



RESEARCH ARTICLE

Simulation-Based Evaluation of Rule-Based Dynamic Difficulty Adjustment for Game-Based Learning Using Synthetic Bot Agents

Isabel Fátima de Jesus¹, Khoirina Wulan Nurbianti^{2*}

¹Administração de Aeroportos e Navegação Aérea de Timor-Leste

²Institut Teknologi Sepuluh November Surabaya, Indonesia

*Corresponding author: 6026261018@student.its.ac.id

Article History

Received: 13 October 2026

Revised: 19 October 2026

Accepted: 20 October 2026

Published: 22 October 2026

Publication Information

DOI: 10.36456/nd6wdk02

Volume / Issue: 3 / 1

Pages: 1-21

Keywords

difficulty calibration; dynamic difficulty adjustment; game-based learning; Monte Carlo simulation; synthetic agents

ABSTRACT

Dynamic difficulty adjustment (DDA) can regulate challenge in game-based learning, but favorable results against a fixed comparator do not necessarily demonstrate controller superiority. We evaluated a prespecified three-success/two-failure controller (DDA-R32) using Expert, Intermediate, and Novice synthetic agents across 3,000 simulation runs (three million interactions). Under the prespecified target tolerance, DDA-R32 increased target-zone occupancy relative to Fixed-5 by 0.6773, 0.3732, and 0.7694, respectively. Mean absolute probability error (MAE) decreased for Expert and Novice but not for Intermediate. A separately frozen exploratory protocol added 10,800 policy runs, tolerance sensitivity analyses, paired precision analyses, 40-state Markov validation, and horizon extensions. A feedback-only Window-20 policy achieved lower MAE than DDA-R32 in all nine profile-by-slope combinations. At the central slope, its MAE was 0.06948, 0.06553, and 0.06818, compared with 0.10377, 0.10493, and 0.09260 for DDA-R32. For the Intermediate profile, widening the target tolerance from 0.100 to 0.105 reversed the occupancy contrast with Fixed-5. The Markov analysis corroborated that the mechanistic balance reference describes adjustment requests rather than a stationary success-rate target. These findings demonstrate metric-dependent performance and limit claims of superiority for the streak controller. No human learning outcomes were measured.

INTRODUCTION

Dynamic difficulty adjustment (DDA) changes an interactive system's challenge in response to observed performance or estimated player state. Recent reviews distinguish heuristic thresholding, performance-driven adaptation, player modeling, and learning-based controllers while emphasizing inconsistent operational definitions and evaluation practices (Mortazavi et al., 2024; Paraschos & Koulouriotis, 2023; Souza et al., 2025). In serious games, similar mechanisms are discussed in relation to learning and rehabilitation objectives, although their outcome measures and participant populations differ from entertainment games (Víteková et al., 2026). This diversity makes it important to separate the operation of a difficulty controller from the downstream experience it might eventually produce.

A goal-centered account of DDA argues for specifying the quantity a game is intended to control rather than presuming that adjustment automatically generates a desirable subjective state (Guo et al., 2024). The present study adopts this narrower engineering position. Its target is not enjoyment, flow, or educational attainment: it is a mathematically defined balance of the next upward and downward adjustment requests and the accuracy with which a simple controller operates near

that balance.

Contemporary approaches demonstrate how much information can be incorporated into the adaptation loop. Facial-expression analysis has been used to estimate perceived difficulty and personalize arcade-game challenge (Blom et al., 2019). A laboratory comparison of affective exergame adjustments, however, illustrates that different control methods need not translate into clear user-experience differences on every outcome (Darzi et al., 2021). These investigations motivate more explicit claims about what a controller has actually measured, even when the available feedback is restricted to binary success and failure.

Player-state modeling has also been integrated into Monte Carlo tree search to diversify DDA agents (Moon et al., 2022), while fuzzy logic has been applied to challenge management in an educational three-dimensional game (Chrysafiadi et al., 2023). Reinforcement-learning formulations define states, actions, and rewards for adaptation in single-player action games (Climent et al., 2024). These methods can respond to richer features than a fixed outcome streak, but the diversity of objectives complicates comparisons. A transparent, prespecified streak controller therefore remains a useful computational baseline whose behavior can be isolated from estimation and training decisions.

The range of possible control objectives is particularly apparent outside conventional single-player action games. Pareto-based adjustment has been proposed for conflicting rehabilitation-exergame objectives (Ajani & Mallipeddi, 2023); mobile vocabulary games have used DDA for learners with autism spectrum disorder (Shohieb et al., 2022); and adaptive rubber-banding has been investigated for racing-game challenge (Mi & Gao, 2022). These systems are not interchangeable with the present Bernoulli model, but collectively demonstrate why algorithm evaluation should state both the target objective and the environment in which performance is measured.

Reliable adaptation also depends on the relationship between a nominal difficulty setting and the task outcomes it produces. Research on mobile puzzle games has combined player analytics with simulated data for difficulty estimation (Kristensen & Burelli, 2024). Recent work has explored the use of large language models in DDA-oriented game design (Aşkın, 2025) and the combination of adaptive difficulty with stealth assessment in collaborative learning (Rjiba et al., 2025). Such approaches highlight the value of richer response models, but they also introduce additional modeling assumptions and information requirements not present in deterministic streak-based rules.

Recent evaluations have compared fixed and adaptive progression strategies within virtual-reality games (Grosu et al., 2026), mapped adaptive techniques in cognitive-rehabilitation VR applications (Guzmán et al., 2024), and validated a framework supporting several adaptation paradigms in serious games (Lima et al., 2026). These human-facing studies supply important motivation, yet outcomes such as presence, preference, or rehabilitation relevance cannot be inferred from simulated binary agents. The present study instead limits its conclusions to controller mechanics and probability calibration under an explicitly stated data-generating process.

In educational systems, combining physiological measures with item-response models has been investigated for personalized difficulty prediction (Arevalillo-Herráez et al., 2023). That work further illustrates a measurement challenge that a synthetic experiment can isolate: proximity of a mean observed success rate to a nominal target need not imply that the controller consistently operates near a suitable level. We therefore examine three separate properties—target-region occupancy, absolute probability-calibration error, and adaptation timing—using an unchanged rule-based controller under three agent profiles.

A hypothetical game-based learning system may present successive practice tasks at discrete challenge levels while recording only completion success or failure. A low-information streak controller could regulate task challenge without a psychometric learner model, yet unnecessary changes could destabilize practice difficulty for learners already near the desired operating region. This motivates studying whether controller-level success indicators agree before exposing students to such a policy. We do not treat the synthetic profiles as empirical student proficiency groups or infer learning benefits from modeled task success.

The study evaluates DDA-R32, which increases difficulty after three consecutive successes and decreases it after two consecutive failures, against a fixed Level-5 comparator and a model-informed Oracle. Five research questions operationalize the frozen objectives: (RQ1) How do target occupancy and probability MAE differ across Expert, Intermediate, and Novice synthetic profiles? (RQ2) How rapidly is the first qualifying adaptation window detected? (RQ3) What transition and reversal behavior accompanies apparent adaptation? (RQ4) What are the paired occupancy and calibration differences relative to Fixed-5? (RQ5) Are the directions of those differences retained when the difficulty-response slope changes? Our contributions are a derived request-balance reference for R32, a prespecified multi-profile computational test, a demonstration of threshold-sensitive occupancy, and a separately frozen supplementary benchmark revealing that a feasible feedback-only alternative can outperform the streak controller in continuous calibration. The supplementary questions ask whether this result persists across profile and slope, whether an independent Markov model reproduces the original policy rewards, and whether late-window performance is stable over longer synthetic horizons. The comparison is deliberately limited to the chosen synthetic model and does not establish instructional effectiveness.

MATERIALS AND METHODS

Prespecified computational protocol

All controller settings, equilibrium definitions, target zones, evaluation windows, hypotheses, and engineering decision thresholds were locked in protocol RES-DDA-001 v1.1 on 12 October 2026, before the official experiments. The design follows the broader logic of transparent simulation reporting, where data-generating mechanisms, estimands, methods, and performance measures should be distinguished (Morris et al., 2019). The reported results follow this frozen baseline, and the unsuccessful Intermediate results were not used to modify the algorithm or select new thresholds. The protocol was prespecified internally, not registered in a public registry.

Synthetic agents and probability model

In each run, one of three synthetic agent profiles generated conditionally independent binary outcomes, given the current difficulty level. Initial success probabilities at difficulty Level 5 were $p_0 = 0.80$ (Expert), 0.50 (Intermediate), and 0.20 (Novice). Agents did not learn, fatigue, acquire memory of prior outcomes, or react strategically to the controller. Under E0, success probability was constant at p_0 regardless of level. Under E1 and E2, probabilities followed the prespecified logistic function:

$$p(L) = [1 + \exp(-(\text{logit}(p_0) - \beta(L - 5)))]^{-1}. \quad (1)$$

Difficulty L ranged from 1 to 10; increasing L reduced success probability. The main sensi-

tivity parameter was $\beta = 0.55$ (E1), with $\beta = 0.45$ and 0.65 as sensitivity conditions (E2). The logistic model is a simulation assumption, not an empirical fit to student or player data. Values $\beta = 0.45, 0.55, 0.65$ form a narrow, symmetric engineering stress test around the nominal slope rather than empirically estimated uncertainty bounds.

