Technological Design and Durable Independent Learning

A Critical Narrative Synthesis and Experimental Protocol on Converting Study Effort into Learning

Lluis Eriksson

Independent researcher · Sweden

Research proposal v3 · 1 November 2026

Conceptual article and proposed experimental protocol. No original participant data have been collected, and the proposed trial has not been conducted.

Abstract

Access to technology can improve performance during practice without producing competence that persists when assistance is removed. This article examines a more specific hypothesis than the general superiority of technology: instructional designs that preserve independent attempts while supporting feedback, retrieval and subsequent application may increase durable independent learning within a fixed study period. A critical narrative synthesis of thirteen research publications, supplemented by two contextual sources on El Salvador, brings together research on learning, computer tutoring, generative artificial intelligence and chess. Findings include benefits, null effects and harms. They do not establish a universal technological advantage or isolate technology as the cause of historical improvements in chess. The proposed framework distinguishes the organization of practice from the cognitive operations learners are intended to acquire. It operationalizes an observable attempt, feedback and reapplication cycle without treating cycle frequency as evidence of learning by itself. A proposed four-arm randomized trial compares active print study, fixed digital practice, constrained adaptive digital practice and direct access to solutions, with common core blocks protecting scheduled syllabus coverage. The primary outcome is unaided performance on novel chess positions four weeks after training. Improved assisted performance without a delayed advantage would not establish durable learning. The contribution is a falsifiable research framework and a proposed protocol, rather than a new empirical finding or a registered study.

Keywords: educational technology; deliberate practice; independent learning; retention; near transfer; chess; experimental design.

1 Introduction

A study hour can contain very different activities. A learner may search for a suitable exercise, attempt a solution, locate feedback, interpret an error or try a related problem. The motivating intuition is that a well-designed tool can reorganize this hour so that more of it contributes to developing competence. The scientific task is to translate that intuition into claims that can be compared and potentially refuted.

The expression effort plus technology does not identify a testable intervention. It leaves the tool, instructional design, comparator and outcome unspecified. It also fails to distinguish obtaining a

correct answer from learning how to produce one. A useful investigation must state which operations the intervention facilitates, which operations remain the learner's responsibility and what competence remains after assistance is removed.

Chess offers a bounded setting in which problems have assessable answers, attempts can be recorded and assisted training can be separated from independent assessment. It also exposes a common inferential error. The existence of engines stronger than humans and highly prepared contemporary players does not show that every use of an engine improves learning, or that studying a book necessarily produces less effective practice.

The principal question is whether constrained adaptive instructional design improves durable independent learning relative to active digital practice with a fixed sequence and the same protected core syllabus. A second question compares fixed digital practice with equally active print study. A third examines direct access to solutions. These comparisons separate components that are often bundled when an interactive tool is compared with an inadequately specified alternative.

2 Method of the Narrative Synthesis

Directed literature searches were conducted on 31 October 2026. An initial background text supplied by the study's initiator served as a starting point for locating publications; it was not treated as independent evidence. English search combinations included intelligent tutoring systems, learning outcomes, learning versus performance, cognitive offloading, productive failure, generative AI learning randomized trial, and chess practice learning efficiency. Searches were extended to conceptual references and relevant positive and negative findings available in 2026. A second supplied text prompted the El Salvador case discussion.

Original studies, meta-analyses, publisher pages and institutional or author repositories were prioritized. Selection was purposive, based on relevance to unaided learning, explicit comparators and practice mechanisms. Thirteen research publications form the core synthesis. Two additional sources provide contextual evidence on El Salvador. Three of the core 2026 publications are discussion papers or preprints hosted by IZA or arXiv; peer review is not assumed for those documents.

Relevant sections of accessible full texts were consulted for Ma, Sinha and Kapur, Kestin, Bastani, Strittmatter, Gallegos, Liu, and Contractor and Reyes. Publisher or institutional records and available texts were consulted for the conceptual foundations. The Southwick finding is taken from the publisher's abstract because the full article was not available for inspection. The El Salvador analysis was available as an accepted-article abstract, with the final formatted text pending at the consultation date.

This was not an exhaustive database search, duplicate screening exercise or formal risk-of-bias assessment. No new meta-analysis was performed. Numerical effects belong to the original publications and should not be compared as if they measured the same outcome under the same conditions. The selection supports the construction and criticism of an argument; it cannot estimate the proportion of favorable findings across the entire literature. Source access and major verification decisions are documented in Appendix A.

3 What the Evidence Supports

3.1 Learning and Performance During Acquisition

Soderstrom and Bjork [1] distinguish changes observed during acquisition from learning inferred through retention or transfer. Momentary fluency is therefore an insufficient indicator of lasting competence. Assisted performance is a legitimate outcome, but it is not interchangeable with independent learning.

Dunlosky and colleagues [2] identify practice testing and distributed practice as techniques with high utility across a range of conditions. These findings help identify operations a tool could support. They do not imply that those operations require a screen: a book containing questions and an appropriate review schedule can implement them. A meaningful comparator should therefore preserve a credible instructional design.

Sweller [3] examines how some problem-solving demands can interfere with acquiring the underlying structure of a task. This perspective cautions against both unnecessary overload and the assumption that all difficulty is beneficial. Complementing this concern, Sinha and Kapur [4] synthesize 53 studies with 166 comparisons. Problem solving before instruction favored conceptual knowledge and transfer, with $g = 0.36$, while the effect on procedural knowledge was not significant. The value of an initial attempt depends on the learning objective and the instruction that follows it.

3.2 Computer Tutoring and Interaction Design

Ma and colleagues [5] analyzed 107 effect sizes involving 14,321 participants. Their abstract reports an advantage of intelligent tutoring systems over textbooks or workbooks of $g = 0.35$ and no statistically significant difference from individualized human tutoring. The effect is a standardized difference, not a percentage increase in learning. The synthesis supports comparative benefits under particular conditions; it does not isolate the effect of every system feature or demonstrate long-term retention for every included intervention.

In university physics, Kestin and colleagues [6] studied 194 eligible students and reported better immediate outcomes with an AI tutor than with active-learning classes. The intervention covered two lessons, and the conditions also differed in pacing and setting. This is favorable evidence for a particular instructional package, not an automatic basis for extrapolation to entire courses or competence maintained over months.

