Retaining Negative Returns when Comparing Heterogeneous Games

Gavin Russell (GA), Sean Spencer (MT), James Bennett (SI)

Abstract

The public BBF comparison includes 186 negative normalized scores and 63 exact zeros. A transformation used for aggregation must retain those observations honestly. An equally game-weighted model gives a raw-score contrast of 1.030 and an inverse-hyperbolic-sine contrast of 0.466, both positive but expressed in different units. Their intervals are 0.480 to 1.581 and 0.300 to 0.633 respectively. This secondary analysis explains why a change of scale can be a declared sensitivity analysis while deletion or arbitrary shifting can silently change the represented outcomes.

1. A scale choice can remove the most relevant cases

Performance distributions can contain values for which a convenient mathematical transformation is not defined. An ordinary logarithm cannot be applied directly to zero or negative scores. Adding a constant changes its interpretation and can make the result depend on a convention selected for convenience. Deleting the incompatible values changes the observation population. Either decision may be defensible in a specific analysis, but it should not occur unnoticed inside preprocessing.

The issue matters to Make a game's interactive evaluation problem (Meng et al., 2026a) because adverse outcomes can be the very observations an interface needs to retain. A rendering study may measure a difference from a baseline or a signed error rather than a strictly positive quantity. Its scoring pipeline should preserve what those values mean. This paper does not analyze output from that renderer. It uses released BBF scores to examine two transparent scales while keeping every finite observation in the comparison.

A benchmark interface supplies a common language for launching tasks, but it does not eliminate differences among the tasks. The Arcade Learning Environment was designed to evaluate general agents across games (Bellemare et al., 2012). Gym offers a common interaction interface (Brockman et al., 2016), while Procgen makes procedural variation an explicit part of the assessment population (Cobbe et al., 2019). These contributions help distinguish software compatibility from evidence of transfer. The BBF release supplies actual repeated-run scores that can be reanalyzed (Schwarzer et al., 2023), whereas the statistical-precipice analysis explains why uncertainty and aggregation choices deserve attention (Agarwal et al., 2021). Together, these sources motivate preserving task-level results. They do not imply that an Atari return is a visual-quality score or that a renderer supports the mechanics of every game available through the same interface. A proposed evaluation should identify its games, initial conditions, and action histories, then state which population the aggregate represents. Additional runs can improve knowledge of a task's outcome distribution without increasing the number of different tasks observed.

Representation choices can change what an interactive system remembers and what it discards. The compressed environment representations in World Models (Ha and Schmidhuber, 2018), latent-imagination approach in Dreamer (Hafner et al., 2019), transformer world model in IRIS (Micheli et al., 2022), and diffusion formulation studied by DIAMOND (Alonso et al., 2024) offer different examples. Their differences matter because an apparently small visual feature can determine a later action or reward. A proposed rendering mechanism should be assessed through controlled observations of that consequence, rather than justified solely through the family of architecture it resembles. It should also distinguish short-horizon reconstruction from the persistence of behaviorally relevant information. No comparison among those architectures is performed by describing their designs. Their role here is to

identify alternative explanations and appropriate baseline questions for a subsequent study. An informative experiment would match available history, training access, action information, and evaluation duration before attributing a difference to one representation.

2. The observations preserved by both specifications

The combined public files contain 2,912 unique configuration-game-seed records over 55 common games. The minimum supplied normalized score is -1.500 and the maximum is 16.706. Among all records, 186 are strictly negative and 63 are exactly zero. These counts make it clear that applying an ordinary logarithm without another decision would not preserve the full dataset. The analysis performs no clipping, zero replacement, or deletion based on score magnitude.

Two weighted regressions use the same records, configuration indicator, game fixed effects, and reciprocal configuration-by-game count weights. The first outcome is the released normalized score. The second is its inverse hyperbolic sine, which is defined over the full real line and compresses large magnitudes. Each configuration has equal total weight within each game. Both covariance estimates cluster by the 55 game labels. The specifications differ in the outcome scale rather than in their sample, group weighting, or available covariates.

Precision and identification are separate properties of a result. A narrow interval can surround an association whose scientific target is unclear. Robust covariance calculations allow particular departures from simple error assumptions, including unequal variance (White, 1980) and dependence within declared clusters (Cameron and Miller, 2015). Their role is to improve uncertainty calculations under those assumptions. They do not establish a causal assignment, guarantee a correct measurement, or expand coverage to conditions missing from the data. The analyst should therefore explain both the averaging population and the source of replication. An evaluation repeated many times on one history may provide information about stochastic variability while supplying little evidence about transfer to another history. Those two sources of uncertainty should remain distinguishable in the final interpretation.