DDA-R32 controller and control policies

DDA-R32 started at Level 5. Three consecutive successes increased difficulty by one; two consecutive failures decreased difficulty by one. A contrary outcome reset the opposing streak counter, and threshold attainment reset both counters, including when a requested adjustment was blocked at Level 1 or 10. At most one level changed per interaction, and adjustments affected only subsequent interactions. Fixed-5 retained Level 5 throughout; the Oracle used complete model information to select the difficulty minimizing $|p(L) - p_{eq}|$ and started directly at that level. Consequently, the Oracle is a theoretical benchmark with an initialization advantage, not a realistically information-matched adaptive policy.

Analytical equilibrium and frozen target zones

For a constant probability p between streak resets, the probability that the next DDA-R32 adjustment is upward is $Q_{\uparrow}(p) = p^3(2 - p)/(1 - p + p^3)$. The request-balance equilibrium solves $Q_{\uparrow}(p_{eq}) = 0.5$, yielding $p_{eq} = 0.604450147809492$. This value describes balance between upward and downward adjustment requests under a constant response probability between streak resets; it is not the mean success probability at stationarity for the bounded, probability-dependent controller, nor a recommended target for human learners. To derive the request probability, denote by $q_N, q_{S1}, q_{S2}, q_{F1}$ the probability that the next request is upward when the current streak stage is neutral, one success, two successes, or one failure. First-step conditioning gives

$$q_N = pq_{S1} + (1 - p)q_{F1}, \quad (2)$$

$$q_{S1} = pq_{S2} + (1 - p)q_{F1}, \quad (3)$$

$$q_{S2} = p + (1 - p)q_{F1}, \quad (4)$$

$$q_{F1} = pq_{S1}. \quad (5)$$

Solving yields $q_N = Q_{\uparrow}(p)$ above. The equality $Q_{\uparrow}(p_{eq}) = 1/2$ concerns the direction of the next requested change after a reset, not the time-weighted marginal success rate at all difficulty levels.

A difficulty level was in the target zone if $|p(L) - p_{eq}| \leq 0.10$. The inclusive tolerance was fixed in advance. Table 1 presents the resulting level memberships, which were verified in the frozen equilibrium contract.

Table 1. Prespecified target-level matrix derived from p_{eq} and the ± 0.10 tolerance.

Sensitivity β	Expert	Intermediate	Novice
0.45	Levels 7–8	Level 4	Level 1
0.55	Levels 6–7	Level 4	Levels 1–2
0.65	Levels 6–7	Level 4	Level 2

Experiment design and randomization

Each profile-by-policy-by-slope cell comprised 100 independent replications of 1,000 sequential interactions. The fixed 1,000-interaction horizon was a computational study budget, not an empirically measured educational session length. The first 500 interactions were excluded from the principal late-window summaries to allow transient adaptation, without asserting that convergence necessarily occurred. The 100 replications per cell were a frozen pragmatic budget; their achieved simulation precision is quantified below rather than retrospectively described as an a priori power calculation. E0 characterized controller mechanics under fixed Bernoulli probabilities; E1 was the confirmatory DDA-versus-Fixed-5 comparison and included Oracle as a reference; E2 repeated the E1 policy comparisons at two additional β values. Table 2 summarizes the experiment design. All data were generated by the Python simulator v0.2.0 using NumPy PCG64 and a master seed of 20261012. Replication-specific streams were constructed from the master seed, profile code, and replication identifier, and common random numbers were reused across corresponding policies and β conditions. Operationally, for each profile and replication, the same 1,000 values $U_t \sim \text{Uniform}(0, 1)$ were generated by PCG64 and reused at the same iteration t by every paired policy and slope. An outcome was recorded as success when $U_t < p(L_t)$; because each policy could occupy a different state-dependent level L_t , matched runs shared random innovations but not necessarily binary outcomes. This coupling reduced extraneous simulation noise without treating correlated policy outcomes as independent observations. Statistical seed 20261012 was used for the reported analysis. The configuration, metadata, and stored output files form an audit trail; detailed metadata are recognized as important to reproducible computational research (Leipzig et al., 2021).

Table 2. Official simulation design (1,000 interactions per run; 100 replications per profile-policy-slope cell).

Condition	Probability model	Policies	Simulation runs
E0	Constant p_0	DDA-R32	300
E1	Logistic, $\beta = 0.55$	DDA-R32 / Fixed-5 / Oracle	900
E2	Logistic, $\beta = 0.45$ and 0.65	DDA-R32 / Fixed-5 / Oracle	1,800
TOTAL	—	100 replications per cell	3,000

Outcome definitions

The primary outcomes were target-zone occupancy and probability mean absolute error (MAE), both evaluated over interactions 501–1,000. Occupancy was the proportion of these 500 interactions at levels satisfying the target criterion. Probability MAE was the mean $|p(L_t) - p_{\text{eq}}|$ across the same window. Both outcomes concern the modeled probability function rather than noisy empirical success alone.

Secondary outcomes included (a) adaptation detection time (ADT), defined as the earliest ending iteration of a moving window of 20 interactions containing at least 16 target-zone hits, with runs censored at interaction 1,000 if never detected; (b) the proportion of the final 500 interactions that actually changed level; (c) direction reversals between successive nonzero difficulty adjustments in this evaluation window; and (d) the proportion of all adjustment requests blocked at level boundaries. ADT indicates the first qualifying window, not permanent convergence.

Confirmatory inference and decision rules

The inferential unit was the complete 1,000-interaction replication, avoiding pseudoreplication. For each of three profiles, two prespecified hypotheses compared DDA-R32 and Fixed-5 in E1: H1, occupancy improvement $\Delta O = O_{\text{DDA}} - O_{\text{Fixed-5}} > 0$; and H2, calibration improvement $\Delta E = \text{MAE}_{\text{Fixed-5}} - \text{MAE}_{\text{DDA}} > 0$. Mean paired effects were tested using two-sided Monte Carlo sign-flip tests with 100,000 draws. Six p values underwent Holm adjustment for familywise $\alpha = 0.05$; contemporary methodological literature discusses control of the familywise error rate when multiple outcomes are tested (Vickerstaff et al., 2019). Reported 95% percentile confidence intervals were computed from 10,000 paired bootstrap resamples; the percentile bootstrap has useful computational simplicity but requires interpretation in light of its known coverage limitations (Rousset et al., 2021). These intervals were not multiplicity-adjusted. Sign-flip testing assumes exchangeability of signs under the null and should not be mistaken for unconditional randomization inference merely because matched runs share random streams.

The two-sided tests were retained because they were the tests specified and implemented in the original analysis, although the conceptual engineering questions anticipated beneficial directions. Two-sided evidence is therefore interpreted as a conservative departure-from-zero check, not as a replacement for the sign and magnitude criteria. Inferences require symmetry/sign-exchangeability of paired differences under the relevant null; shared uniform draws alone do not prove that requirement. Observed paired-difference skewness and Monte Carlo standard errors (MCSEs) are reported to make this conditional inferential interpretation explicit. The six corrected p values are supplementary to effect magnitude and finite-replication precision, particularly where the randomization procedure reaches its numerical floor.

Separately, engineering success required C1: occupancy gain ≥ 0.10 ; C2: MAE reduction ≥ 0.025 ; C3: adaptation detected in $\geq 90/100$ runs; C4: Kaplan–Meier median ADT ≤ 250 ; and C5: positive occupancy and MAE gains at both $\beta = 0.45$ and 0.65 . Kaplan–Meier estimation was used for time to first qualifying adaptation window, with right censoring for runs not detected within 1,000 iterations; the assumptions associated with censoring must be considered when interpreting survival estimators (Cole et al., 2020). Every criterion had to pass for every E1 profile for the overall engineering decision to pass. The frozen values were *engineering acceptance conventions*, not thresholds calibrated to educational gains: C1 corresponds to 50 additional target interactions among the last 500; C2 to 0.025 units of absolute success-probability error; C3 requires detection in at least 90% of replications; and C4 asks that half the replications detect adaptation within the first quarter of the 1,000-interaction horizon. The 16-of-20 rule specifies 80% local target membership for first detection, not sustained convergence; C5 checks only the sign of both changes across a limited slope range. The adopted values represent transparent design tolerances, not empirical minimal important differences or universal benchmarks. They were not adjusted in response to observed performance. Tests, confidence intervals, and engineering criteria therefore answer related but nonidentical questions.