Bastani and colleagues [7] studied almost one thousand mathematics students. A general AI interface improved practice performance but produced a relative 17% reduction against the control in a subsequent unassisted examination. A tutor designed to provide instructional safeguards mitigated that harm without producing a significant positive examination effect. Preventing harm and demonstrating superiority are different findings. Pedagogical intentions cannot be treated as evidence that an intervention works.

3.3 Recent Evidence and Competing Interpretations

Gallegos [8], in an IZA discussion paper, distinguishes encouraging adoption of an econometrics assistant from guiding students toward learning-oriented use. The adoption intervention did not improve the midterm examination. The later guidance intervention improved final-examination performance, with the abstract reporting an effect of 0.21 standard deviations. The comparisons

concern different interventions and assessment occasions. Their value is in separating more use from a different form of use.

Liu and colleagues [9] report a randomized trial of access to course-integrated tutoring involving 2,379 students and 30 instructors. The reported grade reduction of 0.37 standard deviations concerns an exact-match subsample of sections within the same course, containing 1,353 students. It is not a significant academic-loss estimate for the full sample. The full text reports no significant academic harm in the overall analysis and reduced participation in existing learning activities. Reading only the abstract could encourage an overly broad interpretation. This source remains a discussion paper.

Contractor and Reyes [10], in a preprint revised in October 2026, compare AI access during a study and writing assignment, followed by unaided assessments immediately and one week later. They report improved test performance and a smaller subsequent gain. Differences associated with using AI for explanation versus delegating writing are suggestive, but students' chosen use styles were not randomly assigned. Those subgroup patterns do not establish what would happen if the study imposed one style rather than another.

These results are not pooled into a new average. They differ in subject matter, duration, allocation, actual uptake and assessment. Together they justify examining conditions and intervention components. They also identify displacement of existing activities, limited adoption and learner differences as alternatives to a simple account based on tool quality.

3.4 Chess as Evidence and as a Boundary

Southwick and colleagues [11] analyzed a longitudinal cohort of 44,213 Chess.com players. Their publisher abstract reports a 3.61-fold association in learning efficiency for activities aligned with deliberate practice relative to gameplay. The design is observational and is assessed here only through the abstract. It does not establish a causal comparison with books, a universal prescription or a specific conclusion about elite players.

Strittmatter, Sunde and Zegners [12] examined historical professional games using engine-based move evaluation and found performance improvements across cohorts. The study supports discussion of historical performance changes. It does not experimentally isolate the effect of training with technology. Player selection, access to accumulated knowledge and other developments can accompany those changes. Attributing stronger contemporary play exclusively to technology exceeds what this design identifies.

Chess is useful as a laboratory for specific practice. Better chess-problem performance would not demonstrate improved general intelligence, mathematics achievement or school performance. Transfer to another domain requires an appropriate study in that domain.

3.5 El Salvador and the Limits of Attribution

The El Salvador case illustrates the same problem at the system level. The World Bank [14] reported that 1,198 volunteers from 171 schools scored 417 in mathematics, 492 in reading and 462 in science in the July 2026 PISA for Schools assessment. The source explicitly states that the results do not represent the country or the entire program. A comparison with the national average does not measure the progress of these same students. The intervention also combined curriculum, materials, teacher support, management, assessment and technology.

An analysis by Restrepo-Morales and colleagues [15], accepted on 29 October 2026 and available as an abstract, examines how much selection could be compatible with the reported scores without assuming learning gains. It concludes that the available information cannot distinguish strong selection from exceptional improvement, or attribute any improvement specifically to the tutor within the broader reform. Its calculations are not reproduced without access to the final text. This case neither confirms nor refutes the effectiveness of a particular AI brand. It shows why baseline measurement and an appropriate comparator matter. Differences between groups are not converted here into school years recovered.

4 Conceptual Framework and Contribution

4.1 Redistributing Demands While Preserving Learning Operations

Risko and Gilbert [13] describe cognitive offloading as altering processing requirements through actions and external supports. This raises the question of what is externalized. The present design proposal would automate aspects of selecting exercises and organizing reviews while retaining the learner's responsibility for deciding, explaining and retrieving. This is a testable instructional specification, not a natural boundary already demonstrated by research.

An operation is not peripheral to every learning objective. If the objective is planning one's own study, selecting exercises is part of the intended competence. If the objective is tactical calculation, selection can be an additional organizational demand. The classification must be fixed relative to the objective before the experiment; otherwise favorable results could be explained retrospectively as removal of supposedly unnecessary demands.

Mental energy is used only as a motivating metaphor. No single energetic reserve is proposed, and metabolic consumption is not inferred from clicks, fatigue or time. Duration, perceived effort, reported fatigue, persistence and response quality are separate variables. An intervention could reduce fatigue while impairing learning, or increase perceived difficulty while improving retention.

4.2 Observable Practice Cycles

A complete cycle is defined as a recorded sequence containing an independent attempt before solution disclosure, consultation of relevant feedback, and a subsequent unassisted application on a related problem. The cycle is a process unit. An incorrect response still counts as an attempt, and the subsequent application may also fail. This avoids defining productive practice as whatever happens to lead to the desired outcome.

Cycle density is completed cycles per 450 assigned minutes. It is a descriptive interface-behavior measure, not an estimate of learning efficiency. Attempts, initial accuracy, hint use, duration and difficulty are also reported. Error-driven scheduling and difficulty progression make cycle counts endogenous to treatment and performance: more errors can generate more opportunities without more learning. Appendix B specifies initiated episodes, delivered opportunities, completion fractions, one-to-one linkage and end-of-study censoring. Feedback-open duration is an observable exposure proxy, not proof of cognitive consultation.

For measurement, a feedback episode will be linked to the first prespecified related reapplication opportunity. A reapplication will not close multiple earlier cycles. Cycles may span sessions, and their completion date will be recorded. Linking rules and the mapping of related problems will be frozen

after the pilot, before the confirmatory trial. Failure to inspect feedback and absence of a later application will be recorded separately rather than treated as completed cycles.

The proposed mechanism is that instructional design changes time allocation, task selection and the probability of applying a correction again. These changes could support retention, but novelty, dependence, frustration or withdrawal could also occur. Motivation and prior competence may influence both process measures and outcomes. An association between completed cycles and improvement would not establish causal mediation.

