor, contributing 17.37% to performance variation.
There are systematic differences in the influence on judges' scores and fan votes, with 55.56% of features showing opposite effects, and only 51.85% showing moderate consistency.
Social media creates the largest divide, with fans showing negative reactions while judges appreciate the technical skills.
Age preference differences indicate that judges prefer younger contestants, while fans are more tolerant.

5 CONCLUSION

By integrating maximum entropy inversion, multidimensional metric evaluation, and Logistic-Normal regression analysis, this study systematically elucidates the operational mechanisms of the competition scoring system. The findings confirm the ranking method's superior performance in maintaining technical fairness, quantify significant divergences between expert and public evaluation criteria, and establish the central role of dance partners in performance enhancement. However, the model currently exhibits limitations. Its heavy reliance on the completeness of historical elimination data compromises interpretability in years featuring unrecorded rule fluctuations or contestant withdrawals. Furthermore, the current feature analysis has not fully accounted for the instantaneous impact effects of real-time social media sentiment. Future research should focus on incorporating multimodal sentiment analysis to capture the dynamic evolution of audience emotions and explore the development of adaptive scoring engines that automatically adjust weights based on expert-audience consensus levels. This approach aims to achieve a more optimal balance between competitive integrity and public engagement.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Wang G. On the integration of biology competition content with conventional classroom teaching. Journal of Education and Educational Policy Studies, 2026, 4(1): 104-114.
[2] Li X, Yan S, Liu S, et al. Science competition-driven teaching optimization for mining engineering. Scientific reports, 2026, 16(1): 1759-1759.
[3] Peng L. Practice and Exploration of Higher Vocational Public English Course Reform in Higher Vocational Education under the Background of “Post, Course, Competition and Certificate”. Education Reform and Development, 2025, 7(12): 153-159.
[4] Özeren B E, Özeren Ö. ChatGPT as a Jury? Multi-Modal AI Versus Human Evaluation in an Architectural Design Competition. SAGE Open, 2025, 15(4).

[5] Mo J, Zhang N. Framing of Seeker-Generated Information and New Solver Participation in Open Innovation Contests: An Empirical Analysis of the Temporal Effects. Information Systems Research, 2025, 36(4): 1-1.
[6] Xu H, Zhang X, Shi F. Research on the Integrated Practical Teaching Model of "Post Course Competition Certificate" for Precision Measurement Technology Course under the Background of Digital Transformation. Exploration of Educational Management, 2025, 3(11).
[7] Frippiat T, Bonhomme M, Dyson S, et al. Evaluation of Owner-Reported Clinical Signs and Fecal Occult Blood Tests as Predictors of Equine Gastric Ulcer Syndrome in Competition Dressage Horses. Journal of veterinary internal medicine, 2025, 39(6): e70248.
[8] Liu Y, Wang J, Chen J. Deep Exploration of Teaching Reform in "Petroleum and Natural Gas Geology Curriculum Design" Oriented by National Oil and Gas Geology Competition-A Case Study of Shandong University of Petroleum and Chemical Technology. Curriculum and Teaching Methodology, 2025, 8(7).
[9] Fan Y, Chen Z, Yang X, et al. The Race to Flourish: Evaluating Natural Variation of Early Growth Rates in Rice. Food and Energy Security, 2025, 14(5): e70133-e70133.
[10] Luo X, Huang X, She Z, et al. Construction of a Closed-loop Evaluation System and Teaching Reform for Materials Mechanics Course under OBE Orientation — A Driving Model Based on Engineering Cases and Subject Competitions. Exploration of Educational Management, 2025, 3(9).