Separately frozen supplementary exploratory protocol

After an internal simulated review identified potential comparator and threshold artifacts, a separate protocol (RES-DDA-R2-EXP-001 v1.0) was frozen on 19 October 2026 before executing the additional analyses. The original RES-DDA-001 v1.1 confirmatory specification and its 3,000 runs were not revised. The new benchmark generated 10,800 further runs: three agent profiles,

three slopes ($\beta = 0.45, 0.55, 0.65$), 100 replications, 12 new policies, and 1,000 interactions per run. Policies were Fixed-L1 through Fixed-L10, DDA-R22 (request +1 after two successive successes and -1 after two successive failures, resetting the streak after any request), and Window-20. The original DDA-R32 and Oracle trajectories were used for paired reference comparisons and were not counted as new runs. Each fixed policy started directly at its selected level; adaptive policies started at Level 5, which is an explicit initialization asymmetry.

Window-20 uses 20 nonoverlapping binary outcomes per block. After each complete block, it requests a one-level increase if the observed success proportion exceeds $p_{eq} + 0.10$, a one-level decrease if it is less than $p_{eq} - 0.10$, and no change otherwise. Updates affect the next block and levels are clamped to 1–10. It receives neither the true $p(L)$, agent identity, nor β . The reference equilibrium and tolerance are design constants shared with DDA-R32, not secret per-agent model information. Fixed-level minima selected after inspection of true $p(L)$ and the model-informed Oracle are privileged descriptive bounds, not deployable learner-blind comparators.

Common random numbers were paired using PCG64 and `SeedSequence([20261012, profile_code, replication_ID])`; at interaction t , each policy received the identical U_t and recorded success if $U_t < p(L_t)$. Thus randomized streams were shared, but outcomes and level states could differ. Final-500 occupancy and MAE used the original estimands. For each new comparison, the paired effect, SD, MCSE = $s_{\Delta}/\sqrt{100}$, 95% paired t confidence interval and 2,000-draw paired percentile bootstrap confidence interval were calculated. Exploratory interval widths describe stochastic Monte Carlo precision, not uncertainty about learner populations. No newly tested contrast was designated confirmatory, nor was a multiplicity-adjusted superiority claim added post hoc.

Tolerance, stochastic-model and temporal validation

We reevaluated the unchanged original DDA-R32 trajectories for 96 prespecified tolerances τ over approximately 0 to 0.45, including dense values surrounding $\tau = 0.10$, for all nine profile-by-slope cells. Occupancy was recalculated from $|p(L_t) - p_{eq}| \leq \tau$ in interactions 501–1,000; MAE and all policy trajectories were unchanged. An independently coded 40-state Markov chain combined ten levels with four streak states $N, S1, S2, F1$. Its transition kernel used the model-derived Bernoulli probability at each level, streak resets after accepted or blocked requests, and clamping at both bounds. Stationary rewards and exact final-500 expected rewards starting from Level 5/neutral were compared against the replication means and MCSEs. We also examined within-run windows (251–500, 501–750, 751–1,000, final 250, final 500, and full 1,000) across all three slopes, and continued the same random streams to 2,000 and 5,000 interactions at $\beta = 0.55$ to assess final-500 horizon dependence.

Adaptation uncertainty and provenance safeguards

For adaptation detection proportions, 95% Wilson intervals were calculated across 100 replications per E1 profile. For Kaplan–Meier median time to the first qualifying window, 2,000 replication-level bootstrap resamples supplied percentile intervals; nonadapted runs remained right-censored at 1,000. The original two-sided sign-flip tests were preserved as frozen results, with their assumption of null sign-exchangeability explicitly distinguished from the independence assumption behind mean-based Monte Carlo intervals. A finite Monte Carlo sign-flip p value is calculated as $(b+1)/(B+1)$ for $B = 100,000$ draws and b equally or more extreme randomized statistics; multiple identical

adjusted minimum values do not measure identical effect magnitude. The supplementary analysis was executed from a documented deterministic reconstruction that reproduces original aggregate results, not from a byte-verified copy of the original Windows raw exports. The all-row three-million-interaction state reconciliation was not completed in this analysis session, and a Git commit hash was not available in the supplied metadata. These are provenance limitations, not claims of verification.

RESULTS

Mechanistic behavior without probability feedback (E0)

When difficulty had no effect on the success probability, DDA-R32 continued to issue adjustment requests that often accumulated against the level bounds. Mean blocked-request rates were 70.96% for Expert, 39.70% for Intermediate, and 95.33% for Novice. Across E0, mean empirical success rates remained near their designated values (0.8022, 0.4962, and 0.2006, respectively). This negative-control result confirms that boundary attraction can be a property of the controller mechanics alone; E0 cannot establish ability calibration because the modeled probability is insensitive to difficulty.

Main experiment: occupancy and calibration (E1)

Table 3 contrasts the three policies at $\beta = 0.55$. Oracle achieved target occupancy of 100% for all profiles by starting at its selected optimal model level. Fixed-5 occupancy was identically zero because Level 5 belonged to none of the prespecified target sets. DDA-R32 maintained the target level(s) for 67.73% of late interactions in Expert, 37.32% in Intermediate, and 76.94% in Novice. In the Intermediate profile, occupancy substantially exceeded Fixed-5 despite a high frequency of departures from Level 4. Figure 1 displays occupancy by profile, and Figure 2 compares the probability-calibration errors. Importantly, Fixed-5 has occupancy zero by construction under the frozen tolerance; H1 is consequently evidence of entering a defined set, not independent evidence that DDA delivers better continuous calibration.

Table 3. E1 primary outcomes: 100 runs per profile and policy (means across runs).

Profile	DDA occupancy	DDA MAE	Fixed-5 MAE	Oracle MAE
Expert	0.6773	0.10377	0.19555	0.03336
Intermediate	0.3732	0.10493	0.10445	0.02969
Novice	0.7694	0.09260	0.40445	0.03890

Occupancy for Fixed-5 is 0 and for Oracle is 1 in every E1 profile. Smaller MAE indicates better probability calibration.

Paired confirmatory tests

Table 4 reports the six prespecified paired comparisons. Five of the six E1 comparisons showed positive estimated effects that were statistically significant after Holm adjustment. For H1, all three occupancy effects were positive in all 100 matched replications per profile. For H2, Expert and Novice improved, whereas Intermediate produced a near-zero negative mean effect, a confidence interval spanning zero, and an adjusted p value of 0.59052. The Monte Carlo floor for extreme permutation results was about 1/100,001; thus very small p values are reported at their finite-draw resolution rather than as zero.

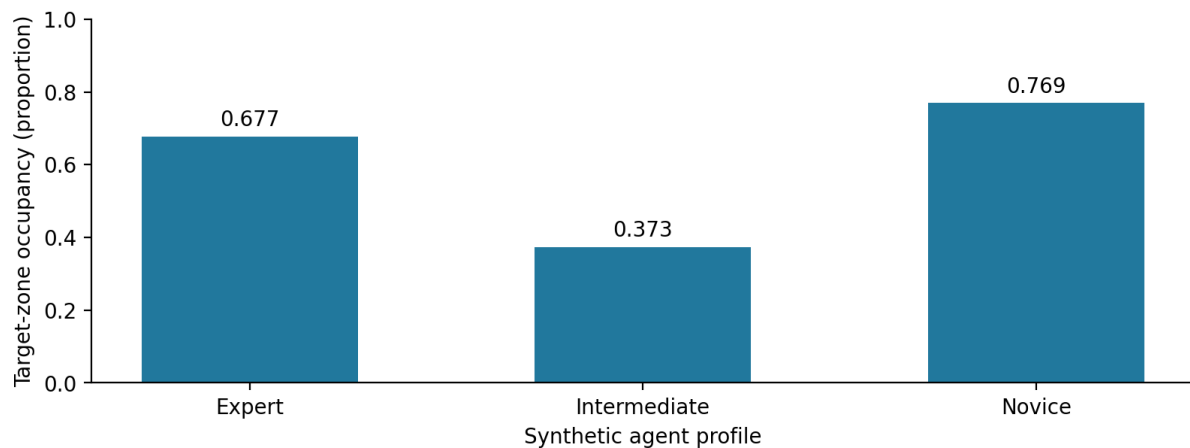


Figure 1. DDA-R32 target-zone occupancy in E1 during interactions 501–1,000; bars represent the mean proportion of evaluated interactions in the target zone.

