Journal of Transfer and Generalization
Volume 1 · Number 1 · 2026

More Recorded Runs, More Weight: A Decomposition of a Pooled Atari Comparison

Julian Walker[1] · Jonathan Mitchell[2] · Gavin Bell[3]

[1] Iceland · [2] Malta · [3] Cyprus

Original publication: November 1, 2026 | Article 571

Abstract

Simply averaging every available run gives different games different weights, and those weights can differ between the compared conditions. In captured results for 55 Atari games, separately pooled BBF and no-reset means differed by 1.3800 normalized-score units. Giving each game equal weight reduced the gap to 1.0305. An exact symmetric decomposition attributed 0.2410 of the discrepancy to different game composition between conditions and 0.1086 to common reweighting of within-game gaps. Both aggregates favored BBF, but their magnitudes answered different questions. Reporting run counts alongside scores made that distinction recoverable.

Analysis of captured public benchmark results. The decomposition changes the population represented by an aggregate; it does not rerun an agent, identify a causal reset effect or evaluate a game renderer.

More Recorded Runs, More Weight: A Decomposition of a Pooled Atari Comparison

> Analysis of captured public benchmark results. The decomposition changes the population represented by an aggregate; it does not rerun an agent, identify a causal reset effect or evaluate a game renderer.

Abstract

Simply averaging every available run gives different games different weights, and those weights can differ between the compared conditions. In captured results for 55 Atari games, separately pooled BBF and no-reset means differed by 1.3800 normalized-score units. Giving each game equal weight reduced the gap to 1.0305. An exact symmetric decomposition attributed 0.2410 of the discrepancy to different game composition between conditions and 0.1086 to common reweighting of within-game gaps. Both aggregates favored BBF, but their magnitudes answered different questions. Reporting run counts alongside scores made that distinction recoverable.

1. An aggregate result can change before any score changes

Suppose a benchmark contains more recorded runs for some games than others. An average over all records then represents a draw from the stored records, not a draw that first selects a game uniformly. If the second condition has a different distribution of records across games, subtracting its pooled mean adds another complication: the two averages no longer describe the same mixture of games.

Our research question is how much this distinction changes a captured Atari comparison and whether the discrepancy comes from different game composition or from weighting the same within-game contrasts differently. The individual scores remain fixed. We examine the accounting of a reported mean, with no resampling selected to create a reversal and no removal of high-scoring games to make the conditions look more similar.

Meng et al., 2026 concerns interactive game rendering as a research paradigm. Our analysis does not test its renderer or convert policy scores into rendering-quality measures. Its connection is methodological: comparisons made with interactive environments require an explicit definition of the collection over which performance is summarized. That definition remains consequential even when every stored score is read correctly.

The empirical example uses the captured BBF and no-reset run files associated with the BBF research context (Schwarzer et al., 2023). The two conditions favor the same direction under both of our main summaries, so the finding is not a claim of a hidden ranking reversal. It is a quantified difference between two aggregate effect magnitudes whose denominators and game weights can be inspected directly.

2. Benchmarks supply tasks as well as observations

The Arcade Learning Environment establishes the task context for Atari comparisons (Bellemare et al., 2012). A game is a meaningful stratum because the numerical score distribution, range and difficulty differ across games. DeepMind Control Suite provides another example of a collection built from distinct control tasks (Tassa et al., 2018). Its results are not added to our data, but it reinforces why a benchmark's task structure should survive the final averaging step.

Research on reliable reinforcement-learning evaluation emphasizes uncertainty and the sensitivity of aggregate summaries (Agarwal et al., 2021). Reproducibility concerns extend to implementation and reporting choices (Henderson et al., 2017). Our narrower issue is the deterministic weighting implied by

unequal numbers of stored runs. Even a perfectly reproducible sum can answer an unintended question when that weighting is left unstated.

World-model studies illustrate the range of capabilities evaluated in interactive research. World Models, DreamerV3 and EfficientZero examine learning or control using modeled dynamics (Ha and Schmidhuber, 2018; Hafner et al., 2023; Ye et al., 2021). DIAMOND, Genie and GameFactory provide additional contexts involving learned environments or visual world construction (Alonso et al., 2024; Bruce et al., 2024; Yu et al., 2025). Those studies motivate explicit evaluation units; none of their model outputs is pooled with the present Atari records.

Software and interface choices are also part of that definition. CleanRL and Gymnasium document implementation and environment conventions (Huang et al., 2021; Towers et al., 2024). A common interface does not imply equal empirical task weights in a result file. Similarly, static-data resources and sequence-based control methods have their own evaluation cohorts (Fu et al., 2020; Chen et al., 2021). The stored row count is a property of the captured evidence, not automatically the probability distribution intended by a benchmark claim.

