Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Growth Bias and False-Positive Learning Claims in AI-Assisted L2 Writing Assessment

Tan Nguyen Huynh Si

nhsytan@gmail.com

University of Phan Thiet https://orcid.org/0009-0000-9801-2760

Research Article

Keywords: generative artificial intelligence, L2 writing, writing assessment, longitudinal validity, Monte Carlo simulation, learning effects, calibration, unaided transfer

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11765598/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

AI-supported writing gains may combine durable independent proficiency, immediate assistance, and changes in measurement conditions. This fully synthetic Monte Carlo study compared six estimators under four primary scenarios (2,000 replications per scenario; n = 600 per replication), expanded sensitivity scenarios (1,200 replications each), and independent missing-data validation. A further 1,200-replication analysis per response mechanism evaluated fitted inverse-probability weighting (IPW) with stacked-sandwich uncertainty; paired-mode calibration was stress-tested under assessment-period effects, directional carryover, and task-by-mode interactions. Under zero true AI-related learning and a mean direct assistance effect of 0.50 standardised score units, assisted ANCOVA estimated the supported-outcome effect as 0.498 and yielded a statistically positive result in every primary replication when mistakenly interpreted as autonomous learning; it was not biased for its legitimate assisted-outcome target. Representative paired calibration reduced endpoint bias to −0.003 (94.7% coverage). In the new interval experiment, fitted-IPW stacked-sandwich coverage for the true unaided-transfer effect was 94.0% under MCAR, 93.8% under MAR, and 24.7% under MNAR; corresponding MNAR effect bias was −0.153. With balanced assessment order, a stipulated 0.20-unit carryover effect still produced +0.103 calibration bias. Conclusions: Supported performance, independent endpoint learning, and subsequent unaided transfer are distinct targets. Valid inference depends on construct alignment, measurement transportability, observability, selection assumptions, and appropriate uncertainty estimation. All effects and operating characteristics are conditional on synthetic data-generating mechanisms, not observed AI effects in classrooms.

1. Introduction

AI-assisted writing systems can revise grammar, increase lexical sophistication, reorganize arguments, and propose more fluent formulations while a learner is composing. From an instructional perspective, these capabilities may be productive: targeted feedback can make linguistic patterns salient, invite revision, and help learners test alternative expressions. From an assessment perspective, however, the same capabilities introduce an additional contributor to the score of the final written product. The resulting measurement problem is especially consequential in studies that interpret observed pretest–posttest score changes as evidence of learner development.

The potential divergence is intensified when an intervention and its outcome measurement both involve AI. In an intervention study, AI feedback can be simultaneously a mechanism of instruction, a component of the task environment, and a contributor to the scored response. If these roles are not separated analytically, an intervention may be judged successful because an outcome completed under more advantageous conditions receives a higher rating. This interpretation may be defensible for a study of AI-supported writing productivity, but it does not resolve whether the same writers could organise, formulate, and revise an unfamiliar text without the tool. A longitudinal label does not by itself supply that missing comparison: an assessment is longitudinal because the same construct is compared over time, not simply because some score is recorded at more than one occasion.

The issue also affects causal language. When participants are assigned to a supported instructional condition, a subsequent score difference can establish an effect of the assigned *package* on the outcome as administered, subject to the design's assumptions. It does not automatically decompose that effect into a change in acquired writing proficiency and a transient contribution of the system. The distinction is especially important when researchers report "learning gains," "proficiency gains," or "retention" without documenting whether AI remained available during the relevant tasks.

The key distinction is between *performance achieved under assistance* and *competence available when assistance is no longer present*. These outcomes need not be equivalent, but neither are they necessarily opposed. AI may improve a current product and, through meaningful engagement with feedback, contribute to subsequent independent learning. Alternatively, an output may improve mainly because the system supplied language or structure that the learner cannot yet generate independently. A valid assessment design must be able to accommodate both possibilities without treating either as established in advance.

Emerging L2 writing research gives substantive reasons to take the distinction seriously. Escalante et al. (2023) compared ChatGPT-generated with human feedback in a longitudinal English-as-a-new-language setting and did not find superior linguistic progress for the AI-feedback group. Yan and Zhang (2024) documented differences in how L2 writers engaged behaviorally, cognitively, and affectively with ChatGPT corrective feedback, underscoring that exposure to feedback cannot be equated with successful learning. More recently, Xie et al. (2026) reported sustained lexical development in an AI-feedback condition using pretest, immediate posttest, and delayed posttest essays; Wu et al. (2026) observed that process-oriented AI feedback was associated with stronger subsequent unaided transfer performance than product-oriented feedback. These studies use different interventions, outcome constructs, samples, and designs. They do not support a universal conclusion that AI either enhances or undermines learning, but they make the interpretation of score changes a concrete empirical and methodological issue.

This interpretive concern is also relevant to Vietnamese university EFL learning. In a six-month study of ChatGPT engagement and perceived cognitive load, Tran et al. (2026) reported proficiency-dependent differences in cognitive demand and within-group pretest–posttest improvements. As the authors acknowledged, their design lacked a non-AI control condition, so these observed gains were not isolated causal effects of ChatGPT. The example reinforces the importance of distinguishing evidence of improvement over time from evidence identifying the reason for that improvement.

The problem is not removed simply by adding a control group. Suppose learners are randomly assigned to receive an AI-supported intervention or a conventional alternative. If the intervention group completes the outcome essay with AI and the comparison group does not, a group difference in posttest scores contains both any true learning effect and the immediate contribution of the tool to the observed text. Randomisation removes systematic pretreatment differences in expectation; it does not convert a treatment-contaminated outcome into a direct measurement of independent proficiency. Likewise,

controlling for baseline writing scores can improve precision, but baseline adjustment cannot separately identify two components that enter the same posttest outcome.

This is an argument about the interpretation and use of scores, not a declaration that assisted writing products are intrinsically invalid. In Kane's (2013) argument-based view of validation, evidence must support the particular interpretation for which scores are used. An AI-assisted essay may be appropriate evidence of *assisted task performance* while providing insufficient evidence for a claim about *unaided proficiency growth*. Messick (1995) similarly emphasized the need to examine the meaning and consequences of score-based inferences. An intervention effect defined on assisted output is a legitimate estimand; the error occurs when it is reported as an intervention effect on autonomous writing competence.

A controlled simulation is suited to this question because the components of observed performance can be specified and the true learning effect can be known. Unlike a classroom study, which usually cannot observe a learner's latent writing ability or the exact number of score units contributed by AI, a simulation can vary these quantities independently. It can then test whether estimators recover the intended parameter, under what conditions confidence intervals attain their claimed coverage, and how frequently a study would report a positive learning effect when no AI-attributable learning occurred. The purpose is methodological evaluation rather than the creation of artificial evidence about learner behaviour.

The argument bears on the design of assessment evidence. Evidence-centred design asks investigators to connect the intended claim, the observable evidence, and the task situation that elicits it (Mislevy & Haertel, 2006). For independent L2 writing proficiency, a polished text created with unrestricted rewriting assistance may contain relevant information about task management and tool use but an insufficiently isolated demonstration of unaided linguistic control. Conversely, excluding AI from every assessment would overlook legitimate questions about performance in an authentic AI-enabled writing ecology. Researchers should select an assessment condition by working backwards from the claim: whether the claim concerns collaborative output quality, autonomous learning, transfer to a novel task, or sustained retention.

The manuscript addresses that choice at the level of statistical operating characteristics. It does not attempt to infer which assistance mechanism predominates in authentic EFL classrooms or to claim that any particular published study used an invalid outcome. Such claims require examination of task instructions, AI access, scoring rubrics, and evidence beyond a published summary. The narrower methodological question is whether, under known synthetic truth, commonly used analyses provide evidence for the specific learning parameter that investigators wish to report.

The present study contributes an explicit longitudinal distinction between the true endpoint learning effect, the subsequent unaided transfer effect, and a later retention effect. It tests three common modes of inference—within-group gain, controlled assisted-posttest analysis, and independent outcome assessment—and evaluates one possible correction using separately collected paired-mode calibration

data. It also examines the circumstances under which that correction fails. The resulting contribution is a conditional map of inferential risk, rather than an unconditional endorsement of a preferred estimator.

2. Conceptual and Analytical Background

2.1. Why product improvement and competence growth can diverge

An observed writing score summarises features of a product that may be generated through several routes. A learner might independently recall and deploy newly acquired structures; use real-time feedback to diagnose an error and revise it; or accept a complete reformulation supplied by AI. All three routes can raise the score of the submitted text. Their implications for subsequent independent performance differ because the latter depends on what the learner can retrieve, adapt, and produce without the external scaffold.

Feedback quality and uptake consequently matter to learning. Yan and Zhang (2024) showed that ChatGPT feedback engagement varied with learner proficiency and technological competence; Wu et al. (2026) connected process feedback with the distribution of pauses, revision decisions, and later unaided writing performance. These findings support treating learning as an outcome requiring evidence beyond a corrected product. At the same time, Xie et al. (2026) found improvements that persisted at delayed testing in lexical measures, illustrating that AI feedback may support acquisition rather than merely decorate a text. A defensible simulation therefore has to include both zero-learning and genuine-learning conditions.

Related methodological studies have treated distinct ways in which the apparent improvement of an AI-revised product may become misleading. A synthetic error-migration framework separated corrected, persistent, migrated, and revision-emergent errors, showing why a reduction in aggregate errors cannot characterise every underlying revision transition (Nguyen, 2026b). A proficiency-compression simulation examined whether increasing overall score quality can reduce information about differences among writers (Nguyen, 2026j), while a diagnostic-masking simulation focused on weaknesses no longer detected in assisted work despite their presence in independent writing (Nguyen, 2026i). Those studies establish different *assessment questions*, not interchangeable measures of learning: transitions among textual errors, between-writer differentiation, and detection of domain weaknesses do not estimate intervention-related growth in independent competence.

The observed assisted posttest outcome may be written schematically as the sum of latent independent performance, an immediate tool contribution, task-specific variation, and measurement error. This decomposition is a model, not a claim that all AI effects are additive in real writing. It is useful because it reveals a basic identification problem: one observed score alone cannot separate two unobserved contributors to that score. Additional assessments, constraints, or assumptions are required.

2.2. The estimand determines the validity claim

Consider three quantities that could each be called an intervention effect. The first is the effect of assignment to AI-supported instruction on an essay produced *with access to the intervention tool*. The second is the effect on independent writing competence at the end of instruction, even if the observed assisted essay is not a clean measure of that competence. The third is the effect on a new writing task conducted without AI sometime after instruction. The quantities answer different questions and can differ even in a perfectly randomised trial.

Prior work on product–process non-identifiability has likewise examined why comparable assisted writing scores can conceal differences in independent competence and learner–AI processes (Nguyen, 2026f). Other Monte Carlo studies have examined the attenuation of stylometric writer-identification signals after AI revision (Nguyen, 2026d) and the instability of learner ranking or selection decisions when assistance changes scores (Nguyen, 2026h). These studies concern identification, diagnostic inference, or relative standing at a measurement occasion. A separate simulation of AI-inflated self-presentation in L2 communication examined how language-mediated proficiency impressions might affect modeled interlocutor expectations and repair requests (Nguyen, 2026g). That communication study does not measure writing development, but it illustrates that AI-mediated output may influence interpretations extending beyond the immediate text. The present study adds a different inferential target: the causal effect on *independent competence over time*.

The present investigation targets the latter two quantities: endpoint learning and unaided transfer, with delayed retention evaluated separately. Importantly, a positive endpoint learning effect does not guarantee an equally large delayed effect. If learners acquire a skill during the intervention but cannot maintain or transfer it, post-intervention performance may decline without implying that no learning occurred at the earlier endpoint. Conversely, an assisted essay can appear markedly better even when the endpoint effect on independent competence is zero. Treating all these quantities as interchangeable introduces estimand drift, a problem distinct from low statistical power.

2.3. Why longitudinal modelling is not automatically sufficient

Repeated measurements provide information about change, but additional time points alone do not identify the direct assistance component if every relevant assessment remains assisted. A time-by-condition coefficient from a mixed-effects model can still describe a treatment-dependent *observed score trajectory* rather than a trajectory in independent competence. A model becomes informative about learning only through its design, its observed anchors, or defensible assumptions about how assistance affects each measurement occasion.

One practical possibility is to assess transfer without access to AI, making the outcome directly aligned with the intended construct. A second possibility is to collect parallel assisted and unaided writing from a separate, representative calibration sample and estimate the immediate mean tool contribution. Subtraction of that contribution from an assisted-outcome treatment contrast identifies the endpoint learning effect under particular conditions. These conditions are demanding: the paired tasks must be

appropriately equated; order or fatigue effects must be neutralised or modelled; the calibration sample must represent the intervention group; and the assistance effect must transport across tasks, raters, and settings. Calibration is therefore a testable design proposal, not a universal statistical fix.

2.4. Validation arguments, tasks, and change over time

The validity question is not whether AI contamination makes a numerical score inaccurate in the abstract. A score can be reliable, predictive, and faithfully calculated while still failing to warrant a particular inference about the writer. Kane (2013) treats validation as scrutiny of an interpretive argument: an observed response is scored; scores are generalised across relevant tasks; and an interpretation or decision is extrapolated to a broader construct. A supported writing score can enter a defensible argument for the capacity to produce a high-quality text with available assistance. When that score is offered as evidence of unaided language development, however, the extrapolation warrants independent support. The simulation isolates one form of this extrapolation failure by treating latent autonomous proficiency and immediate support as distinct constituents of the observed outcome.

The significance of assessment-task design becomes clearer longitudinally. At baseline, a learner may write independently; at the intervention endpoint, the same learner may have access to suggestions, full-sentence rewrites, or content generation; and at a subsequent transfer task, the assistance may be withdrawn. Even identical scoring rubrics do not guarantee comparable construct representation across these occasions. The evaluator may apply the same criteria for grammatical accuracy and coherence, yet the task condition changes whose resources can contribute to those qualities. Consequently, the meaning of a change in score depends on both the score scale and the support condition. A pretest–posttest contrast that ignores the second factor leaves an important part of the interpretation implicit.