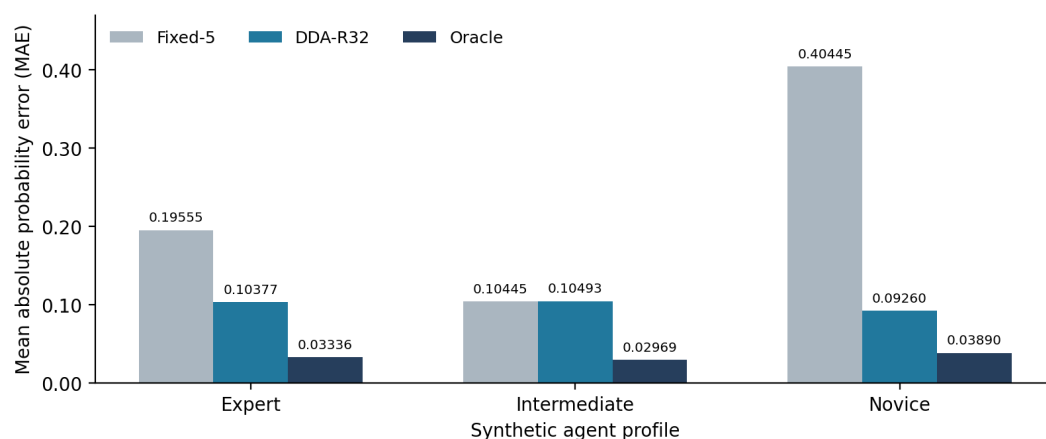


Figure 2. Mean absolute probability error (MAE) in E1 for DDA-R32, Fixed-5 and Oracle. Five-decimal annotations preserve the small Intermediate difference; smaller values indicate better calibration.

Table 4. Prespecified E1 paired hypotheses (n = 100 pairs per row).

Profile	Test	Mean paired Δ	95% paired bootstrap CI	Holm p	Decision
Expert	H1	+0.67730	[+0.66694, +0.68778]	0.00006	Supported
Expert	H2	+0.09178	[+0.09000, +0.09353]	0.00006	Supported
Intermediate	H1	+0.37322	[+0.36470, +0.38182]	0.00006	Supported
Intermediate	H2	-0.00048	[-0.00225, +0.00126]	0.59052	Not supported
Novice	H1	+0.76944	[+0.75810, +0.78072]	0.00006	Supported
Novice	H2	+0.31185	[+0.31026, +0.31340]	0.00006	Supported

H1: DDA minus Fixed-5 occupancy. H2: Fixed-5 minus DDA MAE. Positive effects favor DDA; two-sided sign-flip tests with Holm correction.

Table 5 reports the replication-level Monte Carlo precision of the six E1 paired effects. Each MCSE equals the sample standard deviation of 100 run-level paired differences divided by $\sqrt{100}$; it is not a standard error based on three million independent interactions. For Intermediate H2, the estimate -0.00048 is smaller in magnitude than its MCSE (0.000894), reinforcing that the calibration difference is indistinguishable from zero at the available resolution. The observed skewness coefficients in the official paired summaries ranged approximately from -0.51 to -0.12 across the six contrasts. These descriptions do not establish null sign-exchangeability; the conditional sign-flip p values must therefore be read alongside the effects and MCSEs.

Table 5. Additional Monte Carlo precision audit for E1 paired effects (100 replication pairs per profile).

Profile	Contrast	Mean paired effect	Paired SD	MCSE
Expert	H1 (ΔO)	+0.67730	0.05403	0.005403
Expert	H2 (ΔE)	+0.09178	0.00920	0.000920
Intermediate	H1 (ΔO)	+0.37322	0.04330	0.004330
Intermediate	H2 (ΔE)	-0.00048	0.00894	0.000894
Novice	H1 (ΔO)	+0.76944	0.05790	0.005790
Novice	H2 (ΔE)	+0.31185	0.00819	0.000819

Adaptation detection, stability, and adjustment dynamics

Figure 3 and Table 6 summarize adaptation timing and late-window adjustment dynamics. The Kaplan–Meier median ADT was 28 interactions for Expert, 309 for Intermediate, and 32 for Novice in E1. All Expert and Novice replications met the detection rule, compared with 94 of 100 Intermediate replications. All Fixed-5 replications remained undetected and were censored at iteration 1,000, whereas Oracle met the 20-interaction window criterion by iteration 20. The relatively long Intermediate median indicates that reaching a predominantly Level-4 moving window was substantially harder than reaching the target sets for Expert and Novice.

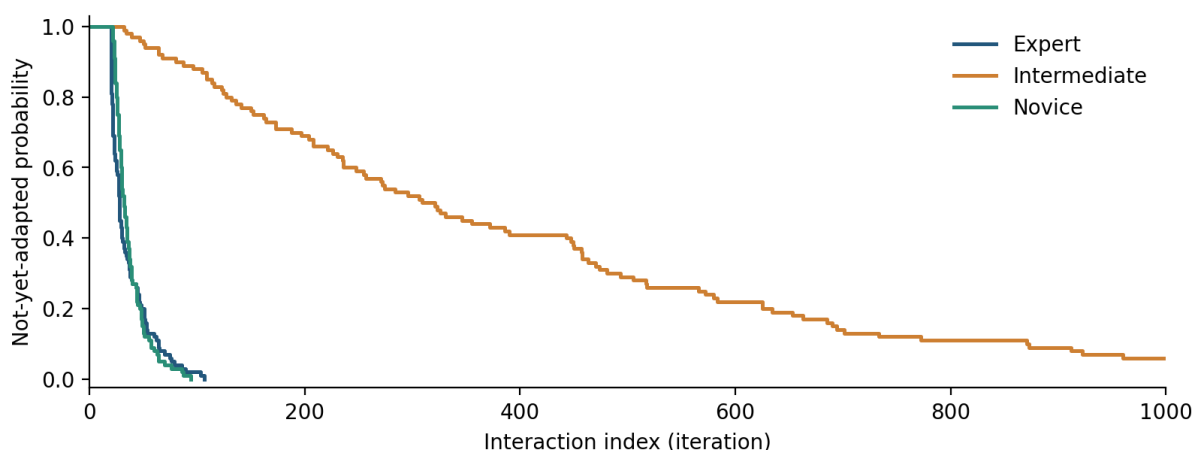


Figure 3. Kaplan–Meier survival of DDA-R32 runs not yet meeting the adaptation-detection criterion in E1.

Across profiles, DDA changed level in approximately 21–24% of the final 500 interactions, with direction reversals in approximately 60–70% of adjacent nonzero moves. These metrics document oscillatory dynamics but were not assigned independent pass/fail thresholds. For Intermediate, the

Table 6. E1 adaptation detection and late-window difficulty fluctuations.

Profile	Adapted runs	KM median ADT	Move rate	Reversal rate
Expert	100/100	28	23.69%	62.07%
Intermediate	94/100	309	23.44%	59.91%
Novice	100/100	32	20.92%	70.15%

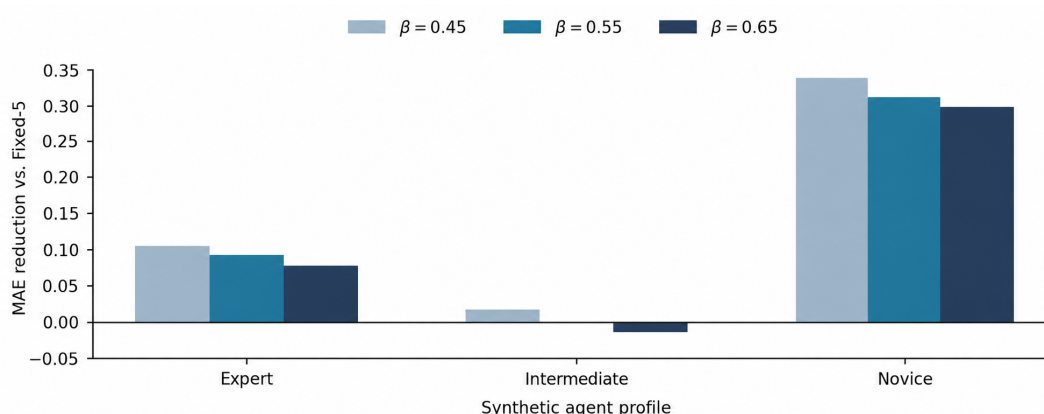
late-window move rate was 23.44% and the reversal rate 59.91%, consistent with low target occupancy despite an aggregate empirical success rate of 0.60241, numerically close to p_{eq} . This apparent proximity does not imply low MAE: opposing probability deviations can cancel in the average success rate while adding under an absolute-error metric.

Sensitivity analysis (E2)

Table 7 and Figure 4 present the sensitivity results. DDA improved occupancy relative to Fixed-5 for all profiles at both secondary sensitivity values; calibration gains, however, were parameter-dependent in Intermediate. At $\beta = 0.45$, Intermediate mean MAE reduction was +0.01544, below the E1 engineering magnitude criterion but positive. At $\beta = 0.65$, it became -0.01403 , violating C5 directional robustness. Expert and Novice retained positive MAE reductions at both values. Adaptation speed was also sensitive to the difficulty-response slope for the Novice profile: median ADT increased to 262 interactions at $\beta = 0.65$, a descriptive sensitivity finding outside the primary E1 C4 decision.

Table 7. E2 sensitivity results (descriptive; 100 runs per policy-condition).

