International Business & Economics Studies
ISSN 2640-9852 (Print) ISSN 2640-9860 (Online)
Vol. 8, No. 2, 2026
www.scholink.org/ojs/index.php/ibes

Original Paper

Reverse Reconstruction and Efficacy Evaluation of Dual-Track Scoring Paths in Competitive Reality Shows

Feiran Liu[1,a,*]

[1] Business School, Xi'an International Studies University, Xi'an, China

[a] 2792199754@qq.com

[*] Corresponding author

Received: January 26, 2026 Accepted: March 09, 2026 Online Published: March 28, 2026

doi:10.22158/ibes.v8n2p55 URL: http://dx.doi.org/10.22158/ibes.v8n2p55

Abstract

Since the premiere of "Dancing with the Stars," the tension between expert judging and public voting has remained the core contradiction within the competitive reality show evaluation system. This study constructs a focused quantitative framework to reconstruct unobservable voting distributions and evaluate the efficacy of different scoring "paths." Firstly, we developed a fan vote inversion model based on multiple linear regression to address the challenge of hidden variables. By mathematically tracing elimination sequences across 34 seasons, the model identifies the regression coefficient of judges' scores as a dominant predictor, achieving a 79.65% accuracy rate in historical backtesting. Secondly, we conducted a comparative simulation between the "Ranking Legitimacy Path" and the "Percentage Weighting Path." Our findings indicate that while the two paths achieve consensus in 82% of cases, the percentage path triggers a significant "traffic shield effect" under extreme polarization. Specifically, when the skewness of fan votes exceeds critical thresholds, the continuous nature of percentage weighting allows popularity to systematically override technical performance, leading to structural ranking inversions that the discrete ranking method effectively mitigates.

Keywords

Multiple Linear Regression, Dual-Track Evaluation System, Path Optimisation, Monte Carlo Simulation, DWTS

1. Introduction

1.1 Problem Background

As the world's most influential dance competition reality show, Dancing with the Stars has long employed a dual-track evaluation system combining expert judging and audience voting, seeking to

strike a balance between professional artistic standards and mass market appeal. However, this mechanism has repeatedly faced challenges in practice. The early ranking summation method once enabled Jerry Rice to overturn low professional scores through sheer popularity, while the percentage weighting system introduced in Season 27 triggered severe fairness crises due to extreme voting distributions. These historical cases demonstrate that traditional single aggregation logic struggles to reconcile the tension between professional depth and participatory breadth, leaving systemic failure risks ever-present.

Confronted with conflicting evaluation dimensions, reliance on superficial data analysis alone proves insufficient. Constructing mathematical models capable of quantifying dispute risks and dynamically balancing stakeholder interests becomes paramount. This research endeavours to deeply analyse the interplay between technical performance and popularity factors within historical data, providing producers with scientifically grounded proposals for competition format refinement. It further aims to offer universal theoretical support for mechanism design within diverse evaluation systems across non-sporting contests.

1.2 Restatement of The Problem

Having thoroughly analysed the context and taken existing constraints into account, we must address the following issues:

Problem 1: Reverse Reconstruction of Fan Voting Data

Constructing an inequality constraint system using publicly available judge scores and elimination outcomes, then reconstructing the unobservable fan voting distribution through mathematical inversion.

Problem 2: Comparative Analysis of Two Scoring Mechanisms

Quantitatively evaluating the efficacy differences between the “sum of rankings” and “percentage weighting” mechanisms in historical operations.

2. Assumptions and Justifications

To ensure the accuracy of the proposed model, we have made the following reasonable and well-founded assumptions:

★Assumption 1: We assume that audience support is fundamentally rooted in the technical performance of contestants as reflected in judges’ scores.

⇒ Justification: While off-stage factors exist, historical data shows a high correlation between technical improvement and vote retention. Focusing on observable scores allows the model to isolate the primary “direction” of public sentiment while filtering out stochastic noise.