The problem does not disappear when the posttest score predicts a delayed score. Correlations between repeated measures may be generated by stable learner proficiency, stable access to assistance, variation in individual use of feedback, or selection into later testing. To support an acquisition claim, investigators need some evidence that a relevant change persists beyond the particular assisted product. An unaided transfer task offers one such source, provided that its genre, difficulty, rubric, and administration conditions fit the intended generalisation. It is not a universal gold standard: a single novel prompt can be noisy, unfamiliar, or mismatched to instruction. The methodological implication is to distinguish *construct alignment* from *statistical precision* and to seek both.

2.5. The difference between an effect on treatment-defined outcomes and an effect on proficiency

An intervention may have multiple legitimate causal effects. Assignment to an AI-enabled writing condition can change the immediate quality of the completed essay, the strategies learners use during composition, the knowledge they retain for later use, and their willingness to write without assistance. None of these effects is logically reducible to the others. In the potential-outcomes vocabulary, an average treatment effect must specify both the intervention contrast and the outcome under

comparison. “Effect of AI on writing” is underspecified because it does not identify whether the outcome is a supported product or an unaided competence measure.

A randomised experiment can consistently estimate the effect of assigned support on an outcome measured under the corresponding assigned conditions. This is a strength, not a defect, of randomisation. The inferential mistake arises later, when the estimated effect on a compound outcome is rhetorically assigned to one of its unobserved constituents. Baseline adjustment likewise cannot distinguish product assistance from proficiency growth when both components influence the dependent variable. It reduces residual variation and adjusts chance baseline imbalance; it supplies no new observation of the counterfactual unaided posttest. This distinction helps prevent the present study from describing ordinary assisted-outcome ANCOVA as a generally defective statistical procedure. Under the specified simulation it performs well for the *supported-outcome* estimand and badly only if judged against the distinct *independent-learning* target.

External paired-mode calibration is one possible attempt to connect the two constructs. By observing the difference between assisted and unaided performance on otherwise comparable tasks, the investigator estimates the contribution attributable to assistance and subtracts it from the supported-outcome contrast. Such a correction trades one problem for additional assumptions: the paired modes must isolate assistance, parallel tasks must be comparable, and the average assistance contribution measured in the calibration sample must transport to the experimental population. If assistance benefits weaker and stronger writers differently, simply drawing a large calibration sample is insufficient when its proficiency distribution differs from that of the treatment group. The present study accordingly evaluates calibration as a conditional method, not as a general solution.

2.6. Feedback engagement, learning transfer, and the ambiguity of a revised text

The distinction between performance support and learning should not be interpreted as an assumption that AI-assisted feedback merely substitutes for learner effort. There are plausible pathways from feedback to learning: noticing a discrepancy between an intended expression and a suggested form, evaluating a revision against a communicative goal, rehearsing the form in subsequent drafts, and retrieving it later without the system. These pathways require learner engagement, and their effectiveness can differ by feedback type, proficiency, and task. Yan and Zhang (2024) examined behavioural, cognitive, and affective engagement with ChatGPT feedback; Wu et al. (2026) linked process-oriented feedback to pausing, higher-level revision, and subsequent unaided writing. Their empirical designs are not interchangeable with this simulation, but they make it inappropriate to impose a universal “no learning” assumption.

At the same time, an accurate correction may bypass the very processes intended for assessment. If a tool supplies a grammatical reformulation that the learner copies without evaluating it, the corrected product no longer uniquely indexes the learner’s independent control of that form. The author’s error-migration study demonstrated, within a synthetic framework, that overall improvement can coexist with

persistence, redistribution, and emergence of errors (Nguyen, 2026b); the product–process non-identifiability study likewise treated similar completed products as compatible with different underlying competence states (Nguyen, 2026f). Neither establishes how frequently such processes occur in real writing, but each motivates caution about substituting final-product quality for a developmental claim.

Qualitative evidence also indicates that noticing and later production should not be conflated. Sasaki and Takeuchi (2026) analysed learners' feedback histories, interviews, and revisions and found that noticed AI-proposed forms were not consistently deployed in later independent writing. Their findings do not estimate a general treatment effect, but they reinforce the importance of an assessment outcome that actually requires unaided retrieval. In the present simulation, the distinction is operationalised as separate supported and unaided score-generating processes rather than inferred from learners' reported experiences.

A delayed unaided assessment is valuable because it makes a narrower claim possible: whether the treatment-group advantage is still expressed in an assessment that does not offer the same immediate help. Yet transfer is not identical to the original learning process. A learner may acquire feedback-specific knowledge that transfers imperfectly to a new genre or task, or a delayed score may reflect ordinary practice and later instruction. In a controlled design, the relevant contrast is therefore a *between-group effect at the designated unaided occasion*, with baseline and task conditions appropriately handled, not merely an increase within the AI group. Delayed assessment adds evidence about durability, but its validity also depends on attrition, comparability, and the choice of a meaningful retention interval.

2.7. Related synthetic studies and the distinct methodological contribution

The related computational publications supply conceptually adjacent but distinct methodological arguments; they are not independent empirical confirmations of this simulation. Proficiency-compression work considers between-writer variance and also reports rank shifts as a secondary outcome (Nguyen, 2026j); the dedicated rank-reversal study places relative ordering and selection instability at the centre of its estimands (Nguyen, 2026h). Diagnostic masking instead concerns whether latent domain weaknesses cease to be detected in assisted products (Nguyen, 2026i), while stylometric signal decay concerns computational writer-identification rather than educational learning (Nguyen, 2026d). Product–process non-identifiability explicitly tests process evidence and noisy unaided anchors for recovering baseline independent proficiency (Nguyen, 2026f), so this paper does not claim that unaided anchoring is a novel concept. It asks a different, time-indexed question: whether an intervention effect on latent proficiency growth can be estimated from supported and unaided outcomes. The AI-inflated persona study concerns modeled expectations and conversational repair, not writing-growth outcomes (Nguyen, 2026g). Because these studies use distinct synthetic generators, their numerical results cannot be pooled as if they were replications of a common population effect.

The relevance of contextual and non-writing publications is narrower. A systematic review of context-based vocabulary teaching concerns learning through meaningful lexical use, providing a background

rationale for assessing generalisation to new writing contexts (Nguyen, 2026e). Research on followership and organisational commitment illustrates the need to distinguish predictive relationships from causal intervention effects (Nguyen, 2026c). A separate synthetic investigation of speech and lexical diversity demonstrates the importance of not collapsing distinct operational outcomes into one broad proficiency label (Nguyen, 2026a). These publications should not be interpreted as evidence for numerical parameters in the present experiment. They are included only to locate the shared methodological problem of mapping measurements to clearly specified constructs.

Bibliographic status matters for the comparison: the rank-reversal manuscript is treated as an author-supplied preprint, and the Research Square identifier requires final verification against an authoritative record before formal journal submission. Its conceptual contrast is used here without importing any unverified numerical claim from that record.

Together, the literature motivates a controlled evaluation of three questions: whether supported outcomes bias estimates of independent learning, whether calibration or unaided testing changes estimator performance, and whether additional design complications weaken otherwise favourable results. The contribution does not depend on demonstrating that AI is harmful. Its value comes from specifying conditions under which researchers may make a warranted claim about different kinds of improvement.

2.8. Research questions

RQ1. Under what combinations of immediate AI contribution and genuine L2 writing development do conventional pretest–posttest analyses overestimate learning gains?

RQ2. How accurately do alternative endpoint and unaided-outcome estimators recover their explicitly defined learning and transfer effects?

RQ3. How do calibration-sample size, measurement noise, and violations of task equivalence or calibration representativeness affect bias, precision, interval coverage, and positive-learning conclusions?

3. Methods

3.1. Design and reporting framework

This study was a fully synthetic Monte Carlo experiment. No learners were recruited, no essays were sampled, and no observed classroom results were represented as study data. The simulation was organised according to the ADEMP framework—aims, data-generating mechanisms, estimands, methods, and performance measures—proposed by Morris et al. (2019). Four primary scenarios were evaluated using 2,000 independent replications per scenario. Each replication contained 600 synthetic learners allocated in equal numbers to an AI-intervention condition ($Z = 1$) and a no-AI control condition

($Z = 0$). Exact balance was imposed to focus the experiment on outcome contamination rather than allocation imbalance.

Each simulated learner was observed at five occasions. *T0* was an unaided baseline assessment. *T1* represented an intermediate task with AI access available only to the intervention condition. *T2* was the supported endpoint assessment. *T3* was an independent near-transfer task completed without AI by both conditions. *T4* was a delayed independent assessment, also completed without AI. The labels represent assessment occasions, not a claim about any specific real-world intervention duration. A separate calibration sample contributed paired assisted and unaided writing performances at the endpoint but was not part of the 600-person trial, which preserved independence between the treatment-effect estimator and the calibration estimate.

The treatment indicator jointly represents assignment to the AI intervention and permission to use AI during the T1 and T2 outcome tasks. This is a deliberately bundled treatment, not a factorial randomisation that independently varies instruction and assessment tool access. The resulting supported-endpoint group contrast identifies an effect of the assigned assessment-and-intervention package on observed performance; the independent-learning component is known only because the generator specifies it, or is estimated under additional calibration assumptions.

The ADEMP components were specified before the add-on sensitivity analyses: the aim was to evaluate estimand alignment and robustness; the data-generating mechanisms were explicit synthetic models; the estimands concerned treatment-attributable latent growth at defined occasions; the methods included unadjusted, controlled, calibrated, and unaided-outcome analyses; and the performance measures summarised repeated-sampling error. Morris et al. (2019) recommend these distinctions because otherwise a simulation can conflate a procedure's valid target with the researcher's preferred interpretation. In the present case, the same supported-outcome estimator has almost no error for its own target and substantial apparent error when wrongly judged against an autonomous-learning estimand.

The original analyses were retained as a frozen baseline. A separate executable script, expanded_sensitivity.py, conducts new computational experiments using its own seed. The extension does not silently alter the four main scenarios, the initial cohort size, or the original 2,000-replication results. Each added scenario contains 1,200 repetitions of an independent 600-learner synthetic cohort, and any estimated calibration effect uses a separate sample of 120. This separation permits direct verification of the previously released operating characteristics and reduces the risk of attributing a result from a changed generator to the original experiment.

Figure 1. Study procedure for the synthetic Monte Carlo programme. The five assessments index simulated occasions; AI access at T1 and T2 was restricted to the intervention condition. Calibration learners were separate from the 600-person trial cohorts, and later robustness conditions were generated independently of the primary experiment.

3.2. Data-generating mechanism

For simulated learner i at assessment occasion t, the latent independent writing-proficiency trajectory was specified by the following data-generating equation:

$\theta_{it} = u_i + (0.12 + s_i)t + Z_i L_i g_t$

Here $u_i \sim N(0, 0.90^2)$ denotes pretreatment latent proficiency, $s_i \sim N(0, 0.065^2)$ is the individual deviation in ordinary growth per assessment occasion, Z_i is the AI-intervention assignment indicator (1 = AI; 0 = control), L_i is the person-specific AI-attributable independent-learning effect at T2, and g_t describes how this effect emerges and persists. The control group follows the same mean ordinary-growth process but has no AI-attributable learning term. These distributions and effects are illustrative generator inputs, not empirically calibrated population estimates.

The observed writing score was generated separately as:

$Y_{it} = \theta_{it} + d_t + \varepsilon_{it} + Z_i \delta_i \times 1(t = 1 \text{ or } 2)$

Here d_t is a common assessment-occasion/task shift; $\varepsilon_{it} \sim N(0, 0.40^2)$ is independently generated score/task error; δ_i is the immediate writing-score increment due to assistance, applied only at T1 and T2; and $1(\cdot)$ is an indicator function. Task offsets at T0–T4 were 0.00, 0.05, − 0.04, 0.07, and − 0.03, respectively. The offsets are common to both trial groups, so their between-group differences cancel in expectation. In the homogeneous primary scenarios, $\delta_i \sim N(0.50, 0.20^2)$ independently of proficiency. Each measurement remains continuous and unbounded; no operational IELTS band or bounded writing-rubric score is simulated.

The stipulated learning trajectory used $g_t = (0, 0.5, 1, r, 0.8r)$ at T0–T4, where r is the scenario-specific retention multiplier. AI-attributable learning is consequently half developed at T1, fully developed at T2, multiplied by r at the unaided T3 assessment, and reduced to 80% of its T3 component at delayed T4. The shape and the labels “near transfer” and “delayed” are simulation conventions; no calendar interval, task genre, or validated cognitive theory of retention was specified.

Table 1 specifies the four primary scenarios. In the heterogeneous scenario, $L_i = 0.20 + 0.12u_i + \eta_i$, with $\eta_i \sim N(0, 0.12^2)$, while $\delta_i = 0.50 - 0.15u_i + \xi_i$, with $\xi_i \sim N(0, 0.16^2)$. Independent η_i and ξ_i are distinct noise terms. In that scenario, lower baseline proficiency tends to accompany larger immediate assistance gains, while higher baseline proficiency tends to accompany greater AI-attributable learning. Heterogeneity changes conditional effects without changing their target-population means of 0.20 and 0.50.

Table 1
Primary simulation scenarios and stipulated effects

Scenario	True endpoint gain	Transfer retained	Delayed retained	AI contribution
A. Performance only	0.00	0.00	0.00	0.50
B. Retained learning	0.20	0.18	0.144	0.50
C. Fading learning	0.20	0.05	0.040	0.50
D. Heterogeneous	0.20	0.15	0.120	0.50 (ability-linked)

Note. Values are population-level standardized effects stipulated by the data-generating mechanism. The delayed effect equals 0.8 × transfer effect.