The contribution is operational: combining an active comparator, observable process measures and delayed independent assessment in a falsifiable protocol. This article does not claim to have discovered deliberate practice, adaptive tutoring or the distinction between learning and performance. Establishing the empirical originality of the combination would require a more exhaustive literature review and original data.

5 Hypotheses and Falsification Criteria

H1, primary. Following training, adaptive digital practice will outperform fixed digital practice on novel, unaided problems four weeks later, adjusting for baseline performance and allocation strata. This comparison evaluates the adaptation package against an active digital comparator.

H2, secondary. Fixed digital practice will outperform active print study on the same delayed outcome. This comparison estimates a difference between specified practice environments, not the universal effect of technology versus its absence.

H3, secondary. Fixed practice requiring a prior attempt will outperform direct solution access on the delayed assessment, even if direct access improves within-session correctness. Effects could differ by experience; observing a solution is not assumed to be inherently harmful.

H4, exploratory and descriptive. Adaptation may change cycle density and opportunity delivery, and these measures may be associated with delayed performance. Because the scheduling policy and learner errors mechanically affect those measures, neither arm differences nor outcome associations identify efficiency or causal mediation. Fatigue is another measured correlate, not an established mechanism.

An advantage confined to practice would not confirm H1. An imprecise null estimate would leave the question open. A narrow interval incompatible with an educationally meaningful advantage would weaken the practical hypothesis. A delayed disadvantage would require revising the design. Satisfaction, exercise counts or a favorable subgroup will not replace the primary outcome after results are known.

6 Proposed Experimental Protocol

6.1 Participants and Preparation

A multicenter trial is proposed for adults who know the rules of chess and perform sufficiently well on a common baseline test to attempt the problem bank. Complete beginners and experts with near-ceiling performance would be excluded. Eligibility thresholds will be calibrated in a separate pilot of 40 participants, ten per condition. The pilot will assess comprehension, duration, accessibility and measurement quality, not confirmatory efficacy.

The feasibility pilot is distinct from an independent assessment-calibration stage targeting 180 complete three-visit records, with up to 225 recruits. Neither sample enters the main trial. Appendix C supplies the rationale, form blueprint, example positions and scoring decisions, reliability and comparability gates, and mandatory pause/recalibrate rules. Before main-trial recruitment, eligibility rules, materials, the algorithm, rubrics and analysis plan will be frozen in a dated public registration. This manuscript is not that registration. No recruitment, ethical approval, committed funding, implemented intervention or validated bank is claimed to exist.

6.2 Conditions and Assigned Study Time

Training will last six weeks, with three supervised 25-minute sessions per week: 450 assigned minutes per participant. Conditions will share learning objectives, a common source bank and expert-reviewed explanations. Training material will remain separate from assessment items. No condition will receive additional individualized tutoring during the sessions.

Condition A, active print study. A workbook and physical board will present a fixed sequence, decision and explanation prompts, solutions in a separate section, and subsequent application problems. Learners will be instructed to record an attempt before consulting the solution. This is an active instructional comparator that recognizes the possibilities of a well-designed book.

Condition B, fixed digital practice. The protected core and fixed supplementary sequence and schedule used in A will be delivered through a digital board. A recorded attempt will be required before feedback opens. Hints and explanations will be equivalent to the workbook. The system will log events and facilitate navigation and representation of variations. The comparison with A includes differences in medium, manipulation and compliance; it does not isolate a single property of screens.

Condition C, constrained adaptive digital practice. B's interface, explanations and attempt requirement are retained. Every session has common core blocks in minutes 0-10 and 18-25 and a supplementary block in minutes 10-18; all arms share the core. C adapts only the supplementary block, using errors to queue next-session and one-week reviews and correct streaks to adjust difficulty. The queue is capped at 24 active entries, expires reviews after two scheduled sessions beyond their due date and cannot displace core blocks. Appendix B defines eight motifs, item records, seeded selection, priorities, eviction, exhaustion and boundary rules. All values are prospective choices requiring implementation and feasibility testing. This deliberate pre-data revision preserves scheduled syllabus exposure while estimating a bundled adaptation package; it does not isolate spacing from difficulty progression.

Condition D, direct solution access. The interface and fixed sequence from B will allow the solution and explanation to be revealed before an attempt. A generative model is not required: reviewed answers will prevent answer accuracy from becoming a second experimental difference. Voluntary attempts will be logged. D estimates the effect of permitting early solution access; it does not guarantee that every participant delegates reasoning.

The exact exposure in C will differ from the fixed sequence because selection is part of its treatment. Exposure to problems, motifs and difficulty will be reported by group. Assigned time is equal; actual mental effort cannot be equated with a timer. Outside practice will be recorded weekly and will not justify exclusion from the primary analysis.

6.3 Allocation and Assessment

Participants will be individually randomized in a 1:1:1:1 ratio using variable block sizes, stratified by center and baseline ability. Someone uninvolved in assessment will prepare the allocation sequence and conceal it until enrollment is complete. Participants will know their environment; outcome raters and the initial analysis will use masked group labels.

Assessments will occur before training, within 72 hours after the final scheduled training session and at a target of 28 days after that scheduled session, irrespective of attendance. The delayed window is days 21-35 inclusive; Appendix C specifies repeat-selection and out-of-window rules. The primary test will use the same paper and board format for every group, without engines, internet, hints or study material. A brief common familiarization with that format will precede baseline assessment.

The primary outcome will be the delayed score on 30 novel positions: 20 involving trained motifs in unseen configurations, and ten involving near application through different combinations of those motifs. Each position will yield one point for the correct decision and one for a separately scored correct tactical explanation, using the closed rubric in Appendix C. An incorrect selected move does not automatically eliminate explanation credit, but credit requires a correct explanation of an accepted move and its critical continuation, without contradictory claims. The two components remain separately reportable. The total will range from zero to 60; item categories will be reported as secondary outcomes. This measures near transfer within chess, not general cognitive ability.