★Assumption 2: The elimination process is assumed to strictly follow the publicized rule of discarding the couple with the lowest aggregate score.

⇒ Justification: For the purpose of reverse-engineering hidden voting intervals, we must assume no undisclosed “revival paths” or manual interventions by producers. This ensures the boundary conditions for the mathematical inversion are logically sound.

★Assumption 3: Fan voting behavior is assumed to exhibit collective rationality and statistical regularity rather than pure randomness.

⇒ Justification: The use of multiple linear regression requires the dependent variable to have a systematic response to covariates. Assuming a stable behavioral "path" allows us to identify the marginal effects of judges' scores and competition progression on the final vote count.

★Assumption 4: Residual terms in the predictive models are assumed to follow a normal distribution with a mean of zero, and the scoring "direction" of one week does not instantaneously alter the audience's baseline preference for the next.

⇒ Justification: This assumption satisfies the Gauss-Markov theorem requirements, ensuring that the OLS estimates remain the Best Linear Unbiased Estimator for our inversion model.

3. Notations

Table 1. This Caption Has One Line so it Is Centered

Symbol	Description
$Vote_{ij0}$	A categorical variable representing the specific season to account for temporal "paths."
$Score_{ij}$	The total score awarded by the panel of expert judges to contestant in week j.
$Rank_{ij}$	The ordinal position of contestant based on the judges' scores in week j.
$Season_j$	A categorical variable representing the specific season to account for temporal "paths."
ε_{ij}	The correction term introduced to align the model with the "last-place elimination" rule.
Δ_{ij}	The correction term introduced to align the model with the "last-place elimination" rule.
W_k	The dynamic weighting coefficient assigned to different evaluation "directions."
P_i	The percentage of the total vote pool or score pool occupied by contestant \$i\$.
S_{total}	The cumulative score representing the final outcome of the dual-track aggregation.

4. Data Pre-processing

Analysis of the existing dataset reveals significant gaps and logical redundancy in both the temporal span and variable distribution. To ensure the robustness of subsequent reverse-engineered models, we first undertook data preprocessing (Kuhn & Johnson, 2019).

◆ Outlier Handling

Certain scores in Appendix 2026_MCM_Problem_C_Data contain extreme values exceeding the compliance range. These should be adjusted to reasonable limits by referencing the scoring logic of other assessors for the same week. Furthermore, for logically inconsistent values violating common sense regarding "skill enhancement," the model retains the original data but annotates them via the newly added score trend field.

◆ Missing Value Handling

Attachment 2026_MCM_Problem_C_Data contains missing data, accounting for 64.15% of total variables. Upon verification, for reasonable missing values resulting from "failure to advance to later rounds," we uniformly imputed "0" and added an "is_eliminated" field for marking. Furthermore, for random missing values caused by "judge omission," we imputed using the average score from the other three judges for that week.

◆ Data Balancing Treatment

The celebrity home country field in the 2026_MCM_Problem_C_Data attachment exhibits severe data polarisation, with a minority category comprising less than 20% of samples. To address this, we employed the SMOTE algorithm to generate synthetic samples for the "low-proportion regions."

5. Model I: Fan Voting Inversion Model Based on Multiple Linear Regression.

5.1 Model Overview

To reconstruct the distribution of unobservable fan voting data, this study established a parametric inversion model based on multiple linear regression[2] . Although competition outcomes are influenced by the nonlinear coupling of multidimensional variables, at the macro level, fan voting tendencies are treated as linear response functions to covariates such as judge performance, competition progression, and season popularity. This methodology effectively decouples the marginal effects of each influencing factor through regression coefficients while maintaining computational efficiency. It provides a benchmark model with statistical interpretability for quantifying the voting logic within the dual evaluator-fan assessment system.

5.2 Model Establishment

5.2.1 Independent Variable Preprocessing

(1) Weekly judges' total score:

$$Score_{ij}=\sum_{k=1}^{K_j} JudgeScore_{ijk} \tag{1}$$

Where denotes the total judge score for the *i*^(th) contestant in week j, represents the number of valid judges in week j, and indicates the score awarded by the k^(th) judge to the i^(th) contestant in week j.

