Journal of Audit and Accountability
Volume 1 · Number 1 · 2026

A Higher Average Is Not Always a More Likely Win: Two Game-Benchmark Objectives

Vincent Russell[1] · Charles Brooks[2] · Evan Anderson[3]

[1] Norway · [2] Portugal · [3] Austria

Original publication: November 7, 2026 | Article 589

Abstract

Both mean score and run-to-run superiority favored BBF in 49 of 55 games, but the matching totals concealed four different game-level rankings. Using every cross-condition run pair, BBF's equal-game probability of superiority was 0.8170 with half credit for ties. Amidar and Skiing favored BBF by mean score while favoring the no-reset variant by this probability; Bowling and Private Eye reversed the other way. The disagreement concerns the objective being summarized, rather than a computational mistake. Benchmark reports should distinguish expected score magnitude from the chance of beating another captured run.

Secondary analysis of released evaluation-run scores. Comparisons concern two empirical run distributions within a game, without new agent training or interactive rendering.

A Higher Average Is Not Always a More Likely Win: Two Game-Benchmark Objectives

> Secondary analysis of released evaluation-run scores. Comparisons concern two empirical run distributions within a game, without new agent training or interactive rendering.

Abstract

Both mean score and run-to-run superiority favored BBF in 49 of 55 games, but the matching totals concealed four different game-level rankings. Using every cross-condition run pair, BBF's equal-game probability of superiority was 0.8170 with half credit for ties. Amidar and Skiing favored BBF by mean score while favoring the no-reset variant by this probability; Bowling and Private Eye reversed the other way. The disagreement concerns the objective being summarized, rather than a computational mistake. Benchmark reports should distinguish expected score magnitude from the chance of beating another captured run.

1. What does a better game result promise?

A player or evaluator can ask at least two questions about a stochastic game system. Which condition produces the higher average score across repeated runs? Which condition is more likely to beat the other if one run is drawn from each? These questions can produce different answers because an occasional large score contributes strongly to the mean but counts as only one winning comparison in a probability summary.

The distinction becomes relevant when benchmark outcomes support claims about interactive systems. Meng et al., 2026 proposes a paradigm for interactive game rendering. A rendering system would require direct visual and interaction evidence, rather than an arcade policy score alone. Nevertheless, any claim that an interactive condition is better needs an explicit performance objective. This study examines that reporting issue using released game-policy outcomes; it does not evaluate rendering quality, user interaction, or the target method.

Our research question asks whether ranking two captured conditions by their mean normalized scores agrees with ranking them by empirical probability of superiority. The data are released BBF and no-reset runs. We compare complete captured distributions within each game, retain exact ties, and give each game equal weight. The intention is not to choose a favorable aggregation after seeing an overall leaderboard, but to expose how two common interpretations of improvement can differ locally.

The contribution is a map of those differences with the observation unit kept clear. Cross-products of runs provide an exact calculation for the finite captured distributions. They do not create thousands of independent experiments. Regression summaries therefore use games as rows, and their uncertainty remains conditional on the recorded run distributions rather than claiming a complete analysis of future training seeds.

2. Score magnitude and pairwise superiority

The BBF work provides the algorithmic context and released run resource (Schwarzer et al., 2023). Work on reinforcement-learning evaluation emphasizes the instability of conclusions drawn from limited experimental evidence (Henderson et al., 2017). Agarwal et al., 2021 further develops statistical reporting for deep reinforcement learning, including probability-oriented comparisons. These concerns motivate separating the estimand from the number of decimals used to report it.

For a game, the mean gap is the BBF average normalized score minus the no-reset average. The superiority quantity is the fraction of all BBF/no-reset score pairs in which the BBF score is larger, plus half the fraction tied. It describes independently selecting one recorded run from each condition with equal probability within that condition. It does not describe a paired experimental seed, because source seed labels are not matched across conditions.