Three disjoint forms will be prepared, with eight motifs represented by two or three single-motif positions each and ten prespecified combined-motif positions per form. Six form orders will be balanced within treatment and center. The separate calibration stage in Appendix C will evaluate difficulty, form/visit effects, reliability, interrater agreement and floor/ceiling behavior under explicit gates; failing a gate pauses the main trial and requires revision and independent recalibration. Equivalent alternative solutions will be validated by an expert and an engine, documenting engine version, configuration and search depth. Engine checks will support legality and evaluation, while the human rubric will assess explanations. Rater agreement will be measured in a random 20% of responses, with discrepancies resolved through masked review.

6.4 Process Measures and Fatigue

Secondary measures will include immediate performance, within-session performance, solution consultation before an attempt, cycle density, persistence and attendance. Fatigue will be reported from zero to ten before and after each session; perceived effort will be reported from one to nine afterward. These are separate self-report indicators, not metabolic measurements or clinical instruments.

Condition A will use timed attempt sheets and sampled observation; B, C and D will use event logs. Lower-resolution print records could bias process comparisons. Accordingly, the principal cycle-density comparison will be C versus B, which shares instrumentation. A sample of sessions in every group will be audited using an equivalent observation protocol and appropriate consent. Inactivity will not automatically be labeled inattention, since it may contain mental calculation.

6.5 Sample Size and Analysis

For initial planning, detecting a standardized difference of 0.30 with 80% power and a two-sided 0.05 test requires approximately 175 complete outcomes per group in a simple independent comparison.

Allowing 20% attrition, 220 allocations per condition are proposed, or 880 participants overall. This normal approximation makes no allowance for a potential gain in precision from baseline adjustment. It is not a protocol simulation or a guarantee of power for interactions, mediation or equivalence. Cluster randomization, if adopted, would require recalculation before recruitment.

The primary analysis will follow intention to treat, comparing C with B in a regression of delayed score on treatment, baseline score, allocation strata and assessment form. Stable supervision cohorts will be established before allocation; cohort-clustered CR2 standard errors with Satterthwaite degrees of freedom will address shared session conditions. Main-trial power must be revised for the planned cohort structure. Appendix C gives the treatment-aware imputation and sensitivity specifications. Results will include the difference in score points, a 95% confidence interval and a standardized difference with an explicitly identified denominator. The primary test will be two-sided. Secondary contrasts B versus A and B versus D will be adjusted together using Holm's procedure. Remaining analyses will be labeled exploratory.

A difference of 0.20 outcome standard deviations is provisionally proposed as the smallest effect of interest, not a validated threshold, and will be translated into points using calibration data. Power is planned for 0.30, not equivalence at 0.20. At 175 complete cases per arm, an unadjusted 95% interval near a zero effect has approximate half-width 1.96 times sqrt(2/175) = 0.210 SD. Thus even an exact zero estimate cannot exclude a 0.20-SD benefit under that approximation. No baseline precision gain is assumed. Cohort dependence requires recalculating power before recruitment. Only an actual sufficiently narrow interval could support an exclusion statement; nonsignificance alone cannot falsify a 0.20-SD benefit.

Attrition and reasons will be reported by group. Under a missing-at-random assumption, 100 treatment-aware chained-equation imputations will include assigned arm, baseline score, arm-by-baseline terms, center, ability stratum, cohort, form, attendance and available assessments, using predictive mean matching for bounded scores. Complete-case, out-of-window and arm-specific delta sensitivity analyses will be reported; Appendix C fixes a 25-cell delta grid for C versus B and corresponding secondary contrasts. Missing participants will not be scored zero, and last observation carried forward will not be the primary method. Specific imputation and sensitivity procedures will be frozen before unmasking outcomes by condition.

Process measures occur after allocation and will not be added to the primary model to remove their effects. Such adjustment could block part of the treatment effect or introduce bias. Relationships among cycles, fatigue and retention will be exploratory and will state their assumptions. Identifying a mechanism would require further randomized components or additional defensible assumptions.

6.6 Ethics and Reproducibility

Implementation requires appropriate ethical review, informed consent, separation from course grading and voluntary withdrawal without penalty. Only relevant data will be collected, using pseudonymous identifiers. Any recordings used for auditing require separate consent and a deletion schedule. The initial proposal concerns adults; extension to minors requires a dedicated protocol.

Registration will include prior exclusion rules, material licenses, software versions, the adaptive algorithm, rubrics, assessment windows and analyses. Authorized materials, a data dictionary and analysis code will be made available. Individual data will be shared only with adequate consent and protection. No commitment is made to disclose information that could identify participants.

7 Discussion

The hypothesis does not require learning to feel easier. A tool could reduce time spent locating feedback while requiring more demanding attempts. Alternatively, it could feel comfortable because it removes the need to reason. Both may appear efficient during practice. Delayed unaided assessment separates their consequences.

An active workbook comparator prevents instructional structure from being attributed automatically to technology. Similar outcomes in A and B, with improvement in C, would favor the adaptation package. An advantage of B over A without an additional advantage of C could point toward representation, navigation or compliance. If D matched or exceeded B, the presumption that mandatory attempts add value in these conditions would require revision. None of these patterns should be relabeled after the fact as automatic confirmation of the framework.

Practical usefulness also depends on producing accurate explanations, reviewing errors, maintaining software and sustaining participation. A small benefit may matter if it expands access at low cost; a laboratory benefit may lose value if it requires intensive supervision. The study will record implementation resources without inferring economic cost-effectiveness from a score difference alone.

The broader claim concerning effort requires another study. A subsequent factorial design could allocate fixed versus adaptive practice and two session durations, such as 15 versus 30 minutes. The interaction would ask whether the effect of greater assigned study time depends on design. Assigned duration would still not equal mental effort. That experiment needs its own power calculation; the proposed sample of 880 does not cover it.

Substantial limitations remain. The synthesis is selective, several recent findings are preprints, and the chess efficiency finding is assessed here through an abstract. Cycle measurement needs validation and could overlook learning through observation. Appendices B-C now provide provisional algorithm and assessment specifications, but the actual item bank and implementation remain to be built and independently checked. The proposed assessment is not calibrated and effectiveness remains unknown. The search does not establish priority for a new general principle.

The defensible contribution is an explicit set of research conditions: compare credible active alternatives, preserve the reasoning operation being learned, observe process without equating it with learning, and allow findings that contradict the initial expectation. Extending any result to languages, mathematics or professional training requires domain-specific replications and relevant outcome measures.