3. The two source files and the count imbalance

The BBF file contains 2,082 run records; the no-reset file contains 830. Both contain the same 55 game labels. We use all of those common games and all captured records. The outcome is the supplied GameScoreNormalized field. Raw game points are not averaged across games, and the released normalized values are not clipped, winsorized or transformed for this analysis.

BBF contributes twenty records for 27 games, sixty for fifteen games, 56 for nine games, 52 for two games, sixteen for one game and eighteen for one game. The no-reset file contains fifteen records for 54 games and twenty for one. These unequal counts are precisely what the pooled calculation converts into unequal task weights. We do not infer why the counts differ or whether any unavailable run would have changed the result.

Within each game, we calculate each condition's mean normalized score and record its count. We do not pair source seed identifiers across files, since shared numeric labels do not by themselves establish matched experimental draws. The present decomposition needs the condition-specific means and counts, not a hypothetical run-level pairing or an estimate of a common random-seed effect.

The captured records do not establish a randomized reset intervention with every other training and selection choice controlled. Accordingly, the score gap is a condition contrast in these files. Calling it the causal effect of resetting would require stronger design and provenance evidence. Our arithmetic remains valid for the captured contrast without making that additional claim.

4. Three quantities behind one apparent difference

An equal-game contrast first subtracts the no-reset mean from the BBF mean within each game and then averages those 55 differences. Every game receives weight one divided by 55. In contrast, a separately pooled contrast averages BBF records using BBF's observed game shares and no-reset records using the no-reset file's shares, then subtracts the resulting condition means.

We connect the two using a symmetric decomposition in the spirit of rate-composition accounting (Kitagawa, 1955). For each game, the midpoint weight is the average of its two condition-specific record shares, and the midpoint score is the average of its two condition means. The pooled contrast equals the midpoint-weighted within-game gap plus the sum of the weight differences multiplied by the midpoint scores.

The latter sum is the game-composition component. It can be nonzero even if a game's two condition means are nearly identical, because a generally high-scoring game can receive more weight in one condition. It is an algebraic composition term, not a mediator, a behavioral mechanism or an estimate of bias relative to an identified causal treatment effect.

We then subtract the equal-game contrast from the midpoint-weighted gap. The remainder is a common-reweighting component: both conditions now use the same game weights, but those weights differ from one divided by 55. The full difference between separate pooling and equal-game averaging is exactly the composition component plus this common-reweighting component.

Weighting arguments require a target quantity rather than a universal preference for larger samples (Solon et al., 2015). Observed record shares here are neither inverse-probability survey weights nor estimated inverse-variance weights. We calculate the aggregate they imply and compare it with an explicitly uniform game distribution. A different intended task population would justify another declared weighting scheme and another numerical estimand.

5. Making the accounting visible in the result table

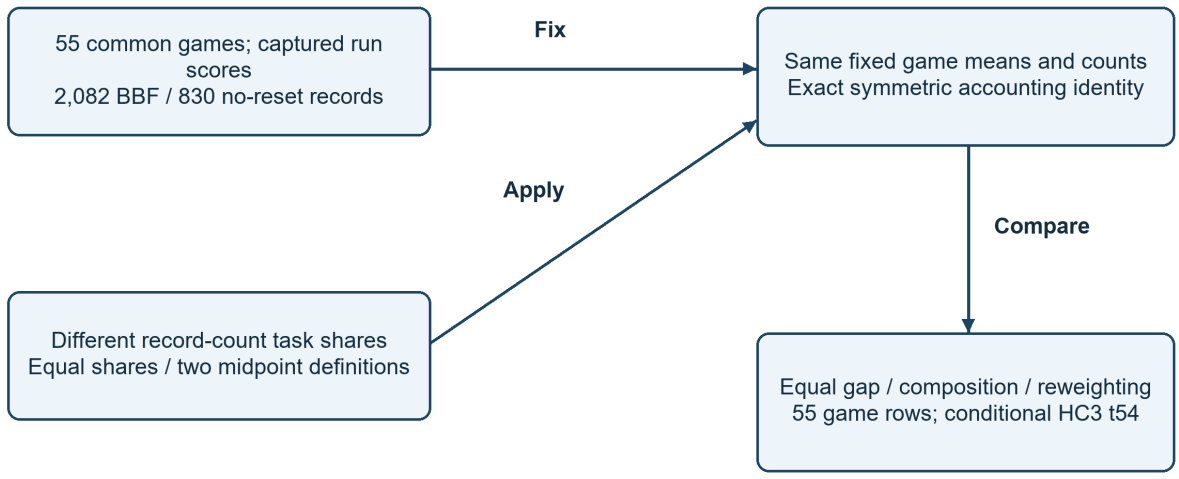


Figure 1. *Fixed condition means and record counts generate separate-pool and equal-game summaries. A symmetric identity separates game composition from common reweighting of the same within-game gaps.*

