Unequal Experimental Coverage in a Released Game Benchmark

Benjamin Fisher (ME), Jack Phillips (BB), Benjamin Marshall (KH)

Abstract

The main BBF score file assigns between sixteen and sixty recorded runs to each of 55 games. Run count and a game's mean score have a Spearman association of 0.238, while the comparison file has nearly constant coverage and an association close to zero. These observations explain why archive-weighted and game-weighted summaries need not agree. They do not reveal why particular experiments were repeated. The article develops a coverage audit for interactive rendering evaluations in which the number of generated outputs should not silently determine the population of environments being described.

1. Repetition is an allocation decision

An experimenter can repeat a task because it is important, uncertain, cheap, difficult, or unexpectedly promising. A public table may later combine several rounds of such decisions. The number of rows associated with a task is therefore an observable allocation, but its interpretation is not self-evident. Giving every row equal weight can let that historical allocation decide what an aggregate means. This may be appropriate for describing the archive and inappropriate for a claim about an equally weighted task collection.

Make a game motivates studying an interface across different interactive settings (Meng et al., 2026a). A renderer with lightweight control plugins could plausibly be assessed on environments that differ in available actions and generation cost. Those differences make an allocation record valuable. The present analysis does not assess its implementation or reconstruct unavailable experimental details. It examines public BBF repetition counts as a concrete example of the distinction between how experiments were allocated and the environment population a final report claims to represent.

Task coverage and experimental repetition add different kinds of evidence. ALE provides diverse Atari tasks for studying agents (Bellemare et al., 2012), and Gym standardizes how an agent interacts with an environment (Brockman et al., 2016). Procgen separately introduces controlled procedural variation (Cobbe et al., 2019). A model that works across interface-compatible tasks has not necessarily demonstrated robustness to new procedural instances or unfamiliar interaction histories. Released BBF score files are useful because their task and run identifiers permit those levels to be distinguished (Schwarzer et al., 2023). Reliable-evaluation research likewise shows why a point estimate can hide variability across a small number of runs (Agarwal et al., 2021). The reporting consequence is to preserve the hierarchy of tasks and repetitions rather than flatten it prematurely. For rendering research, the same hierarchy can include environments, recorded histories, and stochastic generations. An average should name which of these units receives weight. Otherwise, changes in the number of repetitions can masquerade as changes in the distribution of supported environments.

Learned world models connect visual representation to behavior in several ways. World Models explores compressed representations of environment experience (Ha and Schmidhuber, 2018). Dreamer uses latent imagination to support behavior learning (Hafner et al., 2019), IRIS studies transformer-based world modeling (Micheli et al., 2022), and DIAMOND investigates diffusion-based representations in which visual detail can matter to control (Alonso et al., 2024). These works establish a useful set of questions rather than a single interchangeable implementation. What information is retained, which histories condition prediction, and how is a predicted observation connected to an action consequence? A rendering study should state which of these questions it actually tests. Visually plausible outputs alone do not

establish an accurate transition model. Conversely, an agent can obtain a favorable return without reproducing all details relevant to a human viewer. The measurement design should preserve this distinction and avoid translating a result from one endpoint into an unobserved improvement on another.

2. Counting before estimating performance

The main file contains 2,082 unique configuration-game-seed records. Its 55 games receive between sixteen and sixty runs, with a mean of 37.85. The comparison file contains 830 records, with fifteen to twenty runs per game and a mean of 15.09. Both cover the same task labels. Thus their coverage differences concern repetition within a common observed game set, rather than a mismatch in which game names are present. The audit retains this distinction and performs no interpolation of unrecorded runs.

Within each file, the analysis computes a game's mean supplied normalized score and relates it descriptively to that game's run count. The Spearman association is 0.2382 for the main configuration and -0.0171 for the comparison. The latter is particularly limited as an association because counts are nearly constant. Neither number identifies a policy of preferentially repeating successful games. The release could reflect prespecified experiment groups, later additions, or other historical decisions not recorded in these fields. Attributing intent from the association would exceed the available evidence.

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.

The scientific interpretation should specify what would change under the proposed intervention. Holding an observed covariate fixed statistically is not equivalent to manipulating it in an environment. A causal graph helps separate confounders, mediators, and selection mechanisms (Greenland et al., 1999). It can also reveal why controlling for every recorded variable is not automatically defensible. Meanwhile, an evaluation with multiple conditions and endpoints creates a family of opportunities for apparently striking findings. Multiplicity methods require a declared set of comparisons and an appropriate inferential purpose (Benjamini and Hochberg, 1995). Descriptive tables may be useful without testing every cell. When the data do not identify a mechanism, the appropriate contribution can be a sharper question and a transparent comparison rather than an unsupported causal answer.