β	Profile	DDA occupancy	MAE reduction vs Fixed-5	KM median ADT
0.45	Expert	60.76%	+0.10432	44
0.45	Intermediate	34.88%	+0.01544	306
0.45	Novice	53.81%	+0.33707	50
0.65	Expert	71.80%	+0.07643	27
0.65	Intermediate	38.94%	-0.01403	332
0.65	Novice	42.60%	+0.29809	262

**Figure 4.** Sensitivity of MAE reduction relative to Fixed-5 across agent profiles and logistic difficulty-response slopes. Negative values indicate deterioration with DDA-R32.

Threshold sensitivity of the original occupancy outcome

A separate deterministic re-evaluation of the existing E1 trajectories varied only the *measurement rule* $|p(L) - p_{eq}| \leq \tau$; the DDA control law, generated random streams, realized levels, and all original hypothesis decisions remained unchanged. The original DDA-R32 trajectories for all 900 logistic runs across E1 and E2 were reproduced under the documented v0.2.0 simulator and master seed, matching run-level occupancy and MAE within numerical precision. The expanded threshold grid covers all profiles and slopes; its graphical results are presented in the expanded occupancy tolerance analysis below. Table 8 reports sensitivity near the frozen $\tau = 0.100$ boundary; the initial audit was conceived after an internal simulated review and was subsequently broadened under the separately frozen supplementary protocol; it is *exploratory*, not a preregistered alternative test.

Table 8. Exploratory E1 occupancy sensitivity for the Intermediate profile ($\beta = 0.55$; 100 paired trajectories).

Tolerance τ	Target levels	DDA occupancy	Fixed-5 occupancy	DDA minus Fixed-5
0.090	4	0.37322	0.00000	+0.37322
0.100 (frozen)	4	0.37322	0.00000	+0.37322
0.104	4	0.37322	0.00000	+0.37322
0.105	4, 5	0.65282	1.00000	-0.34718
0.110	4, 5	0.65282	1.00000	-0.34718
0.150	3, 4, 5	0.86080	1.00000	-0.13920

The frozen Intermediate Fixed-5 probability is $p(5) = 0.50$, whose absolute distance from the equilibrium is 0.1044501478. Thus a tolerance only 0.0044501478 wider than 0.100 changes the Fixed-5 occupancy deterministically from zero to one. At $\tau = 0.105$, the Intermediate occupancy contrast reverses to -0.34718 , even though no policy behavior changes. At the central slope, Expert and Novice contrasts are unchanged between $\tau = 0.100$ and 0.105 because no relevant additional levels cross the boundary; other tolerance thresholds vary by profile and slope. This discontinuity demonstrates that H1 is strongly contingent on set membership, particularly for Intermediate; continuous MAE is the more discriminating evidence for comparative calibration in this case. The frozen C1–C5 decisions in Table 9 must *not* be recomputed from this exploratory threshold audit.

Prespecified engineering success decision

Table 9. Engineering success matrix using frozen RES-DDA-001 v1.1 criteria.

Rule	Frozen threshold	Expert	Intermediate	Novice
C1	$\Delta\text{occupancy} \geq 0.10$	PASS	PASS	PASS
C2	$\Delta\text{MAE} \geq 0.025$	PASS	FAIL	PASS
C3	$\geq 90/100$ adapted	PASS	PASS	PASS
C4	KM median ADT ≤ 250	PASS	FAIL	PASS
C5	Both effects > 0 at $\beta = .45 / .65$	PASS	FAIL	PASS

As reported in Table 9, engineering success was achieved for Expert and Novice but failed overall because Intermediate did not meet C2 (the required MAE reduction), C4 (median ADT ≤ 250), and C5 (positive MAE improvement under both sensitivity settings). The engineering decision is distinct from statistical support for H1 and H2 and is not revised in response to the results.

Exploratory comparisons with feasible and privileged baselines

The most consequential supplementary finding is that Window-20, a feasible feedback-only adaptive controller, produced a smaller mean probability MAE than DDA-R32 in all nine profile-slope cells (Table 10; Figure 5). Its advantages were not limited to a binary occupancy threshold. At the central slope $\beta = 0.55$, Window-20 MAE was 0.06948, 0.06553, and 0.06818 for Expert, Intermediate, and Novice compared with 0.10377, 0.10493, and 0.09260 for DDA-R32. The paired difference was defined as $MAE_{\text{Window-20}} - MAE_{\text{DDA-R32}}$, so negative values favor Window-20. All three 95% paired t confidence intervals at this slope excluded zero in the Window-20 direction (Table 11). The baseline was not tuned on these observed outcomes within the frozen exploratory protocol, but this benchmark does not establish superiority in educational attainment or in other agent processes.

Table 10. Exploratory MAE benchmarking across all synthetic profiles and difficulty slopes ($n = 100$ paired replications per cell; lower is better).

β	Profile	DDA-R32	DDA-R22	Window-20	Window-20 – R32
0.45	Expert	0.09123	0.13059	0.05543	-0.03580
0.45	Intermediate	0.08901	0.13095	0.04839	-0.04062
0.45	Novice	0.06738	0.13293	0.02322	-0.04417
0.55	Expert	0.10377	0.14368	0.06948	-0.03429
0.55	Intermediate	0.10493	0.14379	0.06553	-0.03941
0.55	Novice	0.09260	0.14200	0.06818	-0.02442
0.65	Expert	0.11912	0.15767	0.08433	-0.03479
0.65	Intermediate	0.11848	0.15480	0.07980	-0.03867
0.65	Novice	0.10636	0.15148	0.07205	-0.03431

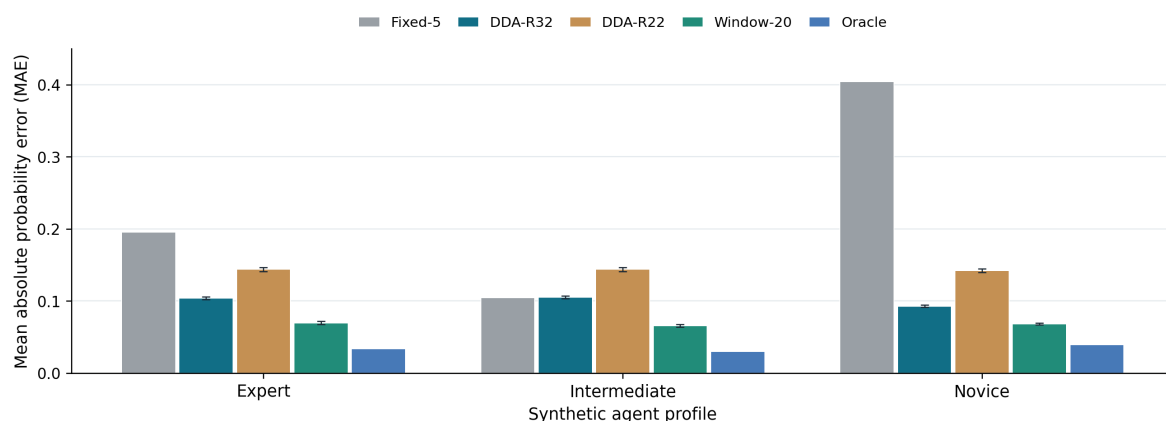


Figure 5. Exploratory MAE comparison at $\beta = 0.55$ with 95% confidence intervals for the mean MAE of each adaptive policy, calculated across independent replications (not intervals for paired differences). Fixed-5 and the model-informed Oracle are nonadaptive and privileged reference points, respectively, not information-matched adaptive baselines. Full results for all ten fixed levels appear in the supplementary files.

DDA-R22 performed worse than DDA-R32 for every profile-slope combination. The ten fixed policies expose the original Fixed-5 selection effect: at $\beta = 0.55$, the lowest-MAE levels chosen *after inspecting true probabilities* were Level 7 (Expert, 0.03336), Level 4 (Intermediate, 0.02969), and Level 2 (Novice, 0.03890). These reproduce the privileged Oracle optima and are not prospective per-learner fixed-level recommendations. Thus, the original Fixed-5 contrast demonstrates

Table 11. Exploratory paired Window-20 minus DDA-R32 MAE effects ($\beta = 0.55$; 100 paired replications per row).

Profile	Mean difference	MCSE	95% paired t CI
Expert	-0.034286	0.001042	[-0.036353, -0.032218]
Intermediate	-0.039408	0.001094	[-0.041578, -0.037238]
Novice	-0.024417	0.001121	[-0.026641, -0.022194]

ability to leave one starting level, whereas the Window-20 comparison provides an information-comparable feedback-only alternative (with a different decision cadence and an explicitly prespecified controller rule).

Expanded occupancy tolerance analysis

Figure 6 displays occupancy as a function of the tolerance across all profiles and slopes. The stepwise dependence reflects the discrete level sets and is not a smooth improvement in continuous calibration. At $\beta = 0.55$, the Intermediate Fixed-5 success probability 0.50 enters the target zone exactly at $\tau = 0.104450147809492$. Moving from $\tau = 0.100$ to 0.105 therefore reverses DDA-minus-Fixed-5 occupancy from +0.37322 to -0.34718 without altering any state transitions. This finding qualifies the importance of the frozen H1 occupancy tests and provides a reason to prioritize continuous MAE when comparing policies with nearly boundary-level probabilities.

