World Journal of Management Science
Print ISSN: 2959-9628
Online ISSN: 2959-9636
DOI: https://doi.org/10.61784/wms3108

LATENT VOTE INVERSION AND VOTING RULE COMPARISON UNDER RULE CONSTRAINTS

XinQiao Wu
Sino-European School of Technology, Shanghai University, Shanghai 200444, China.

Abstract: Using 34 seasons of panel data from Dancing with the Stars, this study develops a rule-constrained latent vote inversion model with entropy regularization to reconstruct unobservable weekly audience vote shares from observed elimination outcomes. Model validation yields a high consistency rate of 99.62%, and an entropy-based metric further reveals that audience preferences are generally diffuse and uncertain. Through counterfactual simulations, we systematically compare rank-based and percentage-based aggregation rules. The results indicate that switching from rank-based to percentage-based rules leads to a 10.34% outcome reversal rate, while the reverse switch only causes a 1.21% reversal, confirming that the percentage-based system significantly amplifies audience influence and favors high-popularity contestants. Mechanistic analysis demonstrates that the percentage-based rule preserves vote magnitude differences, whereas the rank-based rule compresses extreme popularity advantages into ordinal rankings. These findings quantify the systemic biases of different voting rules and offer empirical evidence for designing fairer and more balanced competition mechanisms.
Keywords: Latent vote inversion; Rule constraints; Voting rules; Rank-based system; Percentage-based system

1 INTRODUCTION

Hybrid voting mechanisms that combine professional expert scoring with public audience voting have become a core component of televised talent competitions. However, the opacity of audience vote data, coupled with frequent adjustments to voting rules, often leads to persistent fairness controversies, with unobserved popularity frequently overriding technical merit. Taking the long-running show Dancing with the Stars as a representative case, its historical shifts between rank-based and percentage-based aggregation rules, alongside the confidentiality of viewer votes, create a complex latent variable inference problem. Existing research on voting fairness predominantly relies on direct observable voting data or oversimplified rule simulations[1-2], while few studies systematically reconstruct unobservable audience preferences from elimination outcomes and empirically quantify the systemic biases embedded in different aggregation rules within real competition scenarios.

Scholarly discussions on voting systems and inverse modeling have evolved across multiple disciplines. Classical social choice theory has long examined the rationality, consistency and manipulability of various voting mechanisms, laying the theoretical foundation for evaluating rule design flaws [3-7]. Inverse problem theory and statistical modeling provide robust technical frameworks for inferring hidden variables from indirect observable information. Meanwhile, research on multi-criteria decision-making and composite indicator construction offers analytical tools for comparing the performance of different aggregation methods [8-9]. Despite these advances, most existing studies remain confined to theoretical deductions or hypothetical simulations, lacking large-scale empirical validation using real-world competition panel data, and failing to capture the dynamic interplay between expert evaluation and audience preferences.

To address these research gaps, this paper constructs a rule-constrained latent vote inversion model integrated with entropy regularization. Using longitudinal data spanning 34 seasons of Dancing with the Stars, the study reconstructs weekly audience vote shares consistent with historical elimination outcomes. Furthermore, counterfactual simulation methods are employed to systematically compare rank-based and percentage-based aggregation rules, quantifying their differential impacts on elimination results and audience influence amplification. The marginal contributions of this paper are threefold: first, it develops a novel framework for latent vote reconstruction in hybrid voting systems with unobservable public preferences; second, it empirically quantifies the systemic biases of mainstream voting rules and reveals the inherent imbalance between technical merit and public popularity; third, it provides data-driven insights and practical references for optimizing fair, balanced and stable competition voting mechanisms.

2 REGULARIZED INVERSE PROBLEM FORMULATION FOR FAN VOTE SHARES

This section describes how we estimate weekly fan vote shares from observed judges' scores and elimination outcomes. Because viewer votes are not publicly available, the task is formulated as a rule-constrained inverse problem: we search for vote distributions that are consistent with how eliminations occurred under the show's official rules.