3.3. Target parameters and alignment of outcomes

The primary endpoint estimand is $\tau_2 = E[\theta_{i2}(1) - \theta_{i2}(0)] = E[L_i]$ under the stipulated generator, measuring AI-attributable independent-proficiency change at T2. The transfer estimand is $\tau_3 = rE[L_i]$ at unaided T3, and the delayed estimand is $\tau_4 = 0.8rE[L_i]$ at unaided T4. These are time-specific, population-average contrasts in independent latent competence under intervention and control assignment. Across the four primary scenarios, τ_2 equals either 0 or 0.20; τ_3 and τ_4 differ according to scenario-specific retention.

The separate average effect of assignment on the supported T2 writing score is $\tau_2 + E[\delta_i]$ under the additive generator. This is a valid supported-outcome estimand, not an unbiased estimate of τ_2. Accordingly, when an assisted-outcome method is compared to τ_2, the reported "bias" is an estimand-mismatch diagnostic or a hypothetical error from relabelling its effect as independent learning—not a failure to estimate the assisted-outcome target. The distinction is prespecified and retained in every result and table interpretation.

3.4. Analysis methods

Six estimators were examined. First, the one-group gain estimator was the mean difference between *T2* and *T0* in the AI-assigned condition; its ordinary paired standard error was calculated using the across-learner standard deviation of individual changes. This estimator deliberately omitted the control group and was expected to include ordinary longitudinal development, task shift, and direct assistance.

Second, assisted difference-in-differences (DiD) compared mean *T2–T0* changes in the two assignment conditions. Third, assisted ANCOVA estimated the coefficient for AI assignment in an ordinary least squares regression of the *T2* observed score on AI assignment and the baseline observed score. Both controlled contrasts were anticipated to recover the total effect on assisted output, not independent learning.

Fourth, calibrated ANCOVA subtracted an externally estimated mean direct assistance contribution from the assisted ANCOVA coefficient. Its independent calibration sample consisted of 120 additional

synthetic learners drawn from the same AI-exposed target population. Each contributed two parallel endpoint scores, one assisted and one unaided. Their within-person assisted-minus-unaided difference estimated the mean immediate tool contribution. In the principal scenarios, the tasks were equivalent in expectation and their independent errors had the same variance as trial tasks. Because the calibration sample was separate from the trial sample, the calibrated estimator's standard error was obtained as the square root of the sum of the ANCOVA variance and the paired-calibration mean variance. This calculation would not be valid without adjustment if the same learners contributed to both samples.

For transparency, the assisted difference-in-differences coefficient was calculated as [(mean Y_AI,T2 − mean Y_AI,T0) − (mean Y_control,T2 − mean Y_control,T0)]. The calibrated endpoint estimate was $\hat{\tau}_2$,cal = β_assisted,T2 − δ_cal, where δ_cal is the mean within-calibrator assisted-minus-unaided difference. Its model-based standard error was SE_cal = √(SE²_assisted + SE²_δcal) because trial and calibration samples were generated independently. These algebraic definitions are part of the executable specification and identify the estimates reported in Tables 2 and 3; they are not additional fitted models.

Fifth and sixth, baseline-adjusted ANCOVA used the unaided T3 and T4 scores to estimate their respective transfer and delayed treatment effects. These were evaluated against τ_3 and τ_4, not against the earlier endpoint effect τ_2. A smaller T3 estimate in the fading-learning scenario therefore reflects a genuinely smaller time-specific effect in the generator; it is not a statistical failure of an estimator targeting T3.

Table 2 aligns the six analysis methods with the parameter they were evaluated against, the information required, and the central inferential concern.

Table 2
Methods, targets, and identifying information

Method	Target judged against	Data required	Primary interpretation issue
AI-only gain	T2 independent learning	AI T0 and assisted T2	Also contains ordinary growth, task shift and assistance
Assisted DiD	T2 independent learning	Both groups T0/T2	Cancels common secular trend, not direct AI support
Assisted ANCOVA	T2 independent learning	Both groups T0/T2	Adjusts baseline, not support contamination
Calibrated ANCOVA	T2 independent learning	Assisted ANCOVA + independent paired calibration	Depends on parallel-task equivalence and representative calibration
Unaided T3 ANCOVA	T3 independent transfer	Both groups T0/T3; no AI at T3	Measures transfer rather than immediate endpoint learning
Unaided T4 ANCOVA	T4 delayed independent effect	Both groups T0/T4; no AI at T4	Measures delayed retention, not an immediate effect

The decomposition has an immediate identification consequence. At the supported endpoint, the outcome coefficient identifies the average sum of the true latent treatment effect and the immediate assistance effect. Without further observations or externally justified restrictions, the two components cannot generally be recovered from one observed coefficient. A calibrated estimator becomes identified only by adding a valid estimate of the average assistance contribution; an unaided assessment instead identifies a different time-specific treatment effect. This is why the analysis does not attempt to deduce a latent learning trajectory from a single assisted endpoint score, even when repeated baseline information is available.

All ANCOVA models in the original program use a treatment indicator and the observed baseline score as predictors, with a common coefficient on baseline. The confidence interval is constructed using a normal multiplier of 1.96 and an ordinary homoskedastic OLS residual variance estimator. The calibration sample is statistically independent of the trial, so the squared standard error of the correction is the sum of squared standard errors. Under an internally shared calibration sample, this simple expression would be incorrect unless the covariance between estimates were incorporated. The simulation does not estimate a fully latent longitudinal mixed model and does not claim that one is required to remove immediate assistance effects. Introducing latent-model parameters without an additional identification source would merely reparameterise the same ambiguity.

3.5. Simulation performance measures

For each estimator and each scenario, the simulation estimated mean estimate, bias (mean estimate minus the appropriate true effect), root mean squared error (RMSE), and empirical coverage of a nominal 95% Wald interval calculated as estimate ± 1.96 estimated standard errors. We also calculated the proportion of replications in which the lower endpoint of this interval exceeded zero. This was labelled a *statistically positive learning claim*. Under a true zero effect and a well-calibrated two-sided 95% interval, its nominal positive-tail frequency is approximately 2.5%, not 5%; the remaining 2.5% corresponds to significant negative estimates. Under a positive true effect, the same measure was interpreted as detection probability for that specified effect, not a false-positive rate.

Monte Carlo uncertainty was quantified by the standard error of the simulated mean estimate and the binomial standard error of estimated coverage. For a 95% coverage rate across 2,000 replications, the Monte Carlo standard error is approximately 0.005. Normal Wald intervals were chosen to reflect conventional practice under the deliberately Gaussian data-generating mechanism; these were not offered as a general substitute for sample-size-appropriate or heteroskedasticity-robust inference in empirical studies.

3.6. Sensitivity experiments

Four sensitivity families examined the robustness of the main findings. First, the direct assistance contribution was set to 0.15, 0.30, 0.50, or 0.80 under the zero-learning scenario. Second, the independent calibration sample size varied across 30, 60, 120, and 300 in both the zero-learning and

retained-learning scenarios. Third, the individual-level task-error standard deviation was set to 0.25, 0.40, or 0.70 in the retained-learning scenario.

The final family deliberately violated calibration assumptions. In a task-mismatch experiment, the mean unaided score on the calibration task was elevated by 0.10 or 0.20 standardised units relative to the parallel assisted task for reasons unrelated to AI assistance. Consequently, subtracting the observed paired difference under-estimated the real AI contribution. A separate transportability experiment used the heterogeneous-learning model but shifted mean baseline ability of the calibration sample upward by 0.50 or 1.00 units. Because direct assistance decreased with ability in this scenario, nonrepresentative calibration systematically under-estimated the assistance contribution among the trial participants. The values were stress-test settings, not estimates drawn from published learner populations.

The primary rate of an apparently positive learning claim is calculated as the proportion of replications with a lower two-sided 95% Wald interval bound greater than zero. This rule has a nominal *positive-tail* probability of approximately 2.5% for a correctly calibrated zero-mean effect under the normal approximation. It is not an unrestricted two-sided Type I error rate of 5%. When the treatment genuinely affects supported output but not independent learning, the associated supported-outcome confidence interval is not making a statistical mistake about the supported-outcome coefficient. The false claim occurs when its result is interpreted as evidence of positive independent learning. Accordingly, we use *interpretive false-positive learning claim* for this mismatch and reserve statistical Type I error terminology for correctly specified null tests of their own target parameters.

Monte Carlo error is an additional source of uncertainty. For an estimated mean coefficient across replications, its Monte Carlo standard error equals the empirical standard deviation divided by the square root of the number of replications. For a proportion, the binomial approximation is the square root of p(1-p)/R. With R = 2,000, a true event probability of .025 yields a Monte Carlo standard error of approximately .0035, or 0.35 percentage points; with R = 1,200, this is approximately 0.45 percentage points. Numerical differences smaller than these scales should not be overinterpreted. Coverage is evaluated against the declared independent-learning target in the original tables; that is a purposeful construct-alignment evaluation and must not be mistaken for coverage of the supported-outcome treatment effect.

3.7. Computation and validation

The simulation was implemented in Python with NumPy and pandas. A fixed pseudorandom seed of 20261116 was used. Ordinary least squares was computed from the same design matrix as the stated regression models, with the conventional model-based standard error; paired calibration variances were estimated from the simulated score differences. The code exported complete aggregate results, parameter metadata, publication figures, and one illustrative cohort-level synthetic dataset. The reproducibility package accompanies the manuscript; the illustrative cohort is not the dataset from which all Monte Carlo summaries were obtained. All numeric Results below were transcribed from the executed program outputs.

The construction supplies analytic checks on implementation. In the zero-learning scenario, the expected AI-only gain is 0.12 × 2 − 0.04 + 0.50 = 0.70, while the expected assisted between-group contrast is 0.50. Under equivalent and representative calibration, subtracting 0.50 gives an expected endpoint learning estimate of zero. Similarly, in the retained-learning scenario, the known learning effects are 0.20 at *T2*, 0.18 at *T3*, and 0.144 at *T4*. These values were used as theoretical comparison points, not estimated from the simulated outputs.

3.8. Expanded robustness experiments

The extension examines four departures from the baseline model. They are informative stress tests, not estimates calibrated from real EFL datasets. The first replaces the direct assistance term with a proficiency-dependent nonlinear function, $\delta_i = 0.50 - 0.15u_i + 0.10(u_i^2 - 0.90^2) + \eta_i$, where u_i follows a normal distribution with mean zero and standard deviation 0.90 and η_i has standard deviation 0.16. Its population mean equals 0.50 in the target distribution. The genuine endpoint learning effect is generated as $L_i = 0.20 + 0.12u_i + \xi_i$, with ξ_i normally distributed at standard deviation 0.12; its mean remains 0.20. An independent calibration sample is drawn with the same variance but a proficiency-location shift of 0, 0.5, or 1.0. The calibrated estimate subtracts the unconditional mean paired-mode difference. If the calibrators are shifted relative to the trial learners, the average observed support component need not represent the target population. This is a deliberately simple test of transportability under nonlinear heterogeneity.

The nonlinear calibration mechanism also supplies an analytic expectation check. For a calibrator proficiency shift a, the expected direct assistance contribution changes from 0.50 in the target population to $0.50 - 0.15a + 0.10a^2$. The resulting target-minus-calibration difference is $0.15a - 0.10a^2$, implying reference endpoint biases of 0.00, + 0.05, and + 0.05 for shifts a = 0, 0.5, and 1.0, respectively. These expectations are derived from the stipulated generator and were used to audit, not fit, the corresponding Monte Carlo outcomes.

The second departure introduces within-learner correlation across assessment scoring errors while preserving their marginal standard deviation of 0.40. For a chosen correlation $\rho \in \{0, 0.35, 0.70\}$, errors are generated as $e_{it} = \sqrt{\rho}c_i + \sqrt{1 - \rho}v_{it}$, with independent normal terms c_i and v_{it} each having standard deviation 0.40. These scenarios test whether the point-estimate distinction survives a common writer-specific component in measurement errors. The treatment and learning means match the retained-learning primary scenario, and paired external calibration remains independent. The add-on evaluates empirical bias and root mean squared error rather than extending the original confidence-interval procedure to an untested correlation model.

The third robustness family introduces missing unaided transfer outcomes at T3. Every synthetic learner has a generated T3 score, but only records with response indicator $R_i = 1$ enter complete-case analyses. Under MCAR, $P(R_i = 1) = 0.72$. Under MAR, $P(R_i = 1) = \text{logistic}[1.1 - 0.80\, Z_i(Y_{i2} - Y_{i0}) + 0.15Y_{i0}]$; assignment, baseline, and the supported T2 score are all observed for every simulated learner. Under MNAR, $P(R_i = 1) = \text{logistic}[1.0 - 1.40\, Z_i(Y_{i3} - Y_{i0}) + 0.15Y_{i0}]$; the response probability now depends directly

on the unobserved T3 outcome for nonrespondents. The distinct mechanisms also generate different overall and arm-specific response rates. These are deliberately specified stress tests, not models estimated from actual student attrition.

The initial extension compared full-data ANCOVA, complete-case ANCOVA, and oracle inverse-probability-weighted (IPW) ANCOVA using each individual's true simulated response probability. The oracle estimator isolates selection error when those probabilities are known; it is not a feasible MNAR correction. An independent validation experiment, using seed 20261119 and 1,200 additional replications under each mechanism, also fitted logistic response-propensity models using only observed assignment, Y(T0), and Y(T2). It compared (a) a fully specified observed-data predictor set, including assignment × (T2 − T0), with (b) a reduced model omitting that interaction. Each fitted probability was truncated to [0.025, 0.975]; observed T3 cases were weighted by $R_i/\hat{p}_i$ in baseline-adjusted weighted least squares. No unobserved T3 values entered feasible propensity estimation. The original validation established point-estimator behaviour only; the separately seeded V5 experiment examines interval inference and remains subject to standard modelling assumptions.

Inverse-probability weighting was chosen as a transparent demonstration of observation-mechanism assumptions, not as a universal preferred analysis. Seaman and White (2013) explain that IPW requires careful specification of the response model and attention to unstable weights, and they distinguish it from alternatives such as multiple imputation and augmented weighting. Here the controlled generators permit direct evaluation of a fitted response model against an unavailable oracle benchmark. The new interval experiments further assess whether uncertainty accounting for estimated nuisance parameters behaves as intended.

