

The Price of Fairness

Qualification Reliability in Agentic Poker Tournaments

Henry De Courcy Thompson, FullHouse Research

ABSTRACT

Agentic poker competitions select finalists from noisy outcomes under finite simulation budgets. We study top-six qualification in 60-agent, six-player no-limit hold'em Swiss tournaments using three controlled fields and two mechanism-specific reference estimands. At the baseline of 40 rounds with 400 hands per match, independent reruns change 3.83 to 4.38 of six qualifiers on average. Even in an engineered separated control, six independently validated leaders qualify individually with probabilities of only 0.31 to 0.76. Selection improves at every observed checkpoint from five to 40 rounds, without an observed plateau. At an equal scheduled hand budget, ten 1,600-hand rounds produce 0.42 to 0.87 more changed qualifiers than forty 400-hand rounds. Adaptive Swiss is more consistent than fresh-randomness replay of its realised schedules, although this comparison does not identify adaptation's causal effect. A short-match duplicate package improves ranking consistency across all three fields; improvements in changed qualifiers and regret are individually clear only in the separated control. Most wrong qualifiers lie outside the reference-ambiguous set, and reference uncertainty accounts for little of the measured strength loss. These results favour distributing scheduled simulation across more rounds in the tested setting, while leaving the budget required for stable qualification unresolved.

Introduction

Poker is a notoriously noisy and high-variance game. This is what attracts gamblers and thrill seekers all over the world. However, when evaluating poker agents this gives us a nasty headache; declaring a winner confidently requires a huge amount of simulations. We are therefore motivated, given a limited budget, to optimise our resources and find methods of ranking agents as repeatably and fairly as possible. We examine Swiss tournaments, duplicate play, extending horizons and more over different reference landscapes in order to suggest frameworks for fair and robust qualification.

Related work

Tournament design determines how a limited comparison budget is used. Sziklai, Biró and Csató [1] compare tournament formats under noisy pairwise outcomes and identify repeated comparisons as a possible source of wasted budget. Their result motivates our allocation experiment: should a poker tournament spend its hand allowance on more opponent configurations or on longer matches? Six-player poker adds table-composition effects, persistent stacks and chip-valued outcomes, so their pairwise conclusions require testing in this setting. Hua [2] evaluates partial and complete ranking recovery in simulated Swiss tournaments. Top-k criteria are particularly relevant to a competition that advances only a small part of the field: whole-ranking correlation can conceal errors at the cutoff.

Computer poker has long addressed evaluation variance. The Annual Computer Poker Competition used duplicate poker, including cyclic duplication in multiplayer games [3]. Baseline control variates [4] and AIVAT [5] improve statistical efficiency by removing components of stochastic outcome variance. These methods motivate evaluating how much information a fixed simulation allowance yields. Our outcome is the qualification set produced after noisy results have entered standings and changed future opponents.

The contribution is an empirical study of top-six selection in adaptive, six-player poker tournaments. We compare round allocation, schedule replay and whole-match duplicate packages. Swiss tournaments, duplicate poker, common random numbers and top-k metrics are established methods.

Setting and experimental design

Tournament and agents

Each tournament contains 60 agents; the six highest final standings qualify. The baseline comprises 40 Swiss rounds, each partitioning the field into six-player no-limit hold'em matches with a maximum of 400 hands. Agents begin each match with 100 big blinds (BB). Stacks persist within the match, and busted agents remain absent until the next Swiss round, when stacks reset. Standings sum raw chip deltas across rounds, with the implemented public tiebreaks.

These are research reference semantics derived from the public 2026 engine. They do not establish historical production settings or specify a future FullHouse format. The engine source hash is 4647ee69857e7c9b throughout the frozen study.

The fields draw from 96 synthetic rule-based agents spanning eight archetypes and parameter perturbations. They provide controlled strategic variation. Their weak stack adaptation and limited use of available within-match history and opponent information restrict transfer to sophisticated adaptive agents.

Three fixed fields

The separated control deliberately creates a clear qualification boundary. Six leaders and 54 weaker agents were chosen from pool-calibration data; the candidate was bound to a one-shot validation protocol before validation began. In 4,000 fresh pool-style rounds, the bootstrap probability that the intended leaders formed the top six was 0.998, and the one-sided 95% lower bound on their boundary gap was positive. This validation took place in the 96-agent pool context. A subsequent independent field-relative reference confirmed separation in the final 60-agent field.

The crowded-cutoff field contains three designed leaders and a near-equal crowd around the remaining qualifying positions. The style-diverse field preserves broad strategic coverage across strength levels. Each scenario contributes one frozen field; tournament replications therefore estimate reliability conditional on that field.

Table 1. Controlled fields and final field-reference precision. Gaps are BB per ordinary-equivalent 400-hand match. Top-six stability is the bootstrap probability of retaining the point-estimate reference set, not tournament qualification probability.