2.1 Weekly Vote Share Representation

For each season s and week t, let n_{s,t} denote the number of contestants who actually performed. We represent the unknown fan support by a vector:

$$\mathbf{v}_{s,t}=(v_{1,t},v_{2,t},\ldots,v_{n_{s,t},t}) \tag{1}$$

Where $(v_{i,t})$ is the relative fan vote share of contestant i in that week. These shares satisfy

$$v_{i,t}\geq 0,\ \sum_{i=1}^{n_{s,t}} v_{i,t}=1 \tag{2}$$

Thus, votes are modeled only in relative terms within each week.

Judges' scores are denoted J_{i,t}. When required by the rule format, scores are normalized within the week to remove scale differences across episodes. We denote the normalized value by $\tilde{J}_{i,t}$, which is obtained by dividing each judge's score by the weekly total. The model consistency is shown in Figure 1.

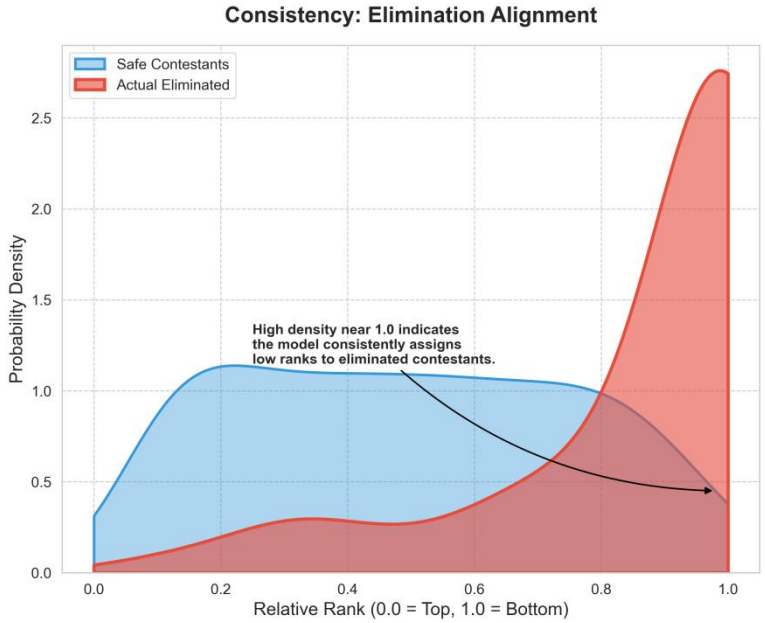

Figure 1 Consistency between Model-Implied Risk Ranking and Actual Eliminations

Eliminated contestants concentrate near the bottom of the model-implied relative ranking, indicating that the rule-constrained formulation aligns inferred vote shares with observed elimination outcomes.

2.2 Rule-Consistency Constraints

The key modeling principle is that the inferred vote shares must produce elimination outcomes consistent with the rule in effect.

Let $(S_{i,t})$ denote the combined score used for elimination decisions. Its definition depends on the voting rule phase.

Rank-based seasons Contestants are ranked separately by judges and fans. The combined score is $S_{i,t}=R_{i,t}^{(J)}+R_{i,t}^{(V)}$, where $R_{i,t}^{(J)}$ is the judges' rank and $R_{i,t}^{(V)}$ is the rank induced by vote shares $v_{i,t}$. A lower combined rank indicates a higher risk of elimination.

Percentage-based seasons Judges' scores and fan votes are combined as proportions: $S_{i,t}=\alpha\tilde{J}_{i,t}+(1-\alpha)v_{i,t}$, where α is the rule weight between judges and audience. The contestant with the lowest combined score faces elimination.

Bottom-two seasons: The two contestants with the lowest combined scores form the bottom-two set $B_{s,t}$. Since the final elimination depends on judges' decisions, which are not recorded in the dataset, the model enforces only that the inferred vote shares reproduce the observed bottom-two membership.