Paired precision and adaptation uncertainty

The original Intermediate Fixed-5 minus DDA-R32 mean MAE effect was -0.000484 with a paired MCSE of 0.000894, underscoring the lack of evidence for calibration improvement in this starting-near-target profile. Figure 7 shows the paired-difference distribution and associated interval. Mean-based paired t intervals and percentile bootstrap intervals serve as robustness summaries under independent seeded replications; the original sign-flip results remain conditional on null sign-exchangeability. The finite-draw sign-flip lower bound and Holm adjustment do not make occupancy contrasts against a structurally zero Fixed-5 target share more substantively meaningful.

Table 12. Adaptation-detection and Kaplan–Meier timing uncertainty for original E1 DDA-R32 runs.

Profile	Adapted / 100	Wilson 95% rate CI	KM median	Bootstrap 95% median CI
Expert	100	[0.963, 1.000]	28	[25, 30]
Intermediate	94	[0.875, 0.972]	309	[236, 390]
Novice	100	[0.963, 1.000]	32	[30, 35]

The confidence intervals quantify variability among synthetic replications, and adaptation detection describes the first qualifying window rather than sustained control; they cannot be transported to learning outcomes or time scales of real instruction.

Independent Markov validation of the frozen controller

The 40-state Markov analysis independently replicated the state transitions and rewards of the original controller for all profile-slope cells. As summarized in Table 13, the exact Markov expectation for the final-500 window agreed with empirical means within 1.90 empirical MCSEs across the nine cells; the full cell-by-cell validation is supplied as supplementary data. The stationary

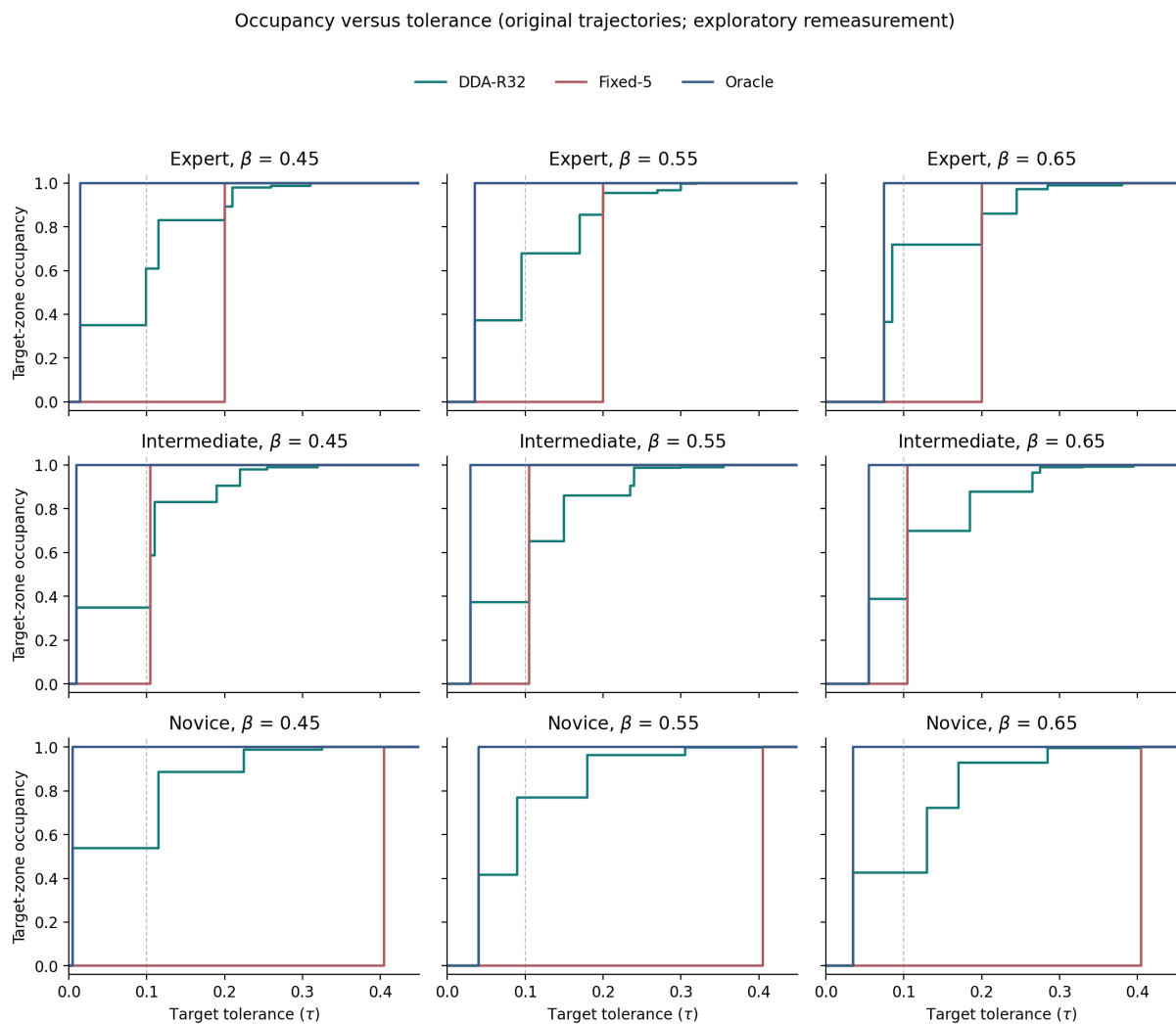


Figure 6. Exploratory target-zone occupancy sensitivity for all nine profile-slope combinations over the predeclared tolerance grid. Curves are stepwise because only discrete difficulty levels are available.

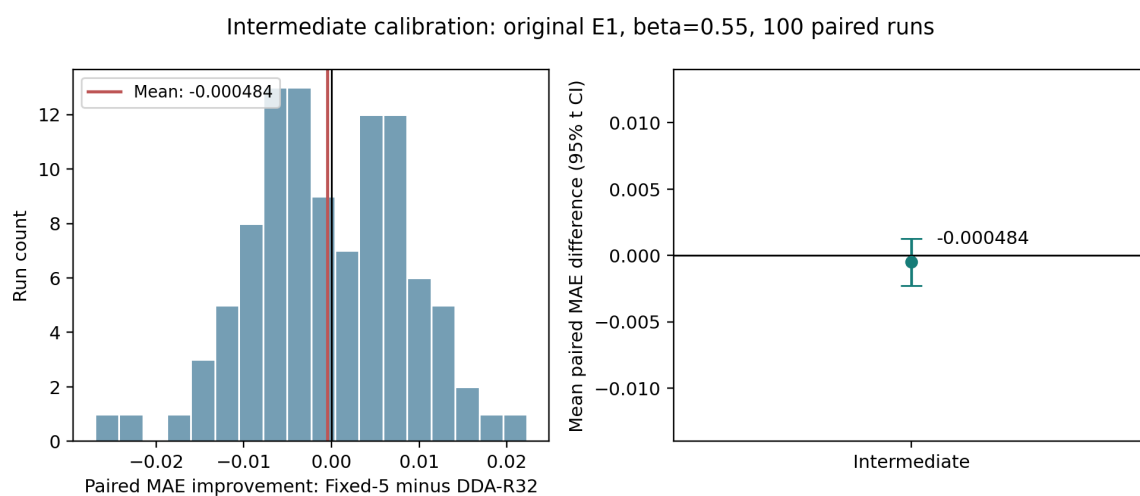


Figure 7. Replication-level Intermediate MAE difference between Fixed-5 and DDA-R32 at the central slope, including a mean-effect uncertainty display. Negative effects do not favor the streak controller.

mean success rate is not generally the request-equilibrium value. At $\beta = 0.55$, it was approximately 0.60419 for Expert, 0.60413 for Intermediate, and 0.57549 for Novice, despite the common $p_{eq} = 0.60445$ used to construct the target. This explicitly separates the local next-request balance from the long-run distribution of a bounded controller.

Table 13. Independent 40-state Markov predicted versus simulated final-500 outcomes, $\beta = 0.55$.