Describing a contrast accurately can be more valuable than giving it an unjustified mechanistic name. A causal interpretation needs assumptions about how exposure and outcome arose, including selection into the analyzed sample. Causal diagrams provide a language for those assumptions (Greenland et al., 1999). If several alternative metrics or partitions are examined, the investigator should also account for the freedom to select whichever result appears strongest. False-discovery control is relevant to explicit collections of statistical tests (Benjamini and Hochberg, 1995), while a descriptive sensitivity analysis should be presented as such. A model can fit the observed rows well and still leave the principal explanation unresolved. Reporting that boundary preserves the value of the observation without overstating what adjustment or a significance threshold has established.

Control and fidelity should be assessed as related but distinct properties. Neural game generation in GameNGen (Valevski et al., 2024) and action-controlled environment generation in Genie (Bruce et al., 2024) bring both into view. Structural similarity (Wang et al., 2004) or a learned image distance (Zhang et al., 2018) may help quantify a reference comparison; a video-distribution metric (Unterthiner et al., 2018) can summarize another aspect of temporal appearance. These measurements do not identify whether the interface interpreted a command correctly. The evaluation record should link an output to the exact control history and to any available state observation. It should also distinguish permitted variation from a broken requirement. A stochastic rendering system need not reproduce every pixel to be useful, but it does need to preserve the aspects of interaction that its claim depends on. Measuring those requirements explicitly makes the relationship between visual quality and behavior open to examination.

3. Same direction, different quantities

The raw-score configuration coefficient is 1.0305, with a standard error of 0.2744 and a 95 percent interval from 0.4804 to 1.5806. The transformed coefficient is 0.4662, with a standard error of 0.0830 and an interval from 0.2998 to 0.6327. The common positive direction is useful as a sensitivity finding. The smaller numeric transformed coefficient is not evidence that an advantage has been reduced by a corresponding percentage. The two coefficients have different units, and the transformation changes how outcome magnitudes contribute to the average.

The fitted R-squared values likewise answer scale-specific questions. Their comparison cannot establish that one outcome representation is scientifically superior. A transformation can make an additive model describe a distribution differently without changing the underlying task outcomes. The scientific reason for using it should therefore be stated: here it supplies an alternative that compresses extremes while retaining signs and zeros. It is not chosen to imitate a percentage effect or to make the numerical result resemble a desired magnitude.

For a rendering benchmark, the same discipline applies when combining errors, preferences, or task scores. A signed metric might encode whether an intervention helped or harmed an outcome. A bounded failure rate might have a meaningful probability scale. A visual distance might not. A transformation should preserve the interpretation required by the study rather than make incompatible endpoints appear comparable. If a proposed total score combines several transformed quantities, its weighting and practical meaning require a separate argument.

An especially important distinction concerns missing values. A failed generation is not necessarily equivalent to a zero score, and a score that cannot be computed is not necessarily evidence of average performance. The evaluation should retain the reason that a measurement is absent and specify the primary treatment before comparing systems. The present archive has no nonfinite values in its normalized-score field, so it does not require such imputation. That favorable property should not be generalized to generation pipelines whose failures occur before scoring.

An evaluation protocol should state what remains unfamiliar at assessment time. Leaving out an observation while retaining its close relatives can test interpolation rather than the broader transfer suggested by the study's wording. Validation methods for structured data emphasize matching the split to temporal, spatial, or hierarchical dependence (Roberts et al., 2017). Separately, parameter selection can bias an evaluation if the same observations guide both choices and reporting (Varma and Simon, 2006). These issues are not solved by calling a split a test set. The record should identify shared sources, the information accessible during development, and the stage at which each design choice was fixed. A result can be informative within a narrow boundary while providing no direct evidence about a more distant deployment population.

A substitute measurement needs evidence about its relationship to the intended judgment. Two scores may correlate strongly while differing systematically in particular conditions. Method-comparison analysis distinguishes association from agreement (Bland and Altman, 1986). When a system additionally reports confidence, calibration asks whether those probabilities correspond to observed outcome frequencies (Guo et al., 2017). These are separate properties: a metric can rank examples usefully while having a distorted numerical scale or unreliable confidence. An evaluation should therefore retain the original target observations when possible, describe the scale of each endpoint, and examine disagreements instead of averaging them away. A convenient automated score becomes more credible when its limitations are measured in the conditions that motivate its use.