Vargha and Delaney, 2000 discusses a superiority effect measure applicable to ordinal discrete or continuous outcomes. We use its probability interpretation without conducting the paper's hypothesis tests. A value above one half favors BBF under the pairwise objective; below one half favors the variant. Equality at one half is retained as a tie and is not assigned arbitrarily to either condition.

The two objectives weight outcomes differently. Mean scores retain the magnitude of large gains and losses. Pairwise superiority depends on ordering, with exact ties split evenly. A rare exceptional score can move a mean appreciably while changing a small fraction of cross-pairs. Conversely, a condition can win narrowly in many pairs while losing a smaller number by large margins. Neither summary alone establishes stochastic dominance at every score threshold.

Measurement-scale choices matter as well (Stevens, 1946). A common positive affine transformation of scores within a game preserves the sign of its mean gap and the pairwise ordering. A general monotone nonlinear transformation preserves pairwise ordering but need not preserve mean comparisons. Across-game averages of normalized score magnitudes also depend on the source normalization denominators. We retain those released units rather than estimating new reference baselines from the same outcomes.

3. Games, runs, and the finite comparison population

The captured files contain 2,082 BBF rows and 830 no-reset rows across 55 common games. Eligibility requires at least ten recorded runs in each condition for a game. All 55 common games pass: BBF counts range from sixteen to sixty and variant counts from fifteen to twenty. The count gate is a declared sample rule, not a claim that ten runs are sufficient for every uncertainty question.

No run is trimmed, winsorized, or discarded for a high score. Every score is finite, and each retained game contributes one pairwise probability and one mean gap. The unequal run counts mean a pooled Cartesian-pair average would give different weights to different games. We instead average the 55 within-game probabilities, so a game with more available runs does not automatically exert more influence on the reported overall objective.

All cross-condition pairs within a game are enumerated. Their total is 31,310, but the same run appears in many pairs. The number is useful for reproducing the finite empirical calculation; it is not the inferential sample size. Enumerating more combinations from the same captured observations cannot supply independent new environment interactions, evaluation episodes, or training seeds.

The source records do not independently establish an identical training budget for every captured run, and their score-normalization provenance is inherited from the release. We do not infer a new compute-efficiency result from row counts. The comparison is consequently an audit of the available outcome distributions, rather than a controlled experiment isolating the causal effect of removing a reset operation.

There is one exact distributional tie in the relevant summaries. All captured Montezuma's Revenge scores in both conditions are zero, yielding mean gap zero and superiority one half. Its 270 cross-pairs are all

ties. This game remains in the denominator of 55; excluding it would change the benchmark population and should not occur silently because it fails to contribute a direction.