8 Conclusion

The literature examined supports investigating how technological design changes practice. It does not establish that adding technology to effort always improves outcomes. The proposed hypothesis predicts a delayed advantage for an adaptive intervention that preserves independent attempts over active fixed digital practice. The trial specifies how that prediction could be examined and how its failure would be recognized. Until data are collected, this work provides an operational framework and proposed protocol rather than empirical proof of superiority.

Declarations

Data and results. No original participant data were collected, and the trial has not been conducted. Empirical numbers belong to the identified publications. Sample sizes and design choices are proposals.

Registration and ethics. There is no prospective registration or ethical approval for the proposed experiment. These must be obtained as appropriate before recruitment.

AI assistance. Codex, an OpenAI AI assistant, supported searches, checks of principal claims, development of the argument and protocol, drafting, translation and document production. The drafting model version was not recorded. Separate internal gpt-6-astra and gpt-6.1-sol reviews and a Google Gemini API gemini-4-argon review of the preceding English PDF informed this revision; original adverse findings are preserved separately. Their assessments do not apply automatically to these changed bytes. Astra and Sol also helped specify the provisional appendices after reviewing the earlier draft; that subsequent involvement must be declared in any new review. Lluis Eriksson supplied the initial idea and instructed submission and publication in English. Assistance does not constitute independent peer review or establish a complete human verification of every reference. The present revision clarifies prospective intervention, measurement and analysis rules; these are proposed design choices, not empirical findings.

Funding and conflicts of interest. The author declares no specific funding for this work and no additional relevant financial relationships with educational or chess platforms. Lluis Eriksson operates AIRR.science, the intended receiving archive; this author-operator relationship is explicitly disclosed. These declarations were confirmed by the author on 1 November 2026. Deposit is not an independent evaluation or an acceptance decision.

References

[1] Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. Perspectives on Psychological Science, 10(2), 176-199. https://doi.org/10.1177/1745691615569000

[2] Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14(1), 4-58. https://doi.org/10.1177/1529100612453266

[3] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

[4] Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. Review of Educational Research, 91(5), 761-798. https://doi.org/10.3102/00346543211019105

[5] Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901-918. https://doi.org/10.1037/a0037123

[6] Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15, 17458. https://doi.org/10.1038/s41598-025-97652-6

[7] Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

[8] Gallegos, S. (2026). Guidance over adoption: Experimental evidence on AI-assisted learning. IZA Discussion Paper 18513. Discussion paper. https://docs.iza.org/dp18513.pdf

[9] Liu, J., Sweet, T., Chen, M. H., Engelberg, J., Masters, M. C., Clark, M., Persaud, A., Lancaster, A., Hollingsworth, J. K., & Rice, J. K. (2026). The effects of course-integrated AI tutoring on student performance and engagement: A randomized university trial. IZA Discussion Paper 18989. Discussion paper. https://docs.iza.org/dp18989.pdf

[10] Contractor, Z., & Reyes, G. (2026). Experimental evidence on the learning impact of generative AI. arXiv:2607.08849v2, revised 4 October. Preprint. https://arxiv.org/abs/2607.08849v2

[11] Southwick, D. A., Harwell, K. W., Wright, G., Olsen, J. A., & Ogles, B. M. (2026). Not all practice is created equal: Longitudinal evidence from over 40,000 chess players. Psychological Science, 37(7). Access limited to the publisher abstract. https://doi.org/10.1177/09567976261452568

[12] Strittmatter, A., Sunde, U., & Zegners, D. (2020). Life cycle patterns of cognitive performance over the long run. Proceedings of the National Academy of Sciences, 117(44), 27255-27261. https://doi.org/10.1073/pnas.2006653117

[13] Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676-688. https://doi.org/10.1016/j.tics.2016.07.002

[14] Cordeiro Guerra, S., & Saavedra, J. (2026, September 10). Una señal alentadora desde El Salvador: un sistema alineado y tecnología al servicio de los aprendizajes pueden acelerar los cambios [An encouraging signal from El Salvador: an aligned system and technology in the service of learning can accelerate change]. World Bank Blogs. Institutional communication, not a causal trial. https://blogs.worldbank.org/es/latinamerica/el-salvador-sistema-alineado-educacion-tecnologia-servicio-aprendizajes-pueden-acelerar-cambios

[15] Restrepo-Morales, J. A., Rodríguez Flores, E. A., Giraldo Betancur, E. A., & Zea Restrepo, F. (2026). How much selection would be enough? Bounding the learning claim of El Salvador's artificial intelligence tutoring pilot. Frontiers in Education. Accepted 29 October; abstract available and final formatted version pending at consultation. https://doi.org/10.3389/feduc.2026.1993600

Appendix A Source Traceability

Consultation date: 31 October 2026. This appendix documents the directed search and major verification decisions. It is not a systematic selection diagram or an exhaustive bibliography.

Foundations [1-3, 13]. APS publisher records were consulted for learning versus performance and study techniques; the Wiley publisher record and the UNSW academic record supported the Sweller reference; UCL's repository supported cognitive offloading.
https://www.psychologicalscience.org/journals/perspectives/1745691615569000/ ;
https://www.psychologicalscience.org/publications/journals/pspi/learning-techniques.html ;
https://www.unsw.edu.au/staff/john-sweller ; https://discovery.ucl.ac.uk/id/eprint/1508770/

Meta-analyses [4-5]. Sinha and Kapur's abstract and findings for conceptual and procedural knowledge were checked separately. Ma's abstract and comparator table were checked; model choices and rounding can differ from abstract values. This manuscript uses the abstract's 0.35 estimate for textbooks and workbooks. https://journals.sagepub.com/doi/10.3102/00346543211019105 ;
https://www.apa.org/pubs/journals/features/edu-a0037123.pdf

Published trials [6-7]. Kestin's eligible sample, two-lesson design and immediate assessment were checked. Bastani's abstract and results were compared, preserving the distinction between a relative

reduction in grades and a percentage-point difference. https://pmc.ncbi.nlm.nih.gov/articles/PMC12179260/ ; https://pmc.ncbi.nlm.nih.gov/articles/PMC12232635/

Recent papers [8-10]. Gallegos's abstract and guidance-intervention results were checked. Liu's Section 5 and exact-match sample description were consulted so that the subsample effect was not attributed to the full sample. Contractor and Reyes's version record and HTML were consulted without using numerical entries whose automated extraction was inconsistent; randomized access was distinguished from chosen use style. Document links appear in the references.