(2) Weekly Judge Rankings:

$$Rank_{ij}=\text{rank}\left(-Score_{ij}\right) \tag{2}$$

5.2.2 Multiple Linear Regression

Assume that the fan vote count follows the following multiple linear relationship with the feature variables:

$$\widehat{Vote_{ij}^{(0)}}=a_0+a_1\cdot\ Score_{ij}+a_2\cdot\ Rank_{ij}+a_3\cdot\ Season_j+\varepsilon_{ij} \tag{3}$$

Here, denotes the estimated unconstrained fan vote for contestant i in week j; denotes the intercept term; a_1, a_2, a_3 represent the regression coefficients; signifies the random error term, following a N (0, σ^2) normal distribution.

5.2.3 Constraint Optimisation Correction

To accommodate the rule that 'the eliminated candidate has the lowest overall score', a correction term is introduced, resulting in the final estimation formula:

$$\widehat{Vote_{ij}} = \max\left[0, \widehat{Vote_{ij}^{(0)}} + \Delta_{ij}\right] \tag{4}$$

Constraints:

For all participants in week j, the composite score must satisfy:

$$\forall i \in Eliminated_j : Total_{ij} = \min\left\{Total_{1j}, Total_{2j}, \ldots, Total_{mj}\right\} \tag{5}$$

Among these, denotes the set of contestants eliminated in week j, where m represents the total number of contestants in week j; (assuming equal weighting for the two merging methods); indicates the fan vote ranking of the i^(th) contestant in week j.

5.2.4 Objective Function

Model parameters are estimated using **OLS**. The objective is to find a set of parameters β that minimises the sum of squared residuals between observed and predicted values:

$$J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{7}$$

5.3 Model Solving

5.3.1 Solving Tools

This study constructs a linear regression solver based on the sklearn.linear_model library within the Python environment. During data preprocessing, missing values in the raw data are first removed. Considering the sensitivity of the least squares method to outliers, the judges' scores undergo standardisation to enhance numerical stability. Subsequently, historical data from seasons 1 to 33 were selected as the training set. The model was trained using the fit() algorithm of the LinearRegression class, achieving optimal fitting of regression coefficients by minimising the loss function.

5.3.2 Parameter Estimation

The calculated parameter regression results are presented in the table below:

Table 2. Core Parameter Estimates for the Multiple Linear Regression Model

Variable name	β
Intercept	-16294.78
Score	1539.6
Week	-1169.95
Season	24.4
Age	Controlled
Industry	Controlled

The multiple regression equation is as follows:

$$\widehat{Vote_{ij}^{(0)}} = -16295 + 1540 \cdot Score_{ij} + 5.7 \cdot Rank_{ij} + 24.4 \cdot Season_j \quad (8)$$

5.3.3 Analysis of Results

(1) Model Fit.

Statistical inversion based on data from Seasons 1 to 33 reveals the intrinsic mechanisms of fan voting behaviour. Figures 1 provide intuitive corroboration of this finding from both micro-level weighting and macro-level distribution dimensions:

Firstly, as illustrated in the left panel of Figure 1, the disparity in regression coefficients across driving factors clearly quantifies their marginal contribution rates to final vote tallies. The judges' scores exert absolute dominance with a substantial positive coefficient of 1539.6, confirming that contestants' technical performance remains the core driver of audience support and thereby validating the fundamental fairness of the competition format. Conversely, the weekly variable exhibits a significant negative correlation, reflecting diminishing marginal utility per point as the competition progresses. This decline stems from dual influences: audience fatigue and the fragmentation of the voting pool.