2.3 Regularization and Identifiability

Because elimination results provide limited information, many vote-share vectors can satisfy the rule constraints. To select a representative solution, we introduce two regularization principles.

Smoothness across weeks. Audience support typically changes gradually. Large week-to-week fluctuations are discouraged by penalizing differences between $v_{i,t}$ and $v_{i,t-1}$.

Avoiding extreme concentration. Highly uneven vote distributions are discouraged unless required by the elimination constraints. This is implemented using an entropy-based term that favors more balanced vote shares[10].

These regularizations do not force a specific outcome but rather guide the solution toward stable, plausible patterns of audience support.

2.4 Optimization Formulation

For each week, we solve an optimization problem that balances rule consistency and regularization:

$$\min_{\mathbf{v}_{s,t}} L_{\text{rule}}(\mathbf{v}_{s,t})+\lambda_1 L_{\text{smooth}}+\lambda_2 L_{\text{entropy}}, \tag{3}$$

subject to

$$v_{i,t}\geq 0, \sum_i v_{i,t}=1 . \tag{4}$$

Here: L_{rule} measures how well the inferred votes reproduce observed eliminations or bottom-two membership, L_{smooth} penalizes abrupt week-to-week changes, and L_{entropy} discourages unnecessary vote concentration. The weights λ_1 and λ_2 control the strength of these regularizations.

2.5 Interpretation of the Estimated Vote Shares

The resulting $v_{i,t}$ values should be interpreted as model-consistent relative audience preferences, not as exact vote counts. They represent one feasible explanation of historical eliminations under the competition rules, chosen to be stable and minimally extreme.

In the next section, we evaluate how consistently these inferred vote shares reproduce elimination outcomes and how certain the model is about these estimates.

3 COMPARISON OF THE RELIABILITY ESTIMATION OF HIDDEN VOTES AND VOTING RULES

3.1 Elimination Consistency Metrics

Validation must rely on whether the model reproduces observed elimination outcomes. For each season-week (s,t), we apply the rule phase phase(s) to combine inferred vote shares $v_{s,t}$ with judges' scores and obtain a model-implied elimination result.

We use two consistency measures aligned with the show's rule structure:

Strong consistency (direct-elimination phases).$\text{Strong}_{s,t}=1\{\hat{E}_{s,t}(v_{s,t})=E_{s,t}\}$, where $\hat{E}_{s,t}(v_{s,t})$ is the eliminated set implied by the aggregation rule.

Weak consistency (bottom-two Judges' save phases). $\text{Weak}_{s,t}=1\{E_{s,t}\subseteq\hat{B}_{s,t}(v_{s,t})\}$, where $\hat{B}_{s,t}(v_{s,t})$ is the model-implied bottom-two set.

Strong consistency is nearly saturated in direct-elimination seasons (S1–27): the model reproduces the eliminated set in 99.62% of weeks, with a recall of 99.67% for eliminated contestants.

In the Judges' save era (S28+), outcomes are only partially determined by votes and scores, so we evaluate whether the actual eliminated contestant falls in the model-implied bottom-two. Under this criterion, weak consistency is 91.90%, reflecting judges' discretion rather than failure of vote inference.

3.2 Certainty of Vote Estimates

Because the inverse problem admits multiple feasible vote configurations, we report two certainty measures to summarize how decisive each week's inferred vote pattern is. Weeks with concentrated support receive higher certainty, while diffuse distributions indicate ambiguity.

Entropy certainty (distribution concentration): $C_{s,t}^{(H)}=1-\frac{-\sum_{i=1}^{n_{s,t}} v_{i,t}\log(v_{i,t}+\epsilon)}{\log(n_{s,t})}$. Higher values indicate that inferred support is concentrated on a few contestants; lower values indicate diffuse support and weaker identifiability.

