Journal of Data Efficiency
Volume 1 · Number 1 · 2026

How Many Released Runs Preserve a Game-Benchmark Ranking?

Scott Hall[1] · Dominic Hall[2] · Matthew Fisher[3]

[1] Italy · [2] Portugal · [3] Spain

Original publication: October 24, 2026 | Article 639

Abstract

Retaining ten rather than two run records per condition increased average agreement with the full-record ranking from 86.73 to 96.33 percent across fifty-four games. The gain was 9.60 percentage points, with a game-level interval of 6.53 to 12.67 points. We independently subsampled released BBF and no-reset records, preserving nested two-, five-, and ten-run budgets in each of 4,000 repetitions. Expected squared error relative to the captured full-record mean gap was calculated exactly. The reference is the available record collection, not true population superiority. This experiment estimates sensitivity to retaining fewer published observations, without conducting new training or establishing a generally sufficient run count.

Finite-record retention experiments on BBF and its no-reset variant, with game-specific references held fixed.

How Many Released Runs Preserve a Game-Benchmark Ranking?

Finite-record retention experiments on BBF and its no-reset variant, with game-specific references held fixed.

Abstract

Retaining ten rather than two run records per condition increased average agreement with the full-record ranking from 86.73 to 96.33 percent across fifty-four games. The gain was 9.60 percentage points, with a game-level interval of 6.53 to 12.67 points. We independently subsampled released BBF and no-reset records, preserving nested two-, five-, and ten-run budgets in each of 4,000 repetitions. Expected squared error relative to the captured full-record mean gap was calculated exactly. The reference is the available record collection, not true population superiority. This experiment estimates sensitivity to retaining fewer published observations, without conducting new training or establishing a generally sufficient run count.

1. A run budget changes the evidence behind a comparison

A game-benchmark difference can be reported from a small number of runs even when a larger released record collection is available. Reducing the retained count makes the estimate more variable, but the practical consequence depends on the score distribution and how close the compared conditions are. A benchmark designer may care whether a retained sample preserves the direction of the larger-record comparison, not only whether its numerical estimate has a smaller mean squared error.

We ask how often two, five, or ten retained records per condition preserve the sign of the captured full-record mean difference. A second endpoint is expected squared estimation error relative to that finite reference. The first question requires a distributional calculation, estimated here through seeded retention draws. The second has an exact finite-population expression. Their relationship should not be confused with a learning curve in training interaction count: neither endpoint changes an agent's training data.

Interactive-game research such as Meng et al., 2026 motivates specifying how performance comparisons in a game setting obtain their evidence. Our observations are released agent scores on an existing benchmark. We generate no playable environment, rendering sequence, or visual output, and do not test that work's interactive-rendering method. The result is a measurement study relevant to planning an evaluation, with no claim about the quality of newly rendered games.

2. Released conditions and a finite ranking reference

The records come from the public material associated with Schwarzer et al., 2023, including a BBF condition and a no-reset variant. The captured files contain 2,082 and 830 run rows respectively, spanning fifty-five common games. We use the supplied normalized-score field and preserve its numeric units. No human or random-play normalization constant is estimated anew, and the CSVs do not independently establish every original training-budget or hardware detail for every row.

Each game has at least ten records in both conditions. Source counts range from sixteen to sixty for BBF and fifteen to twenty for the variant. A budget of ten therefore retains a substantial fraction of some source pools. This is important for the finite-population correction: estimating a mean from ten of fifteen existing records is different from drawing ten new runs from an effectively unlimited stochastic training population.

The complete captured condition means define each game's reference gap. One game, `montezumarevenge`, has an exact full-record tie and is excluded from the directional-ranking cohort because its reference has no winning sign. Fifty-four games remain, with a positive BBF gap in forty-nine and a negative gap in five. The tie is recorded explicitly rather than counted as a clear condition preference or assigned an arbitrary direction.

The benchmark context in Bellemare et al., 2012 and reproducibility considerations in Henderson et al., 2017 and Agarwal et al., 2021 motivate reporting this evidence boundary. A finite released collection can serve as a transparent reference without becoming the true expected performance of an algorithm. The present study asks how reduced retention reproduces that collection. It does not validate the full collection's ranking against an independent population of future training runs.