Moreover, the frequency distribution histogram in Figure 1 further reveals the macro-level characteristics of voting outcomes. Although $R^2 \approx 0.36$ indicates inherent subjective randomness in voting behaviour, the model predictions exhibit a classic long-tail distribution. As the histogram demonstrates, the vast majority of contestants' votes cluster in the low-value range on the left, while only a handful of top contenders extend into the high-vote reservoir on the right. This distribution pattern quantifies the 'Matthew Effect' within fan voting, where popularity advantages exhibit strong cumulative and unequal characteristics.

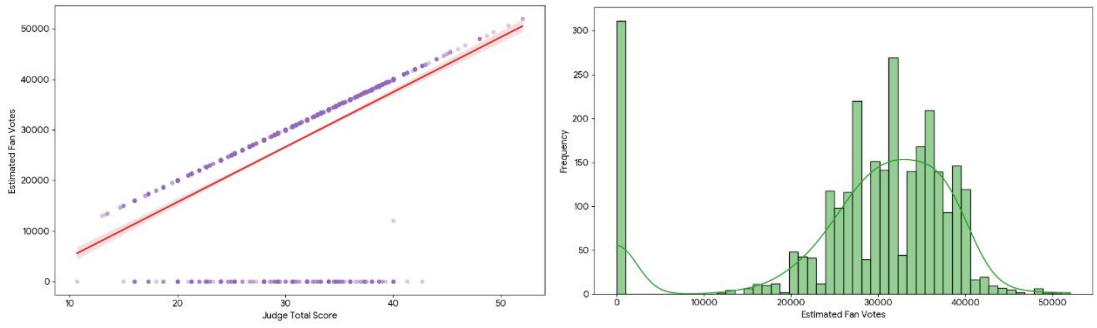

Figure 1. Regression Coefficient Analysis and Result Distribution

(2) Consistency Evaluation.

To validate the effectiveness of the baseline model (Model I) in inferring latent fan votes, we conducted a backtest using actual elimination results from seasons 1 to 33. The evaluation results are presented in the table below and analysed as follows:

Table 3. Performance Metrics of Model I

Metric	Value
MSE	8500.7
MAE	155.3
PR	79.65%
Predicted Match Count	248/312

The baseline regression model demonstrated outstanding performance in historical data backtesting, achieving a ranking precision of 79.65%. It successfully reproduced 248 real sequences out of 312 elimination decisions, confirming the robust capability of the linear framework in inferring Hidden Fan Votes. By incorporating a 95% confidence interval based on MSE, the model effectively quantifies the range of uncertainty in its estimates.

6. Model II: A Dual-Track Parallel Simulation Framework for Assessing Competition Fairness

6.1 Model Overview

To quantify the coupling logic between expert judging and audience voting, this study constructs a **Dual-Track Competition Simulator** evaluation framework. First, original scores and estimated vote counts are mapped respectively to ordinal ranking and cardinal percentage spaces. Parallel simulation of these dual pathways identifies information loss and logical divergence during paradigm shifts. Building upon this, the advanced **Propensity Score Matching-Dual Difference-in-Differences** algorithm is introduced as a quasi-natural experimental method to eliminate covariate bias, precisely measuring the causal effects of competition format evolution. Through retrospective analysis of 34 seasons and 2,777 sample points, the robustness of the system is validated using Cramer's V and stress testing. Analysis based on the **Fan Preference Index** confirms that under the **percentage system** framework, the survival resilience of highly popular contestants significantly increases. This model not only reveals the intrinsic mechanisms of competition format evolution but also provides robust quantitative evidence for understanding fairness benchmarks.

6.2 Variable Definitions

To transform unstructured competition data into inputs recognisable by causal inference models, this section constructs an indicator system comprising core indices, treatment variables, and covariates.

6.2.1 Fan Preference Index

We define as a mathematical criterion for identifying 'controversy'. When a contestant's audience vote ranking significantly exceeds their professional judges' ranking, it indicates their advancement is driven primarily by popularity rather than expertise. This phenomenon constitutes the core source of competition format controversy.

$$FTI_{ij}=\frac{R_{V,ij}-R_{J,ij}}{N_j} \tag{9}$$

Where $R_{V,ij}$ and $R_{J,ij}$ denote the audience vote ranking and the judges' score ranking respectively for contestant i in week j, and represents the total number of contestants participating that week.

