Run-to-Run Dispersion as Evidence about Interactive Evaluation

Adam Lawson (SR), Scott Walters (SI), Cole Gardner (TJ)

Abstract

A configuration with higher average scores can also have more variable outcomes. Across 55 games in the released BBF files, the median within-game standard deviation on an inverse-hyperbolic-sine score scale is 0.151 for the main configuration and 0.103 for the file without resets. The corresponding mean standard deviations are 0.204 and 0.147. This is a descriptive dispersion comparison rather than a causal variance test. It shows why claims about reliable interaction should report the spread of repeated outcomes in addition to their location and should distinguish score variability from rendering instability.

1. Variability is not simply the opposite of performance

A method can move an outcome distribution upward and make it more spread out. That combination can improve expected performance while leaving a user exposed to variable experiences. Conversely, a consistently low-performing system can have very little variance. Calling it reliable merely because its outputs are similar would ignore the requirement it is meant to satisfy. Location, dispersion, and the consequences of low-tail outcomes should therefore be interpreted together.

For a rendering interface, Make a game motivates asking whether repeated interactions behave dependably under declared controls (Meng et al., 2026a). The source's interactive focus does not determine which variability matters: different generated images may be acceptable, while different action consequences may be problematic. This study does not measure repeated outputs from that method. It uses public agent results to illustrate a more general reporting distinction. A higher variance of returns is an observation about scores, not proof that a renderer flickers, loses state, or responds unpredictably to a repeated command.

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.

A large collection of measurements does not guarantee a large collection of independent cases. Repeated observations can share unrecorded acquisition conditions or a common source, making row-level uncertainty misleading. White's covariance estimator addresses heteroskedasticity within a specified regression (White, 1980), while cluster methods address a declared form of grouped dependence (Cameron and Miller, 2015). Neither approach repairs a poorly defined outcome or establishes representativeness of the observed sample. A report should identify the number of groups, their relative sizes, and the reason for selecting the grouping level. When only a few independent groups are available,

apparently precise numerical intervals warrant particular caution. This is a limitation of the evidence structure, rather than a reason to count correlated measurements as fresh observations.

2. Estimating spread within each represented game

The analysis uses 2,082 records in the main BBF release and 830 in its comparison file without resets. Both contain 55 game labels. Run counts vary from 16 to 60 per game in the main file and from 15 to 20 in the other. A within-game sample standard deviation is computed separately for every configuration-game group. This avoids letting differences between games be mistaken for repeated-run variability within a game.

Two score scales are retained. The supplied normalized score preserves the release's relative scoring convention. An inverse hyperbolic sine transformation compresses large positive values while remaining defined for zero and negative scores. Standard deviations on these scales have different meanings, so their absolute magnitudes should not be compared with each other. The purpose is to examine whether a broad descriptive pattern appears under two declared representations. No transformation is chosen by searching for a particular configuration ranking.

A compact latent state is a hypothesis about which information an environment requires. World Models demonstrates one route to learning such a representation (Ha and Schmidhuber, 2018), and Dreamer connects learned latent dynamics to behavior (Hafner et al., 2019). IRIS and DIAMOND explore further world-model choices using transformers and diffusion respectively (Micheli et al., 2022) (Alonso et al., 2024). This literature makes information preservation a testable concern. If a renderer loses an object or confuses an action, the explanation could involve representation, history selection, conditioning, or downstream use. A meaningful comparison should separate those possibilities with matched inputs and explicit outcomes. The existence of a sophisticated latent representation does not establish that every relevant state variable survives compression. Nor does the ability to produce a convincing next image establish correctness over a long interaction. Those claims require observations at the corresponding temporal scale and a record of what the model was allowed to know.

A proposed explanation should be distinguishable from alternative explanations using observations the study can actually collect. Diagramming the relationships among selection, exposure, measurement, and outcome makes that requirement explicit (Greenland et al., 1999). The diagram does not establish its own correctness, but it prevents adjustment choices from remaining invisible. The same transparency is needed when investigators inspect many endpoints before choosing a narrative. Controlling false discovery concerns the behavior of a specified testing family (Benjamini and Hochberg, 1995); it is not a general certificate of scientific validity. Primary endpoints, secondary descriptions, and future mechanism tests should be named separately. Readers can then assess whether an attractive pattern was predicted, explored, or merely noticed after the available comparisons were examined.

