Resource Allocation and Optimization
Volume 1 · Number 1 · 2026

More Training Steps Do Not Substitute Automatically for Broader Level Exposure: A Procgen Protocol Comparison

Cameron Richardson[1] · Trevor Gardner[2] · Benjamin Lewis[3]

[1] Switzerland · [2] Ireland · [3] New Zealand

Original publication: October 31, 2026 | Article 656

Abstract

Can a longer run over a small level set catch up with a shorter run sampling broadly? Public easy-mode Procgen logs offer a descriptive comparison, although their worker roles differ. Unrestricted-level online returns near twenty-five million training steps exceeded the unrestricted evaluation returns of fifty-million-step, two-hundred-level training in all sixteen game averages. The mean normalized contrast was 0.2531, with a working interval of 0.1322 to 0.3741. The analogous fifty-million-step contrast was 0.3415. These source-protocol differences caution against treating additional repeated exposure as an automatic replacement for breadth; they do not identify the causal benefit or computational price of level diversity.

Observational comparison of archived learning protocols. Source training-step counts are compared directly; wall time, total actor work, and fresh policy evaluation are unavailable.

More Training Steps Do Not Substitute Automatically for Broader Level Exposure: A Procgen Protocol Comparison

> Observational comparison of archived learning protocols. Source training-step counts are compared directly; wall time, total actor work, and fresh policy evaluation are unavailable.

Abstract

Can a longer run over a small level set catch up with a shorter run sampling broadly? Public easy-mode Procgen logs offer a descriptive comparison, although their worker roles differ. Unrestricted-level online returns near twenty-five million training steps exceeded the unrestricted evaluation returns of fifty-million-step, two-hundred-level training in all sixteen game averages. The mean normalized contrast was 0.2531, with a working interval of 0.1322 to 0.3741. The analogous fifty-million-step contrast was 0.3415. These source-protocol differences caution against treating additional repeated exposure as an automatic replacement for breadth; they do not identify the causal benefit or computational price of level diversity.

A resource question hidden inside a training budget

Training resources are often summarized by a single count of environment steps. That count is important, but it does not say how many distinct situations the learner encounters. Fifty million interactions with a recurring set of layouts need not provide the same experience as twenty-five million interactions drawn broadly from a procedural environment. A resource statement that omits the exposure distribution can therefore conceal a substantial difference in what the learner has had an opportunity to practice.

This study asks how the online return of unrestricted-level training near twenty-five and fifty million recorded training steps compares with unrestricted-level evaluation after fifty million steps of fixed-two-hundred-level training. We use publicly released Procgen logs and treat the comparison as an empirical description of two source protocols. We do not assume the protocols differ only in level diversity, and we do not convert their step counts into an estimate of computational efficiency.

The underlying reinforcement-learning framework in Sutton and Barto, 2018 makes the environment distribution part of the task. Cobbe et al., 2019b supplies sixteen procedurally generated games with source protocols for broad sampling and finite-level generalization. Cobbe et al., 2019a previously examined how limited training environments relate to broader performance. Together, these sources motivate asking whether the amount of repeated interaction can stand in for the breadth of interaction, rather than treating every environment step as informationally interchangeable.

For an interactive rendering system, the issue is especially concrete. Meng et al., 2026 propose an interactive game-rendering paradigm. Such a system may make new environments inexpensive to present, but its rendering function alone does not establish that a learner actually sees a broad distribution. Our analysis concerns the latter exposure question. It contains no rendering experiment, no implementation of the target method, and no claim about its frame quality or response latency.

Breadth, repetition, and the observable contrast

An unrestricted training protocol can encounter new layouts as interaction proceeds. A finite protocol repeatedly samples from its assigned set and may improve substantially within that set. Additional repetitions could strengthen transferable skills, strengthen configuration-specific behavior, or do both. The balance is an empirical question. A larger nominal step budget alone does not identify which route dominates.

The proposition examined here is that broad exposure at an earlier training budget can be associated with better broad-distribution recorded performance than a longer finite-exposure run. Association is the appropriate wording because source worker roles and return collection differ. We test neither a universal sample-complexity ordering nor a counterfactual statement about what the same fitted policy would have achieved under an alternative history.

Algorithms such as Schrittwieser et al., 2019 integrate learning and planning in ways that change the informational value of interaction. Ye et al., 2021 explicitly addresses limited-data performance, while Hafner et al., 2023 studies broad control competence through learned world models. Their differing designs reinforce why a step count is only one resource measure. None is the learner rerun in this manuscript, and their published achievements are not imported as evidence for our particular source contrast.