For response indicator R_i and response probability p_i, the weighted analysis uses observation weight R_i/p_i; feasible weighting replaces p_i with an estimate based on fully observed data. The fitted logistic model includes an intercept, assignment Z_i, baseline Y_{i0}, and $Z_i \times (Y_{i2} - Y_{i0})$. This fully observed predictor set contains the programmed MAR mechanism, whereas the MNAR mechanism depends on missing T3 outcomes. Oracle MNAR weights remain an unreachable benchmark. The original V3 validation reported point-estimator bias and RMSE; the separate V5 experiment below computes variance and interval properties for the same type of fitted estimator without altering the previous outputs.

The fourth departure isolates the task-comparability problem in external calibration. A difficulty mismatch of 0.20 standardized units systematically shifts the assisted-minus-unaided paired difference when everyone receives the same task ordering. A counterbalanced synthetic calibration assigns half the paired assessments to the positive task difference and half to the negative difference, so its mean task contribution is zero by construction. Comparing the two schemes distinguishes the bias from systematic ordering with the cost of additional within-sample variation. Counterbalancing is not guaranteed to resolve carryover, practice, or prompt-by-treatment interactions in authentic assessments; the modeled advantage is conditional on a symmetric additive task effect.

3.9. Additional reviewer-directed inference and calibration experiments

To evaluate uncertainty for fitted-propensity IPW, a new independent run generated 1,200 synthetic cohorts per MCAR, MAR, and MNAR condition (n = 600; random seed 20261120). The observation mechanisms, retained-learning data generator, and observed-data logistic predictors matched the earlier missingness validation, but the cohorts were newly generated. For each cohort, a logistic regression estimated observation probability $\hat{p}_i$ from $X_i = (1, Z_i, Z_i[Y_{i2} - Y_{i0}], Y_{i0})$, and weighted ANCOVA estimated the T3 assignment contrast using $W_i = (1, Z_i, Y_{i0})$ and weight $R_i/\hat{p}_i$. As in the earlier experiment, fitted probabilities were bounded to [0.025, 0.975] for numerical stability; clipping was inactive in the generated cohorts according to the output diagnostics.

The primary uncertainty estimator used joint M-estimation scores $U\alpha_i = X_i(R_i - \hat{p}_i)$ for the logistic nuisance parameter and $U\beta_i = (R_i/\hat{p}_i)W_i(Y_{i3} - W_i\beta)$ for the weighted outcome coefficients. The joint empirical bread used the derivative of both score systems, including the effect of estimated logistic coefficients on outcome weights. A sandwich covariance was constructed from the empirical outer products of individual stacked scores, with the treatment-coefficient square root used as the standard error. Intervals were $\beta Z \pm 1.96SE$. For comparison, a fixed-weight sandwich treated the fitted weights as known and omitted the nuisance-estimation contribution. Formula and implementation were checked against numerical finite differences of the stacked estimating equations (maximum Jacobian discrepancy 5.5×10^{-12} in the test cohort). These are large-sample Wald intervals; no finite-sample exactness is asserted.

The second new experiment examined the limits of counterbalanced paired calibration (1,200 replications per design; m = 120 independent calibrators; seed 20261121). A symmetric assessment-period effect increased performance on whichever task occurred second by 0.10 units. The balanced design assigned exactly half of calibrators to assisted-first and half to unaided-first ordering; a comparator used assisted-first for everyone. Directional carryover $c \in \{0, 0.10, 0.20\}$ improved the subsequent unaided score only when AI assistance occurred first. Thus the assisted-minus-unaided paired difference included − c among assisted-first calibrators, giving an expected calibrated-endpoint bias of c/2 in the balanced design, but 0.10 + c in the fixed-order design. A separate task-by-mode perturbation $h = \pm 0.10$ shifted the paired assistance contribution regardless of order, predicting calibrated-endpoint bias − h even with counterbalancing. These are explicitly stipulated measurement violations, not observed learner carryover.

The two independent reviewer-directed experiments add operating-characteristic checks rather than altering earlier results. The fitted-IPW analysis reports empirical standard deviation, mean estimated standard error, 95% interval coverage, Monte Carlo coverage uncertainty, and the frequency of strictly positive lower confidence limits. The calibration analysis reports mean estimate, bias, RMSE, Monte

Carlo standard error, and the analytically predicted bias. The response-model target remains $\tau_3 = 0.18$ in the retained-learning scenario; the calibration target remains $\tau_2 = 0$ in the performance-only scenario.

3.10. Reproducibility, code-level checks, and analysis boundaries

The main and extension programs are independent and reproducible from stated fixed seeds (20261116 and 20261117, respectively). The package distinguishes primary_results.csv, sensitivity_results.csv, and expanded_sensitivity_results.csv. These original files remain unchanged. The independently seeded fitted-response experiment (20261119) and reviewer-directed interval and calibration experiments (20261120 and 20261121) add new files instead of overwriting prior numeric outputs. Separate diagnostic files report retention, propensity-score distributions, the frequency of truncation, and calibration-design operating characteristics.

An additional source of validity is an analytic expectation check: in the no-learning scenario with mean direct assistance 0.50, the supported-outcome contrast should be approximately 0.50, while the independent-learning effect should be zero. In the retained-learning scenario, the supported-outcome effect should be approximately 0.70, while true endpoint learning is 0.20 and unaided transfer is 0.18. Under representative nonlinear calibration, the mean of the centred quadratic term is zero; under a shifted calibration distribution, its expected value increases by the squared shift. These expectations give independent checks on sign and magnitude rather than relying only on whether the program executes without error.

The add-on file missingness_validation_v3.py provides the original feasible MAR point-estimator validation, while v5_inference_robustness.py performs the subsequent stacked-sandwich uncertainty and directional calibration-carryover experiments. None changes the source V1–V4 random draws. The response models are refitted separately within each synthetic cohort; V5 interval variances include estimation of the logistic response model. These tests are post hoc methodological sensitivity analyses, not preregistered or empirical classroom findings.

Reproduction requires Python 3.10 or later and compatible NumPy, pandas, and Matplotlib versions. The main simulation, expanded sensitivity, feasible-missingness, and reviewer-directed scripts are simulation.py, expanded_sensitivity.py, missingness_validation_v3.py, and v5_inference_robustness.py, with seeds stated above. The illustrative_synthetic_cohort.csv contains one generated cohort, not the 2,000 or 1,200 replicate datasets pooled together. The archive includes scripts, aggregate outputs, diagnostics, and an explicit execution order; any public repository record remains to be approved by the author.

The original generator fixes exact arm sizes (300 per condition), imposes a common task offset at each occasion, and treats the synthetic latent proficiency scale as continuous and unbounded. These conventions make the experiment transparent but are not the same as a validated operational scoring rubric. The extension preserves the standardized analysis scale to avoid confounding source-scale

changes with robustness mechanisms. No bootstrap confidence intervals, multilevel latent-state estimator, or multiple-imputation analysis was introduced merely for completeness; those methods would require additional implementation and verification beyond the calculations actually reported here.

4. Results

4.1. Primary estimator performance

Table 3 contains the primary Monte Carlo results across all four scenarios. The observed means followed the theoretical contrasts closely, indicating that the implementation recovered the expected decomposition between ordinary development, direct assistance, and independently retained learning. Most importantly, the difference between conventional assisted-outcome analyses and methods targeting independent writing was not a minor matter of model specification. It was a difference in the quantity being measured.

Table 3
Operating characteristics across primary Monte Carlo scenarios

Sc.	Estimator	True	Mean	Bias	RMSE	95% cover	Pos. claims
A	AI gain	0.000	0.698	0.698	0.699	0.0	100.0
A	Assisted DiD	0.000	0.499	0.499	0.501	0.0	100.0
A	Assisted ANCOVA	0.000	0.498	0.498	0.501	0.0	100.0
A	Calibrated ANCOVA	0.000	-0.003	-0.003	0.073	94.7	2.5
A	Unaided T3 ANCOVA	0.000	-0.001	-0.001	0.046	96.0	1.9
A	Unaided T4 ANCOVA	0.000	-0.001	-0.001	0.050	94.2	2.6
B	AI gain	0.200	0.901	0.701	0.702	0.0	100.0
B	Assisted DiD	0.200	0.702	0.502	0.504	0.0	100.0
B	Assisted ANCOVA	0.200	0.702	0.502	0.504	0.0	100.0
B	Calibrated ANCOVA	0.200	0.201	0.001	0.071	95.6	79.5
B	Unaided T3 ANCOVA	0.180	0.183	0.003	0.047	94.8	97.7
B	Unaided T4 ANCOVA	0.144	0.146	0.002	0.049	95.2	84.7
C	AI gain	0.200	0.901	0.701	0.702	0.0	100.0
C	Assisted DiD	0.200	0.702	0.502	0.504	0.0	100.0
C	Assisted ANCOVA	0.200	0.702	0.502	0.504	0.0	100.0
C	Calibrated ANCOVA	0.200	0.201	0.001	0.073	95.0	78.6
C	Unaided T3 ANCOVA	0.050	0.051	0.001	0.047	95.1	20.5
C	Unaided T4 ANCOVA	0.040	0.042	0.002	0.048	95.9	13.0

Note. N = 600; 2,000 replications per primary scenario. Coverage and positive claim rates are percentages; true learning effect = 0 in Scenario A. Positive claims with nonzero true effects are detection probabilities. “95% cover” evaluates the stipulated independent-competence target, not the valid causal effect on supported output.

Sc.	Estimator	True	Mean	Bias	RMSE	95% cover	Pos. claims
D	AI gain	0.200	0.899	0.699	0.700	0.0	100.0
D	Assisted DiD	0.200	0.699	0.499	0.501	0.0	100.0
D	Assisted ANCOVA	0.200	0.699	0.499	0.501	0.0	100.0
D	Calibrated ANCOVA	0.200	0.201	0.001	0.073	94.8	79.0
D	Unaided T3 ANCOVA	0.150	0.150	0.000	0.047	95.5	87.1
D	Unaided T4 ANCOVA	0.120	0.119	-0.001	0.051	94.8	66.1

Note. N = 600; 2,000 replications per primary scenario. Coverage and positive claim rates are percentages; true learning effect = 0 in Scenario A. Positive claims with nonzero true effects are detection probabilities. “95% cover” evaluates the stipulated independent-competence target, not the valid causal effect on supported output.

The nominal 95% coverage of calibrated and unaided estimates was generally close to 95% under the maintained simulation assumptions, while coverage of assisted-outcome estimates for the *independent learning* target was effectively zero at the chosen direct-assistance magnitude. This was not because the assisted analyses fitted the observed scores poorly. They were estimating an effect on supported performance and were compared, intentionally and transparently, to an independent-competence estimand that those scores did not identify.

4.2. No true learning: conventional analyses produce positive learning claims

In Scenario A, assignment to AI-supported instruction generated zero additional independent-proficiency growth at all assessments ($\tau_2 = \tau_3 = \tau_4 = 0$). Nevertheless, the AI-group T2-minus-T0 observed gain averaged 0.698 standardized units. Its difference from τ_2 is 0.698, although this within-group contrast was never a controlled learning-effect estimator. Adding the control group removed ordinary common growth but not immediate assistance: assisted DiD averaged 0.499 and assisted ANCOVA 0.498, both close to the valid supported-outcome target of 0.50. All 2,000 replications returned positive intervals for these supported-outcome contrasts; treating those results as learning effects would therefore produce 100% positive mislabelled learning claims under this generator.

The independent calibration estimator averaged − 0.003, with RMSE = 0.073, empirical 95% coverage = 94.7%, and a positive-claim rate of 2.5%. Unaided transfer ANCOVA averaged − 0.001, RMSE = 0.046, coverage = 96.0%, and a positive-claim rate of 1.9%. Delayed unaided ANCOVA showed a similar pattern. The control of false learning claims was therefore restored by designs that measured or explicitly removed the direct assistance component, provided their identifying assumptions held.

Figure 2 displays mean estimated effects, typical model-based uncertainty, and the scenario-specific theoretical target effects. The comparison shows that the large assisted-outcome estimates persisted across all scenarios, whereas calibrated and unaided outcomes tracked the relevant independent-competence targets.

4.3. Genuine retained learning: corrected and unaided methods recover different valid effects

In Scenario B, the endpoint AI-attributable independent learning effect was 0.20, with 90% of that contribution retained at unaided *T3* and 72% at delayed *T4*. The AI-only gain averaged 0.901 and assisted ANCOVA averaged 0.702; both markedly exceeded the endpoint learning effect because of ordinary development in the first case and direct tool contribution in both cases. These estimates would be defensible as descriptions of supported outcomes, but not as estimates of unaided competence gains.

Calibrated endpoint ANCOVA averaged 0.201 (bias = 0.001, RMSE = 0.071, coverage = 95.6%). Its rate of statistically positive learning claims was 79.5%. Unaided transfer ANCOVA averaged 0.183 against a true transfer effect of 0.180 (coverage = 94.9%, positive-claim rate = 97.7%). Delayed ANCOVA averaged 0.146 against 0.144 (coverage = 95.2%, positive-claim rate = 84.7%). These detection probabilities differed because their estimands, assessment-noise structures, and implied standard errors differed. A higher detection probability for transfer ANCOVA did not establish superiority for estimating the earlier endpoint effect.

4.4. Fading learning and heterogeneity

Scenario C had the same genuine endpoint learning effect as Scenario B (0.20), but the retention parameter was reduced to 0.25. Calibrated endpoint ANCOVA still recovered the immediate learning effect (mean = 0.201; coverage = 95.0%). Unaided transfer ANCOVA instead recovered the smaller true transfer effect of 0.050 (mean = 0.051; coverage = 95.1%), and delayed ANCOVA recovered 0.040 (mean = 0.042; coverage = 95.9%). Statistically positive claims were detected in 78.6% of calibrated endpoint analyses but only 20.5% of unaided transfer analyses and 13.0% of delayed analyses. The contrast shows why a defensible endpoint learning conclusion and weak evidence of durable transfer can coexist. Describing these as contradictory findings would confuse different target effects.