For the table, Model 1 is an intercept regression on the 55 within-game score gaps. Models 2 and 3 regress the corresponding game-level composition and common-reweighting contributions on an intercept. Each contribution is multiplied by 55 before fitting, so its equal-game mean equals the aggregate component. This scaling is bookkeeping, not an expansion of the number of independent observations.

All three regressions use HC3 standard errors and t54 intervals, with no controls. Intercept-only R-squared is zero apart from machine precision. Heteroskedasticity-consistent variance addresses unequal dispersion of the game-row quantities under a working regression calculation (White, 1980). It does not model uncertainty in the captured run counts or the condition means estimated from those finite records.

Table 1. *Equal-game gap and scaled aggregate components, N=55 game rows, HC3 covariance, t54 confidence intervals. The latter two coefficients sum to the pooling-minus-equal discrepancy. Their marginal intervals cannot be added to obtain an interval for that discrepancy.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	1.0305 (0.2743)	[0.4806, 1.5804]	55
M1: fit	R-squared = 0.0000	-	-	55
M2	Mean outcome/contrast	0.2410 (0.1147)	[0.0111, 0.4709]	55
M2: fit	R-squared = 0.0000	-	-	55
M3	Mean outcome/contrast	0.1086 (0.0767)	[-0.0452, 0.2625]	55
M3: fit	R-squared = 0.0000	-	-	55

Model labels: M1 = Equal-game captured score gap; M2 = Different game-composition contribution; M3 = Common versus equal-game weighting contribution.

The component rows also share global weight normalizers, so the working intervals do not establish general repeated-sampling coverage. The mathematical identity itself is exact for the captured files and does not need a significance test. We retain conventional interval summaries to show game-level variation while distinguishing that variation from uncertainty about an unobserved population of future tasks or runs.

6. Pooling enlarges the observed gap

The separately pooled BBF mean is 1.832150 and the no-reset mean is 0.452115, producing a gap of 1.380035. With equal game weight, the condition means are 1.484269 and 0.453813, producing a gap of 1.030456. The 0.349579 difference between those gaps is substantial relative to the equal-game contrast, even though both favor the same condition.

Model 1 estimates the equal-game gap at 1.030456, with standard error 0.274281 and interval [0.480555, 1.580357]. Its unit is the supplied normalized-score unit. It is not a percentage-point accuracy difference, a win probability or a rendering-quality gain. Measurement-scale distinctions matter when a benchmark headline moves between these different forms of evidence (Stevens, 1946).

The symmetric composition component is 0.240955, with working standard error 0.114671 and interval [0.011054, 0.470856]. The midpoint-weighted within-game gap is 1.139080. Relative to equal game weight, common reweighting adds 0.108624, with standard error 0.076749 and interval [-0.045249, 0.262496]. These two additions recover the pooled gap to floating-point precision.

The first addition is larger than the second in this resource. That does not mean it is more causally important. It says that the stated symmetric accounting assigns more of the discrepancy to different game mixtures than to applying one shared nonuniform mixture to the within-game gaps. Another reference decomposition could allocate an interaction differently, so the declared midpoint convention belongs with the reported numbers.

Krull illustrates the composition distinction particularly clearly. Its BBF and no-reset means are approximately 6.28124 and 6.07398, a modest within-game gap relative to either level. Yet it has 56 BBF records and fifteen no-reset records. Its scaled composition contribution is about 2.99843, while its scaled common-reweighting contribution is only 0.04905. A high common score level can matter through unequal task shares even when the within-task contrast is small.

Breakout combines a large within-game contrast with unequal shares. Its means are approximately 12.33640 and 1.16731, again with 56 and fifteen records. Its scaled composition and common-reweighting contributions are about 3.27716 and 2.64329. These are game-row contributions multiplied by 55; each contributes one fifty-fifth of that displayed row quantity to the aggregate coefficient. We state the scaling to prevent confusing a large row contribution with the total discrepancy.

7. A transparent aggregate is more useful than a preferred headline

Extreme contributions deserve inspection, but deleting them would change the target collection. Influence diagnostics and robust-estimation theory explain why a mean can respond strongly to particular observations (Cook, 1977; Hampel, 1974). Here the high-scoring games remain included. Their role is disclosed rather than removed after observing that they enlarge the pooled difference.