Field	Experimental role	Matched reference rounds	Rank-6/7 gap	Reference top-six stability
Separated control	Engineered clear boundary	250	23.83	0.999
Crowded cutoff	Compressed qualification boundary	500	0.61	0.205
Style diverse	Broad style and strength coverage	500	2.02	0.366

The separated control replaced an original spread field whose apparent separation failed independent reference evaluation. A split-sample replacement also failed held-out validation. Both failures remain in the provenance record; neither field enters the publication matrix. Appendix A describes the replacement decisions.

Reference estimands and metrics

The static reference, μ_{field} , is expected performance against random partitions of the fixed field under the specified match semantics. We estimate it using six-lane whole-match duplication with cyclic seating and common chance streams. Scores are normalised to one ordinary-equivalent 400-hand match. Uncertainty resamples complete matched rounds jointly across agents, preserving dependence within a round.

The Swiss reference, μ_{swiss} , is expected final performance under repeated operation of the baseline adaptive mechanism, estimated from 200 complete tournaments per field. It includes the opponent distribution generated by standings. These estimands can disagree in multiplayer games; neither defines context-free poker skill. μ_{field} is the primary strength reference. Agreement with μ_{swiss} measures convergence to that tournament mechanism rather than normative fairness.

For two independent qualifier sets Q_1 and Q_2 , changed qualifiers equals $6 - |Q_1 \cap Q_2|$. It counts replaced qualifying places, ranges from zero to six, and is half the symmetric difference. Top-six overlap is its complement. Wrong qualifiers relative to a reference counts selected agents outside that reference's point-estimate top six.

Qualification regret measures strength loss:

$$\text{Regret}(Q; \mu) = \sum_{i \in Q^*} \mu_i - \sum_{i \in Q} \mu_i,$$

where Q^* is the reference-optimal six. μ_{field} regret is the sum of six agents' reference match strengths, in BB per ordinary-equivalent match. It is not chips lost during the tournament. Compare regret across designs within a field; its scale depends on field strength dispersion, so raw regret does not rank the difficulty of different fields.

Supporting measures include per-agent $P(\text{qualify})$, the number of agents with $P(\text{qualify})$ between 0.10 and 0.90, Spearman and Kendall correlations, rank displacement, boundary strength loss and inversion diagnostics. Reference ambiguity is distinct: an agent is reference-ambiguous when its bootstrap probability of reference top-six membership lies strictly between 0.05 and 0.95. Clear misqualifications exclude swaps within this reference-ambiguous set.

Experiments and work accounting

Experiment A evaluates prefixes at five, ten, 20, 30 and 40 rounds of the same baseline trajectories. Prefix observations are dependent and add no simulation.

Experiment B compares $R10 \times 1600$, $R20 \times 800$ and $R40 \times 400$, where R denotes rounds and the second number the hand cap per match. All allocate 16,000 scheduled hands per agent, or 160,000 table hands per tournament. They have equal scheduled maximum hand budgets, while played hands differ because matches can end early. They do not establish equal compute cost.

Experiment C replays each baseline tournament's realised table schedule once under fresh poker randomness, without feeding replay results into subsequent pairings. Adaptive and frozen outcomes are paired by source tournament. The frozen schedule inherits information from the adaptive source's outcomes; this is a source-schedule counterfactual, not a separately generated nonadaptive schedule or a clean causal intervention on feedback.

Experiment D compares ordinary $R40 \times 400$ with $R7 \times 400 \times 6$ and $R40 \times 67 \times 6$. Six duplicate lanes rotate seats cyclically and share chance streams where meaningful, with separate stacks, agent state and decision randomness. Bots cannot observe lane identity. Aggregation occurs after complete lanes. The packages schedule respectively 1.050 and 1.005 times baseline work and play approximately 5% to 7% more hands. They change round count or match length alongside duplication, so this is an operational package comparison.

Replication and statistical protocol

Six designs across three fields yield 18 cells, each with 50 replications: 900 matrix tournaments. The μ_{swiss} sample adds 200 independent tournaments per field. Final μ_{field} references use 250 matched rounds for the separated control and 500 for each other field. The crowded reference still triggered its extension criterion at the 500-round ceiling; its residual uncertainty is retained.

Reported intervals are 95% tournament-bootstrap intervals for levels, paired-bootstrap intervals for contrasts, and Wilson intervals for qualification proportions. Pairing uses shared replicate identifiers, or source tournaments in C. Additional intervals propagate joint μ_{field} reference uncertainty. All contrasts are first design minus second; negative error contrasts favour the first design.

The matrix extension rule allowed 50 to 100 replications if an identity-metric interval had half-width above 0.30, or a pair-specific rerun changed-qualifier contrast interval spanned both -0.25 and $+0.25$. Regret did not trigger extension. After the first matrix audit, the contrast clause was clarified to govern rerun changed qualifiers. The original broader interpretation also included wrong-vs-reference contrasts and returned four EXTEND_TO_100 verdicts. The clarified interpretation returned 18/18 LOCK_AT_50. Appendix C preserves the four-cell sensitivity; those extensions were not run.