Loss certainty (solution stability): $C_{s,t}^{(L)}=\exp(-\tilde{L}_{s,t})$, where $\tilde{L}_{s,t}$ is a normalized inverse-optimization loss. This serves as a stability diagnostic and is interpreted comparatively rather than probabilistically. These certainty measures reflect relative informativeness rather than statistical confidence in a point estimate, consistent with inference in partially identified models.

Empirically, entropy-based certainty is generally low (mean 0.0416), indicating that inferred fan support is often diffuse rather than dominated by a single contestant. Only 12.1% of contestant-week cases show highly concentrated support (certainty > 0.1), while 41.6% fall into very low-certainty regimes (certainty < 0.01).

3.3 Sensitivity to Regularization Parameters

We stress-test the two regularizers that control interpretability—λ_{judge} and λ_{smooth}—using a one-factor sweep and a small interaction grid. The resulting consistency and certainty metrics are highly stable (Figure 2).

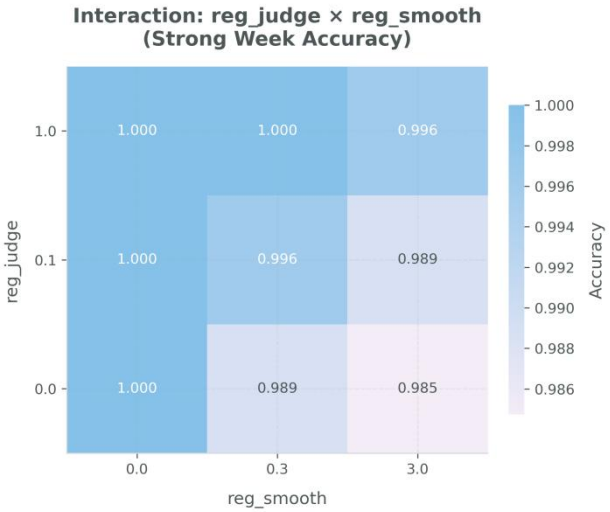

Figure 2 Week-level Strong Consistency under Different Regularization Settings (Judge-Alignment Weight Vs. Temporal Smoothness Weight)

With robustness established, we next interpret low-certainty weeks as regimes where the rules reveal limited information about fan support.

3.4 Interpretation of Uncertain Weeks

The certainty patterns above allow us to identify weeks where the voting system provides limited information about fan support. Three mechanisms recur:
Score compression: limited dispersion in judges' scores weakens directional information about $v_{s,t}$.
Rule-driven partial observability: bottom-two seasons reveal only a risk set, not the final decision rule.
Structural irregularities: double eliminations or non-elimination weeks weaken constraint sharpness.
These weeks represent low-identifiability regimes of the voting system. Importantly, when certainty is low, changes in aggregation rules can more easily alter elimination outcomes. Because many weeks fall into low-certainty regimes, small changes in aggregation rules can more easily alter elimination outcomes.

3.5 Counterfactual Outcome Differences under Rule Swaps

For each season–week (s,t), we recompute the elimination outcome under both aggregation rules using the same inferred fan vote shares $\hat{v}_{s,t}$ and judges' scores. A week is defined as a flip if the eliminated contestant (or, in judges-save phases, the bottom-two risk set) differs between the two rules.
The swap experiment reveals a clear asymmetry. Switching from the rank-based system to the percentage-based system changes elimination outcomes in 10.34% of weeks, whereas the reverse swap changes outcomes in only 1.21% of weeks. This indicates that outcome determination under the percentage-based rule is more responsive to differences in inferred fan support, while the rank-based rule produces more stable results under the same underlying vote patterns.
To better understand where these rule-induced flips occur, we examine how flip frequency varies with the joint configuration of judges' score share and inferred fan vote share. Figure 3 shows that flips are concentrated in regions where judge scores and fan support diverge, rather than in uniformly strong or uniformly weak performances.