Statistical testing guidance also separates effect size from a thresholded declaration of significance (Wasserstein and Lazar, 2016). We do not search across games for individually significant contrasts or run an unreported family of comparisons. If future work did test many game-specific hypotheses, multiple-testing procedures would become relevant (Benjamini and Hochberg, 1995). The present main result is the exact change in an aggregate definition, accompanied by conditional descriptive intervals.

There is a further reason to preserve within-game contrasts. Consider adding the same game-specific constant to both condition means. The within-game gap is unchanged, and so are its equal-game and common-weighted averages. The separately pooled gap can nevertheless change because the added constant receives different weights in the two conditions. The symmetric composition term records exactly that difference. This algebra does not change any source score in our analysis; it clarifies why condition-specific mixtures are sensitive to score levels as well as condition differences.

A positive game-specific rescaling has a different consequence. It preserves the direction of that game's contrast but changes its magnitude and its influence on a cross-game average. The released human-normalized scale supplies one explicit convention for comparing games. Our conclusions are conditional on that supplied scale. We neither pool raw points nor claim invariance to arbitrary choices of per-game normalization, since the latter would be mathematically false for a mean-score aggregate.

Equal game weight also does not mean that a game with sixteen records has the same precision as one with sixty. It gives those games equal importance in the target summary despite different information about their respective means. Estimating a sampling distribution that includes finite-run uncertainty would require an additional design, such as a justified hierarchy of task and run resampling. The present fixed-file decomposition addresses target weights first and leaves that uncertainty requirement explicit.

The decomposition cannot tell a reader which aggregate matches their practical priorities. An application that genuinely samples stored runs has one distribution; a benchmark intended to represent each game equally has another. Classification-performance comparisons likewise depend on the decision objective rather than a single universally correct scalar (Elkan, 2001). Predictive-scoring theory reinforces the need to name the quantity being optimized (Gneiting and Raftery, 2007), without making our normalized policy score a probabilistic scoring rule.

Generalization claims would require a validation population beyond these selected games (Roberts et al., 2017). Outcome-driven selection of a favored aggregation rule would introduce an additional evaluation choice that needs protection from overfitting (Varma and Simon, 2006). We therefore present both observed summaries and the identity connecting them. Neither is silently promoted to the uniquely valid definition of overall performance.

The reusable record contains game counts, condition means, midpoint weights and both component columns. Dataset documentation supports interpreting those fields (Gebru et al., 2021), and provenance makes their source traceable (Wilkinson et al., 2016). A reader can reconstruct either headline from the same captured evidence. That recoverability is the practical result: changing the number of stored runs changes the represented game mixture unless the intended task weights are explicitly restored.

References

1. 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.
2. Yu, Jiwen, et al. "GameFactory: Creating New Games with Generative Interactive Videos." *arXiv preprint arXiv:2501.08325*, 2025. https://arxiv.org/abs/2501.08325.
3. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
4. Bruce, Jake, et al. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.
5. 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.
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. 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.
8. Kitagawa, Evelyn M.. (1955). Components of a Difference Between Two Rates*. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1955.10501299.
9. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
10. Fu, Justin, et al. "D4RL: Datasets for Deep Data-Driven Reinforcement Learning." *arXiv preprint arXiv:2004.07219*, 2020. https://arxiv.org/abs/2004.07219.
11. Solon, Gary; Haider, Steven J.; Wooldridge, Jeffrey M.. (2015). What Are We Weighting For?. Journal of Human Resources. https://doi.org/10.3368/jhr.50.2.301.
12. 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.
13. 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.
14. Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.
15. Ye, Weirui, et al. "Mastering Atari Games with Limited Data." *arXiv preprint arXiv:2111.00210*, 2021. https://arxiv.org/abs/2111.00210.
16. 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.
17. 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.
18. Tassa, Yuval, et al. "DeepMind Control Suite." *arXiv preprint arXiv:1801.00690*, 2018. https://arxiv.org/abs/1801.00690.
19. 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.
20. Alonso, Eloi, et al. "Diffusion for World Modeling: Visual Details Matter in Atari" *arXiv preprint arXiv:2405.12399*, 2024. https://arxiv.org/abs/2405.12399.
21. 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.

22. Chen, Lili, et al. "Decision Transformer: Reinforcement Learning via Sequence Modeling." *arXiv preprint arXiv:2106.01345*, 2021. https://arxiv.org/abs/2106.01345.

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. 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.

26. 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.

27. 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.

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

30. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.