4. Research model and regression quantities

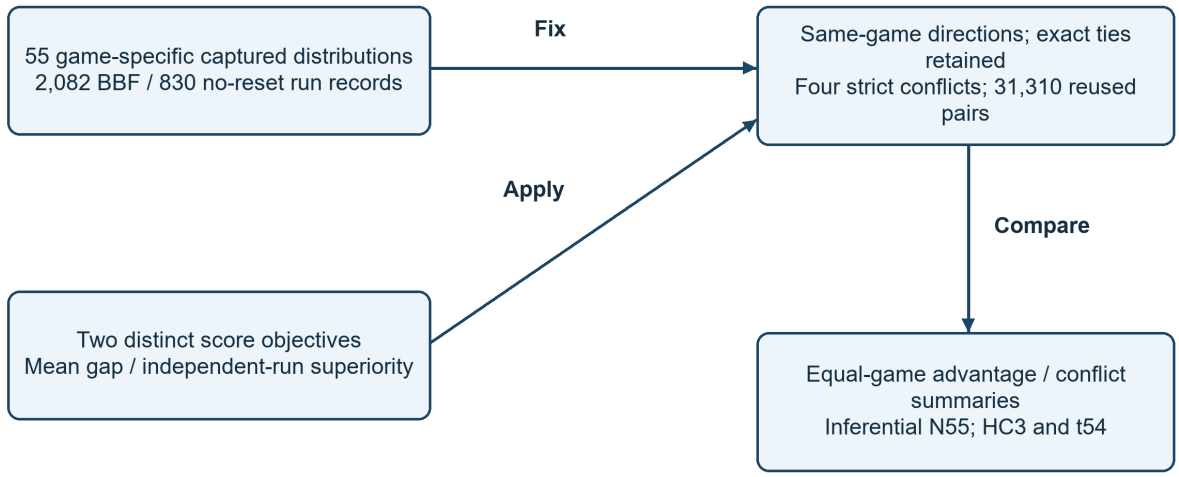

Figure 1. *The same captured game distributions feed a mean-gap objective and a pairwise-superiority objective. Their directions are compared within game before equal-game summaries are estimated.*

Model 1 regresses each game's superiority probability minus one half on an intercept. Its coefficient measures the equal-game average advantage over an even pairwise contest. Model 2 regresses a strict-conflict indicator on an intercept. A conflict is one only when the mean gap and centered superiority probability have nonzero, opposite signs. Exact ties are not called conflicts, and the denominator includes all eligible games.

Both models use 55 rows, HC3 standard errors, and t54 confidence intervals. There are no covariate controls or game fixed effects because each row is already a game-level comparison. Intercept-only R-squared is zero apart from numerical rounding. The regressions summarize variation across the selected games; they do not treat individual Cartesian pairs as independent rows.

Robust covariance methods address specified error-variance and dependence concerns, not the representativeness of an arbitrary benchmark (White, 1980; Cameron and Miller, 2015). These intervals condition on each game's captured score distributions. They do not include uncertainty from drawing new runs within games, and related game mechanics could produce dependence between rows that HC3 does not model. A broader population statement would need a justified game-sampling and run-resampling design.

Table 1. *Equal-game intercept regressions, N=55, HC3 covariance and t54 intervals. M1 is a probability difference relative to 0.5; M2 is the proportion of all eligible games with strict direction conflict. No significance stars are used, and no Mann-Whitney or stochastic-dominance test is reported.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.3170 (0.0265)	[0.2640, 0.3701]	55
M1: fit	R-squared = 0.0000	-	-	55
M2	Mean outcome/contrast	0.0727 (0.0357)	[0.0012, 0.1442]	55
M2: fit	R-squared = 0.0000	-	-	55

Model labels: M1 = Equal-game probability of superiority minus one half; M2 = Fraction of games with strictly opposed mean and win-probability directions.

5. Matching overall counts hide four reversals

The mean normalized gap across games is +1.030456 in the source units. Mean gaps favor BBF in 49 games, favor the variant in five, and tie in one. Superiority probabilities also favor BBF in 49, favor the variant in five, and tie in one. These identical marginal counts would conceal the disagreement if the report omitted the identities of the games on either side.

The equal-game superiority probability is 0.817044. Model 1 therefore estimates a centered advantage of 0.317044, with standard error 0.026473 and 95% interval [0.263969, 0.370119]. Adding one half gives the corresponding interval for the mean probability under the same working assumptions. This is the chance objective averaged across the selected games, not a claim that BBF wins 81.7% of every game's runs.

Four games have strictly opposed directions. Amidar has a mean gap of +0.004469 but superiority 0.486667; Skiing has a mean gap of +0.157691 but superiority 0.493333. Both favor BBF by score magnitude and the variant by pairwise probability. Bowling has a mean gap of -0.006065 with superiority 0.573333, and Private Eye has -0.013731 with superiority 0.505000, reversing in the other direction.

Model 2 estimates a conflict fraction of 0.072727, or four of 55, with standard error 0.035665 and interval [0.001224, 0.144231]. The interval is a small-sample linear working summary, not a coverage guarantee for an unlimited game population. Several of the conflicting probabilities are close to one half, so the recorded directions should not be promoted into confident population rankings without additional run-level uncertainty analysis.