Environment breadth also differs from importing prior data or language-guided knowledge. Guss et al., 2019 supplies a benchmark built around human demonstrations, Fan et al., 2022 uses a broader open-ended task ecosystem, and Wang et al., 2023 studies an agent that accumulates reusable skills. These are alternative sources of experience structure, not interchangeable replications of a procedural-layout experiment. They help locate the exposure question without expanding the empirical population beyond the sixteen games analyzed here.

The acquired logs and the two worker roles

The local source is the archived repository in OpenAI, 2026, fixed to commit 1a2ae2194a61f76a733a39339530401c024c3ad8. We acquired all one hundred forty-four easy-mode CSV files: separate training and evaluation logs for fixed-two-hundred-level runs, and online training logs for unrestricted-level runs. Each family covers sixteen games and three source run labels. The acquisition also preserves the license, explanatory files, source constants, and file-level cryptographic checks.

The fixed-level regime uses a training worker and an evaluation worker. Source code gives the evaluation worker an unrestricted level parameter and zero weight in the PPO gradient update. The broad regime instead trains online with unrestricted level sampling. Thus, both outcomes concern the broad source level distribution operationally, but one comes from a noncontributing evaluation worker and the other from a worker whose experience contributes to learning.

This distinction is a material limitation. Online returns can reflect exploration behavior, contemporaneous policy updates, and the episode history of the training worker. The broad training log is not a separately collected fixed-policy test suite. Conversely, the finite regime's evaluation worker is not a complete list of independently held-out level identifiers: the acquired files do not supply those identities. We describe the source roles exactly rather than recast both series as identical test-set evaluations.

The source curves contain 3,051 updates, with recorded training-step increments of 16,384. Every series used here shares that grid. The twenty-five-million window consists of one hundred updates ending at 25,001,984 steps; its first record is at 23,379,968. The final window spans 48,365,568 through 49,987,584

steps. These fixed windows reduce dependence on one last logged value without pretending to remove all dependence from the rolling episode buffers.

Measuring performance across unlike reward scales

Each game's return is transformed by subtracting the lower reference value in the source easy-game constants and dividing by that game's supplied reference range width. We do not clip the result. The anchors facilitate aggregation across unequal raw reward scales; they are neither estimated extrema nor a claim that a normalized score of one has the same utility for all games.

The measurement distinctions in Stevens, 1946 are directly relevant to that transformation. An affine reward normalization changes units without establishing a common substantive value scale. Carroll et al., 2006 further motivates caution where noisy measured quantities stand in for latent constructs. Here, return is an observed task score, while breadth of useful competence is a broader interpretation that the records cannot measure completely.

Shannon, 1949 offers a general distinction between the quantity of transmitted symbols and information relative to a distribution. We use that distinction conceptually, without calculating entropy or claiming that a novel level contains a fixed number of informative bits. The empirical variables remain source step counts and logged returns. A statement about information gained per step would require additional measurements absent from the archive.

For each regime, game, and source run, we calculate the normalized window mean. We average the three runs within game before comparing regimes. The run labels identify files; they do not prove that identical random seeds generate matched trajectories across protocols. The final analysis therefore uses differences between game-level regime means rather than a claimed paired-seed randomized design.

A transparent two-contrast analysis

The first contrast subtracts the finite regime's final unrestricted evaluation mean from the unrestricted regime's twenty-five-million online mean. It asks whether the broad regime was already ahead at roughly half the source training-step budget. The second contrast makes the analogous comparison using the unrestricted regime's final window. Both contrasts retain exactly the same sixteen games and the same final finite-regime reference.

We estimate each equal-game mean through intercept-only OLS. HC3 standard errors are used with t15 intervals, following the heteroskedasticity-robust logic of White, 1980. N is sixteen game rows. No controls, time trends, or fixed effects are included; adding covariates after aggregation would not recover a randomized diversity effect. R-squared is zero because each regression estimates a mean rather than a predictive relationship.

The cluster-inference discussion in Cameron and Miller, 2015 helps explain why many correlated observations do not automatically provide many independent comparisons. Our thousands of log records are reduced before estimation. Newey and West, 1987 addresses serial covariance in time-series settings, but we do not add a time-series correction to a regression whose observations are already game summaries. Temporal dependence remains embedded in each summary and is part of the conditional scope of the working intervals.