Scenario D included heterogeneity tied to baseline proficiency. The average direct assistance effect remained approximately 0.50, while the average true learning effects were 0.20 at endpoint, 0.15 at transfer, and 0.12 at delay. Calibrated endpoint ANCOVA averaged 0.201 (coverage = 94.8%), unaided transfer ANCOVA 0.150 (coverage = 95.6%), and delayed ANCOVA 0.119 (coverage = 94.8%). Thus, under representative calibration and the specified linear observation model, heterogeneity of the form generated here did not itself invalidate average-effect estimation. These aggregate results should not be mistaken for accurate recovery of individual learning trajectories, which was not a study aim.

4.5. Magnitude of assistance and false-positive conclusions

Table 4 and Fig. 3 show the results when the true learning effect was fixed at zero while the direct assistance contribution varied. Even with a relatively modest mean assistance effect of 0.15, assisted ANCOVA produced statistically positive independent-learning claims in 88.7% of replications. When the mean assistance effect was 0.30 or greater, the corresponding rate reached 100% at the simulated sample size. The calibrated endpoint and unaided transfer analyses remained near the nominal 2.5% positive-tail frequency across the tested values. These patterns reinforce that a sample-size increase alone cannot eliminate systematic estimand contamination; it can instead make a contaminated contrast increasingly precise.

Table 4
False-positive learning claims as AI direct assistance increases

Mean AI direct effect	Assisted ANCOVA bias	Assisted positive %	Calibrated positive %	Unaided T3 positive %
0.15	0.149	88.7	2.4	2.1
0.30	0.298	100.0	2.5	2.1
0.50	0.501	100.0	2.7	2.1
0.80	0.801	100.0	2.9	2.2

Note. True AI learning effect is zero in all rows. The nominal positive-tail rate under a two-sided 95% interval is 2.5%.

4.6. Calibration sample size and scoring noise

The accuracy of calibration depended on the information available to estimate the direct tool contribution. Under the retained-learning scenario, increasing the external calibration sample from 30 to 300 reduced the calibrated endpoint RMSE from 0.121 to 0.057 and increased the frequency of statistically positive endpoint-learning claims from 39.3% to 92.8%. Bias remained negligible across these sample sizes; the improvement was primarily one of precision and detection probability. By comparison, the unaided transfer estimator did not use calibration data and therefore had nearly unchanged performance across the calibration-size manipulations. Table 5 shows the pattern for both the no-learning and retained-learning scenarios.

Table 5
Calibration sample size and endpoint estimator performance

Calibration n	Scenario	Bias	RMSE	95% cover	Positive %
30	No learning	-0.004	0.119	94.5	2.5
30	Retained learning	-0.000	0.121	94.4	39.3
60	No learning	0.002	0.092	94.5	2.6
60	Retained learning	-0.001	0.094	94.6	59.2
120	No learning	0.002	0.069	95.7	2.4
120	Retained learning	-0.000	0.072	95.2	78.3
300	No learning	-0.002	0.059	95.8	1.9
300	Retained learning	-0.000	0.057	95.2	92.8

Note. The true endpoint effect is 0.00 in the no-learning case and 0.20 in the retained-learning case. Positive % represents false positive claims only in the former.

Figure 4 illustrates the diminishing precision gains as calibration sample size increased. Larger calibration samples did not solve the problems explored in the assumption-violation analysis below; they merely estimated a potentially wrong calibration contrast with greater precision.

When scoring/task-error standard deviation increased from 0.25 to 0.70 in the retained-learning scenario, calibrated endpoint RMSE increased from 0.048 to 0.117 and its detection probability fell from 98.8% to 39.0%. Unaided transfer RMSE increased from 0.032 to 0.076, with detection probability declining from 100% to 66.6%. This sensitivity confirms that independent outcome assessment is not immune to low measurement precision: it addresses the specified assistance contamination but still requires reliable tasks and ratings.

4.7. Failure of calibration when its assumptions are violated

The final sensitivity experiments produced the clearest qualification of the calibration proposal. In the no-learning scenario, a mean calibration task advantage of 0.10 in favour of the unaided version caused the calibrated endpoint estimator to have a bias of 0.099, interval coverage of 71.6%, and a false positive-learning claim rate of 28.4%. Increasing this mismatch to 0.20 produced bias = 0.200, coverage = 20.8%, and positive-claim rate = 79.2%. These failures arose because the paired difference mixed direct AI assistance with non-AI task differences.

The transportability experiment exposed a related problem. When AI's direct assistance contribution decreased with learner ability but the calibration sample's mean ability exceeded that of the trial population by 0.50 standardised units, calibrated endpoint bias rose to 0.078 and coverage fell to 80.1%.

A shift of 1.00 produced bias = 0.151 and coverage = 45.1%. The true endpoint learning effect remained 0.20; the corrected estimate was inflated because an unrepresentative calibration sample understated the assistance experienced in the target group. These results are summarized in Table 6.

Table 6
Sensitivity to violations of calibration assumptions

Violation	Size	Bias	95% cover	Positive %
Calibration task mismatch	0.00	0.001	94.0	3.1
Calibration task mismatch	0.10	0.099	71.7	28.3
Calibration task mismatch	0.20	0.200	20.8	79.2
Calibration ability shift	0.00	-0.001	95.1	78.6
Calibration ability shift	0.50	0.078	80.1	97.2
Calibration ability shift	1.00	0.151	45.1	99.9

Note. The task-mismatch experiment assumes no true AI learning, so its positive rates are false-positive rates. The ability-shift experiment assumes true endpoint learning = 0.20, so its positive rates are detection frequencies, which can be inflated by bias.

Figure 5 compares positive-learning claim rates under the principal scenarios. The figure should be read alongside the relevant true effect at each assessment occasion: in the retained and fading scenarios, the unaided transfer procedure tested a distinct, later effect from calibrated endpoint ANCOVA.

4.8. Nonlinear assistance and calibration transportability

The nonlinear assistance experiment maintained an average true endpoint learning effect of 0.20. With a representative calibration sample (shift = 0), unconditional calibration recovered that effect: mean bias was + 0.0005 standardized units (RMSE = 0.0752), while unaided transfer ANCOVA had bias + 0.0020 relative to the true T3 transfer effect of 0.15. When calibrators were drawn from a population shifted upward by 0.5 latent units, the mean calibrated endpoint bias increased to + 0.0554 (RMSE = 0.0913). A larger shift of 1.0 produced bias + 0.0484 (RMSE = 0.0872). The slightly nonmonotonic pattern is not evidence that a greater sample shift is safer: it follows the specified combination of a negative linear assistance gradient and a positive quadratic term. A mean calibration adjustment can therefore fail even with many paired observations if the relationship between proficiency and assistance differs across samples.

The trial's assisted ANCOVA remained misaligned with independent proficiency throughout these conditions, with bias approximately + 0.50. Conversely, the T3 unaided outcome remained close to its declared target because the change in calibration-sample composition did not affect the trial's unaided assessment. This contrast is a particularly useful stress test: it shows the dependence of the externally corrected model on transportability without suggesting that external calibration is intrinsically invalid. A

richer conditional model might improve transportability if its predictors were adequately measured and its functional form correct, but the extension did not estimate or validate such a model.

Table 7
Expanded robustness: nonlinear support and calibration-sample shift

Calibrator shift	Estimator	Truth	Mean	Bias	RMSE
0.0	Calibrated endpoint	0.200	0.200	0.000	0.075
0.0	Unaided transfer	0.150	0.152	0.002	0.049
0.5	Calibrated endpoint	0.200	0.255	0.055	0.091
0.5	Unaided transfer	0.150	0.149	-0.001	0.049
1.0	Calibrated endpoint	0.200	0.248	0.048	0.087
1.0	Unaided transfer	0.150	0.149	-0.001	0.048

Note. Independent 1,200-replication experiments, n = 600 in each trial and m = 120 paired calibrators. Calibration estimand = 0.20; unaided T3 estimand = 0.15. Shift is in latent-ability units.

4.9. Correlated scoring errors

Introducing positive within-writer correlation in assessment errors did not remove the supported-outcome versus independent-learning distinction. Assisted ANCOVA remained approximately 0.50 units above the endpoint learning target under all three error-correlation conditions. Calibrated ANCOVA showed small mean bias, ranging from + 0.0034 at ρ = 0 to − 0.0012 at ρ = 0.70. Unaided transfer ANCOVA also remained close to its true transfer effect of 0.18. The principal change was in empirical precision: the unaided T3 RMSE declined from 0.0475 under independent errors to 0.0300 at ρ = 0.70, conditional on the same regression and measurement-generation structure.

This reduction should not be read as proof that correlated errors generally increase statistical power. Here the baseline score and follow-up score share a stable error component, and adjustment for baseline helps account for that component. Alternative forms of correlation, particularly drift that is specific to intervention, tasks, or raters, may instead weaken model adequacy. These scenarios evaluate point estimates under one convenient shared-component model; they do not test an arbitrary covariance process or guarantee valid ordinary-OLS confidence intervals under every dependence structure.

Table 8
Expanded robustness: within-writer scoring-error correlation

Within-writer ρ	Estimator	Bias to learning target	RMSE
0.0	Assisted endpoint	0.501	0.504
0.0	Calibrated endpoint	0.003	0.070
0.0	Unaided transfer	0.002	0.048
0.35	Assisted endpoint	0.501	0.503
0.35	Calibrated endpoint	0.003	0.068
0.35	Unaided transfer	-0.002	0.038
0.7	Assisted endpoint	0.500	0.501
0.7	Calibrated endpoint	-0.001	0.062
0.7	Unaided transfer	0.000	0.030

Note. The assisted endpoint estimates a supported-outcome effect; its apparent bias is relative to the different independent-learning estimand. Error correlations use a shared normal component with fixed marginal error SD = 0.40.

4.10. Missing transfer assessments and the limits of correction

Missingness produced the largest sensitivity in unaided transfer estimation. Under MCAR, complete-case ANCOVA showed mean bias − 0.0021 relative to the true transfer effect of 0.18, with RMSE 0.0551 compared with 0.0478 for the full-data benchmark. The retention rate was approximately 72.1% overall and 72.0% in the AI arm. The increase in RMSE primarily reflected loss of observations, not systematic selection in the observed outcome. The oracle weighting estimator was numerically equivalent to complete-case ANCOVA because every case received the same constant inverse response probability.

Under the MAR mechanism, fewer treated outcomes were observed (approximately 58.7%; approximately 66.8% overall). Complete-case ANCOVA had bias − 0.0523 and RMSE 0.0778. Its bias arose because the probability of remaining in the analytic sample depended on an observed assisted-posttest change associated with later independent scores, a variable omitted from the ordinary complete-case ANCOVA model. Weighting by the correctly specified true observation probabilities reduced bias to − 0.0037, although its RMSE (0.0588) remained larger than that of the full-data estimator. Missingness qualified as MAR *relative to the declared observed history*; it was not equivalent to MCAR or automatically harmless for a complete-case analysis that omitted the selection predictors.

Under MNAR selection, complete-case bias reached − 0.1864 and RMSE 0.1953; approximately 53.1% of treated and 63.0% of all outcomes remained observed. Weighting with the unobservable true response probabilities reduced bias to − 0.0019 in the simulation, but this number is an **oracle benchmark**, not a

feasible correction from standard observed data. A researcher would generally need unverifiable sensitivity assumptions, external tracing data, a justified selection model, or additional design information to approach this performance. The experiment therefore supports a stronger reporting requirement: the claim that a delayed unaided effect has been measured should be accompanied by an analysis of who returned for testing and what assumptions support inferences about those who did not.

The missingness mechanisms also have distinct treatment-arm retention. In the newly independently generated validation cohorts, mean overall retention was 72.0% (MCAR), 66.8% (MAR), and 63.1% (MNAR); corresponding AI-assigned-arm rates were 72.1%, 58.8%, and 53.2%, while control-arm rates were 71.9%, 74.9%, and 73.0%. These different selection rates must be reported when interpreting cross-mechanism comparisons; MCAR, MAR, and MNAR were not constructed to have identical marginal dropout rates.

Table 9
Expanded robustness: missing T3 outcomes (true transfer effect = 0.18)

Mechanism	Estimator	Bias	RMSE	Treated observed %
MCAR	Full data	-0.001	0.048	72.0
MCAR	Complete case	-0.002	0.055	72.0
MCAR	Oracle IPW	-0.002	0.055	72.0
MAR	Full data	-0.002	0.046	58.7
MAR	Complete case	-0.052	0.078	58.7
MAR	Oracle IPW	-0.004	0.059	58.7
MNAR	Full data	0.001	0.046	53.1
MNAR	Complete case	-0.186	0.195	53.1
MNAR	Oracle IPW	-0.002	0.072	53.1

Note. The true T3 effect is 0.18. The oracle method uses *true synthetic observation probabilities*, including probabilities depending on unobserved T3 in the MNAR condition; it is not an implementable MNAR solution from observed data alone. Coverage was not computed for these extensions.

4.11. Feasible inverse-probability weighting and model specification

Table 10 reports the independent 1,200-replication validation, with a known true T3 effect of 0.180 standardized units. Under MCAR, all three observed-data approaches had negligible bias. Under MAR, complete-case ANCOVA had bias − 0.0508, whereas IPW based on a logistic response model containing the fully observed assignment-by-T2-change predictor reduced bias to − 0.0015 (RMSE 0.0596). An underspecified logistic model omitting that predictor retained bias − 0.0514, demonstrating that merely

labelling a procedure “IPW” is not evidence that it corrects selection. The correctly specified oracle analysis in the original stress test remains a separate benchmark.

Under MNAR, complete-case bias was − 0.1870; estimated observed-data IPW with the richer MAR-style model still had bias − 0.1519, and the reduced model had bias − 0.1906. Oracle MNAR weighting nearly recovered the truth only because its simulated probabilities used unobserved T3 values; these were not available to the feasible models. The richer estimated model therefore improved but did not identify the population transfer effect under outcome-dependent response. The comparison isolates the limits of observed-history weighting and the need for explicit MNAR sensitivity analyses or additional tracing information.

