Does the Training Opponent Matter? Four Preregistered Studies of a Compact Pokémon Card Game Policy

Olayinka Vaughan
Wesleyan University
yvaughan@wesleyan.edu
ORCID: https://orcid.org/0009-0005-8933-5396

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

***Abstract*—Agents for collectible card games are often trained against a single opponent, and a more varied set of opponents is expected to generalize better to unseen decks. We test this for a 2,881-parameter option-scoring policy in the Pokémon Trading Card Game environment CABT, on deck families held out from training. Four preregistered studies, each with 30 training seeds per arm and 7,200 held-out games, compare a fixed opponent pool with a diverse one. In two studies the policy imitates a rule-based teacher whose demonstrations are collected against each pool. In the other two it learns from its own wins and losses. The difference in held-out win rate is -1.22, $+1.40$, -1.60 and $+0.15$ points, and the 90% interval of each difference lies inside the preregistered margin of five points. The first study could not have found an effect: the student copied its teacher on at least 99.8% of held-out decisions. With a teacher that reads the board, agreement falls to about 61% and the pools remain equivalent. Learning from play moved the policies by about the size of evaluation noise at 400 games and learned clearly at 2,000 games from random weights, where the margin is met only just. Both budgets are far below those of competition entries.**

***Index Terms*—Pokémon Trading Card Game, imitation learning, opponent diversity, equivalence testing, preregistration**

I. Introduction

Collectible card games combine hidden information, a large and changing set of legal actions, and decks that change what good play looks like. Work on competition-level agents trains them against populations of opponents, and several authors argue that the variety of opponents seen in training determines how far an agent generalizes [1], [2]. For small agents that must run within a fixed inference budget, it is cheaper to ask whether such variety pays off before building a population.

We ask one question: does the set of opponents used during training change a compact policy's win rate on decks it never saw? The policy is a 2,881-parameter network that scores the legal options at each decision in the Pokémon Trading Card Game environment CABT [3]. We test two ways a pool of opponents can enter training. In imitation, the pool only decides which game states a teacher visits while its decisions are recorded. In learning from play, the pool supplies the wins and losses the policy learns from.

We make three contributions.

1) A preregistered protocol for small studies in CABT: the plan is frozen before the run, the training seed is the unit of analysis, intervals come from a cluster bootstrap, and a margin of five points is fixed in advance so that a null result can be stated with a bound.
2) Four controlled comparisons, each with 30 seeds per arm, that find the fixed and diverse pools equivalent within that margin.
3) An account of why the first comparison was forced, and a second teacher that removes the cause. We also state where the learning-from-play comparisons are limited.

We searched arXiv and the public write-ups of the 2026 competition and found no earlier controlled test of this question, for imitation learning or for card games.

II. Related work

Card games as a testbed. Collectible card games combine hidden information with a branching factor that changes with the game state and is often large. Hoover et al. [4] note that these games are relatively understudied in AI research despite their popularity, and lay out the challenges Hearthstone poses. For a competition-level agent, Xi et al. [5] describe an end-to-end policy and an optimistic smooth fictitious play method for Legends of Code and Magic, and report winning both the official LoCM 1.5 track and the bonus LoCM 1.2 track of the COG 2022 competition. The Pokémon Trading Card Game has received less attention. PTCG-Bench [6] evaluates large language model agents on it. The 2026 Kaggle Pokémon TCG AI Battle, run on the same CABT engine, produced public write-ups from many learned agents. Several top entries trained by reinforcement learning from self-play or against a league of frozen opponents [7], [8]. One team distilled a gradient-boosted teacher into a network and notes that the teacher's games widen the training states but inherit the teacher's weaknesses [9]. Another reports that every hand-written rule or mechanical teacher it tested cost win rate [7]. We study a different agent class, a 2,881-parameter option scorer that runs in well under a millisecond, and we test the opponent pool directly.