Chess [11-12]. Southwick's sample and efficiency association were taken only from the publisher abstract; the full article was not audited. Strittmatter's accessible full text was consulted to distinguish historical performance trends from causal identification of technology. https://www.psychologicalscience.org/journals/psychological-science/09567976261452568/ ; https://pmc.ncbi.nlm.nih.gov/articles/PMC7959507/

El Salvador [14-15]. Participation, scores and limitations were checked against the World Bank communication. The Frontiers abstract supplied the acceptance status and the analysis's stated conclusion. The PISA for Schools technical report was not directly verified. No school-year equivalents or numerical comparisons with Germany are used. These sources supplement the thirteen research publications in the core synthesis.

Scope exclusions. Calculator studies and professional writing-productivity studies mentioned in the initial background text were left outside the core synthesis to keep its focus on independent learning and practice design. Exclusion was not interpreted as a negative result.

Appendix B Provisional Intervention And Recorded-Cycle Specification

Status and Scope

This deliberate pre-data protocol revision protects syllabus coverage against review crowd-out. Every parameter below is proposed, not validated. This specification describes a candidate implementation for the adult chess trial; no tagged problem bank, software implementation or validated difficulty scale currently exists. Expert review, an independent feasibility pilot and a versioned final specification are prerequisites to recruitment. The assessment bank remains separate and inaccessible to training software.

B1 Finite taxonomy and source-bank records

The provisional syllabus contains eight tactical motifs: (1) fork or double attack; (2) pin; (3) skewer; (4) discovered attack or discovered check; (5) removal or deflection of a defender; (6) attraction or decoy; (7) interference or clearance; and (8) mating net or back-rank weakness. Each training position has one primary motif and optionally one secondary motif from this list. Two qualified chess reviewers independently assign tags and reconcile disagreement before the bank is frozen. Untagged positions are excluded.

Each record contains immutable item_id, family_id, position in FEN, side to move, primary and secondary motifs, difficulty level 1-3, source and license, accepted first moves and continuations, decision and explanation rubrics, reviewed explanation, permitted hints, and expert-review status. Records also contain engine name/version/configuration/search limit and analysis output, reviewer decisions, estimated completion time, and a content checksum. Document alternative solutions.

Family_id groups closely related configurations; exact item reuse and related-family reuse are distinguished in reporting.

A separate frozen mapping assigns each item eligible reapplication items. A related item must share the primary motif and require a fresh decision in a different position; mere replay of the disclosed solution is ineligible. Mapping edges receive expert approval and identify whether the target is same-level or harder. The bank manifest states available counts by motif, level and family. The feasibility gate requires sufficient eligible items to run the full candidate schedule without relying on unplanned reuse; failing this gate requires expanding the bank and repeating affected checks before registration.

B2 Common delivery and fixed sequence

There are 18 supervised 25-minute sessions; training ends at session 18. Each session reserves minutes 0-10 for a common core, minutes 10-18 for supplementary practice, and minutes 18-25 for a second common core. All groups receive the same session-specific core items, prompts and reviewed content. The first core targets motif 1+((session-1) modulo 8); the second targets motif 1+(session modulo 8). Each core contains an index item and a mapped application. Index difficulty is level 1 in sessions 1-6, level 2 in 7-12 and level 3 in 13-18. Every motif therefore has scheduled exposure independent of errors. Report actual completion.

Generate the core manifest before recruitment using a published seeded permutation within each motif-level cell and approved application mapping. Core and supplementary inventories are disjoint. Reserve every future core item against adaptive selection. At a block boundary save the response, close the item and open the next block; no response grace extends a block. Record unfinished items and feedback not reached. Never delay a core block to finish reviews or add training after session 18. Missed core blocks are not reassigned.

A and B require a decision and explanation before disclosure; A records attempts on timed sheets and B enforces field completion. An incorrect response counts; blank or explicitly declined responses do not. D permits immediate solution disclosure and records voluntary attempts. In A, B and D, the supplementary block follows a common frozen index/application sequence, cycling motifs 1-8 and levels 1-3. Participants resume their supplementary cursor across sessions. C changes supplementary selection only. C versus B therefore estimates this constrained adaptation package, not unrestricted personalization or a single mechanism.

B3 Adaptive state and deterministic selection

C shares B's interface and attempt requirement. For each motif maintain level L=1, correct streak S=0, unused inventory and review entries. A first move is correct if it matches the reviewed accepted-move set; explanation scoring does not drive adaptation. Every C response, including core responses, updates state: a correct move increments S; two consecutive correct moves raise L by one, capped at 3, and reset S. An error resets S without lowering L and queues reviews for the next scheduled session and the first scheduled session at least seven calendar days later. Both use source difficulty; coincident due dates merge. Errors on reviews may generate new entries.

Only the eight-minute supplementary block can serve the queue. Missing sessions create overdue entries, not replacement minutes. A queued review expires after two scheduled sessions beyond its due session, whether attended or not. Retain expired entries in the audit log. Cap the active queue at 24 entries. On overflow evict the entry with the latest due session, then latest source-feedback

timestamp, then lexically greatest opportunity_id. Thus nearer obligations survive; eviction is a recorded protocol event, not a completed review.

At every supplementary item boundary:

1. Expire entries, enforce the cap and exclude obligations beyond session 18 from active selection while retaining their records.
2. Select due or overdue entries by earliest due session, earliest source-feedback timestamp, source item_id, then opportunity_id. No early review is served.
3. Select an unused approved related item at source difficulty. Break ties by a seeded hash of participant study_id, opportunity_id and item_id. If unavailable, try one level lower, then one higher, using approved mapping edges only. If no item exists, retire the entry as unavailable and continue.
4. If no due review is serviceable, select a fresh supplementary index item, rotating motifs 1-8 at their current L. Resolve ties by the seeded hash; try one lower then one higher level when the cell is empty. Never draw from core inventory or substitute an unreviewed item.
5. Stop at minute 18 even with outstanding obligations. If supplementary inventory is exhausted, permit quiet inspection of disclosed explanations until the next core block; record remaining time and count no new cycles.