The point of these examples is not that the means are erroneous. They answer a different question from pairwise probability. A user who values a rare large score may rationally prefer a condition with a higher mean; a user who values winning a typical independent contest may prefer the other. Cost-sensitive decision analysis provides a general reason to define the objective before selecting a system (Elkan, 2001), rather than declaring one aggregate universally correct.

The probability is not simply a comparison of the two medians either. It integrates the relative ordering of every recorded score against every score from the other condition. Calling it a typical-run win can be convenient, but the precise interpretation remains independent uniform draws from the two finite distributions. It does not promise that the median BBF run exceeds the median variant run, or that the advantage holds after conditioning on a particular seed or game state.

An aggregate of 0.817 also leaves the game-selection policy unresolved. It corresponds to choosing an eligible game uniformly and then drawing its two condition runs. Choosing games in proportion to play time, user demand, or recording count would define another target. Those weights were not measured here. Equal-game weighting is an explicit benchmark convention, rather than evidence about how frequently a player encounters each condition in practice.

Method-comparison work encourages attention to disagreement on named scales (Bland and Altman, 1986; Lin, 1989). Here we inspect direction agreement instead of computing a correlation between two measures with different units. A strong global correlation could coexist with precisely the four local reversals relevant to a game-specific choice. The reporting implication is to retain the game-level estimates alongside any overall summary.

6. Boundaries for interactive-system interpretation

Game environments and world-model systems present different evaluation objects. Procedural environments emphasize variation across generated levels (Cobbe et al., 2019), while GameNGen and Genie concern generated interactive environments (Valevski et al., 2024; Bruce et al., 2024). GameFactory explicitly addresses controllable game-video generation (Yu et al., 2025). The released score comparisons contain no generated frames, so they cannot establish image fidelity, controllability, or interaction latency for these systems.

World-model research also separates representation learning from the policy scores produced through an environment. World Models and DreamerV3 provide examples of learned internal dynamics in decision systems (Ha and Schmidhuber, 2018; Hafner et al., 2023), while MuZero and EfficientZero explore model-based planning and data-efficient learning (Schrittwieser et al., 2019; Ye et al., 2021). These examples support stating which system component is being evaluated, rather than treating an arcade score as a direct test of a rendering model.

The evaluation interface itself deserves a record. Gymnasium and CleanRL support explicit environment and implementation practices (Towers et al., 2024; Huang et al., 2021). Offline settings such as D4RL and return-conditioned sequence modeling introduce other distributions and objectives (Fu et al., 2020; Chen et al., 2021). We do not mix their results into this cohort. Their relevance is to the scope of an evaluation claim: the same word, performance, can refer to materially different tasks and information conditions.

For probabilistic predictions, scoring-rule theory provides a principled relationship between a stated target and its evaluation loss (Gneiting and Raftery, 2007). Our superiority probabilities are empirical distribution summaries, not model forecasts whose calibration is scored. Keeping that distinction avoids claiming that a probability-oriented ranking is automatically a proper scoring rule or a complete account of risk.

Selection procedures can add further optimism if they choose the objective or method after inspecting the benchmark. Validation-design concerns (Roberts et al., 2017) and model-selection bias (Varma and Simon, 2006) apply to that choice as well as to fitted parameters. We report both predeclared summaries and all four conflicts, rather than selecting whichever objective favors a condition in each game. Statistical interpretation should retain uncertainty and practical meaning beyond a thresholded p-value (Wasserstein and Lazar, 2016).

7. A reproducible answer with a defined denominator

The saved game-level table contains run counts, means, pairwise wins and ties, probabilities, and conflict flags. Clear dataset records support inspecting the source and exclusions (Gebru et al., 2021), while reusable identifiers and documentation support verification (Wilkinson et al., 2016). Broader data and evaluation pipelines can carry hidden dependencies (Sculley et al., 2015); preserving source snapshots matters even when the final calculation is a simple cross-product of scores.