Opponent diversity and generalization. Lanctot et al. [1] show that policies learned by independent reinforcement

learning can overfit to the policies of the other agents seen in training, and propose best responses to mixtures of policies. Strouse et al. [2] argue that a diverse set of training partners is the crux of generalizing to new partners in cooperative play. Both train by reinforcement learning from play, as our studies C and D do. In the first two, the opponent affects only which states the demonstrations visit. For generalization more broadly, Cobbe et al. [10] separate training and test sets of procedurally generated environments, and Kirk et al. [11] survey zero-shot generalization. In the competitive Pokémon video game, VGC-Bench [12] provides behavior cloning and self-play baselines and reports that, as the number of team configurations grows, the best single-team algorithm does worse and is more exploitable but generalizes better to unseen teams. We follow the practice of keeping the test deck families out of training, with a disjointness check between the splits.

Imitation under distribution shift. The learner's own mistakes lead it to states the demonstrator did not visit, which breaks the assumption of identically distributed data. Ross et al. [13] describe this compounding of errors and propose DAgger, which gathers trajectories with the current learned policy. DART [14] instead injects noise into the supervisor while it demonstrates, so that the demonstrations cover recoveries. Belkhale et al. [15] formalize data quality through distribution shift and show that state diversity is not always beneficial and that its effect depends on action divergence and transition diversity. Our first two studies vary the demonstration states in a similar way, through the opponent, and the teacher agreement we report measures how close the learner stays to the demonstrator.

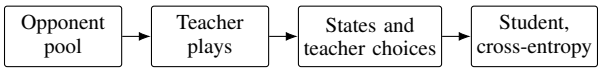
Statistical practice. Henderson et al. [16] document how environment non-determinism and variance in the methods make reinforcement learning results hard to interpret without significance measures. Agarwal et al. [17] show that point estimates from a few training runs can mislead and recommend interval estimates. The CABT engine shuffles with a system entropy source, so no game can be replayed from a seed. We therefore treat the training seed as the unit and fix the analysis before each run.

III. Environment, policy and agents

Environment. We use CABT 1.0.0 through kaggle-environments 1.33.0 [3], [18], with its 1,431-card catalog pinned by SHA-256 hash. An agent receives only its own observation, with the opponent's hand masked, and returns the indices of the options it chooses. The engine seeds its shuffles from the system entropy source, so games are not replayable from seeds and we keep a record of every game instead.

Policy. The policy is a two-layer network, 88 inputs to 32 hidden units to one logit, with 2,881 parameters, written in NumPy. The 88 inputs are 72 features of the board (turn, prizes, hand and deck sizes, HP, attached energy and status for the active Pokémon and five bench slots of both sides) and 16 features of one option (its CABT option type, attack damage, whether it targets the active Pokémon or the bench, and its position in the list). The network scores every legal option, and the policy plays the highest score at test time.

(a) Imitation, studies A and B



(b) Learning from play, studies C and D

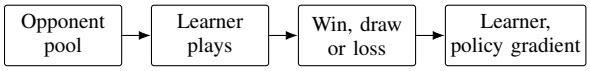


Fig. 1. Where the opponent pool acts. In imitation it decides which states the teacher visits. In learning from play it supplies the outcomes the policy learns from.

TABLE I
THE FOUR STUDIES. EACH HAS 30 TRAINING SEEDS PER ARM.

	Learner	Teacher	Start	Budget per checkpoint
A	imitation	heuristic	random	2,000 decisions, 400 steps
B	imitation	board	random	2,000 decisions, 400 steps
C	REINFORCE	none	study B	400 games, 50 updates
D	REINFORCE	none	random	2,000 games, 250 updates

Teachers. *HeuristicAgent* ranks options by type and attack damage: attacks first, then attaching, evolving, playing a card and ending the turn. It never reads the board. *BoardAgent* reads it. It takes a knockout when one is on offer, evolves, keeps two Pokémon on the bench, attaches energy where it brings an attack closest, retreats an active Pokémon that cannot attack or would be knocked out next turn, and values searched and discarded cards by the board. It uses only its own observation and the public card catalog.