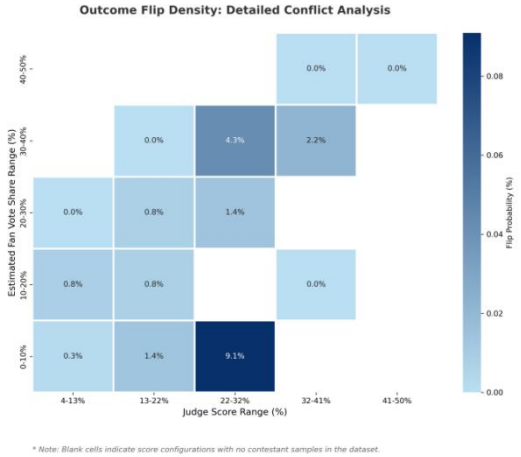

Figure 3 Outcome Flip Density under Rule Swaps, by Judges' Score Share and Inferred Fan Vote Share

Each cell shows the percentage of contestant-weeks where elimination differs between the rank-based and percentage-based rules. Blank cells indicate no observed samples. These structural patterns motivate a closer look at whether one rule systematically favors audience-driven survival in such disagreement regions.

3.6 Directional Bias in Fan Vote Influence

To assess whether one rule systematically favors audience support, we examine contestants whose fate differs in flip weeks. In these weeks, one contestant is eliminated under one rule but survives under the other; we interpret the survivor as being "saved" by that rule.

Contestants saved by the percentage-based rule have a higher average inferred fan vote share (18.5%, n=10) than those saved by the rank-based rule (12.4%, n=10). This difference shows that, when the two systems disagree, the percentage-based rule is more likely to retain contestants with stronger audience backing. In contrast, the rank-based rule is less influenced by extreme fan popularity in these marginal situations.

These results provide quantitative evidence that the percentage-based system places greater effective weight on audience support relative to the rank-based system. The directional preferences are shown in Figure 4.

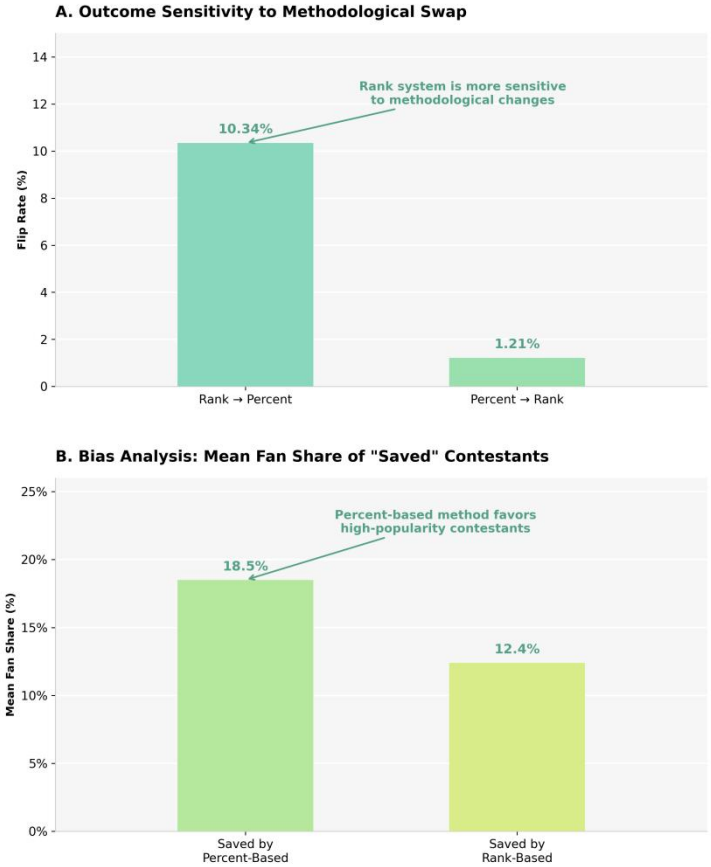

Figure 4 Counterfactual Elimination Outcomes under Rule Swaps, Holding Inferred Fan Vote Shares.A. Proportion of weeks with differing eliminations between rank-based and percentage-based aggregation. B. Average inferred fan vote share of contestants whose survival depends on the aggregation rule.