Table 10
Independent validation of feasible weighting for missing T3 outcomes

Mechanism	Observed-data method	Bias	RMSE	MCSE
MCAR	Complete case	+ 0.0019	0.0560	0.0016
MCAR	Estimated IPW: full observed model	+ 0.0021	0.0551	0.0016
MCAR	Estimated IPW: reduced model	+ 0.0020	0.0560	0.0016
MAR	Complete case	-0.0508	0.0783	0.0017
MAR	Estimated IPW: full observed model	-0.0015	0.0596	0.0017
MAR	Estimated IPW: reduced model	-0.0513	0.0789	0.0017
MNAR	Complete case	-0.1870	0.1963	0.0017
MNAR	Estimated IPW: full observed model	-0.1519	0.1631	0.0017
MNAR	Estimated IPW: reduced model	-0.1906	0.1997	0.0017

Note. All values are standardized score units; 1,200 newly generated replications per mechanism and 600 learners per replication. MCSE is the Monte Carlo standard error of the mean estimated effect, not the per-study standard error. “Full observed model” includes assignment, baseline, and assignment × (T2 − T0); the reduced propensity model omits the interaction. IPW uses estimated observation probabilities from measured predictors and is not generally valid under MNAR.

4.12. Counterbalancing task differences in calibration

The final robustness experiment compared a fixed-order calibration mismatch with an exactly balanced synthetic design. Under a systematic 0.20-unit task difference, the unbalanced correction yielded + 0.2006 bias and RMSE 0.2145 for a zero learning effect. Counterbalancing the sign of that additive difference across calibrators reduced bias to + 0.0006 with RMSE 0.0760. The result arises because the fixed task difference aliases the direct assistance contribution in an unbalanced design, whereas a symmetric allocation has a mean difference of zero by construction.

These findings support counterbalancing as a design principle, not as a guarantee of equivalent assessment modes. Actual writers may perform differently because of order, task familiarity, prompt content, genre, or interaction between task and tool use. If such effects are asymmetric or nonadditive, simple sign reversal does not guarantee valid calibration. A robust validation study would additionally assess prompt equivalence, inter-rater consistency, counterbalancing fidelity, and the extent to which assistance operates similarly across the intended population.

Table 11
Expanded robustness: systematic versus counterbalanced calibration task effects

Calibration design	True	Mean	Bias	RMSE
Fixed-order mismatch	0.000	0.201	0.201	0.214
Exactly balanced	0.000	0.001	0.001	0.076

Note. Zero-learning condition; systematic ± 0.20 task effect, 1,200 repetitions, 120 calibrators. The counterbalanced design imposes equal positive and negative task contributions; findings are conditional on an additive symmetric task difference.

4.13. Confidence-interval validity for fitted inverse-probability weights

Table 12 reports a new 1,200-replication inference experiment for each response mechanism. With a true unaided T3 effect of 0.180, fitted-IPW point-estimator bias was + 0.0020 under MCAR, − 0.0021 under MAR, and − 0.1530 under MNAR. The stacked-sandwich mean standard error tracked empirical spread fairly closely: 0.0545 versus 0.0553 under MCAR and 0.0573 versus 0.0587 under MAR. Nominal 95% interval coverage was 94.0% (Monte Carlo SE 0.0069) under MCAR and 93.8% (0.0069) under MAR. Both lie within approximately two Monte Carlo standard errors of 95%. These findings establish approximate interval calibration for these correctly modelled synthetic response processes, not across arbitrary MAR mechanisms or sample sizes.

Under MNAR, the mean fitted-IPW point estimate was 0.0270 despite a true transfer effect of 0.180. The estimated uncertainty remained close to the estimator's empirical spread, yet stacked-sandwich coverage fell to 24.7% (Monte Carlo SE 0.0124). The undercoverage therefore reflects non-identification and systematic bias, not simply an omitted standard-error correction. Treating fitted weights as fixed gave 94.6%, 94.3%, and 26.8% coverage under MCAR, MAR, and MNAR, respectively; these comparisons should not be generalised to the relative performance of the variance estimators outside this particular data-generating model. Importantly, neither version repaired the MNAR bias.

Table 12
Interval performance of feasible IPW for the T3 treatment effect

Missingness	Variance	Bias	Mean SE	95% cover	MCSE
MCAR	Joint	+ 0.0020	0.0545	94.0%	0.69 pp
MCAR	Fixed	+ 0.0020	0.0554	94.6%	0.65 pp
MAR	Joint	-0.0021	0.0573	93.8%	0.69 pp
MAR	Fixed	-0.0021	0.0592	94.3%	0.67 pp
MNAR	Joint	-0.1530	0.0572	24.7%	1.24 pp
MNAR	Fixed	-0.1530	0.0590	26.8%	1.28 pp

Note. True T3 effect = 0.180; 1,200 newly generated replications per mechanism (n = 600). MCSE is the simulation standard error for reported coverage, in percentage points (pp). Joint = stacked-sandwich variance jointly estimating propensity and outcome coefficients; Fixed = sandwich variance treating fitted response weights as fixed. Confidence intervals are normal Wald intervals; they do not identify unobserved MNAR selection.

4.14. Directional carryover and task-by-mode interactions in paired calibration

Table 13 tests assessment-order effects that could not be represented by the earlier symmetric task-difference experiment. When there was no directional carryover, the balanced-order correction had bias − 0.0025, whereas the all-assisted-first condition had + 0.1024 bias, consistent with the programmed 0.10 second-task period advantage. With carryover of 0.20, bias reached + 0.1027 even under balanced ordering and + 0.2989 when assistance always occurred first. These estimates closely match the prespecified analytic biases of + 0.10 and + 0.30. Counterbalancing eliminates the symmetric period advantage in expectation but does not cancel directional transfer of benefit from the assisted task to the later unaided task.

An additional mode-by-task shift of + 0.10 or − 0.10 produced calibrated endpoint biases of − 0.0969 and + 0.1014, respectively, despite balanced order. Such shifts alter the paired difference used to estimate direct assistance and do not vanish merely because learners are equally assigned to orders. Therefore, paired calibration requires more than counterbalancing: investigators must also justify absence or magnitude of carryover, compare prompt conditions, and evaluate whether assistance has the same meaning on the calibration task and the targeted endpoint. The simulation identifies the direction and magnitude of failure only under its stipulated additive deviations.

Table 13
Calibration bias under assessment-period effects, carryover, and task-by-mode differences

Calibration order	Carryover c	Mode–task h	Bias	Expected bias
Balanced	0.00	+ 0.00	-0.003	+ 0.000
Assisted first	0.00	+ 0.00	+ 0.102	+ 0.100
Balanced	0.10	+ 0.00	+ 0.050	+ 0.050
Assisted first	0.10	+ 0.00	+ 0.202	+ 0.200
Balanced	0.20	+ 0.00	+ 0.103	+ 0.100
Assisted first	0.20	+ 0.00	+ 0.299	+ 0.300
Balanced	0.00	+ 0.10	-0.097	-0.100
Balanced	0.00	-0.10	+ 0.101	+ 0.100

Note. Performance-only scenario (true endpoint learning effect = 0); 1,200 replications per design, 120 calibrators. A 0.10-unit period effect favoured the second task. Carryover improves the unaided score after assisted-first administration; h shifts the paired difference irrespective of order. “Expected bias” is analytically derived from each synthetic generator.

5. Discussion

5.1. The source of inflation is an interpretive mismatch, not simply a weak estimator

The main finding is that the interpretation of an assisted posttest effect as an independent learning effect can be systematically wrong even when the comparison is randomised, baseline scores are adjusted, and conventional interval calculations are mathematically well behaved. In the zero-learning scenario, the observed supported score captured an authentic causal effect of assignment on a task where AI was available. That effect was large and precisely estimated. The invalid step would be to rename it as growth in independent writing proficiency.

This distinction matters for how studies explain improvement. If the intervention group receives feedback during instruction but is assessed under the same unaided conditions as the control group, the assessed contrast can reflect learning, transfer, or remaining group differences subject to the usual design assumptions. If the intervention group is assessed with AI while the comparison group is not, the contrast contains a tool contribution that is part of the outcome-production process. The statistical procedure cannot infer how much of that contribution the learner internalised unless additional information or assumptions are provided. The present simulation isolates this mechanism under unusually favourable conditions for standard statistical analysis; it does not rely on allocation bias, differential attrition, or gross model misspecification to generate misleading claims.

The results extend validity arguments about score interpretation (Kane, 2013; Messick, 1995) into the practical setting of longitudinal AI-assisted writing research. A score is not invalid merely because technology assisted the performance. It becomes inadequately supported evidence when used for an interpretation that depends on a different, unobserved construct. Formative assessments of collaborative writing may properly value the combined learner–tool performance. Placement, certification, independent writing development, or retention claims typically require a different interpretation and therefore different evidence.

5.2. What calibration can and cannot establish

Under the simulation's identifying assumptions, an external paired-mode calibration sample removed the average direct assistance component from the supported endpoint contrast, producing nearly unbiased estimates of endpoint learning. The advantage is conceptual as well as statistical: the estimate attempts to measure the difference between supported and unsupported performance rather than assuming it away. This is potentially useful where a full trial must assess supported performance for instructional reasons but researchers still wish to recover a population-average independent-learning effect at the endpoint.

However, the violation experiments show that calibration is an additional measurement study, not a correction that can be applied automatically after the fact. Parallel task equivalence requires more than superficially similar essay prompts. Topic familiarity, rubric alignment, writing time, order effects, rater severity, or prompts with unequal idea-generation demands can alter the paired difference. If the calibration sample is drawn from more proficient learners and the AI's immediate contribution differs across proficiency, the mean difference will not necessarily transport to the original study. In such cases, larger calibration samples can make a biased adjustment appear more statistically convincing.

One methodological implication is that researchers using calibration should design it before data collection. Random counterbalancing of task order and assistance condition, multiple prompts, common scoring rubrics, independent rating, and stratified or appropriately weighted sampling can make the calibration assumptions more plausible. Where assistance depends on proficiency, a calibrated function conditional on baseline ability may be more appropriate than a single mean subtraction, provided it can be estimated with sufficient precision. These strategies are proposals for empirical validation; they were not all implemented in the present simulation and should not be described as demonstrated solutions.

The additional carryover experiment exposes a sharper boundary of the paired design: equal allocation to assessment order cancels a symmetric period effect but not an influence of earlier AI use on later independent performance. A carryover process and a mode-by-task interaction change what the within-person difference estimates. Consequently, calibration cannot be treated as a simple laboratory subtraction unless task comparability and the absence—or explicit modelling—of mode-specific carryover are justified. Even an externally representative sample and a nominally balanced schedule can give a precise but biased estimate of the assistance component.

5.3. Independent transfer assessments supply a different form of evidence

Unaided transfer and delayed assessments avoided the contamination mechanism imposed on the assisted endpoint. Their estimates tracked the correct effects across both the zero-learning and genuine-learning scenarios. This is not a surprising mathematical outcome given the data-generating model: with assistance explicitly removed at these occasions and treatment assignment randomised, the outcome was aligned with independent competence. Its importance lies in showing how the validity claim changes with assessment design.

The difference between retained and fading learning highlights a second interpretive risk. In both scenarios, the immediate endpoint learning effect was 0.20; only the fraction available at transfer changed. Calibrated endpoint analysis recovered the same immediate effect under both scenarios. An unaided follow-up, by design, assessed what remained later. A study using only a supported immediate posttest could not empirically distinguish these patterns without additional assumptions. Conversely, a study using only a late delayed test could underestimate the size of an initially acquired but partially forgotten effect if its conclusions were described as applying to the original endpoint.

Empirical AI-feedback studies have increasingly incorporated transfer or delayed assessments, but their operationalisations vary. For instance, Wu et al. (2026) examined a supervised no-AI transfer task after comparing feedback modalities, whereas Xie et al. (2026) evaluated lexical development across repeated writing assessments. The simulation clarifies why reports should state whether assistance was present at each measurement occasion and whether the stated inference concerns supported output, immediate independent learning, delayed transfer, or durability. These are not interchangeable dependent variables.

The interval-validation extension clarifies a further distinction between the quality of a point estimate and the quality of its confidence interval. Under the stipulated MCAR and correctly modelled MAR mechanisms, fitted-propensity weighting coupled with joint sandwich variance produced approximately nominal coverage. When response instead depended on the unavailable T3 score, the resulting interval was narrow relative to its systematic bias and rarely covered the true effect. Correct uncertainty for an identifiable model does not confer identification under a misspecified missingness mechanism (Seaman & White, 2013).

5.4. Implications for future L2 writing studies

The methodological findings translate into three priorities for future study design. First, investigators should define the target effect before deciding how to score the posttest. If the intended outcome is the benefit of a learner working with AI, a supported assessment is suitable. If the intended outcome is improvement in independent L2 writing, the principal endpoint should include direct unaided evidence or a separately justified model of the assistance contribution. An apparent gain cannot be assumed to represent independent competence simply because a pretest was administered.

The content of independent assessment should also reflect the learning mechanism being claimed. For example, a systematic review of context-based vocabulary teaching examined how meaningful language use, retention, and contextual transfer figure in EFL vocabulary development (Nguyen, 2026e). That review was not an investigation of AI-assisted writing and cannot identify an AI treatment effect. It does, however, provide a pedagogical reason to evaluate whether an apparent lexical improvement generalises to a new unaided context rather than merely repeating a tool-supported formulation.

Second, study reports should distinguish the timing and availability of technological support. Reporting only “pretest” and “posttest” leaves readers unable to determine whether differences were produced under comparable conditions. An assessment schedule that explicitly identifies supported instructional tasks, unaided outcome tasks, and delayed follow-up makes both causal and measurement arguments more transparent. It also reduces the chance that an effect on assisted product quality will be described as durable language learning.