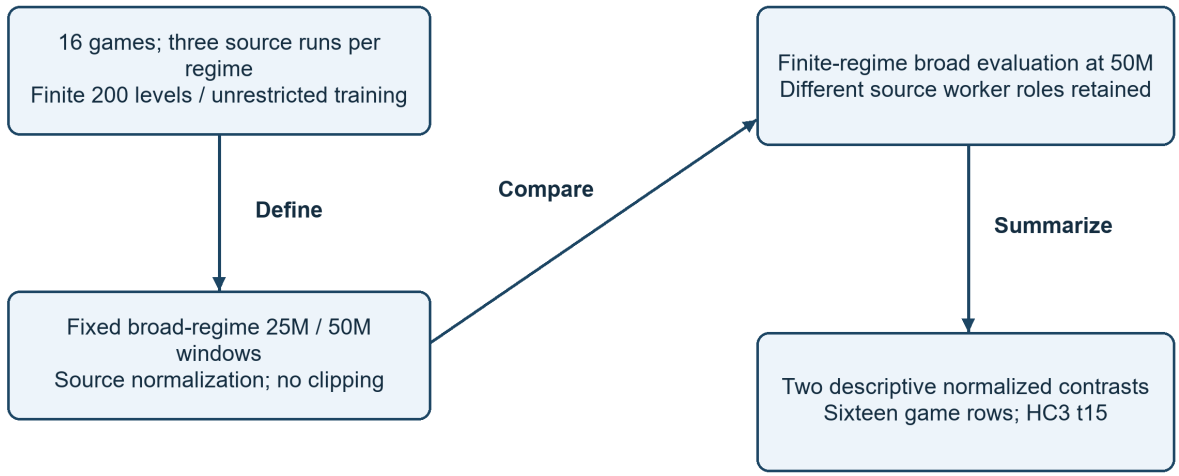


Figure 1. *Two source protocols, three fixed window summaries, and sixteen game-level contrasts. Different worker roles prevent interpreting the arrows as a causal intervention on level diversity.*

Broad-exposure source returns were higher at both budgets

At the final finite-regime budget, the equal-game unrestricted evaluation mean was 0.389590. The unrestricted regime's online mean was 0.642715 near twenty-five million training steps and 0.731106 near fifty million. These are normalized source-record means. They are not independent re-evaluations of the original policies, and they do not include uncertainty from a fresh episode sample.

The half-budget contrast averaged 0.253125, with a standard error of 0.056741 and a working interval from 0.132185 to 0.374066. The equal-budget contrast averaged 0.341516, with a standard error of 0.067481 and an interval from 0.197685 to 0.485348. Both contrasts were positive in every one of the sixteen game averages. That consistency is descriptive evidence about the supplied collection, not a guarantee for any newly generated game.

Table 1. *Unrestricted-regime online return minus finite-regime unrestricted evaluation return, after game-specific normalization. Sixteen game rows; intercept-only OLS, HC3 covariance, t15 intervals, no controls or fixed effects. The reference is the finite regime's final hundred-update window in both columns.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.253125 (0.056741)	[0.132185, 0.374066]	16
M1: fit	R-squared = 0.0000	-	-	16
M2	Mean outcome/contrast	0.341516 (0.067481)	[0.197685, 0.485348]	16
M2: fit	R-squared = 0.0000	-	-	16

Model labels: M1 = Unrestricted regime near 25 million steps minus finite-regime evaluation near 50 million steps; M2 = Unrestricted regime near 50 million steps minus finite-regime evaluation near 50 million steps.

The difference between the two reported mean contrasts is about 0.088391 normalized units, reflecting further progress within the unrestricted regime. We have not fitted a separate interval for that increment and do not infer its uncertainty by comparing the two displayed intervals. More importantly, a favorable

half-budget comparison does not show where the finite regime might eventually catch up. The source archive ends near fifty million steps and supplies no observation of longer finite-level training.

Because every supplied game has a positive contrast at both budgets, replacing equal-game weights with any fixed nonnegative game weights that sum to one would preserve the observed direction. This is a statement about the finite table, not an estimated robustness probability. The magnitude would still depend on the weights and normalization anchors, and uncertainty would require an analysis matching that alternative target. It is useful to separate this limited directional stability from broader claims that the result would survive new games, different algorithms, or matched fresh evaluation episodes. Those extensions change the evidence rather than merely reweight it.

What resource efficiency can and cannot mean here

The records show that more source training steps on the finite level menu did not erase the observed difference from unrestricted online training. They do not show that broad training required half as much computation. Williams et al., 2009 distinguishes hardware performance constraints from high-level operation counts. Mytkowicz et al., 2009 likewise demonstrates how apparently innocuous measurement conditions can distort computational comparisons. We have no matched hardware session, energy meter, aggregate actor-work count, or comparable wall-time accounting for these protocols.

Source training steps exclude the informational and computational role of evaluation workers as a comparable deployment cost. A broad worker contributes to learning; a zero-gradient worker still runs environments to produce an evaluation stream. Equal logged step budgets therefore do not imply equal total sampling expenditure. The result should be described through its explicit budget label instead of upgraded into a claim of measured computational savings.