Decks and opponents. Each deck family has 4 + 4 Pokémon, 4 + 4 Trainers and 44 basic Energy cards. The training families are water-kyogre, grass-bulbasaur and lightning-pikachu. The held-out families are psychic-mewtwo and colorless-snorlax, and every evaluation game is a mirror match. A tuning family, fire-ponyta, was used only in the pilot for study C. The *fixed* pool is the engine's first-option agent in every training game. The *diverse* pool draws the heuristic, the first-option agent or a random agent at each game. The same three agents are the evaluation opponents, so the held-out axis is decks and not opponents.

IV. Studies and statistics

Table I lists the four studies. Each has a plan frozen before its run, in machine-readable form that the runner refuses to execute unless it is marked frozen. Both arms of a study share the architecture, budget, decks and evaluation, so the pool is the only difference.

Imitation (A and B). The teacher plays one of three training decks against the pool, and its first choice at each decision with more than one option becomes a training example. Each checkpoint trains on 2,000 decisions (1,600 for training and 400 for validation) for 400 Adam steps.

Learning from play (C and D). The policy samples each choice from the softmax of its option scores, plays a full game against the pool from a random seat and deck, and receives $+1$, 0 or -1 for a win, draw or loss. REINFORCE updates it after every batch of 8 games with the batch mean outcome as the baseline and Adam at learning rate 0.003, for 400 games. Both arms start from the board-teacher checkpoint of the same seed, so the pool is again the only difference. A pilot chose the learning rate from 0.001, 0.003 and 0.01. It used four seeds outside the study and scored 60 games per checkpoint on the tuning deck. At 0.003 the pilot gained 11.7 points (fixed pool) and 5.0 (diverse pool) over the starting checkpoints, and the other two rates gained about nothing. On the held-out decks the gain did not carry over (Section V), so the pilot was too small to select a rate reliably. Study D gives learning more room: the policy starts from random weights and plays 2,000 games (250 updates) at learning rate 0.03. An 800-game pilot from the imitation start (two seeds per arm) had found no learning at any of the three rates, and a pilot from random weights was the only one in which training scores rose, at 0.03. The plan for D records that the design changed after the pilots and before any study game.

Evaluation. Each checkpoint plays 120 held-out games: 2 decks $\times$ 3 opponents $\times$ 5 evaluation seeds $\times$ 2 pairs, with seats reversed inside each pair. A draw counts as half a win. The checkpoint score is the mean of its win rates on the two held-out decks.

Statistics. The unit is the training seed. The estimate is the mean checkpoint score of the diverse arm minus that of the fixed arm. The interval is a percentile cluster bootstrap that resamples seeds within each arm, with 10,000 resamples. A difference is reported when the 95% interval excludes zero. Equivalence is reported when the 90% interval lies inside ± 5 points, the bootstrap form of two one-sided tests at the 5% level. Otherwise the result is inconclusive. The margin was chosen before any data, at less than half the 12-point gap between the two held-out decks in an earlier run.

V. Results

Table II gives the preregistered result. In all four studies the 95% interval includes zero and the 90% interval lies inside the margin, so the pools are equivalent within five points. Fig. 2 shows every seed. The standard deviation across seeds inside one arm is 3.8 to 10.8 points, larger than every difference.

Study A was forced. The students matched the heuristic teacher on 99.8% to 100.0% of held-out decisions with more than one option. The heuristic scores an option by its type and damage, and the student's features hold both, so any demonstrations teach the same rule and the pool has nothing to change. Study A therefore says little about the pool. We built study B to remove this cause.

Study B. With the board-reading teacher the students agree with it on 51.6% to 69.3% of held-out decisions (mean 60.8% for the fixed arm and 61.8% for the diverse arm). They no longer copy the teacher, and the pools are still equivalent:

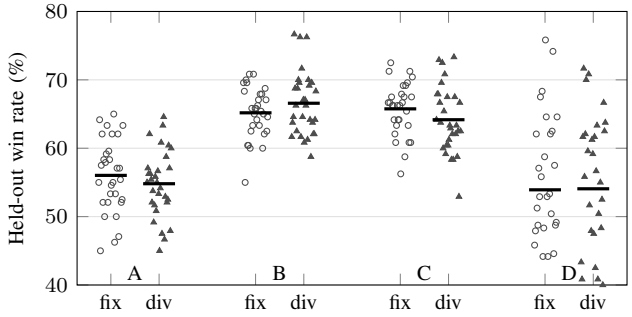


Fig. 2. Held-out win rate of every checkpoint. Fix is the fixed pool (open circles), div the diverse pool (filled triangles), and the bar is the arm mean. A and B imitate a heuristic and a board-reading teacher, and C and D learn from play (C from the imitation checkpoints, D from random weights). Points are spread sideways by seed order only.

the difference is $+1.40$ points with a 95% interval of $[-0.67, +3.58]$. This is the main result of the paper.

Study C. Learning from play moved the policies very little. Against the checkpoint each seed started from (mean score 65.18%), the change was $+0.6$ points for the fixed arm (17 of 30 seeds up) and -1.0 for the diverse arm (11 of 30), with a standard deviation of 5.1 and 5.9 points. A 120-game score alone has a standard error near 4.5 points, so these changes are about the size of evaluation noise. The mean training score also stayed flat, from 58.0% to 57.9% for the fixed arm and 69.4% to 69.7% for the diverse arm. The equivalence in C is therefore weak evidence about the pool. It shows no effect at a budget of 400 games, and does not show that learning from play is unaffected by its opponents. Study D gives learning more room.

Study D. With 2,000 games from random weights the policies learned. The mean training score rose from 44.7% to 49.9% for the fixed arm and from 54.5% to 63.2% for the diverse arm, and it rose in 25 and 25 of 30 seeds. The held-out scores, 53.92% and 54.07%, are below the imitation students of study B and close to the heuristic. The seed-to-seed spread is about 10.6 points, more than twice that of the imitation studies. The 90% interval, $[-4.29, +4.51]$, lies inside the margin with little room, and a margin below 4.5 points would have left the result inconclusive. As an unplanned observation, the arms differ by deck in opposite directions: -6.4 points on psychic-mewtwo (95% interval $[-13.9, +1.2]$) and $+6.8$ on colorless-snorlax ($[-2.2, +15.7]$). Both intervals include zero, and the two gaps cancel in the primary score.

Decks and opponents. Table III splits the win rates by held-out deck and by evaluation opponent. The deck matters far more than the pool. Averaged over the two arms, the colorless-snorlax deck scores 16.0, 30.1, 29.7 and 5.9 points above psychic-mewtwo in studies A to D, against pool differences of at most 1.60 points. With two held-out decks, a deck effect cannot be separated from deck-specific play, and we do not model it.

Baselines. Table IV gives four reference agents under the same 120-game protocol. Win rates for the same agent differ

TABLE II

DIFFERENCE IN HELD-OUT WIN RATE, DIVERSE MINUS FIXED POOL, IN PERCENTAGE POINTS. MEANS ARE CHECKPOINT SCORES OVER 30 SEEDS PER ARM. INTERVALS ARE CLUSTER-BOOTSTRAP INTERVALS OVER TRAINING SEEDS. EQUIVALENCE USES THE MARGIN OF ± 5 POINTS.

Study	Fixed mean (%)	Diverse mean (%)	Difference	95% interval	90% interval	Result
A	56.03	54.81	-1.22	$[-3.78, +1.33]$	$[-3.40, +0.90]$	equivalent
B	65.18	66.58	$+1.40$	$[-0.67, +3.58]$	$[-0.33, +3.21]$	equivalent
C	65.76	64.17	-1.60	$[-3.69, +0.61]$	$[-3.40, +0.25]$	equivalent
D	53.92	54.07	$+0.15$	$[-5.04, +5.29]$	$[-4.29, +4.51]$	equivalent

TABLE III

MEAN WIN RATE (%) BY HELD-OUT DECK AND BY EVALUATION OPPONENT, OVER 30 SEEDS PER ARM.