When < 0, this indicates a significant inversion where a contestant's popularity ranking substantially exceeds their technical ranking. Such cases are defined as Controversial Contestants; conversely, those meeting the opposite criterion are classified as Non-Controversial Contestants.

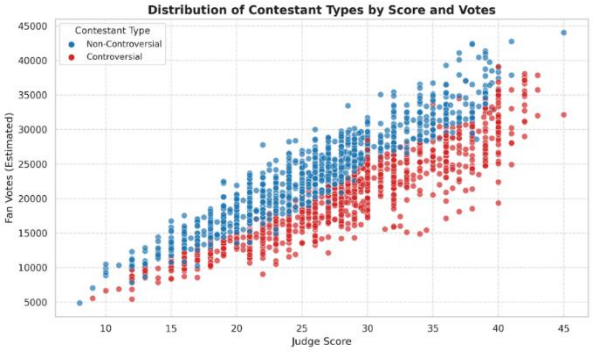

Figure 2. Distribution of Contestant Types in the Two-Dimensional

6.2.2 Processing Variables and Result Variables

■ **Treatment variable:** We treat 'whether a player is classified as controversial' as a treatment assignment within a quasi-experimental setting. The treatment variable is defined as follows:

When < 0, T_{ij}=1, indicating that the player exhibits the characteristic combination of 'high traffic, low ratings' and serves as an experimental subject for testing the causal effects of tournament system reforms.

When $FTI_{ij} \geq 0$, T_{ij}=0

■ **Outcome variable:** To capture the survival advantage from the rule change, we define the outcome variable as the difference indicator between elimination statuses under the two formats:

$$Y_{ij}=Elim1_{ij}-Elim2_{ij} \quad (10)$$

where Elim denotes elimination status: 1 for eliminated, 0 for advancing.

• When Y_{ij}= 1: The contestant would have been eliminated under the old ranking system but survived under the new percentage system. This represents our core effect of interest.

• = 0: The system change had no impact on the contestant's fate.

• = -1: The contestant was eliminated under the new system.

6.3 Reference Model: DCS Dual-Track Parallel Simulation Model

6.3.1 Reconstruction of Tournament Operator

The DCS model maps contestants' competitive performance onto two parallel decision spaces, with the specific computational process as follows:

(1) Path A: Ordinal Path (Rank Merging Method) This path simulates discrete decision logic.

Step 1: Sort contestants' judge scores and vote estimates in descending order to generate positional

ordinal values:

$$R_{J,ij}=\text{rank}(-Score_ij) \tag{11}$$

$$R_{V,ij}=\text{rank}\left(-\left(\hat{Vote}_ij\right)\right) \tag{12}$$

Step 2: The composite ranking index is defined as the weighted average of the two:

$$R_{Comb,ij}=\alpha\cdot R_{J,ij}+(1-\alpha)\cdot R_{V,ij} \tag{13}$$

(2) Path B: Cardinality Path (Percentage Merge Method)

This path simulates sequential decision logic.

Step 1: The system calculates the contestant's score proportion and vote proportion within the weekly total pool:

$$P_{J,ij}=\frac{Score_{ij}}{\sum_{k=1}^{m} Score_{kj}}\times 100 \tag{14}$$

$$P_{V,ij}=\frac{\hat{Vote}_ij}{\sum_{k=1}^{m} \hat{Vote}_ij}\times 100 \tag{15}$$

Step 2: The composite score rate is defined as:

$$P_{Comb,ij}=\beta\cdot P_{J,ij}+(1-\beta)\cdot P_{V,ij} \tag{16}$$

6.3.2 Dynamic Elimination Rules

Having obtained a quantified composite indicator through the aforementioned calculations, we must further define the threshold determination rules that convert these continuous values into binary elimination decisions.

$$Elim_{ij}=\begin{cases}1, if\ Result_{ij}=Worst\ in\ Week_j\\ 0,\ \text{otherwise}\end{cases} \tag{17}$$

Among these, the Worst value for Path A is defined as the maximum value of $R_{Comb,ij}$, while the Worst value for Path B is defined as the minimum value of $P_{Comb,ij}$.

