

Tournament-Based Performance Evaluation and Systematic Misallocation: Why Forced Ranking Systems Produce Random Outcomes

Source: arxiv | **Author:** Jeremy McEntire | **Published:** December 06, 2025 | **PMS Relevance:** PMS RELATED | **Type:** Research Study | **Geography:** Global

Source URL: <https://arxiv.org/abs/2512.06583>

Executive Summary

This paper addresses the structural validity of tournament-based performance evaluation systems that mandate fixed distributional outcomes, commonly known as forced ranking or stack ranking, widely used in technology and corporate environments. The authors argue that such systems produce systematic classification errors independent of implementation quality, not merely as a result of poor execution. Using agent-based simulation across 994 engineers distributed across 142 teams of seven, the study demonstrates that random team assignment alone generates a 32% error rate in termination and promotion decisions, attributable purely to team composition variance. Under conditions modelling differential managerial capability, error rates rise to 53%, with false positives and false negatives each exceeding correct classifications. The authors further argue that cross-team calibration — a commonly proposed corrective — converts evaluation into influence contests driven by managerial persuasion rather than employee merit. Multi-period dynamics are shown to incentivise risk-averse behaviour and accelerate high-performer attrition through adverse selection. The paper concludes that forced ranking persists not because it aligns incentives effectively, but because it satisfies institutional demands for a demonstrable, auditable process, despite producing outcomes statistically indistinguishable from random allocation.

Key Insights

1. Random team assignment alone, absent any managerial bias, generates a 32% misclassification error rate in forced ranking termination and promotion decisions due to team composition variance.
2. Under realistic conditions modelling differential managerial capability, classification error rates reach 53%, meaning false classifications outnumber correct ones.
3. Cross-team calibration processes — frequently proposed as a remedy to forced ranking's inconsistencies — structurally transform evaluation into influence contests where persuasive managers secure favourable outcomes for their reports independent of actual merit.
4. Multi-period dynamics produce adverse selection: as employees observe random or merit-independent outcomes, risk-averse behaviour increases and high performers are disproportionately likely to exit the organisation.
5. The paper argues forced ranking persists not through demonstrated incentive alignment but through satisfying institutional and legal demands for a formalised, auditable process.

Practical Takeaways

- Organisations using fixed distributional requirements in performance evaluation face an inherent structural error rate driven by team composition variance, a factor largely invisible to standard process audits.
- The proposed alternative — delegating evaluative judgment to managers with hierarchical accountability — is identified in the paper as the theoretically efficient solution, though the authors note it cannot be fully formalised within the legal and coordination constraints that originally motivated forced ranking.

Critical Analysis

Strengths: The paper is methodologically transparent, providing specific simulation parameters (994 engineers, 142 teams of 7) that allow scrutiny and replication. The logic connecting team composition variance to misclassification is theoretically sound and novel in its formalisation. The critique of calibration sessions as influence contests is a substantive institutional observation rarely modelled quantitatively. **Limitations:** The study is entirely simulation-based; no empirical validation against real organisational data is presented, which limits the external validity of the specific error rate figures (32%, 53%). The modelling assumptions — particularly those governing 'differential managerial capability' — are not fully specified in the abstract and could materially affect outcomes. The paper does not engage with potential counterarguments, such as evidence that forced ranking successfully identifies and removes persistent low performers in some contexts. The conclusion that outcomes are 'indistinguishable from random allocation' is a strong claim that is demonstrated within the model's parameters but not across varied organisational settings. **Potential bias:** The framing and conclusions are strongly adversarial toward forced ranking systems, and no scenario is modelled in which such systems produce net positive allocation outcomes, suggesting the study may be designed to critique rather than neutrally evaluate the mechanism.

Organizational Practice

The article models forced ranking or stack ranking systems in which employees are mandated to be distributed across fixed performance categories, with fixed percentages terminated (e.g., bottom 15%) and promoted (e.g., top 15%). Cross-team calibration sessions, in which managers advocate for their direct reports' rankings relative to peers on other teams, are also modelled as a common organisational practice used to address inconsistency across teams.

Strategic Implications

If the simulation's findings are directionally valid, organisations relying on forced distribution systems face a structural source of talent misallocation that is undetectable through conventional process quality reviews. The finding that error rates increase under managerial capability variance points to team assignment and managerial quality as underappreciated confounders in performance outcomes. The modelled adverse selection dynamic — high-performer exit driven by perceived outcome randomness — has implications for retention strategy in competitive talent markets. The persistence of forced ranking despite its modelled inefficiencies raises questions about whether performance management systems are optimised for talent allocation or for organisational legitimacy and legal defensibility.