The simulation programme is frozen. The original Notion plan allowed longer horizons if needed, but the final protocol retained five to 40 rounds. No 60-, 80- or 120-round results exist. This post-results horizon decision limits the conclusions to the observed range.

Results

Forty rounds leave substantial qualification instability

Independent baseline reruns change 3.83, 4.32 and 4.38 of six qualifying places in the separated-control, crowded-cutoff and style-diverse fields respectively (Table 2). Mean overlap is therefore only 2.17, 1.68 and 1.62. The 200-tournament Swiss-reference samples corroborate this instability, with changed-qualifier estimates of 3.82, 4.19 and 4.43. These are separate samples, not alternative estimates from the same matrix cell.

Table 2. Baseline R40×400 reliability, 50 tournaments per field. Brackets give 95% tournament-bootstrap intervals. Regret uses μ_{field} .

Measure	Separated control	Crowded cutoff	Style diverse
Changed qualifiers	3.83 [3.59, 4.07]	4.32 [4.12, 4.50]	4.38 [4.18, 4.56]
Wrong qualifiers	2.74 [2.50, 3.00]	3.52 [3.26, 3.80]	3.80 [3.58, 4.08]
Regret, BB	122.9 [110.3, 137.9]	93.7 [81.2, 105.3]	94.3 [85.5, 103.7]
Rerun Spearman	0.43 [0.40, 0.46]	0.37 [0.34, 0.40]	0.43 [0.41, 0.45]
Qualification-ambiguous agents	19 [16, 22]	25 [20, 27]	24 [20, 26]

Instability persists in the separated control despite its independently confirmed reference gap. In the 200-tournament sample, the six validated leaders' individual qualification probabilities range from 0.31 to 0.76 (Figure 5). Across the 50 baseline and 200 Swiss-reference tournaments, the exact reference top-six set occurs in only two of 250 outcomes. This pooled count describes those baseline-format outcomes; it does not use all 900 matrix tournaments as its denominator.

Reference uncertainty has much smaller strength cost. Reference self-regret evaluates the point reference's fixed top-six set under each resampled reference, relative to that resample's optimal six. Its bootstrap median [2.5th, 97.5th percentile] is 0.0 [0.0, 0.0], 2.7 [0.0, 10.7] and 1.5 [0.0, 9.8] BB. These distributions do not establish exact reference identity in the crowded fields, but their losses are small alongside baseline

tournament regret of 93.7 to 122.9 BB. Propagating reference uncertainty leaves the baseline conclusion intact.

Selection improves across the observed round horizon

Every observed checkpoint improves the reported identity and strength measures in all three fields (Figure 1). From five to 40 rounds, wrong qualifiers fall from 4.18 to 2.74 in the separated control, 4.58 to 3.52 in crowded cutoff and 4.56 to 3.80 in style diverse. Regret falls from 224.8 to 122.9, 168.6 to 93.7 and 161.0 to 94.3 BB respectively. Rerun changed qualifiers also decline, from 5.04 to 3.83, 5.13 to 4.32 and 5.12 to 4.38.

Point estimates continue improving between 30 and 40 rounds, with no observed plateau within the preregistered 40-round horizon. Adjacent checkpoints share trajectories and were not tested separately; monotone point estimates do not imply a statistically established gain at every increment. The results do not identify how many rounds would reach a chosen reliability target.

More rounds outperform ten longer matches

At equal scheduled hand budget, R10×1600 is worse than R40×400 on both changed qualifiers and regret in every field (Figure 2; Table 3). Changed qualifiers increase by 0.87 [0.60, 1.14], 0.42 [0.12, 0.69] and 0.52 [0.25, 0.79]. Regret increases by 65.7 [47.2, 84.9], 29.6 [12.8, 46.2] and 41.7 [27.0, 57.1] BB.

R20×800 is closer to baseline. Its changed-qualifier contrasts are 0.17 [−0.18, 0.51], −0.07 [−0.39, 0.24] and 0.34 [0.05, 0.62]. Only style diverse excludes zero; all three regret intervals include zero. These intervals do not demonstrate equivalence.

Played-hand fractions are 54.2% to 59.0% for R10×1600, 78.3% to 81.2% for R20×800, and 94.1% to 94.6% for R40×400. A post hoc comparison matches R10×1600 against baseline prefixes with similar played hands. Its deficit persists in the separated control on changed qualifiers and regret, and in style diverse on regret and wrong qualifiers. Crowded cutoff is inconclusive. Early termination therefore does not fully account for the allocation result in every field, but the evidence cannot attribute the remaining difference solely to opponent breadth.

Table 3. Primary changed-qualifier contrasts. Brackets give paired-bootstrap 95% intervals. Negative values favour the first design. The short-duplicate crowded-cutoff upper bound is +0.0009 before rounding and includes zero.