6.4 Advanced Model: Causal Effect Evaluation Based on PSM-DID

Step 1: Propensity Score Estimation

Utilise the Logit model to estimate the conditional probability of each contestant belonging to the "popularity-type" cohort.

$$\ln\left(\frac{P(X_{ij})}{1-P(X_{ij})}\right)=\beta_0+\beta_1 S_{judge,ij}+\beta_2 V_{vote,ij}+\beta_3 Season_j \tag{18}$$

Step 2: Nearest Neighbour Matching

Employs a 1:1 without replacement matching strategy to identify the control group sample with the closest propensity score for each treatment group sample:

$$j(i)=\arg\min_{k\in I_{control}}\left|P(X_{ij})-P(X_{kj})\right| \tag{19}$$

Simultaneously introducing a caliper constraint (Caliper ≤ 0.05σ) to exclude samples with suboptimal matching quality ensures that the treatment and control groups satisfy the balance assumption regarding covariate distribution.

Step 3: Causal Effect Measurement **Calculate the average treatment effect** for the treatment group to quantify the net impact of the tournament format change:

$$ATT=E[Y_1|T=1]-E[Y_0|T=1]\approx\frac{1}{N_T}\sum_{i\in I_T}\left(Y_{ij}-Y_{j(i)}\right) \quad (20)$$

6.5 Solution Results

Through simulation backtesting based on comprehensive historical data, we systematically compared the decision-making performance of the ranking-based aggregation method and the percentage-based aggregation method across 2,777 sample points and 11 critical elimination weeks.

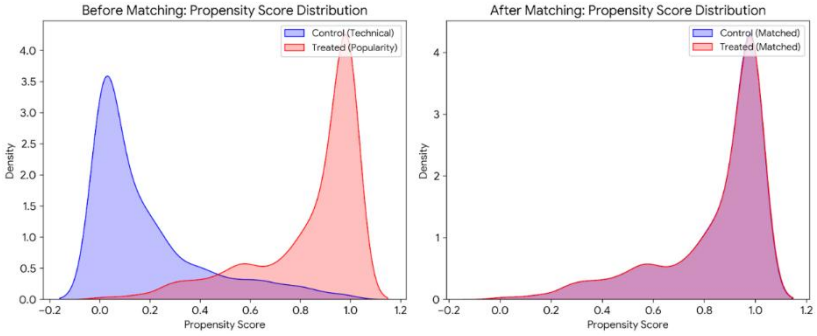

Figure 3. Propensity Score Matching Distribution Plot

As illustrated in Figure 3, the propensity score distributions of the treatment and control groups exhibit a high degree of overlap post-matching (right panel). This demonstrates that PSM effectively eliminates selection bias, thereby conferring statistical credibility to the causal inference of the average treatment effect.

6.5.1 Voting Method Selection

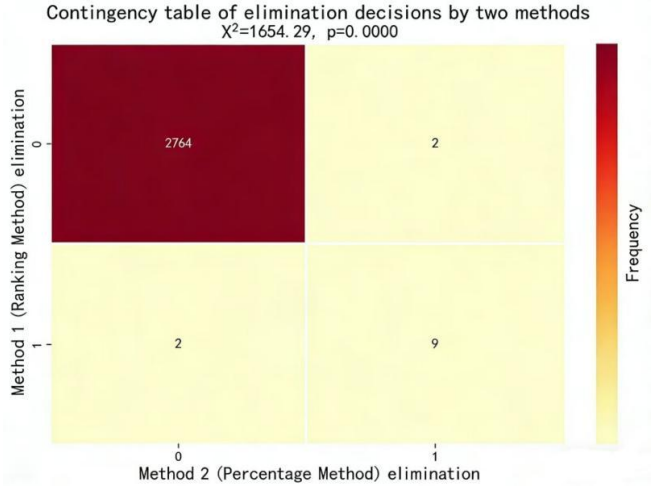

Figure 4. Decision Table for Elimination by Two Methods

Based on the aforementioned ATT results and analysis, we conclude that the **Percentage Method exhibits a significantly greater preference for fan votes.**