		Held-out deck		Opponent		
Study	Arm	Mewtwo	Snorlax	Heuristic	First	Random
A	fixed	48.3	63.7	48.5	58.3	61.3
A	diverse	46.5	63.1	48.5	56.5	59.3
B	fixed	49.9	80.4	59.3	66.2	70.0
B	diverse	51.7	81.4	61.9	66.0	71.9
C	fixed	50.5	81.0	62.0	67.9	67.4
C	diverse	49.7	78.6	58.5	66.1	67.9
D	fixed	54.3	53.6	47.5	58.0	56.2
D	diverse	47.8	60.3	48.2	57.7	56.3

TABLE IV

WIN RATE (%) OF REFERENCE AGENTS ON THE HELD-OUT DECKS, 120 GAMES EACH, SAME PROTOCOL AS THE STUDENTS. STUDY A DID NOT INCLUDE THE BOARD AGENT.

Study	Board	Heuristic	First	Random
A	n/a	56.7	39.6	41.7
B	74.6	58.3	52.1	53.8
C	75.0	57.1	45.4	50.4
D	70.4	53.8	49.2	44.2

by up to 8 points between studies, which shows how noisy 120 games are. Students in studies B and C score between the board teacher and the other three agents, and the learned policies of study D score near the heuristic.

VI. DISCUSSION

The pool did not change held-out play in any of the four designs, within five points. We see three readings, and the data do not separate them. First, in imitation the pool changes only which states the teacher visits, and the teacher's choice is a function of the state, so a student that fits the teacher well on any reasonable sample of states ends up close to the same function. Second, in learning from play, the budget may be too small for the opponent to matter. Study D raised it fivefold, from random weights, and found the same, but 2,000 games is far below the roughly 20 million used by one competition entry [19]. Third, the evaluation opponents are the same three agent types as the diverse pool, so the diverse pool is not out of distribution at test time. The comparison asks about unseen decks only.

Study history. The studies were not our first runs. Earlier three-seed runs of study A gave differences of -3.88 (not logged), -0.55 and $+6.95$ points. The last came after we fixed a defect in which the teacher always played the water-kyogre deck whatever the training deck. The 30-seed run gave -1.22 points. The seed-level analysis and the preregistered plans followed from that spread. Study A also used a feature encoder whose bench flag never fired, which we corrected before studies B and C. The pilots for studies C and D used seeds outside the studies.

VII. LIMITATIONS

- The decks are simplified and every held-out game is a mirror match. The 30-seed reports do not store replays, so we measured end reasons in a separate audit of 600 games: four study B checkpoints (seeds 2001 and 2002, both arms) and the board teacher, 120 games each, which gives 300 games per deck. On colorless-snorlax, 266 of 300 games ended with a player having no Pokémon left, with a median of 12 turns. On psychic-mewtwo, 205 ended by deck-out, with a median of 57 turns. Only 34 and 89 games on the two decks ended by taking all prize cards, the usual way to win, so most games end by another route.
- The margin of five points is a judgment, and a reviewer may prefer another. The 95% intervals rule out differences larger than 5.29 points in all four studies, and say nothing about smaller ones. In study D the 90% interval fits inside the margin with little room.
- Studies C and D used small budgets and learning rates chosen in pilots of two to four seeds on one deck. Study C started both arms from an imitation checkpoint, and study D learned from random weights and reached only about the strength of the heuristic. Neither speaks to a budget in the millions of games.
- One architecture and one environment were used, with three simple opponents. Nothing here is evidence about population training at the scale of the competition entries.
- The study measures win rate against three fixed agents. It does not measure exploitability or play against humans.

VIII. CONCLUSION

Across four preregistered designs, training against a diverse opponent pool did not change a compact policy's win rate on unseen decks by more than about five points. The first design could not have shown an effect, the second removes

that cause, the third is limited by a small budget, and the fourth lets learning from random weights run for 2,000 games and still finds no effect, though only just inside the margin. A stronger test of learning from play needs far more games per checkpoint than we could run.

Reproducibility. Plans, reports, checkpoints and code are available in a public repository.[1] The CABT engine cannot replay games from seeds, so the per-game records in the reports are the record of every game.