The ability to accept commands changes the interpretation of a generated sequence. GameNGen's interactive setting (Valevski et al., 2024) and Genie's action-controllable worlds (Bruce et al., 2024) make this distinction visible. A command-response pair needs to be judged for its consequence as well as its visual continuity. Image structure (Wang et al., 2004), learned perceptual correspondence (Zhang et al., 2018), and video-distribution similarity (Unterthiner et al., 2018) supply useful but incomplete observations. The evaluator should explain whether a reference trajectory exists and whether the same initial condition can be replayed. Without those records, a visually different outcome might reflect a valid stochastic alternative or a failure to obey control. The experiment must distinguish those interpretations through its task specification. Matching output length alone is insufficient, because a longer sequence can contain more opportunities for a delayed or persistent control error. The reported endpoint should remain connected to the history that made the interaction meaningful.

3. What the allocation changes about a result

With equal game weighting, each represented game has weight 1/55, approximately 1.82 percent. In a pooled main-file mean, a game with sixty records contributes 60/2,082, approximately 2.88 percent, while a game with sixteen contributes about 0.77 percent. The difference is substantial even though both belong to the same nominal benchmark. A pooled mean is therefore a mixture of performance and the released repetition pattern. It is not invalid; it answers the question posed by that mixture.

The score consequences are visible in the saved release: the main file's pooled mean normalized score is 1.832, compared with 1.484 when games receive equal weight. The comparison file changes from 0.452 to 0.454. This asymmetry is consistent with allocation mattering more in the main archive. It should not be translated into a corrected universal measure or a causal explanation of the reported advantage. Equal weighting is a declared alternative population, not a recovery of a naturally true distribution over games.

A rendering evaluation could make the choice explicit from the beginning. If the intended claim concerns an equally selected environment, it could allocate a minimum number of histories to every environment and preserve a balanced aggregate. If the intended claim concerns user exposure, weights could follow an independently measured traffic distribution. Extra generation devoted to diagnosing a difficult environment should remain in the record without automatically changing the target population. A useful release would identify diagnostic repetitions separately from the primary assessment allocation.

There is a further distinction between missing repetition and missing coverage. Reweighting can change the influence of an observed game, but it cannot create evidence for an absent mechanic, an unavailable plugin, or a command sequence never tested. A report should therefore describe both the number of cases per represented task and the rules defining the represented task universe. This prevents a detailed repetition table from being mistaken for a complete account of the system's supported scope.

A validation split defines which form of transfer has been observed. Related observations can share subjects, scenes, prompts, or acquisition conditions, so randomly splitting rows may leave important information on both sides of the boundary. Structured cross-validation addresses these dependencies (Roberts et al., 2017). Model selection introduces another boundary: choosing a representation, threshold, or tuning parameter using the final assessment data makes the resulting error estimate optimistic (Varma and Simon, 2006). The appropriate held-out unit follows the intended future use. Generalization to a new frame, a new trajectory, and a new subject are not equivalent claims. A study should record those units before fitting or choosing a scoring rule, and preserve an assessment set that has not become part of iterative development.

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. Interpreting playability evidence

A recorded trajectory combines environment dynamics with the decisions that selected its actions. Reinforcement-learning terminology makes those components distinguishable (Sutton and Barto, 2018). When only fixed historical experience is available, the available action distribution becomes an important limitation (Levine et al., 2020). Conservative Q-learning addresses a specific part of that limitation in value estimation (Kumar et al., 2020). A generated interface inherits a related evaluation challenge: successful replay does not establish correct response to a new intervention. The scientific question should therefore identify the allowed actions, their timing, and the relevant consequences. An unsupported input should be recorded as unsupported rather than removed from the denominator without explanation. The archive can still support descriptive comparisons, provided the outcome is not renamed a proof of arbitrary control. Evidence about general intervention requires observing interventions, not merely increasing the number of frames generated along familiar histories.

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.

A scientific release should make its most consequential choices recoverable. Persistent identifiers and explicit metadata support reuse of observations and results (Wilkinson et al., 2016), while attention to dependencies helps prevent an evaluation from changing unnoticed as software evolves (Sculley et al., 2015). Relevant choices include the source version, preprocessing, accepted input set, metric implementation, and weighting convention. The record should also separate an observed output from a proposed future test. Saving that distinction prevents a plan from later being mistaken for a completed experiment. A small collection of well-linked records can be more useful than a large folder of unexplained scores. The important property is whether an independent reader can reconstruct what was measured and why its reported number has that meaning.

5. Conclusion

Experimental allocation has an observable relationship with the composition of a benchmark average. The BBF release shows unequal repetition and a noticeable change under balanced task weighting, but it does not identify the historical reasons for that allocation. Interactive rendering studies can make stronger claims by declaring their target population, recording why additional cases were generated, and preserving task-level outcomes. The important choice is not whether every average must use equal weights. It is whether the weights describe the scientific question the report actually asks.

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.