The protected assessment population should be defined before investigators inspect detailed failures. Once a collection of cases guides repeated revisions, it has become development information even if its filenames still carry an evaluation label. Cross-validation for dependent data treats the grouping of observations as a substantive design decision (Roberts et al., 2017). Research on model selection likewise shows why tuning and final error estimation should not reuse the same information (Varma and Simon, 2006). A practical record therefore distinguishes fitting data, diagnostic examples, validation data, and the final external assessment. Where only a fixed public table is available, a descriptive analysis remains legitimate, but it should not be renamed a held-out prediction experiment without generating and assessing predictions under a protected split.

3. What the dispersion summaries reveal and leave unresolved

On the transformed scale, the median within-game standard deviation is 0.1510 for the main configuration and 0.1034 for the comparison. Their means are 0.2042 and 0.1470, and their maxima are 0.8965 and 0.6728. On the untransformed normalized scale, the median standard deviations are 0.1789 and 0.1121. The main file is therefore more dispersed under these particular summaries even though other analyses of the release show higher typical outcomes. That coexistence is the principal finding.

The numbers do not establish that the underlying algorithm has a larger intrinsic variance under identical conditions. The archived experiments differ in repetition counts, and file labels do not provide a complete randomized assignment or every evaluation detail. A sample standard deviation itself has uncertainty that depends on sample size and distribution shape. Comparing its medians across games is a useful descriptive operation, but it is not a fully specified test of equal outcome distributions or an estimate of a causal change in variance.

For an interactive generator, a stronger repeated-output study would begin from a fixed initial history and a clearly specified input sequence. It would then record both permitted visual variation and errors in required consequences. If the system intentionally supports multiple plausible continuations, variation alone should not be counted as failure. The relevant question is whether all allowed continuations preserve the constraints that define the task. A protocol could report a distribution of visual outcomes conditional on satisfying those constraints, alongside the rate of constraint violation. This avoids rewarding a static but unresponsive output for its low variance.

The evaluation should also vary initial histories and environments. Repeating one history many times improves knowledge of its stochastic behavior but does not establish transfer to another history. The benchmark needs to retain those levels separately. A useful design might allocate equal repetition within selected histories and then balance the histories across environments. The resulting variance decomposition would distinguish instability within a case from differences in case difficulty. Those observations are not contained in the score table analyzed here and remain a proposal for direct rendering assessment.

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.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a

target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

4. Interpreting playability evidence

An interactive benchmark should explain how its histories were obtained. In reinforcement learning, the behavior policy and the environment jointly determine which observations appear (Sutton and Barto, 2018). Offline-learning research highlights why fixed histories leave some actions and states weakly supported (Levine et al., 2020); conservative value learning provides a concrete example of handling a related estimation difficulty (Kumar et al., 2020). For a renderer, this means that a favorable replay score may be specific to the histories an archive happens to contain. A stronger test would deliberately vary a control while preserving the relevant initial evidence. It would then measure whether the expected consequence occurred and whether visual quality remained acceptable. Such a study requires a valid reference or task rule. Without that reference, a secondary analysis can describe available coverage and score behavior but cannot establish unrestricted correctness of the generated world's dynamics.

A system's reliable operating region may be narrower than the collection of inputs an interface accepts. Selective prediction separates performance on accepted cases from the fraction of cases accepted (Geifman and El-Yaniv, 2017). Reporting only the former can hide a substantial loss of coverage. Uncertainty estimates may also change under distribution shift (Ovadia et al., 2019), so familiar-condition confidence is not automatically evidence about new subjects, scenes, or control histories. A useful evaluation records rejected, ambiguous, and unsupported cases alongside successful ones. This makes it possible to distinguish limited but honest capability from a system that answers broadly with unreliable certainty. Any proposed deferral rule still requires assessment on its intended input population; it is not validated merely by being described.

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

Repeated outcomes in the main BBF file are more dispersed by the reported within-game summaries, even though its score location is more favorable. This result is neither a contradiction nor a direct measure of visual instability. It demonstrates why dependable interaction needs an endpoint that distinguishes permitted diversity from consequential failure. Future renderer studies should preserve repetitions, histories, and environments as separate levels, then connect variation to the task's actual requirements rather than interpreting a single variance statistic as a complete reliability judgment.

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.