3.7 Mechanistic Explanation of Rule-Induced Differences

The observed asymmetry arises from how the two aggregation rules transform score information. The percentage-based system preserves magnitude differences: large fan vote advantages can offset weaker judges' scores because fan support enters with its full quantitative spacing. By contrast, the rank-based system reduces both judges' scores and fan votes to ordinal ranks, compressing extreme differences and limiting the influence of very large fan vote margins.

Consequently, the percentage-based rule tends to amplify audience-driven advantages, whereas the rank-based rule moderates extremes and produces more consensus-oriented elimination outcomes. This structural distinction explains why rule choice has the strongest effect in weeks where fan and judge signals diverge. Such outcome sensitivity to aggregation mechanics rather than preference shifts is consistent with theoretical results in institutional voting rule design and information aggregation [5, 8].

4 CONCLUSIONS

This paper focuses on latent vote inversion in hybrid voting systems of televised talent competitions, using 34 seasons of Dancing with the Stars as the research object to construct a rule-constrained latent vote inversion model with entropy regularization. Key findings include high consistency between the model and actual elimination outcomes (99.62% strong consistency in direct-elimination seasons, 91.90% weak consistency in Judges' save seasons), diffuse audience support (average entropy-based certainty: 0.0416), and asymmetric impacts of rule swaps (10.34% outcome reversal when switching to percentage-based rules, 1.21% for the reverse). Practically, the model provides a quantitative tool for unobservable public votes in similar televised competition scenarios; its conclusions offer preliminary data references for voting rule design in analogous public evaluation activities, with generalization to broader fields requiring further verification.

This study has limitations: 1) Data is limited to Dancing with the Stars, whose specific mechanisms limit the model's generalization; 2) Only two voting rules are analyzed, lacking complex rule comparisons; 3) The model relies on strong assumptions (stable audience voting, no endogeneity) and is highly dependent on program rules, requiring re-calibration for rule changes; 4) Dynamic audience preferences and external factors are unconsidered. Future research will expand samples, include complex rules, optimize the model with time-varying parameters, and use experimental methods to verify causal relationships.

COMPETING INTERESTS

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

REFERENCES

[1] Bugeja M, da Silva Rosa R, Shan Y, et al. Governance consequences of shareholder dissent in director elections: Evidence from a purely majority voting system. Pacific-Basin Finance Journal, 2026, 97: 103119.

[2] Castro V, Martins R, Sakurai S N. Turnout and Invalid Voting in Brazilian Municipal Elections: A Runoff Voting System Tale. Scottish Journal of Political Economy, 2026, 73: e70044.

[3] Pandit V, Cutrone J. Evaluating fairness of voting systems: simulating violations of arrow's conditions. Theory and Decision, 2025.

[4] Kirsch W, Toth G. Optimal weights in a two-tier voting system with mean-field voters. Social Choice and Welfare. 2026, 66: 953-993.

[5] Pilon D. Are Canadian Voting System Reform "Trade-Offs" Really Trade-Offs? Operationalizing Voting System Values and Assessing the Evidence. Canadian Public Policy, 2024, 50(3).

[6] Gibbard A. Manipulation of Voting Schemes: A General Result. Econometrica. 1973, 41: 587-601.

[7] Satterthwaite M. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. Journal of Economic Theory, 1975, 10: 187-217.

[8] Petróczy D G. Optimising the decision threshold in a weighted voting system: the case of the IMF's Board of Governors. Economics of Governance, 2026, 27: 17.

[9] Gibbs T, Oreský J, Hong SW. Building bridges for consensus via alternative voting methods. Innovation: The European Journal of Social Science Research, 2024, 37: 1120-1147.

[10] Jaynes E T. Information Theory and Statistical Mechanics. Physical Review, 1957, 106: 620-630.