First design minus baseline or comparator	Separated control	Crowded cutoff	Style diverse
R10×1600 – R40×400	0.87 [0.60, 1.14]	0.42 [0.12, 0.69]	0.52 [0.25, 0.79]
R20×800 – R40×400	0.17 [−0.18, 0.51]	−0.07 [−0.39, 0.24]	0.34 [0.05, 0.62]
Adaptive – frozen replay	−0.80 [−1.06, −0.56]	−0.38 [−0.65, −0.11]	−0.33 [−0.59, −0.09]
R40×67×6 – ordinary	−0.45 [−0.79, −0.13]	−0.31 [−0.62, +0.0009]	−0.17 [−0.46, 0.09]
R7×400×6 – ordinary	−0.12 [−0.46, 0.21]	−0.49 [−0.78, −0.20]	0.14 [−0.15, 0.42]

Adaptive Swiss is more consistent than frozen replay

Adaptive minus frozen changed qualifiers is negative in every field, with all three intervals excluding zero (Figure 3; Table 3). Rerun rank correlations also favour adaptation. μ_{field} regret contrasts are -54.2 $[-74.3, -36.5]$ BB in the separated control and -21.2 $[-40.4, -1.6]$ BB in crowded cutoff. In style diverse, the contrast is -11.6 $[-24.7, 1.1]$ BB and includes zero, as does the wrong-qualifier contrast.

The measured result is better repeatability under adaptive Swiss than fresh play on schedules generated by its source runs. Because those schedules depend on source outcomes, the comparison does not isolate feedback's causal contribution. In particular, it cannot establish that adaptation improves strength recovery in all three fields merely because it improves rerun consistency in all three.

Short duplicate matches improve consistency more uniformly

R40×67×6 increases rerun Spearman from 0.43 to 0.54, 0.37 to 0.51 and 0.43 to 0.49 across the three fields (Figure 4). Changed qualifiers and regret have favourable point contrasts in every field. Only the separated control excludes zero on both measures: -0.45 $[-0.79, -0.13]$ qualifiers and -17.3 $[-33.0, -3.6]$ BB. Crowded-cutoff regret is -16.6 $[-33.8, 1.3]$ BB; style-diverse regret is -8.2 $[-20.1, 2.6]$ BB. The crowded changed-qualifier interval includes zero despite rounding to an upper bound of 0.00.

The R7×400×6 package has less uniform results. It improves changed qualifiers in crowded cutoff by -0.49 $[-0.78, -0.20]$, while the other fields' intervals include zero. Short duplication is consequently more consistent in direction across fields, without dominating the long-lane package in every field or metric.

Reference-calibration diagnostics suggest a possible explanation. Roughly 60% of bot-lanes bust during 400-hand persistent-stack matches, whereas busts are rare in 67-hand lanes. Changes in stacks and participation weaken alignment between duplicate trajectories. Shorter lanes may preserve more useful cancellation from shared chance streams. This mechanism was not directly tested. Figure 4's Spearman panel was selected after outcomes were seen, so the primary qualification contrasts remain alongside it.

Most identity errors have reference strength cost

Among wrong baseline qualifiers, the share outside the reference-ambiguous set is 100% in the separated control, approximately 80% in crowded cutoff and 79% in style diverse (Figure S1). Thus, uncertainty among near-equal reference contenders does not account for most admissions outside the point top six. At tournament level, wrong-qualifier count correlates with regret at approximately 0.90, 0.79 and 0.79 respectively.

The two reference rankings also agree closely: μ_{field} versus μ_{swiss} Spearman is 0.978, 0.970 and 0.980. Their top-six sets overlap by six, six and five agents, with the style-diverse disagreement inside the ambiguous boundary. The reported pairwise uncertainty diagnostic finds no significantly discordant ordering; this does not establish identical estimands. Together with low reference self-regret and reference-propagated intervals, these checks support substantial tournament error under either reference view.

Discussion and operational implications

Qualification requires evaluating both repeatability and reference strength loss. A stable set could still be systematically wrong, while frequent swaps among near-equals could have little strength cost. In these experiments, both concerns are material: reruns replace several qualifying places, and most wrong admissions fall outside the reference-ambiguous set.

The clearest allocation decision is to favour forty 400-hand rounds over ten 1,600-hand rounds when the constraint is scheduled hands. Additional rounds offer new table compositions and opportunities to update standings, while extending a match repeatedly samples its current opponent set. This interpretation is plausible, but Experiment B changes exposure, feedback opportunities and early-termination behaviour together. It does not identify their separate contributions.

Adaptive pairing contributes useful consistency in the tested replay comparison. Duplicate play needs a more conditional recommendation: the short-lane package improves whole-ranking consistency across fields, but qualification gains remain uncertain in two fields and its actual played work is higher. Organisers should assess duplication using selection outcomes and measured work in their own setting.

Operational decision table. These recommendations apply to the tested research setting; they do not specify future competition rules.