Profile	Markov occ.	Sim. occ.	Markov MAE	Sim. MAE	Stat. mean success
Expert	0.67345	0.67730	0.10473	0.10377	0.60419
Intermediate	0.37678	0.37322	0.10419	0.10493	0.60413
Novice	0.77319	0.76944	0.09218	0.09260	0.57549

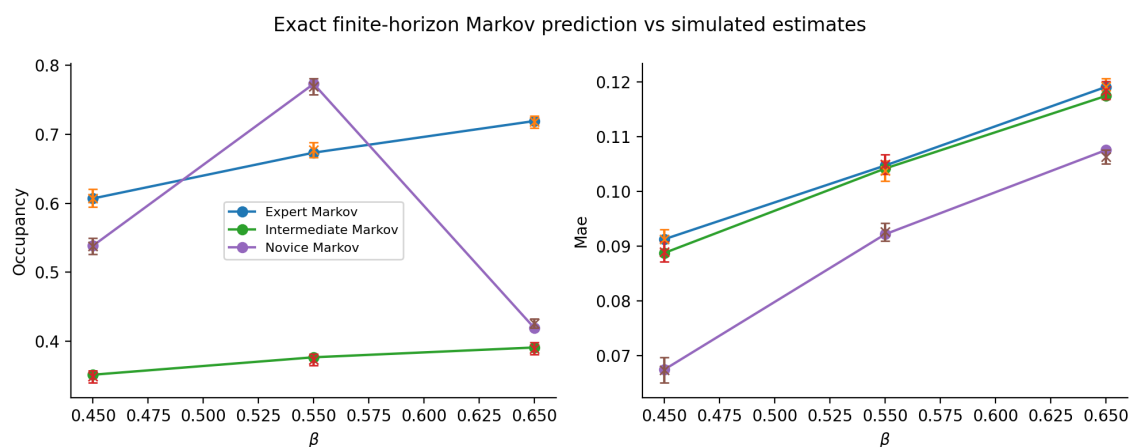


Figure 8. Independent Markov prediction and empirical DDA-R32 simulation agreement across profiles and difficulty-response slopes. Reported simulation uncertainty is at the independent-replication level.

Window and horizon sensitivity of original DDA-R32

Late-window diagnostics across all slopes and time segments are supplied in the supplementary outputs. At $\beta = 0.55$, extending the exact original random-stream prefixes to 2,000 and 5,000 interactions yielded the final-500 summaries in Table 14. The means were broadly near their independent stationary Markov counterparts, although this numerical stability is conditional on fixed independent Bernoulli agents and the unchanged probability curve. In particular, the Intermediate MAE remained close to 0.104–0.105 instead of exhibiting unambiguous convergence toward lower error.

Table 14. Exploratory final-500 horizon sensitivity for original DDA-R32, $\beta = 0.55$ (100 continued seeded streams per profile).

Profile	MAE-1000	MAE-2000	MAE-5000	Occ.-1000	Occ.-2000	Occ.-5000
Expert	0.10377	0.10468	0.10615	0.67730	0.67230	0.66880
Intermediate	0.10493	0.10478	0.10446	0.37322	0.37512	0.37610
Novice	0.09260	0.09235	0.09148	0.76944	0.77120	0.77830

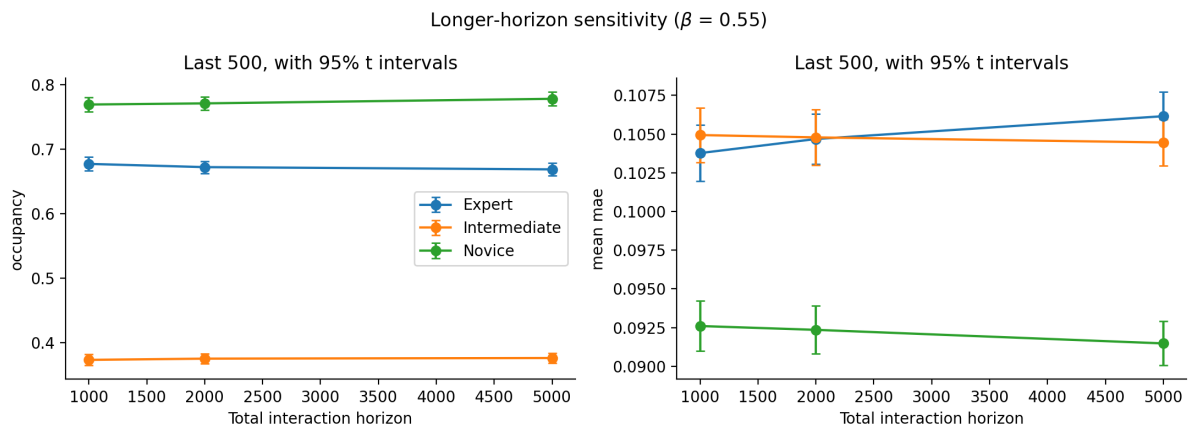


Figure 9. Horizon sensitivity of final-window DDA-R32 MAE and occupancy across 1,000, 2,000, and 5,000 interactions at $\beta = 0.55$.

DISCUSSION

How the stronger adaptive baseline changes the claim

The supplementary benchmark changes the substantive interpretation of the frozen study. DDA-R32 improved continuous MAE against Fixed-5 for Expert and Novice under the nominal model, but this contrast did not establish that its streak rule was comparatively efficient. Window-20, using only outcome feedback and the same initialization level, achieved lower MAE than DDA-R32 in every profile-slope condition; at the nominal slope, all three 95% confidence intervals for the paired MAE differences excluded zero. DDA-R22's poorer performance shows that simply shortening streaks is not a universal remedy. These findings support the narrower claim that DDA-R32 offers a transparent proof-of-mechanism and a baseline for calibration studies, not a preferred controller for synthetic or real educational environments. Window-20's better performance is itself conditional on the tested 20-outcome block, $p_{eq} \pm 0.10$ feedback rule, logistic response curves, and 1,000-step evaluation budget; neither its transferability nor superiority against other feasible algorithms has been established.

Metric dependence, threshold artifacts, and comparator selection

The original occupancy result is partly a property of the target definition and the fixed comparator. Fixed-5 lies outside every frozen target set, whereas the Oracle selects a target level using true model information. Consequently, H1 against Fixed-5 is structurally favorable and should not be presented as general evidence of superior control. The sharply reversed Intermediate effect at $\tau = 0.105$ demonstrates that a discontinuous occupancy indicator can change its ranking without any change in agent response or controller action. Continuous MAE retains information about distance to the analytical reference and is a more discriminating evaluation here, although the reference itself remains a model-defined engineering quantity rather than validated pedagogy. The favorable Window-20 contrast on continuous MAE also shows that the metric disagreement observed for Intermediate was not an unavoidable limitation of binary feedback; it arose under the specific R32 update logic.

What is balanced by the analytical reference

The derived $Q_{\uparrow}$ expression identifies the point where, at a constant success probability, successive adjustment requests are equally likely to point upward or downward. It does not assert that a bounded adaptive controller spends equal time on either side of this probability or that its marginal success probability converges to p_{eq} . The 40-state analysis makes this distinction explicit and agrees with simulated reward means. Its stationary Novice success rate at $\beta = 0.55$ (approximately 0.57549) demonstrates that a single request-equilibrium reference cannot be interpreted as the actual controller-wide stationary outcome. Horizon extensions reduce concern that the final-500 observations are an isolated transient, but the same synthetic assumptions are shared by the Markov and Monte Carlo calculations.

Inference and uncertainty without overclaiming

Each replication is one independent inferential unit; the individual interactions are temporally dependent through controller state and cannot be treated as independent trials. Paired CRN contrasts reduce Monte Carlo variation but do not, by themselves, justify sign-flip randomization inference. The original corrected p values were kept as defined by the frozen protocol, while new confidence intervals for paired mean effects provide a robustness assessment based on independent replication draws and approximate mean normality. The simulations are not a random sample of students, tasks, or institutions; small Monte Carlo standard errors cannot compensate for structural model misspecification. Where binary occupancy differs from an identically zero comparator, substantive interpretation is especially sensitive to the predefined tolerance.

Implications for game-based learning

Consider a proposed aviation safety practice game in which students progress through ten task-difficulty levels and each task records a binary successful or unsuccessful hazard-identification response. A controller could assign the next scenario difficulty from this sparse feedback without access to personal histories or proprietary engagement measures. The present results suggest that an implementer should evaluate both distance from an explicitly justified operating target and the temporal burden of frequent level changes, and should benchmark at least one feasible feedback-only smoother or windowed controller. These are concrete *design and evaluation requirements*, not demonstrations of improved comprehension, motivation, retention, safety performance, or fairness among learners. A real educational intervention requires validation of the question bank, learner-model assumptions, target probability, and instructional outcomes before deployment.

Limitations, reproducibility, and next steps

External validity remains limited to fixed-proficiency, nonlearning Bernoulli agents and a specified ten-level logistic difficulty model. Engineering thresholds C1–C5 are transparent but not educationally validated minimal important differences. Window-20 and R22 were selected in a separately frozen exploratory protocol following an internal mock review; the original study's confirmatory protocol, hypothesis family, and failure decision were not changed. The supplementary comparisons were not preregistered confirmatory tests, and their nominal confidence intervals do not account for exploration of additional potential controller families. The reconstruction-derived aggregate checks and 900 original logistic DDA trajectory replays are strong computational consistency tests but do not establish that every row of the original official three-million-interaction raw export

has been independently reconciled. A release-level source commit was unavailable in the supplied metadata; likewise, live public availability and bitwise equality of external deposits were not independently verified in this analysis. Independent replication should therefore begin with the original deposited exports and verify checksums, code release, and raw-log-to-summary concordance (Leipzig et al., 2021; Morris et al., 2019). Follow-on educational studies should compare feedback-only adaptive alternatives, participant learning and fairness outcomes, and empirically calibrated difficulty targets before offering student-facing deployment recommendations.