Third, the inferential value of multiple assessments depends on their measurement properties. An independent task may be aligned with the target construct yet too unreliable to support a precise estimate. A paired calibration task may be statistically precise yet invalid because its modes are not equivalent. In practice, task design, rubric consistency, measurement invariance across conditions, rater reliability, and missing-data processes will all matter. Good longitudinal analysis must be paired with good assessment engineering.

5.5. Relationship to the broader research literature

The findings should not be used to dismiss empirical results showing the value of AI feedback. Escalante et al. (2023) found no differential linguistic progress in one context, while Xie et al. (2026) reported persistent lexical gains under another design; Wu et al. (2026) reported stronger unaided transfer following process-oriented feedback. Those findings differ in settings, feedback mechanisms, writing constructs, and statistical comparisons. The present study neither reproduces their datasets nor adjudicates their substantive effects. Instead, it clarifies a design-level condition necessary for interpreting such findings: the outcome must correspond to the construct and time point claimed.

The related simulation series shares a concern with the interpretation of AI-mediated language products, but the target parameters differ. Error migration concerns corrected and newly emerging errors (Nguyen, 2026b); proficiency compression and a separate rank-instability study both examine changing differentiation or relative standing, with explicit overlap in their use of rankings (Nguyen, 2026j, 2026h); diagnostic masking concerns domain weakness detection (Nguyen, 2026i); authorship-signal decay concerns stylometric re-identification (Nguyen, 2026d); and product–process non-identifiability concerns recovering independent proficiency from product, process, and anchor evidence (Nguyen, 2026f). The present study adds a longitudinal treatment-effect estimand, but not the first argument that unaided measures matter. No effect size from these distinct simulations is used to calibrate, validate, or replicate the numerical results reported here.

Two other publications illustrate a more general methodological boundary while remaining substantively outside the present AI-writing domain. A study of followership and teachers' organisational commitment differentiated predictive associations among organisational constructs without claiming experimentally identified effects (Nguyen, 2026c). A synthetic bilingual speech analysis operationalised speech output and lexical diversity as separate dependent variables when investigating a different exposure and two languages (Nguyen, 2026a). Neither paper provides evidence about AI-assisted writing or longitudinal learning; their relevance here is limited to a shared methodological requirement that an estimate be interpreted in relation to its defined outcome and research design.

The simulation also does not imply that all aided writing measures are deficient. Writing with technological assistance can be a meaningful target construct in its own right, particularly in professional or educational settings where competent use of tools is part of authentic performance. The methodological objection concerns the attribution of assisted performance to independent proficiency without a valid bridge from one construct to the other. This distinction is central to assessment validity and should remain explicit even as AI systems become more integrated into normal writing practice.

5.6. Limitations and scope of inference

This study has several substantive limitations. The first is the deliberately simple additive observation model. In reality, AI assistance can interact nonlinearly with proficiency, task genre, prompt quality, degree of textual incorporation, motivation, and rater interpretation. It may affect different components of writing differently and may alter the underlying learning trajectory itself. The heterogeneous scenario tested one ability-dependent pattern, not the full range of interactions plausible in authentic writing. Accordingly, the numerical rates reported here are operating characteristics conditional on the specified model and parameter values, not real-world prevalence estimates.

Second, the four primary simulations imposed random assignment with exact arm balance, common occasion offsets, independent normally distributed task errors, and complete follow-up. The additional experiments intentionally relaxed the independent-error and complete-follow-up assumptions; thus the entire programme must not be described as attrition-free. None of the generators models nonrandom class assignment, treatment nonadherence, interference, multi-rater drift, ceiling effects, or clustered classrooms. The new MAR validation shows that even an operational correction depends on a sufficiently informative response model; the MNAR results underscore that observed-history weighting cannot generally identify the missing outcomes without additional assumptions. These findings do not establish the robustness of conventional ANCOVA or calibration outside the tested generators.

Third, the transfer-retention pattern was specified in advance as a deterministic multiplier on the endpoint learning contribution. The simulation therefore demonstrates how estimands differ when retention differs but does not validate a particular theory of forgetting or generalisation. More realistic cognitive models could allow knowledge consolidation, differential transfer across genres, or delayed benefits from feedback after the immediate intervention.

Fourth, the calibration estimator was an intentionally transparent mean correction using an independent external paired sample. The sensitivity results show the dangers of unequal tasks and selected samples but do not compare more sophisticated alternatives such as conditional calibration, multilevel latent-state modelling, or test-equating procedures. Such methods may improve validity only with additional measurement information and assumptions. Their performance should be established in separate simulations before being recommended as replacements.

The original feasible-propensity experiment was limited to point-estimator operating characteristics, but the separately executed reviewer-directed analysis now evaluates joint-sandwich interval coverage for fitted IPW. This added experiment yielded approximately nominal coverage under the specified MCAR/MAR generators and substantial undercoverage under MNAR. These results do not validate intervals under arbitrary model misspecification, clustered sampling, weak positivity, heavy-tailed outcomes, small calibration groups, or other untested mechanisms. The analysis still does not compare multiple imputation, augmented IPW, or pattern-mixture sensitivity models. Even with MAR, inference requires observed selection predictors, adequate propensity overlap, a correctly specified response model, and a compatible baseline-adjusted outcome target.

Finally, this is a computational experiment. No participant responses, essay texts, proficiency ratings, or real AI outputs were analysed. The study cannot establish whether actual AI writing tools contribute 0.15, 0.50, or 0.80 standardized score units, whether true gains of 0.20 are typical, or how often empirical publications conflate the relevant estimands. Parameter settings were selected to expose mechanisms and sensitivity, not fitted to a population. Independent empirical validation is required before making claims about the frequency, magnitude, or generalisability of these effects in language classrooms.

5.7. What the expansion changes about the study's argument

The original experiment established a clear decomposition: a randomised comparison can validly estimate supported output while failing to identify the latent learning effect. The extensions clarify when proposed remedies remain defensible. Representative and equivalent paired calibration reduced endpoint mismatch, but nonlinear selection, asymmetric carryover, and task-by-mode perturbations introduced bias. Balanced order removed a symmetric period effect, but neither directional carryover nor prompt-specific assistance. This is a stronger conclusion than simply recommending counterbalancing: the correction must be demonstrably transportable and free of unmodelled sequential influences. On the follow-up side, correctly specified observed-data propensity modelling yielded approximately unbiased point estimates and near-nominal confidence coverage under the synthetic MAR mechanism, while no variance correction restored validity under MNAR.

The original oracle results and the separately seeded feasible-weighting validation are consistent but not the same simulation run. Small differences in complete-case bias across the two files are ordinary finite-replication differences and are reported with the correct source. Missingness was introduced at T3 only; potential missing baseline and T2 values were not modelled. The severe MNAR biases are therefore

consequences of the exact outcome-dependent response equation, not estimates of real-world dropout bias.

Missing-data validation adds an independent limitation to the strategy of conducting an unaided follow-up test. A construct-aligned assessment can still produce a selected group contrast. Under the stipulated MAR mechanism, a feasible logistic-response IPW analysis was approximately unbiased when it included the measured assignment-by-supported-gain predictor; omitting that predictor did not improve on complete cases. Under MNAR, both feasible models remained biased, even though oracle probabilities could recover the target in simulation. Thus construct alignment answers what the score measures, while the missingness analysis asks whether observed scores represent the target randomised cohort. A valid claim needs support on both dimensions; the present experiments do not identify an empirical MNAR mechanism.

The extension also qualifies a possible over-reading of earlier theoretical criticism. An assisted posttest is not inherently an inferior outcome. If a university or employer wants to know whether students can produce better documents while using AI responsibly, the assisted-outcome effect is directly relevant. The error addressed here lies in changing the outcome's substantive label without changing its evidence. A transparent study might report two co-primary outcomes—AI-enabled productivity and independent unaided transfer—precisely because the intervention may improve both or either. Presenting both estimates would make the pedagogical trade-off visible instead of forcing a single ambiguous notion of writing improvement.

5.8. Design implications for empirical validation

Future empirical research could test the simulation's qualitative predictions through a randomised intervention with assessments deliberately organised by support condition. At baseline, learners would complete an unaided task rated with a validated rubric. During instruction, intervention learners might receive process-oriented AI feedback and control learners comparable non-AI instruction or feedback. At the endpoint, both an AI-permitted task and an independently completed task would be administered, with prompts and order counterbalanced. A subsequent unaided transfer task would use a new prompt of comparable genre and difficulty, and a delayed assessment would address retention. Direct knowledge of writing processes, such as keystroke-based revision events, would help interpret mechanisms, although process traces alone would not substitute for an independent outcome.

Such a design should be registered before results are known and should identify separate primary estimands for assisted performance and independent proficiency. Researchers would need to ensure that treatment-group differences in task conditions do not become confused with effects of instruction, and they should document adherence to the permitted-assistance protocol. Calibration, when used, should preferably draw from the target population or explicitly model effect modification; a paired-mode sample should be sufficiently large to estimate relevant strata without assuming away heterogeneity. Assignment, access, and task administration should be distinguishable in the recorded data. If a single

class supplies each condition, randomisation at the learner level may not eliminate classroom interference, and clustering must be reflected in the design and analysis.

The scoring strategy would also require validation. Rater blinding is difficult when AI-written features reveal the condition, and common rubrics can operate differently when language is supplied externally. Multiple raters, anchor texts, prompt-equating procedures, and checks on measurement invariance could test whether an apparent longitudinal difference reflects a common scale. For dimensions such as lexical richness, corpus metrics may offer repeatability but remain sensitive to length and genre. The lexical-development analysis by Xie et al. (2026) illustrates the substantive value of longitudinal linguistic measures; the present simulation complements such work by making the causal parameter and support condition explicit, not by disputing its reported empirical findings.

Attrition planning should precede recruitment. Follow-up participants may differ from nonparticipants in proficiency, engagement, attendance, or satisfaction with AI feedback. Investigators should record predictors of response, report arm-specific retention and propensity-score distributions, evaluate the plausible missingness process, and use uncertainty estimates that reflect estimation of response probabilities. The new stacked-sandwich results show why this last step matters under an identifiable observed-data mechanism, and the severe MNAR undercoverage demonstrates why it is not sufficient on its own. Investigators should prespecify missingness assumptions and credible sensitivity analyses instead of presenting observed-history IPW as a universal correction (Little & Rubin, 2019; Seaman & White, 2013).

These considerations yield a practical assessment-design audit for empirical reports: specify the intended construct and time-indexed estimand; record which condition could use AI at every task; establish prompt, genre, scoring-rubric, timing, and rater comparability; distinguish supported outcomes from unaided transfer and retention; document assignment and follow-up response; demonstrate calibration representativeness, counterbalancing, and possible carryover; describe how inference accounts for fitted response probabilities; and make sensitivity to nonignorable missingness explicit. Each substantive conclusion should refer to the exact measurement and observation conditions analysed. These are proposed validity safeguards, not educational interventions proven effective by the synthetic experiment.

5.9. Scope, transparency, and priorities for further simulation research

The findings are conditional on a compact data-generating model that deliberately makes the phenomenon analysable. Independent ability is represented by one continuous latent dimension, even though academic writing includes interacting linguistic, rhetorical, strategic, and genre-specific capacities. Ordinary growth is linear with person-level slope variation; treatment-related learning enters through a stipulated time profile, and support operates as an additive component of scored output. These choices are not empirical estimates of how generative AI transforms writing. A stronger follow-on simulation could include multivariate latent traits, thresholded or bounded rubrics, teacher-classroom

effects, prompt-level random effects, selective tool uptake, and differential item functioning. Such complexities should be introduced to test particular inferential claims, not to make the generator cosmetically more realistic.

A further design limitation follows from the bundled treatment indicator: the simulation does not separately randomise AI use during instruction and at endpoint assessment. A future factorial experiment could vary these two conditions independently and compare supported productivity with unaided proficiency. However, an empirical study would still need to verify task and scoring comparability; the current simulated decomposition should not be read as a validated causal decomposition for real AI writing systems.

The absence of a fitted latent-state comparator is a further limitation. Mixed-effects models, latent growth curves, and longitudinal item-response models may improve precision or accommodate heterogeneity; none can automatically separate an unobserved immediate AI contribution from a simultaneous change in independent skill without identifying restrictions or additional measurement evidence. A defensible next comparison would use multiple independent writing tasks, explicit measurement-invariance tests, and repeated calibrated support conditions to assess when a richer model acquires genuine identifying information. The current paper does not imply that a simple calibrated ANCOVA is statistically optimal; it supplies a transparent benchmark whose assumptions can be examined directly.

The interpretation of statistical precision also requires care. A large and stable bias in an assisted outcome can make the estimate look persuasive even while the corresponding learning claim is unwarranted. Conversely, a noisy but aligned independent outcome may produce weak evidence about a real effect. The correct response to imprecision is not to re-label a supported outcome as proficiency, but to improve the design—more comparable tasks, larger cohorts, repeated independent measurements, better calibration, or a more efficient prespecified estimator. The original study's external-calibration sample-size experiment provides one example: reducing uncertainty about the direct support contribution improves RMSE under transportability, yet cannot remove systematic bias caused by incompatible prompts or selected calibrators.

Finally, reproducibility does not confer external validity. Publishing code allows reviewers to verify that a programmed estimator behaves as reported under specified conditions; it cannot establish that the programme's parameter values represent any population of real EFL learners. This manuscript therefore makes a methodological claim about the relationship between measurement conditions and causal interpretation, supported by controlled simulation rather than observational evidence. Any educational recommendation must be framed as a design implication to test in actual settings. The strongest practical contribution is a decision framework: define the construct, identify the outcome and time point, measure its relationship to support, assess the observation mechanism, and report uncertainty appropriate to the claimed estimand.

The computational audit also has clear boundaries. Random assignment and baseline adjustment do not establish that unaided and AI-assisted writing scores measure the same construct; measurement invariance, rater agreement, adherence to support conditions, task comparability, and the selection process for delayed assessments require empirical validation. In particular, feasible response modelling can address the stipulated observed-data MAR condition only under appropriate model specification and positivity; the MNAR oracle cannot be operationalised when its required outcome values are unobserved. Reviewers should distinguish reproducing code from validating its assumptions about real learners or writing practices.