No exact position is repeated. Publish seed, hash algorithm and replay examples. Report core offered/completed counts by motif, review-block time, queue load, expiration, eviction, unavailable items, fallback and exhaustion. The protected blocks prevent reviews from displacing scheduled syllabus coverage; they do not guarantee equal completed exposure or effort.

B4 Recorded cycles, opportunities and censoring

A source episode requires an independently recorded attempt before disclosure and a feedback-open event. In digital groups, feedback-open means the reviewed explanation panel was visibly displayed for at least five cumulative seconds while the session was active. This is an interface exposure proxy, not proof of attention, comprehension or cognitive consultation. Panel timestamps, dwell time and premature closure are retained. Print records use a corresponding declared consultation event with sampled observer checks; their lower resolution prevents treating print and digital cycle counts as interchangeable.

Each eligible source episode receives at most one designated first reapplication opportunity. Its mapping and intended session are recorded when feedback qualifies. In C this is the first scheduled review following an error; following a correct source response it is the first later approved related item actually selected. B, A and D use the next mapped application in the fixed sequence. The additional one-week review is reported as maintenance practice rather than a second closure of the original cycle. A response may close one earlier source cycle and start a new source episode of its own; it cannot close multiple earlier cycles. If multiple sources qualify, link the oldest source-feedback timestamp, breaking ties by item_id; record both identifiers.

A cycle closes when a valid unassisted attempt is submitted at the designated opportunity before its feedback opens, regardless of correctness. Disclosure before the application attempt prevents closure. No qualifying later application means an incomplete cycle. Report incomplete episodes separately as feedback not opened, no mapped item available, opportunity offered but not attempted, session missed, queue expiration or eviction, or follow-up outside the assigned training window. Session-18 errors and reviews due after session 18 receive no extra training. They remain administratively right-censored at training end, never silently converted into failures or completed cycles.

Report completed cycles per 450 assigned minutes; also report source episodes, mapped opportunities due within training, opportunities actually offered, eligible independent application attempts, and completed/offered and completed/due fractions. Sources whose first possible application lies beyond session 18 are excluded only from the due-opportunity denominator and shown separately. Attendance loss and unavailable inventory remain visible. Report distributions by source session, motif and difficulty, plus time to application. Error-driven review eligibility makes cycle counts endogenous to performance and treatment. H4 is descriptive: neither density nor these denominators identifies learning efficiency or causal mediation. More errors can generate more opportunities without more learning.

B5 Pilot whitelist, freeze and audit

The independent 40-person pilot may change only: wording and accessibility; five-second feedback-proxy threshold within 3-15 seconds; bank additions/removals and corrected expert tags; difficulty assignments; motif ordering; consecutive-correct threshold from two to three; initial level from 1 to 2; and deterministic fallback ordering. Block boundaries, core reservations, queue cap/expiry/eviction and no-grace rule are fixed outside this whitelist. Changes require written feasibility or measurement reasons and a dated log. Review due dates, eight-motif syllabus, primary contrast, session count/duration, outcome definition and cycle one-to-one rule are outside this whitelist. A necessary change to those elements requires an explicitly revised protocol and fresh review before registration, not an undisclosed pilot adjustment.

The pilot does not select settings by the largest delayed treatment difference. Before main-trial recruitment, publish the final manifest, algorithm/pseudocode, parameter file, fixed sequence, mapping, proxy rules, denominator definitions, replay examples and amendment log, with licenses or controlled-access arrangements where required. Freeze their checksums with the registration. Later safety or correctness repairs require versioned amendments, disclosure of affected participants and prespecified handling; they cannot be described as the unchanged registered intervention.

Appendix C Proposed Assessment Blueprint And Pretrial Validation Rules

Status and Scope

Every threshold below is a proposed design decision, not a validated educational standard or a calibration result. No participants have been recruited and no assessment has passed these criteria. Materials, scoring rules, the adaptive algorithm and the statistical plan must be frozen and publicly dated before main-trial randomization. The primary construct is unaided tactical decision-making and explanation on unfamiliar chess positions, including near transfer within chess. It does not encompass general intelligence, tournament strength or all forms of independent learning.

C1 Separate feasibility and calibration stages

The feasibility pilot comprises 40 participants, ten per condition. It examines instructions, accessibility, session logistics, software operation and process recording. Its purpose is neither estimating treatment efficacy nor selecting the arm that appears most successful. Its participants will not enter the calibration study or confirmatory trial.

A separate assessment-calibration sample will provide 180 complete three-visit records from adults spanning the intended eligibility range. Recruit up to 225 to allow approximately 20% incomplete

records; report all recruited participants and missing assessments. No participant receives the experimental training. Each completes all three assessment forms, seven days apart, without item feedback, assigned equally to the six possible form orders: 30 complete records per order. This yields 180 observations per item and 540 form administrations. At an item success probability of 0.50, the elementary independent-person standard error is approximately 0.037. That calculation motivates item screening precision; it is not efficacy power or proof of adequate precision for every reliability statistic. Repeated observations within a person must remain linked in all analyses.

Visit effects will be estimated separately from form effects. Eligibility and difficulty decisions must use the intended recruitment population, not a post hoc subgroup selected to make reliability acceptable. Incomplete records will inform feasibility reporting and sensitivity checks; they will not be silently replaced in the analysis. Failure to obtain the planned complete records triggers a documented feasibility review before proceeding.

C2 Item allocation and administration

Prepare three disjoint forms, each containing 30 positions, for baseline, immediate and delayed assessment. Each form contains 20 single-motif items across eight prospectively named syllabus motifs: four motifs receive three items each and four receive two each. Freeze the same allocation across forms before calibration. The remaining ten combine motifs, using the same prespecified pairings and approximate difficulty bands across forms. A motif-by-form blueprint will identify the decision, explanation target, intended difficulty, accepted alternatives and training-bank separation of every item. No assessment position may occur in training, and superficial rotations or color reversals of training positions are excluded.

All forms use paper diagrams and a physical board, the same instructions and a 60-minute limit, subject to accessibility arrangements fixed before randomization. The feasibility pilot must establish that this administration is practicable. Record completion time and unanswered items. An unanswered item on an otherwise administered test receives zero; a wholly absent visit is missing, not zero. No hints, engines, internet or item-level feedback are permitted. Balance the six form orders within treatment and center. The calibration data must support the proposed order counterbalancing rather than assuming that different forms are interchangeable.