3. Nested retention with independent condition samples

Using seed 20261107, we generate 4,000 repetitions for each eligible game. Each repetition independently permutes the BBF and variant source rows. The first ten entries of each permutation define the largest retained sample; the first five and first two supply the smaller budgets. Samples are without replacement within a condition. Nesting makes budget contrasts use related retained observations, without equating a seed label across the two conditions.

The source seed identifiers are not paired across conditions. We therefore do not subtract outcomes with the same apparent position or numeric seed and describe that as a controlled paired training run. Pairing occurs at the game level and through the declared nested retention draws, while the two source-condition samples remain independent. The selected run-record indices are data-retention decisions, not newly executed episodes or new neural-network initializations.

At each budget we subtract the retained variant mean from the retained BBF mean and compare its sign with the full-record gap. Agreement means the signs match. Estimated exact ties do not count as agreement with a nonzero reference direction. Averaging the 4,000 indicators gives a finite-pool sign-agreement estimate for that game and budget. We then average across games equally, so games with more published runs do not automatically dominate the headline result.

Implementation and environment concerns represented by Huang et al., 2021 and Towers et al., 2024 explain why these retained records cannot stand in for a new controlled evaluation. We do not rerun either implementation. Systems concerns in Sculley et al., 2015 likewise motivate preserving missing operational details rather than filling them with a plausible configuration. The actual intervention is removal of observed records from a retained-score calculation, with all source outcomes fixed.

4. Exact expected squared error and simulated sign agreement

For sampling without replacement, the variance of a retained condition mean equals its sample variance multiplied by one over the retained count minus one over the full pool count. Independence between the two condition draws makes their variances add for the mean-gap estimate. That sum is the exact expected squared error relative to the full-record mean gap. The expectation is over the declared retention mechanism, not an unknown population of newly trained agents.

Increasing the retained count lowers this expected error whenever the source pool has positive variance. That direction follows from the formula; it is not an empirical law discovered by the experiment. The observed source distributions determine the magnitude and its variation across games. A lower mean squared error also does not mathematically require every game's sign-agreement probability to increase

monotonically at every budget. Sign is a thresholded property of a distribution, not the same endpoint as squared error.

Error-measure distinctions in Hyndman and Koehler, 2006 and predictive-scoring principles in Gneiting and Raftery, 2007 support treating these outcomes separately. Agreement concepts in Bland and Altman, 1986 and Lin, 1989 do not establish interchangeability between a numerical gap estimate and a binary ranking decision. We calculate no additional agreement coefficient from those methods. The directly defined sign agreement and finite squared-error expectation are the executed endpoints.