REFERENCES

[1] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Perolat, D. Silver, and T. Graepel, "A unified game-theoretic approach to multiagent reinforcement learning," in *Advances in Neural Information Processing Systems*, 2017.

[2] D. J. Strouse, K. R. McKee, M. Botvinick, E. Hughes, and R. Everett, "Collaborating with humans without human data," in *Advances in Neural Information Processing Systems*, 2021.

[3] PTCGABC, "cabt engine documentation, version 0.1.0," https://matsuoinstitute.github.io/cabt/, 2026, accessed 15 November 2026.

[4] A. K. Hoover, J. Togelius, S. Lee, and F. de Mesentier Silva, "The many AI challenges of Hearthstone," *KI - Künstliche Intelligenz*, vol. 34, no. 1, pp. 33–43, 2020, published online 2019.

[5] W. Xi, Y. Zhang, C. Xiao, X. Huang, S. Deng, H. Liang, J. Chen, and P. Sun, "Mastering strategy card game (Legends of Code and Magic) via end-to-end policy and optimistic smooth fictitious play," 2023.

[6] D. Hua, Y. Sun, R. Huang, C. Wang, F. Gao, and Y. Yang, "PTCG-Bench: Can LLM agents master Pokémon trading card game?" in *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026, pp. 36 084–36 099. [Online]. Available: https://aclanthology.org/2026.findings-emnlp.1689/

[7] G. Baldon and Squapale, "Reading the cards: a deck-agnostic policy trained by a league of its own exploiters," Kaggle write-up, 7th place, Pokémon TCG AI Battle Challenge, 2026, https://doi.org/10.34740/kaggle/w/112729.

[8] KawattaTaido, "Seek inspiration, not imitation," Kaggle write-up, 6th place, Pokémon TCG AI Battle Challenge, 2026. [Online]. Available: https://www.kaggle.com/competitions/pokemon-tcg-ai-battle-challenge-strategy/writeups/new-writeup-1787373273295

[9] Team Boss's Orders Are All You Need, "[24th place solution] a population-based reinforcement learning ecosystem," Kaggle write-up, Pokémon TCG AI Battle Challenge Simulation, 2026. [Online]. Available: https://www.kaggle.com/competitions/pokemon-tcg-ai-battle/writeups/24th-solution-a-population-based-rl-ecosystem

[10] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman, "Quantifying generalization in reinforcement learning," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97, 2019, pp. 1282–1289.

[11] R. Kirk, A. Zhang, E. Grefenstette, and T. Rocktäschel, "A survey of zero-shot generalisation in deep reinforcement learning," *Journal of Artificial Intelligence Research*, vol. 76, pp. 201–264, 2023.

[12] C. Angliss, J. Cui, J. Hu, A. Rahman, and P. Stone, "VGC-Bench: Towards mastering diverse team strategies in competitive Pokémon," in *International Conference on Autonomous Agents and Multiagent Systems*, 2026.

[13] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 15, 2011, pp. 627–635.

[14] M. Laskey, J. Lee, R. Fox, A. Dragan, and K. Goldberg, "DART: Noise injection for robust imitation learning," in *1st Conference on Robot Learning*, 2017.

[15] S. Belkhale, Y. Cui, and D. Sadigh, "Data quality in imitation learning," in *Advances in Neural Information Processing Systems*, 2023.

[16] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *AAAI Conference on Artificial Intelligence*, 2018.

[17] R. Agarwal, M. Schwarzer, P. S. Castro, A. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," in *Advances in Neural Information Processing Systems*, 2021.

[18] Kaggle, "kaggle-environments, version 1.33.0," https://github.com/Kaggle/kaggle-environments, 2026.

[19] R. Biswas, "[20th place] learning to play Pokémon TCG," Kaggle write-up, Pokémon TCG AI Battle Challenge Simulation, 2026. [Online]. Available: https://www.kaggle.com/competitions/pokemon-tcg-ai-battle/writeups/20th-place-learning-to-play-pokmon-tcg

[1] https://github.com/Builder106/tcg-cabt-lab