C3 Binary rubric and provisional examples

Each item contributes two separately recorded binary scores. Decision = 1 for a legal move on the frozen accepted-move list; otherwise 0. Explanation = 1 only when the response correctly states the item-specific tactical reason and, where required, the critical continuation or excluded defense; otherwise 0. Scoring is independent, not conditional on decision credit: an incorrect selected move may coexist with explanation = 1 only if the explanation itself correctly identifies the required tactic and continuation for an accepted move. An explanation falsely claiming that the incorrect selected move executes that tactic receives 0. Contradictory claims that negate the required reason fail that component. Accept equivalent notation, verbal square descriptions and semantically equivalent explanations; spelling and prose quality are not targets. Total scores range from 0 to 60. Decision and explanation subtotals will also be reported.

Example 1, mate in one, White to move. FEN: 7k/8/5KQ1/8/8/8/8/8 w - - 0 1. Prompt: Give a move that checkmates immediately and explain why the king cannot escape or capture the checking piece. Decision: Qg7# (g6g7) = 1; any other move = 0. Explanation: the queen on g7 checks h8, controls g8

and h7, and is protected by the king on f6, preventing capture on g7. Equivalent complete verbal descriptions are accepted; merely writing 'checkmate' is insufficient. This position has only one mating first move under exhaustive legal-move enumeration.

Example 2, forcing knight fork, White to move. FEN: 4k3/7q/8/3N4/8/8/8/4K3 w - - 0 1. Prompt: Give a forcing move that permits capture of the queen on the next White move, and explain the sequence. Decision: Nf6+ (d5f6) = 1. Explanation: the knight checks the king on e8 and attacks the queen on h7; Black must move the king, after which Nxh7 captures the queen. Accept an equivalent description or any legal illustrative king reply followed by Nxh7. Naming a fork without identifying both targets and the capture is insufficient. Every legal Black reply after Nf6+ permits that capture.

Both examples were checked with python-chess 1.11.2 for position validity, move legality and the stated immediate continuations. These are provisional rubric illustrations, not calibrated assessment items. No engine evaluation or expert approval is claimed. Before use, an independent chess expert must review alternatives, wording and construct suitability; engine version, settings and search depth must be archived for actual bank validation.

C4 Proposed acceptance gates and failure decisions

Require each form's mean to lie between 15 and 45 points, with no more than 10% of participants scoring 0-6 or 54-60. Review every component with success below 0.10 or above 0.95, and every item with a negative corrected item-total association. Such flags require documented content review rather than automatic deletion solely to increase reliability.

For the 30 item totals, require coefficient alpha at least 0.75, with a person-bootstrap 95% lower confidence bound of at least 0.65, separately for each form. These are screening gates, not proof of unidimensionality or validity. Estimate form differences using a participant-intercept model with form and visit effects. Require the 90% confidence interval for each adjusted pairwise form difference to lie within plus or minus 0.20 pooled calibration standard deviations. Freeze that pooled standard deviation for subsequent planning and sensitivity analyses.

Independently double-score a random 20% of complete forms, stratified by form, visit and ability band, before adjudication. Require at least 90% exact agreement on each component, explanation kappa at least 0.70 with lower 95% bound at least 0.60, and total-score absolute-agreement ICC at least 0.90 with lower bound at least 0.80. Use participant-level resampling for uncertainty. Preserve pre-adjudication ratings; adjudication cannot manufacture agreement.

If any gate fails or its interval is too imprecise, pause. Diagnose the failure, revise the bank, rubric or calibration design, and publish a dated amendment. Changed forms require fresh independent calibration under the same acceptance rules. Do not repeatedly tune thresholds against the same participants. Main-trial randomization begins only after all gates pass; any changed outcome variance, cohort dependence or eligibility range also requires a revised power calculation.

C5 Analysis and timing commitments

The initial efficacy calculation targets a 0.30-SD difference, not equivalence at the provisional 0.20-SD smallest effect of interest. With 175 complete outcomes per arm and no baseline precision gain or clustering, a standardized difference near zero has an approximate 95% interval half-width of 1.96 times sqrt(2/175) = 0.210 SD. Thus even an estimate exactly at zero would not exclude a 0.20-SD benefit under that approximation. Baseline adjustment may narrow intervals and cohort dependence

may widen them; neither improvement nor equivalence power is assumed. Retaining 220 allocations per arm is compatible with a superiority study, but a nonsignificant result cannot by itself falsify a 0.20-SD benefit. Any stronger exclusion claim must follow from the actual interval, not the planned sample size.

The delayed visit target is day 28 after each participant's final scheduled training session, whether attended or missed; days 21-35 inclusive constitute the allowed window. Use the in-window assessment closest to day 28 if a documented repeat is necessary. Out-of-window assessments are retained for sensitivity analysis but treated as unavailable for the primary time-window estimand. This avoids shifting follow-up according to treatment-related attendance.

Primary ANCOVA includes randomized treatment, baseline score, allocation strata and categorical delayed-assessment form. Establish stable supervision cohorts before allocation and retain their identifiers for all participants. Use cohort-clustered CR2 standard errors with Satterthwaite degrees of freedom, acknowledging shared session conditions even with individual randomization. Reassess power for the planned cohort structure before recruitment; do not assume independence merely because allocation is individual. Repeated secondary outcomes require participant-level dependence as well. Record cohort departures for sensitivity analysis without excluding participants.

Use 100 treatment-aware chained-equation imputations, incorporating assigned arm, baseline score, center, ability stratum, cohort, form, attendance and available assessments; include arm-by-baseline terms. Predictive mean matching preserves the bounded score range. Combine primary estimates with Rubin's rules, retaining the specified complete-data covariance procedure. Document diagnostics and Monte Carlo stability before interpretation.

For departures from missing at random, shift imputed delayed scores by arm-specific deltas of -0.50, -0.20, 0, +0.20 and +0.50 frozen calibration standard deviations, bounded to 0-60. For C versus B, report the complete 25-cell grid of paired shifts; use the corresponding paired grids for secondary contrasts. Report where conclusions change, alongside complete-case and out-of-window sensitivity analyses. These shifts are explicit sensitivity assumptions, not estimated truths about missing participants.