Decision	Evidence-based recommendation	Remaining limit
Minimum stable round count	Do not treat 40×400 qualification as stable in these fields	No reliability target or horizon beyond 40 was evaluated
Longer matches or more rounds	Prefer 40×400 over 10×1600 at equal scheduled hands	R20×800 is closer; equal compute was not tested
Retain adaptive pairing	Adaptive Swiss is more consistent than its frozen replay	This counterfactual does not isolate a causal effect
Use duplicate play	Short lanes are a candidate for improved consistency	Qualification gains and overhead require field-specific assessment
Allocate extra budget near a crowded cutoff	Distinguish tournament uncertainty from reference boundary ambiguity	No universal extra-round budget is supported

Limitations

These results are conditional on three fixed fields built from synthetic heuristic agents. They do not estimate how independently sampled competition fields would behave, and limited opponent modelling and stack adaptation restrict external validity. The separated control is deliberately engineered and should be read as a sanity check, not as a typical field. μ_{field} and μ_{swiss} are finite, mechanism-specific references; exact boundary identity remains uncertain in the two harder fields, and qualification regret is only comparable within a field because strength spreads differ.

Experiments C and D also answer narrower questions than their labels might suggest. The frozen-schedule comparison replays schedules generated by adaptive Swiss, while the duplicate packages change match structure as well as duplication. Runtime data are too affected by sleep and throttling for strong CPU-efficiency claims. Finally, one matrix-extension clause was clarified after the initial audit: the broad reading

would have extended four cells, while the final interpretation locked all 18 at 50; both are reported. The study stops at the preregistered 40-round horizon, so it does not identify a budget that guarantees stable qualification.

Conclusion

In these controlled six-player poker fields, forty 400-hand Swiss rounds leave substantial uncertainty about top-six qualification, including in an independently validated separated control. Forty shorter matches outperform ten longer matches at equal scheduled hand budget. Adaptive pairing improves consistency relative to frozen replay, and short duplicate packages offer additional, qualified gains. The study supports these allocation choices while leaving the simulation budget required for stable qualification unanswered, but certainly large.

References

- [1] Sziklai, B. R., Biró, P. and Csató, L. (2022). The efficacy of tournament designs. *Computers & Operations Research*, 144, 105821. **Publisher record.**
- [2] Hua, C. (2017). The Swiss Tournament Model. Wharton Research Scholars undergraduate thesis, University of Pennsylvania. **Institutional manuscript.**
- [3] Bard, N., Hawkin, J., Rubin, J. and Zinkevich, M. (2013). The Annual Computer Poker Competition. *AI Magazine*, 34(2), 112-114. **Publisher record.**
- [4] Davidson, M., Archibald, C. and Bowling, M. (2013). Baseline: Practical Control Variates for Agent Evaluation in Zero-Sum Domains. *AAMAS*. **Author publication record.**
- [5] Burch et al. (2018). AIVAT: A New Variance Reduction Technique for Agent Evaluation in Imperfect Information Games. *AAAI*. **Publisher record.**

Appendix A. Field-construction provenance

The original spread field was selected for an apparent pool-calibration rank-6/7 gap of 16.8 BB. Independent field-relative evaluation found only 0.71 BB. Selecting on noisy estimates caused regression to the mean; the field was retired with its evidence preserved. A split-sample spread_v2 candidate, selected on even pool rounds and validated once on odd rounds, failed: P(leaders form the top six) was 0.208 against the required 0.95. It was not retried.

The separated-control replacement was a post hoc design decision, followed by preregistration of its own construction and one-shot validation before new validation data. The six intended leaders passed fresh pool-style validation: their boundary gap was 12.36 BB with one-sided 95% lower bound 4.38 BB; the largest non-leader reference top-six probability was 0.0015. Independent final field-relative estimation then confirmed a 23.83 BB gap. Validation-context and final-field gaps are different quantities and should not be conflated.

Appendix B. References and frozen evidence

Table 4. Final reference checks. Self-regret reports the bootstrap median and percentile interval, not a mean with that interval.

Check	Separated control	Crowded cutoff	Style diverse
μ_{field} matched rounds	250	500	500
μ_{swiss} tournaments	200	200	200
Reference rank correlation	0.978	0.970	0.980
Reference top-six overlap	6	6	5
μ_{field} self-regret, BB	0.0 [0.0, 0.0]	2.7 [0.0, 10.7]	1.5 [0.0, 9.8]

Reference self-regret uses 1,000 joint resamples. The reporting cleanup changed the displayed statistic from mean to median while preserving the resamples and frozen analysis. For the separated control, 999 of 1,000 resamples have zero regret; a displayed mean could therefore sit outside a [0, 0] percentile interval. Means remain in the evidence sheet without that interval attached.

The canonical handoff is commit c933747; simulations were frozen at db45b56 and the publication package at e222c9a. The dossier manifest records input checksums. Stored runs passed digest, zero-sum, seating and replay checks. No simulation was run for this draft.

Appendix C. Extension rules and researcher decisions

The separated-control field reference locked at 250 rounds. Crowded cutoff and style diverse extended to the preregistered 500-round ceiling; crowded cutoff still met its extension criterion there. Swiss references extended from 100 to the 200-tournament ceiling in all three fields under their qualification-probability precision rule.