CONCLUSION

The frozen RES-DDA-001 v1.1 simulation showed that DDA-R32 entered the designated target zone more often than Fixed-5 for all three synthetic profiles, while the continuous MAE benefit was present for Expert and Novice but not Intermediate. Because the comparator had zero occupancy by construction, and a minor tolerance change reversed Intermediate's occupancy contrast, this finding cannot establish general controller superiority. More decisively, a separately frozen supplementary exploratory benchmark identified a feasible feedback-only Window-20 policy with lower MAE than DDA-R32 in all nine profile-slope combinations. Independent 40-state Markov calculations corroborated the original controller's modeled dynamics, distinguishing request-balance equilibrium from its stationary success behavior; longer-horizon simulations suggested broadly stable model-conditional patterns. The study's key contribution is therefore a reproducible method for evaluating metric dependence and profile-specific limitations of a simple streak controller, rather than evidence that DDA-R32 is optimal for game-based learning. No student learning was measured. Future research should independently validate original raw data provenance, compare further feasible controllers under educationally justified targets, and evaluate effectiveness and fairness with actual learners before instructional deployment.

DECLARATIONS

Funding

This research received no external funding.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Ethics Approval and Informed Consent

Not applicable. This study exclusively simulated synthetic Bernoulli agents and did not recruit human participants, involve animals, or process identifiable personal data.

Data Availability Statement

The original synthetic simulation dataset and available associated statistical outputs are referenced by the author-deposited Zenodo record <https://doi.org/10.5281/zenodo.23340479>. Simulator source and reproducibility materials are referenced by the project GitHub repository <https://github.com/dcypranoto/sbe-dda>. Separately frozen exploratory benchmark summaries, tolerance curves, Markov validation outputs, and analysis scripts accompany this manuscript's supplementary package. Public-repository availability and consistency of the supplementary files have not been independently verified, and the original Windows raw exports were not independently

byte-matched during the supplementary analysis. All observations are synthetic; no student or human-participant data were collected.

AI Use Disclosure

ChatGPT (OpenAI) assisted with manuscript drafting, language editing, proofreading, and the preparation of computational analysis scripts and graphics. The authors reviewed the AI-assisted outputs and retain full responsibility for the scientific analyses, calculations, sources, interpretations, and submitted content.

REFERENCES

- Ajani, O. S., & Mallipeddi, R. (2023). Pareto-based dynamic difficulty adjustment of a competitive exergame for arm rehabilitation. *International Journal of Human-Computer Studies*, 178, Article 103100. <https://doi.org/10.1016/j.ijhcs.2023.103100>
- Arevalillo-Herráez, M., Katsigiannis, S., Alqahtani, F., & Arnau-González, P. (2023). Fusing ECG signals and IRT models for task difficulty prediction in computerised educational systems. *Knowledge-Based Systems*, 280, Article 111052. <https://doi.org/10.1016/j.knosys.2023.111052>
- Aşkın, O. (2025). Designing a large language model-based AI system for dynamic difficulty adjustment in digital games. *Acta Infologica*, 9(1), 183–207. <https://doi.org/10.26650/acin.1670469>
- Blom, P. M., Methorst, S., Bakkes, S., & Spronck, P. (2019). Modeling and adjusting in-game difficulty based on facial expression analysis. *Entertainment Computing*, 31, Article 100307. <https://doi.org/10.1016/j.entcom.2019.100307>
- Chrysafiadi, K., Kamitsios, M., & Virvou, M. (2023). Fuzzy-based dynamic difficulty adjustment of an educational 3D-game. *Multimedia Tools and Applications*, 82, 27525–27549. <https://doi.org/10.1007/s11042-023-14515-w>
- Climont, L., Longhi, A., Arbelaez, A., & Mancini, M. (2024). A framework for designing reinforcement learning agents with dynamic difficulty adjustment in single-player action video games. *Entertainment Computing*, 50, Article 100686. <https://doi.org/10.1016/j.entcom.2024.100686>
- Cole, S. R., Edwards, J. K., Naimi, A. I., & Muñoz, A. (2020). Hidden imputations and the Kaplan-Meier estimator. *American Journal of Epidemiology*, 189(11), 1408–1411. <https://doi.org/10.1093/aje/kwaa086>
- Darzi, A., McCrear, S. M., & Novak, D. (2021). User experience with dynamic difficulty adjustment methods for an affective exergame: Comparative laboratory-based study. *JMIR Serious Games*, 9(2), Article e25771. <https://doi.org/10.2196/25771>
- Grosu, M.-A., Nicola, S., & Ciocârlie, H. (2026). Dynamic difficulty adjustment in a VR game: Comparing fixed progression with two adaptive progression strategies. *Electronics*, 15(17), Article 3929. <https://doi.org/10.3390/electronics15173929>
- Guo, Z., Thawonmas, R., & Ren, X. (2024). Rethinking dynamic difficulty adjustment for video game design. *Entertainment Computing*, 50, Article 100663. <https://doi.org/10.1016/j.entcom.2024.100663>
- Guzmán, D. E., Rengifo, C. F., Guzmán, J. D., & Garcia Cena, C. E. (2024). Virtual reality games for cognitive rehabilitation of older adults: A review of adaptive games, domains and techniques. *Virtual Reality*, 28, Article 92. <https://doi.org/10.1007/s10055-024-00968-3>
- Kristensen, J. T., & Burelli, P. (2024). Difficulty modelling in mobile puzzle games: An empirical study on different methods to combine player analytics and simulated data. *International Journal of Computer Games Technology*, 2024, Article 5592373. <https://doi.org/10.1155/2024/5592373>
- Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. *Patterns*, 2(9), Article 100322. <https://doi.org/10.1016/j.patter.2021.100322>
- Lima, R., Branco, D., Lobo, P., & Bermúdez i Badia, S. (2026). Flow optimizer framework: Validation of a dynamic difficulty adjustment system for serious games. *Games for Health Journal*. <https://doi.org/10.1177/2161783X251414444>
- Mi, Q., & Gao, T. (2022). Adaptive rubber-banding system of dynamic difficulty adjustment in racing games. *ICGA Journal*, 44(1), 18–38. <https://doi.org/10.3233/ICG-220207>

- Moon, J.-Y., Choi, Y.-J., Park, T.-H., Choi, J.-D., Hong, J.-H., & Kim, K.-J. (2022). Diversifying dynamic difficulty adjustment agent by integrating player state models into Monte-Carlo tree search. *Expert Systems with Applications*, 205, Article 117677. <https://doi.org/10.1016/j.eswa.2022.117677>
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. <https://doi.org/10.1002/sim.8086>
- Mortazavi, F., Moradi, H., & Vahabie, A.-H. (2024). Dynamic difficulty adjustment approaches in video games: A systematic literature review. *Multimedia Tools and Applications*, 83(35), 83227–83274. <https://doi.org/10.1007/s11042-024-18768-x>
- Paraschos, P. D., & Koulouriotis, D. E. (2023). Game difficulty adaptation and experience personalization: A literature review. *International Journal of Human-Computer Interaction*, 39(1), 1–22. <https://doi.org/10.1080/10447318.2021.2020008>
- Rjiba, A., Cheniti-Belcadhi, L., & Kasperuniene, J. (2025). Adaptive difficulty and stealth assessment in collaborative game-based learning. *Computer Applications in Engineering Education*, 33(6), Article e70102. <https://doi.org/10.1002/cae.70102>
- Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2021). The percentile bootstrap: A primer with step-by-step instructions in R. *Advances in Methods and Practices in Psychological Science*, 4(1), 1–10. <https://doi.org/10.1177/2515245920911881>
- Shohieb, S. M., Doeniyas, C., & Elhady, A. M. (2022). Dynamic difficulty adjustment technique-based mobile vocabulary learning game for children with autism spectrum disorder. *Entertainment Computing*, 42, Article 100495. <https://doi.org/10.1016/j.entcom.2022.100495>
- Souza, C. H. R., Nascimento, D. F., Berretta, L. O., & Carvalho, S. T. (2025). Solutions for dynamic difficulty adjustment in digital games: A systematic literature review. *Entertainment Computing*, 55, Article 101041. <https://doi.org/10.1016/j.entcom.2025.101041>
- Vickerstaff, V., Omar, R. Z., & Ambler, G. (2019). Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC Medical Research Methodology*, 19, Article 129. <https://doi.org/10.1186/s12874-019-0754-4>
- Víteková, L., Eichhorn, C., Pirker, J., & Plecher, D. A. (2026). Dynamic difficulty adjustment in serious games: A literature review. *Information*, 17(1), Article 96. <https://doi.org/10.3390/info17010096>