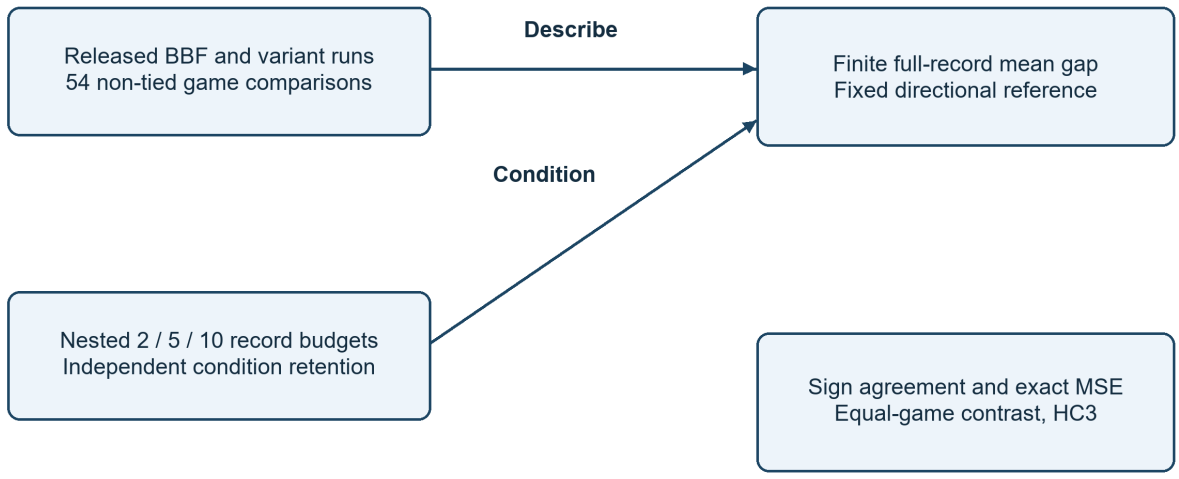


Figure 1. Independent condition samples at nested run counts estimate a fixed game-specific reference gap; sign agreement is simulated and squared-error expectation calculated exactly.

5. Game-level summary models

Each eligible game contributes a ten-minus-two sign-agreement difference and a ten-minus-two expected squared-error difference. Intercept-only ordinary least-squares models estimate the equal-game averages. There are fifty-four regression rows, not thousands of independent games created by resampling. The Monte Carlo repetitions improve estimation of a finite game's sign probability but do not increase the number of distinct benchmark tasks.

We use HC3 covariance and normal-reference intervals under a hypothetical game-draw interpretation, following the heteroskedasticity-consistent approach in White, 1980. Dependence considerations in Cameron and Miller, 2015 remain a limit because benchmark games can share design and score characteristics. There is no claim that HC3 establishes a random sample of all interactive games or incorporates uncertainty in the original training process. The fits summarize variation across the selected released game collection.

Table 1. Ten-minus-two retained-record contrasts, with one observation per game.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.0960 (0.0157)	[0.0653, 0.1267]	54
M1: fit	R-squared = 0.0000	-	-	54
M2	Mean outcome/contrast	-0.3130 (0.0940)	[-0.4974, -0.1287]	54
M2: fit	R-squared = 0.0000	-	-	54

Model labels: M1 = Ten-minus-two retained-run sign agreement; M2 = Ten-minus-two exact finite-record expected squared error.

The mean-only regressions have zero R-squared. Their intervals are different from simulation error in a per-game sign-agreement estimate. For 4,000 independent retention repetitions, the largest Bernoulli standard-error bound for one estimated agreement fraction is about 0.007906. We preserve that numerical approximation separately. It is not a certification of the original ranking or a substitute for uncertainty across training seeds not represented in the captured pools.

6. The observed budget curve

Equal-game sign agreement is 0.867291667 at two retained runs per condition, 0.925175926 at five, and 0.963310185 at ten. The ten-minus-two gain is 0.096018519, with HC3 standard error 0.015668926 and interval 0.065307987 to 0.126729050. In percentage-point terms, the average gain is approximately 9.60, with interval 6.53 to 12.67. These are agreements with the finite reference, not probabilities that BBF is truly superior.

The remaining disagreement at ten records is informative. A larger retained count does not reproduce every finite reference comparison in every draw. Close gaps and variable outcomes can still change the estimated direction. We do not remove such games from the result or present ten as a guaranteed sufficient count. The tied reference game was excluded for a distinct definitional reason, before asking whether a sampled sign reproduces a nonzero direction.

Exact expected squared gap error averages 0.363503125, 0.128726168, and 0.050467182 in squared released-normalized-score units at the three budgets. The ten-minus-two change is -0.313035944, with standard error 0.094042090 and interval -0.497355054 to -0.128716834. The magnitude is source-scale-dependent. It cannot be read as a reward improvement, a change in rendering quality, or a direct percentage of game performance explained by the larger budget.

Expected squared error is calculated, not estimated from 4,000 draws. The simulated squared errors are retained as a diagnostic alongside the exact quantities, without replacing the latter in the primary endpoint. This separation avoids describing Monte Carlo noise as a scientific discrepancy in a known finite-population identity. It also makes the role of resampling clear: approximation is required for the chosen sign probability here, while the squared-error expectation has a direct calculation.

7. Score variation and alternative evaluation targets

Influence reasoning in Cook, 1977 motivates looking at games whose large score variation contributes heavily to squared-error averages. We retain every eligible non-tied game. Robust-estimation traditions in Huber, 1964, Hampel, 1974, and Rousseeuw, 1984 would support a different summary of across-game extremes, but no robust fit is substituted for the declared mean. A trimmed or rank-based endpoint could answer another question and would need its own explicit definition.