Mean score direction and independent-run superiority direction agree for most games here, but disagree in four. Equal numbers of overall favorable games do not imply agreement about which games favor a condition. The practical reporting requirement is consequently specific: name the objective, weight games explicitly, retain ties, and do not count reused run pairs as fresh experiments. These steps make the benchmark comparison reviewable while leaving claims about new games, training seeds, and interactive rendering to evidence that actually measures them.

References

1. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
2. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
3. Schwarzer, Max; Obando-Ceron, Johan; Courville, Aaron; Bellemare, Marc; Agarwal, Rishabh; Castro, Pablo Samuel. (2023). Bigger, Better, Faster: Human-level Atari with human-level efficiency. arXiv:2305.19452. https://arxiv.org/abs/2305.19452.
4. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
5. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.
6. Towers, Mark; Kwiatkowski, Ariel; Terry, Jordan; Balis, John U.; De Cola, Gianluca; Deleu, Tristan; Goulão, Manuel; Kallinteris, Andreas; Krimmel, Markus; KG, Arjun; Perez-Vicente, Rodrigo; Pierré, Andrea; Schulhoff, Sander; Tai, Jun Jet; Tan, Hannah; Younis, Omar G.. (2024). Gymnasium: A Standard Interface for Reinforcement Learning Environments. arXiv:2407.17032. https://arxiv.org/abs/2407.17032.
7. Chen, Lili, et al. "Decision Transformer: Reinforcement Learning via Sequence Modeling." *arXiv preprint arXiv:2106.01345*, 2021. https://arxiv.org/abs/2106.01345.
8. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.
9. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
10. Cobbe, Karl, et al. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.
11. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
12. Meng, Zijie; Che, Jinming; Wei, Bingcai; Cao, Xixin. (2026). Make a game: A novel paradigm for interactive game rendering. ICASSP 2026, 1026-1030. https://doi.org/10.1109/ICASSP55912.2026.11461927.
13. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

14. Hafner, Danijar, et al. "Mastering Diverse Domains through World Models." *arXiv preprint arXiv:2301.04104*, 2023. https://arxiv.org/abs/2301.04104.

15. Bruce, Jake, et al. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.

16. Huang, Shengyi, et al. "CleanRL: High-quality Single-file Implementations of Deep Reinforcement Learning Algorithms." *arXiv preprint arXiv:2111.08819*, 2021. https://arxiv.org/abs/2111.08819.

17. Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.

18. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

19. Vargha, András; Delaney, Harold D.. (2000). A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. Journal of Educational and Behavioral Statistics. https://doi.org/10.3102/10769986025002101.

20. Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

21. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

22. Ye, Weirui, et al. "Mastering Atari Games with Limited Data." *arXiv preprint arXiv:2111.00210*, 2021. https://arxiv.org/abs/2111.00210.

23. Valevski, Dani, et al. "Diffusion Models Are Real-Time Game Engines" *arXiv preprint arXiv:2408.14837*, 2024. https://arxiv.org/abs/2408.14837.

24. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

25. Henderson, Peter; Islam, Riashat; Bachman, Philip; Pineau, Joelle; Precup, Doina; Meger, David. (2017). Deep Reinforcement Learning that Matters. arXiv:1709.06560. https://arxiv.org/abs/1709.06560.

26. Agarwal, Rishabh, et al. "Deep Reinforcement Learning at the Edge of the Statistical Precipice" *arXiv preprint arXiv:2108.13264*, 2021. https://arxiv.org/abs/2108.13264.

27. Schrittwieser, Julian, et al. "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model." *arXiv preprint arXiv:1911.08265*, 2019. https://arxiv.org/abs/1911.08265.

28. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

29. Fu, Justin, et al. "D4RL: Datasets for Deep Data-Driven Reinforcement Learning." *arXiv preprint arXiv:2004.07219*, 2020. https://arxiv.org/abs/2004.07219.

30. Yu, Jiwen, et al. "GameFactory: Creating New Games with Generative Interactive Videos." *arXiv preprint arXiv:2501.08325*, 2025. https://arxiv.org/abs/2501.08325.