Matrix clause 2 originally received a broad audit that included wrong-vs-reference contrasts and returned four EXTEND_TO_100 verdicts. After outcomes were available, the clause was clarified to use pair-specific rerun

changed-qualifier contrasts, yielding 18/18 LOCK_AT_50. No four-cell extension was simulated. The sensitivity is a stopping-rule sensitivity rather than an alternative set of tournament results.

Other outcome-informed decisions were the field replacement, matched-played-hands descriptive comparison, Spearman panel choice and retention of the 40-round horizon. The failed spread_v2 attempt was not repeated. The frozen package analyses every stored tournament in the matrix and Swiss-reference samples, while retaining retired-field evidence for provenance.

Appendix D. Frozen figures

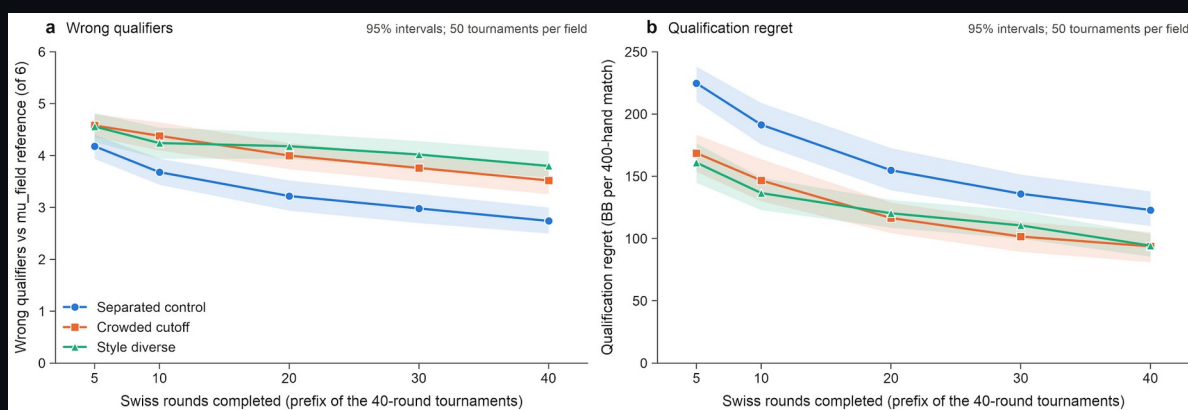


Figure 1. Reliability by round prefix. Identity error and reference strength loss improve at each observed checkpoint in the three fixed fields. Prefixes share baseline trajectories; adjacent gains were not separately tested. The figure supplies no evidence beyond 40 rounds.