4. From a task contrast to a controlled study

The distribution of available behavior constrains what can be inferred from recorded interaction. The reinforcement-learning framework distinguishes the information in a trajectory from the policy that produced it (Sutton and Barto, 2018). Offline reinforcement learning makes the limitations of fixed experience explicit (Levine et al., 2020), and conservative Q-learning studies one way to address out-of-distribution action evaluation (Kumar et al., 2020). The corresponding lesson for rendering is methodological: an archive of familiar actions is not a complete intervention experiment. If commands are chosen by an existing policy, some states and responses may remain poorly represented. An evaluation should document this support and avoid treating absent cases as successful extrapolation. Proposed counterfactual tests should compare matched initial histories where feasible and retain whether a valid response was produced. Such evidence can strengthen a limited claim without implying that the system models every possible action or game mechanic.

Difficult cases can disappear at several stages: data eligibility, generation, measurement, and final scoring. Each exclusion changes the population represented by the reported result. Coverage-aware evaluation makes those decisions visible (Geifman and El-Yaniv, 2017). A confidence score should also be checked when the input distribution changes, because uncertainty under shift is an empirical property rather than a guarantee of the model interface (Ovadia et al., 2019). An informative record preserves failures to produce an output and failures to obtain a valid measurement. Treating either as absent data without explanation can inflate apparent reliability. A proposed method should state its unsupported conditions and test the boundary between accepted and rejected cases as carefully as the quality of its successful outputs.

Reproducibility begins with preserving which scientific objects are being linked. Stable identifiers and reusable metadata help connect inputs, reference observations, model versions, and reported outcomes (Wilkinson et al., 2016). The scoring pipeline also has dependencies that can change the interpretation of a result: preprocessing, representation checkpoints, thresholds, and aggregation code can all matter (Sculley et al., 2015). A compact release should retain those decisions rather than only the final mean. Where source data cannot be redistributed, an access description and hashes can still identify the analyzed inputs. This record allows later users to distinguish a change in the underlying system from a change in how the system was measured. Reproducibility is therefore part of interpretation, not simply a final packaging step.

5. Conclusion

Both declared scales preserve the full BBF score sample and give a positive equally weighted comparison, while emphasizing magnitudes differently. The analysis establishes scale sensitivity without claiming that the coefficients are interchangeable. An interactive evaluation should document transformations, retain meaningful negative and zero outcomes, and distinguish an observed adverse value from an unavailable measurement. Those choices determine which scientific cases remain represented before any model or uncertainty calculation is applied.

References

Meng, Zijie, et al. "Make a game: A novel paradigm for interactive game rendering." *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026.

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.

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.

Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

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.

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.

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.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018. https://doi.org/10.1038/sdata.2016.18.

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

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.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

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.

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.

Cobbe, Karl, et al.. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.

Brockman, Greg, et al.. "OpenAI Gym" *arXiv preprint arXiv:1606.01540*, 2016. https://arxiv.org/abs/1606.01540.

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

Hafner, Danijar, et al.. "Dream to Control: Learning Behaviors by Latent Imagination" *arXiv preprint arXiv:1912.01603*, 2019. https://arxiv.org/abs/1912.01603.

Micheli, Vincent, et al.. "Transformers are Sample-Efficient World Models" *arXiv preprint arXiv:2209.00588*, 2022. https://arxiv.org/abs/2209.00588.

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.

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

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

Zhang, Richard, et al.. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

Unterthiner, Thomas, et al.. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

Schwarzer, Max, et al.. "Bigger, Better, Faster: Human-level Atari with human-level efficiency" *arXiv preprint arXiv:2305.19452*, 2023. https://arxiv.org/abs/2305.19452.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed., MIT Press, 2018. http://incompleteideas.net/book/the-book-2nd.html.

Levine, Sergey, Aviral Kumar, George Tucker, and Justin Fu. "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems." *arXiv*, 2020, arXiv:2005.01643. https://arxiv.org/abs/2005.01643.

Kumar, Aviral, Aurick Zhou, George Tucker, and Sergey Levine. "Conservative Q-Learning for Offline Reinforcement Learning." *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1179-1191. https://arxiv.org/abs/2006.04779.