Forecast-evaluation guidance in Hyndman and Koehler, 2006 is useful in requiring the score and horizon to be named. Gneiting and Raftery, 2007 similarly stresses alignment between a measure and its decision target. Our two windows answer a budget-indexed descriptive question. They cannot establish which protocol is preferable for an application with a different exposure distribution, reward utility, or cost structure.

Connections without algorithm substitution

Levine et al., 2020 distinguishes learning from fixed data from ongoing interaction. Fu et al., 2020 provides benchmark datasets for the former, and Chen et al., 2021 studies sequence modeling of trajectories. Our learner histories were generated through online source training; reading their logs later does not convert the original learning into offline reinforcement learning. That distinction matters when describing what kind of experience breadth was available.

For generated environments, Yu et al., 2025 and Valevski et al., 2024 concern game-oriented world modeling. Micheli et al., 2022 investigates transformer world models, and Bruce et al., 2024 addresses generative interactive environments. These studies motivate questions about which situations a model can generate or simulate. The present comparison instead records a difference in policy outcomes under two exposure protocols. It cannot tell whether a generative environment preserves the same task distribution or whether its visual diversity creates useful control diversity.

Limits and research use

The benchmark is a supplied set rather than a probability sample of games, and three source runs per condition leave considerable training variability unexplored. The study is secondary and exploratory; no preregistered causal experiment is claimed. The working intervals condition on the source curves and on the chosen normalization. They do not encompass original training reproducibility, alternative algorithm settings, or the scientific independence of multiple possible analyses of the archive.

Preserving those conditions follows the documentation approach in Gebru et al., 2021 and the reuse principles in Wilkinson et al., 2016. The analysis files retain every game-level mean, source window, and contrast. A stronger experiment would randomize or systematically vary level counts while matching worker roles, total sampling, architecture, and fresh evaluation episodes. Until then, these results support a bounded observation: additional repetition over the finite menu did not substitute for broad exposure in the captured source records, even when the latter was assessed at an earlier training-step budget.

References

1. 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.
2. 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.
3. Ye, Weirui, et al. "Mastering Atari Games with Limited Data." *arXiv preprint arXiv:2111.00210*, 2021. https://arxiv.org/abs/2111.00210.
4. 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.
5. Wang, Guanzhi, et al. "Voyager: An Open-Ended Embodied Agent with Large Language Models." *arXiv preprint arXiv:2305.16291*, 2023. https://arxiv.org/abs/2305.16291.
6. Hafner, Danijar, et al. "Mastering Diverse Domains through World Models." *arXiv preprint arXiv:2301.04104*, 2023. https://arxiv.org/abs/2301.04104.
7. Shannon, C.E.. (1949). Communication in the Presence of Noise. Proceedings of the IRE. https://doi.org/10.1109/jrproc.1949.232969.
8. 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.
9. Cobbe, Karl; Klimov, Oleg; Hesse, Chris; Kim, Taehoon; Schulman, John. (2019a). Quantifying Generalization in Reinforcement Learning. Proceedings of the36th International Conference on Machine Learning, PMLR97:1282-1289. https://proceedings.mlr.press/v97/cobbe19a.html.
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. Cobbe, Karl, et al. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.
12. Yu, Jiwen, et al. "GameFactory: Creating New Games with Generative Interactive Videos." *arXiv preprint arXiv:2501.08325*, 2025. https://arxiv.org/abs/2501.08325.
13. Guss, William H., et al. "MineRL: A Large-Scale Dataset of Minecraft Demonstrations." *arXiv preprint arXiv:1907.13440*, 2019. https://arxiv.org/abs/1907.13440.
14. OpenAI. (2026). train-procgen: released easy-mode training and evaluation logs, commit1a2ae2194a61f76a733a39339530401c024c3ad8 [Archived repository; snapshot acquired8November2026]. https://github.com/openai/train-procgen/tree/1a2ae2194a61f76a733a39339530401c024c3ad8.
15. 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.
16. Bruce, Jake, et al. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.

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

18. 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.

19. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

20. 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.

21. Mytkowicz, Todd; Diwan, Amer; Hauswirth, Matthias; Sweeney, Peter F.. (2009). Producing wrong data without doing anything obviously wrong!. Proceedings of the 14th international conference on Architectural support for programming languages and operating systems. https://doi.org/10.1145/1508244.1508275.

22. Williams, Samuel; Waterman, Andrew; Patterson, David. (2009). Roofline: an insightful visual performance model for multicore architectures. Communications of the ACM. https://doi.org/10.1145/1498765.1498785.

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

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

25. Fan, Linxi, et al. "MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge." *arXiv preprint arXiv:2206.08853*, 2022. https://arxiv.org/abs/2206.08853.

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

27. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. "Measurement Error in Nonlinear Models." 2006. https://doi.org/10.1201/9781420010138.

28. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

29. Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

30. 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.