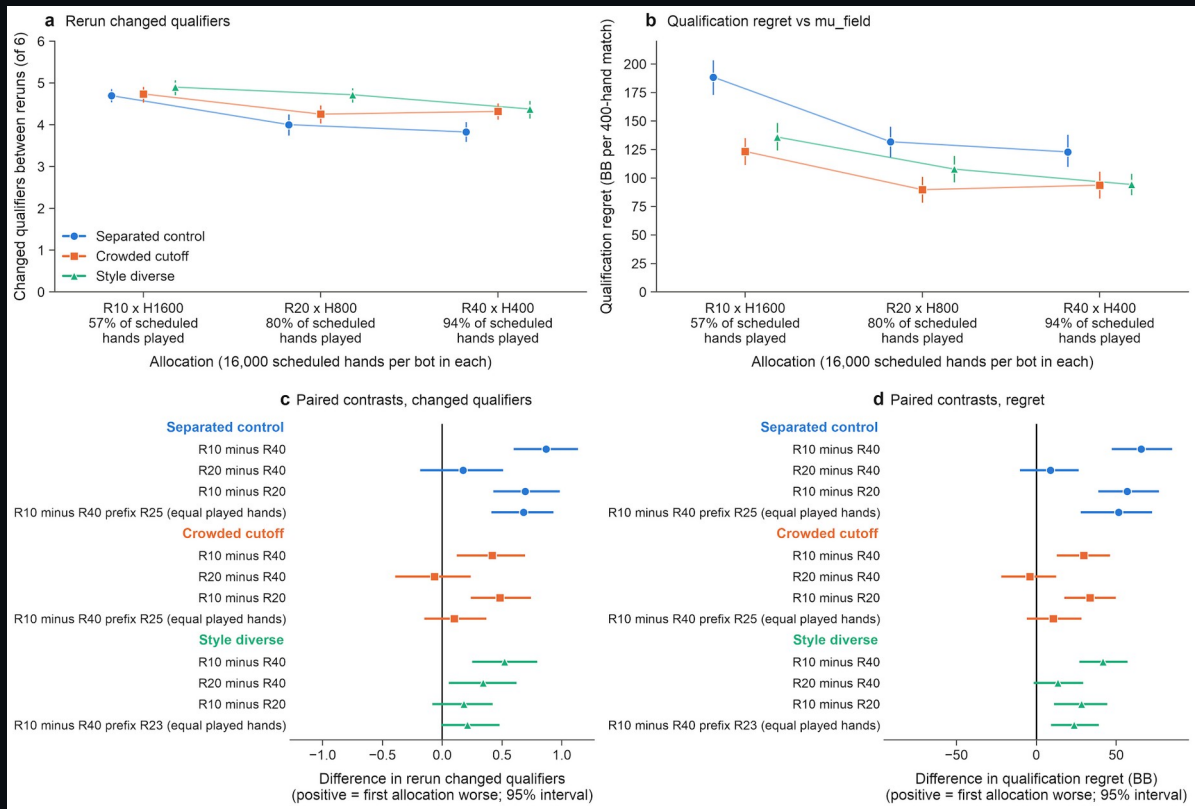


Figure 2. Allocation at equal scheduled hand budget. R10×1600, R20×800 and R40×400 each schedule 16,000 hands per agent. Played work differs through early termination. Matched-played-hands comparisons are post hoc descriptive checks.

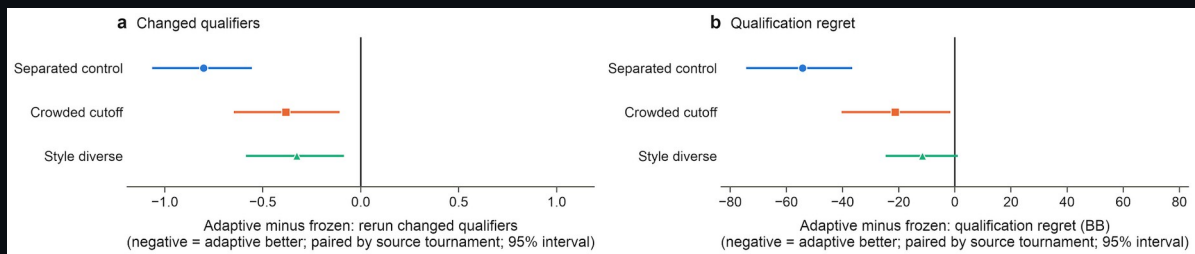


Figure 3. Adaptive Swiss versus frozen replay. Error contrasts are adaptive minus frozen; negative values favour adaptation. Frozen schedules inherit adaptive-source outcomes and are replayed under fresh randomness. Changed-qualifier intervals exclude zero in every field; μ_{field} regret does not in style diverse.