The return-estimation context in Chen et al., 2021 and the offline-benchmark perspective represented by Fu et al., 2020 further illustrate that evidence requirements depend on the target. Neither contributes additional observations. Agent-learning approaches such as Schrittwieser et al., 2019 and Ye et al., 2021 also operate at a different stage from retaining published outcomes. The current record budget cannot be converted into their training sample efficiency or used to compare their algorithms without actually collecting compatible results.

Decision-cost reasoning in Elkan, 2001 is relevant if a researcher wants to trade evaluation stability against expense. We measure record count, not wall time, GPU cost, environmental interaction, or personnel effort. Actual costs can differ between conditions or games, and the source table does not provide them. A favorable retention curve alone therefore cannot identify the most economical new experiment or justify a fixed run budget for a different benchmark.

8. Validation and interpretation limits

The source collection defines both the reference and the samples retained from it. This is intentional finite-record estimation, not independent held-out validation of a winner. Model-selection concerns in Varma and Simon, 2006 would arise if the same record collection were used to select an algorithm and to certify its final advantage. Validation principles in Roberts et al., 2017 also caution against assuming that the same run budget performs similarly across unseen games, training regimes, or observation conditions.

Interpretation following Wasserstein and Lazar, 2016 emphasizes the effect sizes and their target rather than turning an interval into a general recommendation. Multiplicity considerations in Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 would become important in a broad search over many ranking criteria, game subsets, and budget grids. We report the three declared retained counts and the equal-game contrast. We do not search for a budget or subset that makes the source condition appear more consistently favorable.

The full-record reference itself remains uncertain as an estimate of expected algorithm performance. Its available run counts differ across conditions and games, and the CSVs do not resolve every provenance question about training. Preserving this limitation is preferable to inventing missing experimental detail. The completed result describes how a known record-retention process behaves on the released scores, with a distinct question about future newly generated runs left open.

9. Attach the record budget to the reported comparison

The experiment finds that retaining more released observations improves average agreement with a fixed finite ranking, while leaving some disagreements at the largest tested count. The exact squared-error calculation explains one direction of that change, and the sign experiment quantifies a separate consequence relevant to comparative claims. These observations support reporting run counts and retained-record sensitivity with a benchmark difference, without establishing a universal minimum for interactive-game research.

Documentation under Gebru et al., 2021 records the source-condition pools, exact tied-game exclusion, seed ordering, independent condition draws, nested budgets, and finite-reference definition. Preservation following Wilkinson et al., 2016 retains game-level probabilities, exact expectations, and reproducible resampling code. These materials allow the stated sensitivity curve to be recomputed. Determining a generally sufficient training-run budget or evaluating an interactive rendering system requires additional independent evidence beyond this released-record experiment.

References

1. 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.
2. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.
3. 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.
4. 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.
5. Ye, Weirui, et al. "Mastering Atari Games with Limited Data." *arXiv preprint arXiv:2111.00210*, 2021. https://arxiv.org/abs/2111.00210.
6. Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.
7. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
8. 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.
9. 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.
10. Bellemare, Marc G., et al. "The Arcade Learning Environment: An Evaluation Platform for General Agents" *arXiv preprint arXiv:1207.4708*, 2012. https://arxiv.org/abs/1207.4708.
11. Fu, Justin, et al. "D4RL: Datasets for Deep Data-Driven Reinforcement Learning." *arXiv preprint arXiv:2004.07219*, 2020. https://arxiv.org/abs/2004.07219.
12. 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.
13. 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.
14. 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.
15. 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.
16. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
17. 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.
18. Chen, Lili, et al. "Decision Transformer: Reinforcement Learning via Sequence Modeling." *arXiv preprint arXiv:2106.01345*, 2021. https://arxiv.org/abs/2106.01345.
19. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
20. 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.
21. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
22. 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

23. 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.

24. 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.

25. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.

26. Benjamini, Yoav and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.

27. 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.

28. 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.

29. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.

30. 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.