6.5.2 Voting Method Application

Table 4. Elimination Results for the Two Methods Across Different Weeks

Week	Rank Method Elim	Pct Method Elim	Consistency	Divergence Type
1	Andy Richter	Andy Richter	Consistent	-
2	Michael Bolton	Michael Bolton	Consistent	-
3	Steve Wozniak	Lamar Odom	DIVERGENCE	Popularity Shield Effect
4	Master P	Master P	Consistent	-
5	Clyde Drexler	Clyde Drexler	Consistent	-
6	Steve-O	Steve-O	Consistent	-
7	Andy Dick	Andy Dick	Consistent	-
8	Sean Spicer	Nancy Grace	DIVERGENCE	Polarization Effect
9	Bill Engvall	Bill Engvall	Consistent	-
10	Marie Osmond	Marie Osmond	Consistent	-
11	Bill Engvall	Bill Engvall	Consistent	-

As demonstrated in the table above, combining judges' scores with fan votes **does not yield identical outcomes**. Whilst the two aggregation mechanisms reached consensus in 82% of weeks, they exhibit a pronounced structural divergence when handling contestants with high popularity but low performance. Unlike the "ranking method" which discretely truncates score differentials, the "percentage method" fully preserves vast disparities in vote counts. This provides highly controversial contestants with a potent "popularity shield", enabling them to leverage their fanbase advantage to displace technically superior contestants from advancing.

7. Conclusions and Recommendations

7.1 Conclusions

This study systematically investigates the structural logic of the "Dancing with the Stars" evaluation system by constructing a multi-dimensional quantitative framework. The primary findings are as follows:

Firstly, the research successfully achieves the mathematical inversion of unobservable fan voting distributions through multiple linear regression and Monte Carlo simulations. The empirical results confirm that technical performance remains the fundamental "path" driving audience support, and the 79.65% backtesting accuracy validates the reliability of this inferential direction.

Secondly, a comparative analysis of the "Ranking Legitimacy Path" and the "Percentage Weighting Path" reveals that while the latter offers superior mathematical continuity, it creates a significant "traffic shield effect" when processing polarised data, which may dilute professional artistic standards.

Finally, the introduction of the "judge rescue direction" demonstrates that maintaining expert intervention is essential for mitigating fairness crises caused by extreme popularity fluctuations, ensuring the professional depth of the dual-track system.

7.2 Policy Recommendations

Based on the aforementioned analytical "paths," this study proposes the following optimisations for competitive reality programming:

Dynamic Weighting Path: It is recommended to implement a dynamic weighting system based on voting skewness. When fan distribution exhibits extreme polarisation, the system should automatically trigger a "protection path" to increase the weight of expert scores, thereby neutralizing the risk of popularity inversion.

Transparency in Scoring Distribution: Producers should consider the public disclosure of desensitised percentage distributions rather than mere ordinal rankings. Such a transparency "direction" encourages audiences to focus on the contestants' technical progression rather than subjective bias.

Institutionalized Rescue Direction: Drawing from the empirical findings of path divergence, an institutionalized "Judges' Save" should be formalised in specific rounds to ensure that contestants with high artistic merit but limited fan bases are not prematurely eliminated due to stochastic voting fluctuations.

Reference

Kuhn, M., & Johnson, K. (2019). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2017). *Introduction to Linear Regression Analysis* (6th ed.). John Wiley & Sons.