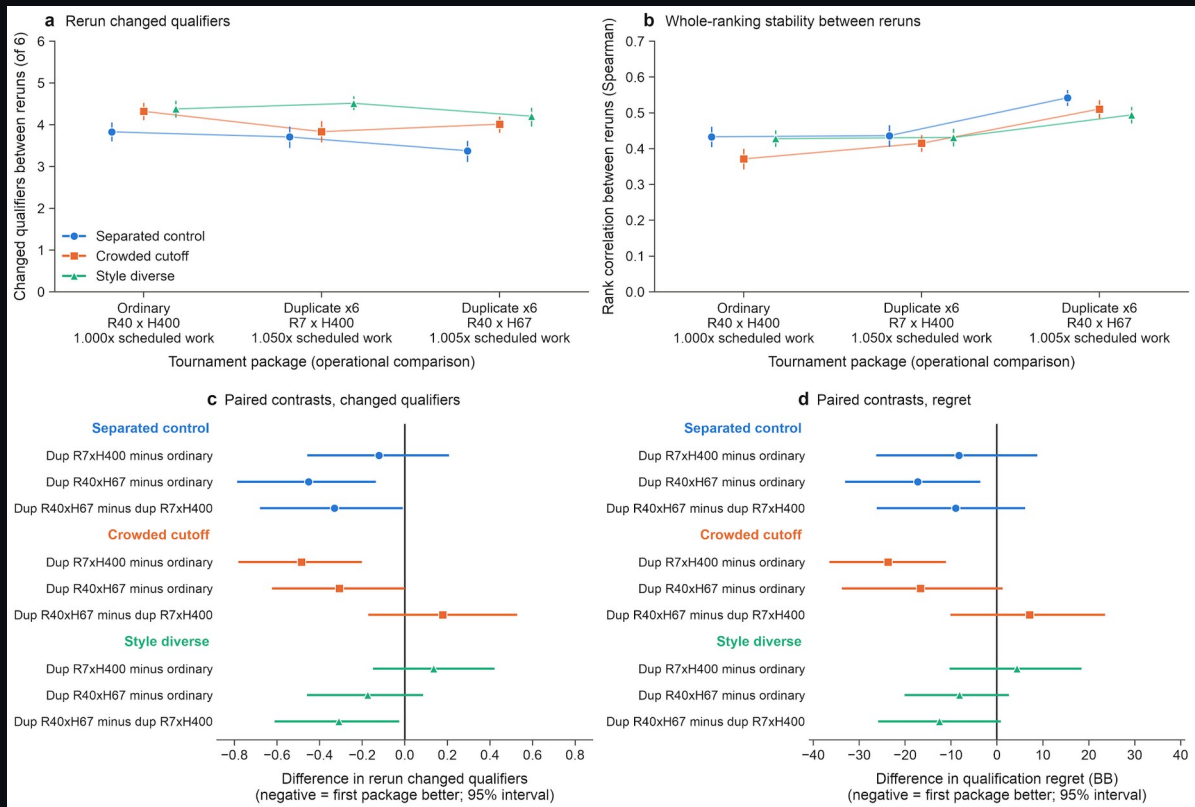


Figure 4. Duplicate tournament packages. R7×400×6 and R40×67×6 are operational comparisons with ordinary R40×400. Both play more hands. The ranking-correlation panel was chosen after outcomes were available. Short-package changed-qualifier intervals include zero in crowded cutoff and style diverse.

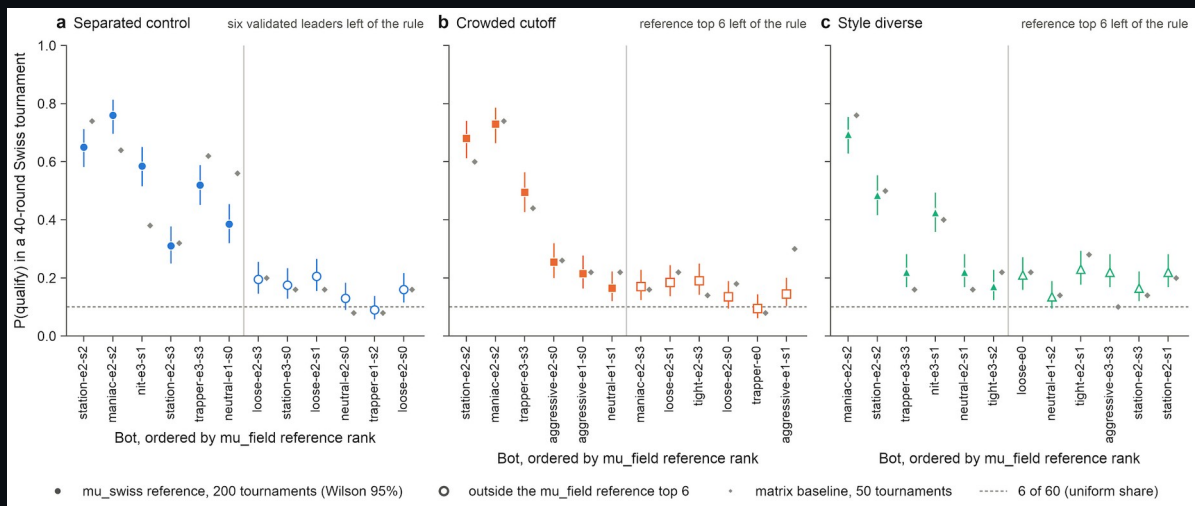


Figure 5. Qualification probability near the cutoff. Probabilities and Wilson intervals use 200 baseline-format tournaments per field. μ_{field} rank and reference membership uncertainty describe a separate random-partition reference. Validated separated-control leaders qualify with individual probabilities of 0.31 to 0.76.

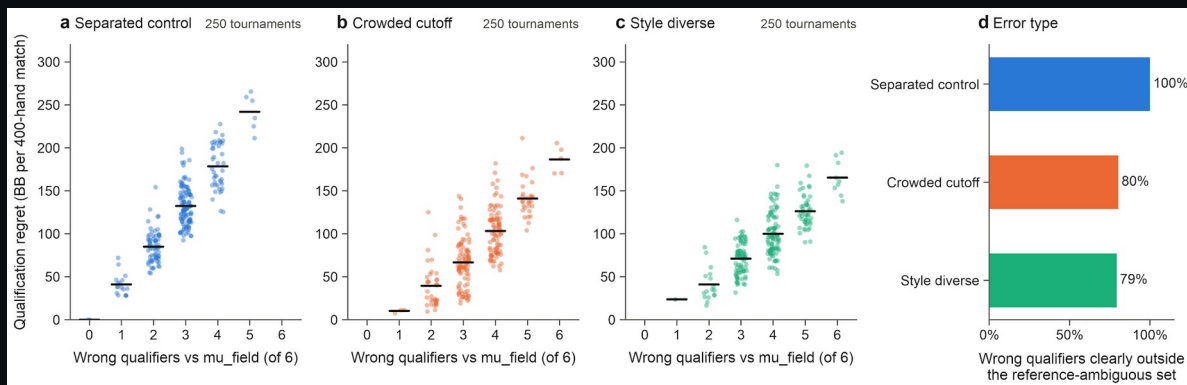


Figure S1. Identity error and strength loss. Baseline wrong admissions mostly fall outside the reference-ambiguous set. Associations between error count and regret are descriptive within each fixed field.

Appendix E. Supplementary data

Full per-agent qualification probabilities are provided in Table A1 in the accompanying frozen dossier.

The accompanying field diversity audit records archetypes, empirical styles and parameter diversity.