6. Conclusion

Longitudinal AI-assisted writing assessment requires explicit separation between supported product quality, endpoint independent-proficiency growth, and later unaided transfer. In the primary simulations, assisted-posttest comparisons precisely estimated supported outcomes but generated erroneous learning-labelled interpretations when immediate assistance was ignored. Representative paired calibration recovered endpoint learning under restrictive task and transport conditions; subsequent stress tests show that selection, directional carryover, and task-by-mode interactions can violate those conditions even when order is balanced. An unaided outcome instead measures a later time-specific effect, but its validity also depends on who completes that assessment. Fitted-propensity IPW with joint-sandwich uncertainty produced near-nominal coverage under the programmed MCAR/MAR mechanisms and severely inadequate coverage under MNAR. These results motivate construct-aligned assessment, explicit causal estimands, calibrated measurement conditions, attrition planning, and transparent sensitivity reporting. Numerical performance is conditional on computer-generated assumptions and is not an empirical estimate of AI-related writing development.

Declarations

Data provenance. All observations in this study were generated by the disclosed simulation algorithm. No empirical participant data were created or analysed; the synthetic dataset supplied with the reproducibility package is illustrative and must not be used or described as actual learner data.

Ethics. Human-subjects review and consent were not applicable because the study involved no human participants and no personal data.

Code and data availability. The reproducibility package accompanying this manuscript contains the original and expanded simulation code, independent replay checks, aggregate Monte Carlo results, fitted-missingness and interval-validation scripts, calibration-carryover stress tests, numerical diagnostics, seven manuscript figures (including the study-procedure diagram), and an illustrative synthetic cohort. The program files and generated CSV outputs are local deliverables pending author-approved deposition to a permanent repository; no public DOI or archival deposit is claimed.

Use of AI in preparation.

AI assistance was used in developing manuscript text and computational materials. The author remains responsible for independent verification of code, analyses, citations, and final submission disclosures in accordance with venue policy.

Funding

and competing interests. These declarations must be confirmed by the author against the target venue's requirements before formal submission; no declaration has been fabricated.

References

1. Escalante J, Pack A, Barrett A (2023) AI-generated feedback on writing: Insights into efficacy and ENL student preference. Int J Educational Technol High Educ 20 Article 57. https://doi.org/10.1186/s41239-023-00425-2
2. Kane MT (2013) Validating the interpretations and uses of test scores. J Educ Meas 50(1):1–73. https://doi.org/10.1111/jedm.12000
3. Little RJA, Rubin DB (2019) *Statistical analysis with missing data*(3rd ed.). Wiley. https://doi.org/10.1002/9781119482260
4. Messick S (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol 50(9):741–749. https://doi.org/10.1037/0003-066X.50.9.741
5. Mislevy RJ, Haertel GD (2006) Implications of evidence-centered design for educational testing. Educational Measurement: Issues Pract 25(4):6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x
6. Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Stat Med 38(11):2074–2102. https://doi.org/10.1002/sim.8086
7. Nguyen HST (2026a) *Beyond pronunciation: Alcohol, speech output, and lexical diversity in L1 Vietnamese and L2 English*[Preprint]. SSRN. https://doi.org/10.2139/ssrn.7536862
8. Nguyen HST (2026b) *Correction or redistribution? A simulation-based error-migration framework for AI-assisted EFL writing revision* [Preprint]. https://doi.org/10.2139/ssrn.7575258. SSRN
9. Nguyen HST (2026c) Followership as a predictor of primary school teachers' organisational commitment in Phan Thiet, Vietnam. J Leadersh Qual Educ 1(1):1–25. https://doi.org/10.65956/leqa.2026.52
10. Nguyen HST (2026d) *How much of the writer remains? Monte Carlo validation of stylometric authorship signal decay under AI-assisted L2 revision* [Preprint]. https://doi.org/10.2139/ssrn.7575939. SSRN

11. Nguyen HST (2026e) The importance of context-based vocabulary instruction in EFL settings: A systematic review. Univ Phan Thiet J Sci 4(2):109–126. https://tapchikhoahocupt.vn/index.php/uptjs/article/view/78
12. Nguyen HST (2026f) *Same score, different competence? A study of product-process non-identifiability in AI-assisted L2 writing assessment* [Preprint]. https://doi.org/10.2139/ssrn.7576659. SSRN
13. Nguyen HST (2026g) *Understanding the AI-inflated persona paradox in L2 communication through computational modeling and inferential validation* [Preprint]. https://doi.org/10.2139/ssrn.7634078. SSRN
14. Nguyen HST (2026h) *When better writing changes the order: A Monte Carlo study of rank reversals and selection instability in AI-assisted L2 writing assessment* [Preprint]. https://doi.org/10.21203/rs.3.rs-11309235/v1. Research Square
15. Nguyen HST (2026i) *When better writing hides weakness: A Monte Carlo study of diagnostic masking in AI-assisted L2 writing assessment* [Preprint]. https://doi.org/10.2139/ssrn.7583178. SSRN
16. Nguyen HST (2026j) *When better writing reveals less about the writer: A simulation study of proficiency compression under AI-assisted EFL revision*[Preprint]. SSRN. https://doi.org/10.2139/ssrn.7575478
17. Sasaki A, Takeuchi O (2026) Noticed but not used: The gap between AI feedback and its delayed application in L2 academic writing. Asian J Engl Lang Teach 35(4):80–100. https://doi.org/10.65961/AJELT-2026-4-005
18. Seaman SR, White IR (2013) Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 22(3):278–295. https://doi.org/10.1177/0962280210395740
19. Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin
20. Tran TQL, Nguyen HST, Ngo TBH (2026) ChatGPT engagement and proficiency-dependent cognitive load among Vietnamese university EFL students. In *Proceedings of the International Conference on Educational Sciences and Foreign Language Teaching: Empowering Learners in the Digital Global Era: Insights, Innovations and Impact*(pp. 374–379). Bach Khoa Publishing House
21. Wu J, Mohd Kasim Z, Abd Razak FN (2026) From product scoring to online language processing: Associations of AI-supported process feedback with pausing and revision trajectories in L2 writing. Front Psychol 17:1951427. https://doi.org/10.3389/fpsyg.2026.1951427
22. Xie L, Li LX, Li H, Yuan Z, Fang Z, Qiao S (2026) Developing lexical richness: A longitudinal learner corpus analysis of L2 writing development under AI and teacher written corrective feedback conditions. Int J Appl Linguistics 36(4):3493–3506. https://doi.org/10.1111/ijal.70151
23. Yan D, Zhang S (2024) L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. Humanit Social Sci Commun 11:1086. https://doi.org/10.1057/s41599-024-03543-y

Figures

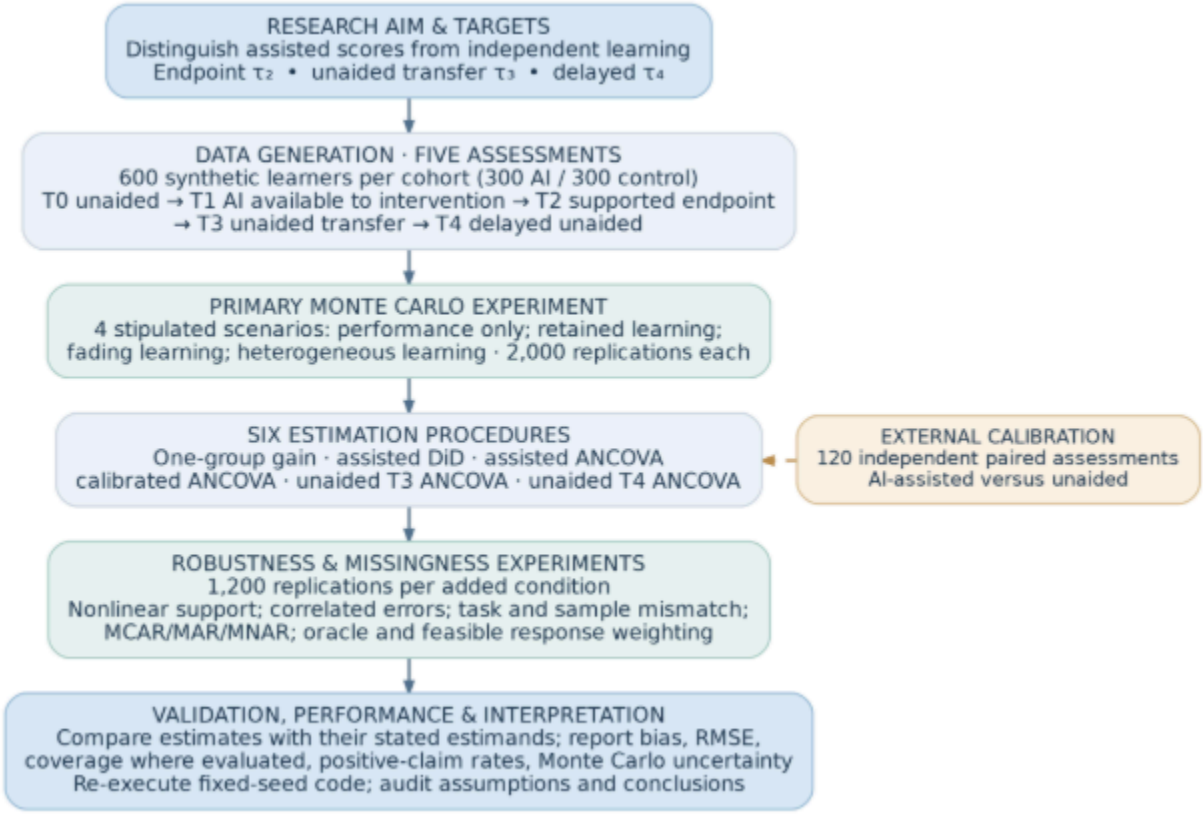


Figure 1

Study procedure for the synthetic Monte Carlo programme. The five assessments index simulated occasions; AI access at T1 and T2 was restricted to the intervention condition. Calibration learners were separate from the 600-person trial cohorts, and later robustness conditions were generated independently of the primary experiment.

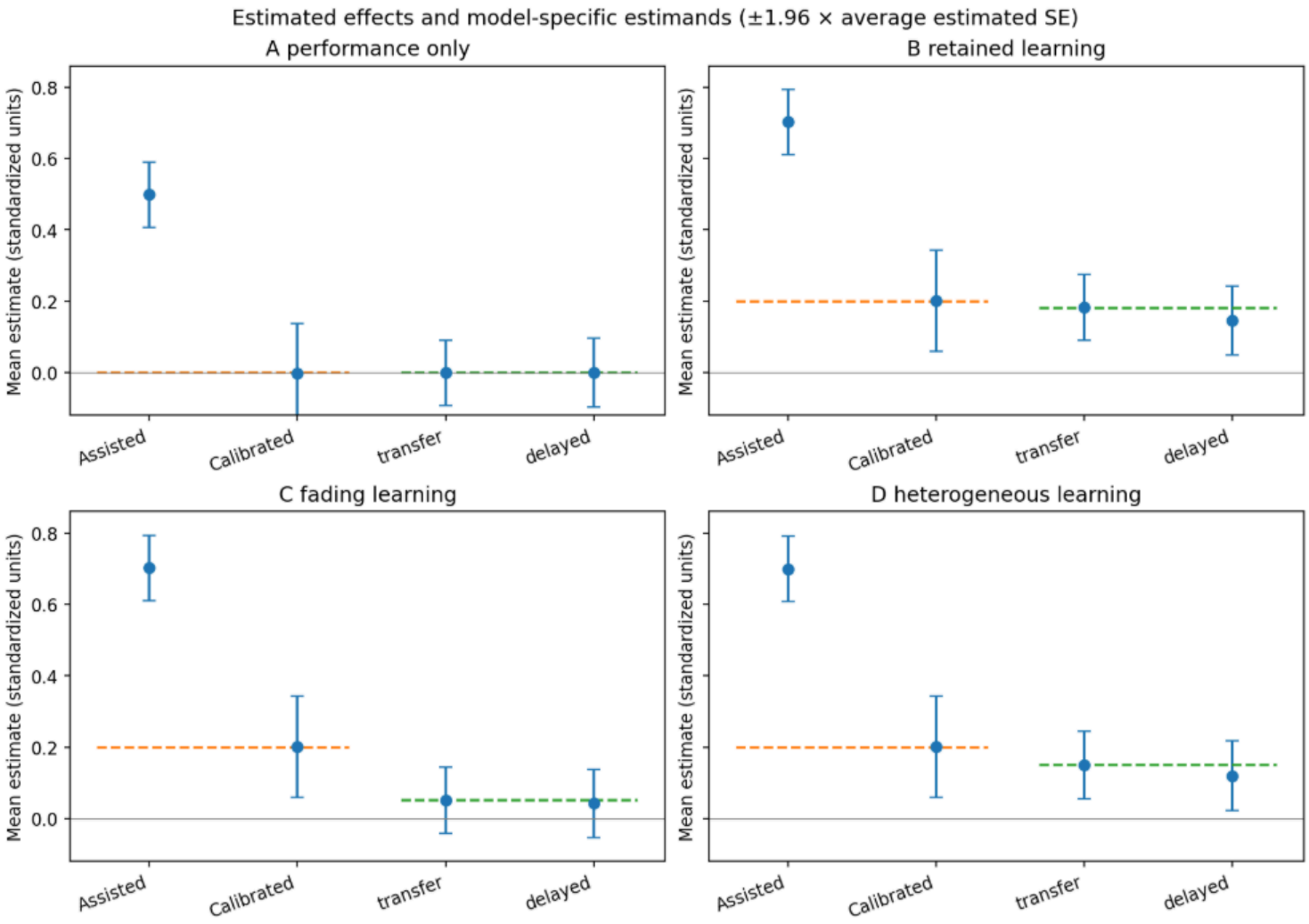


Figure 2

Mean estimated effects by method and scenario. Error bars show ±1.96 times the mean model-based standard error; dashed lines indicate theoretical endpoint and transfer effects.

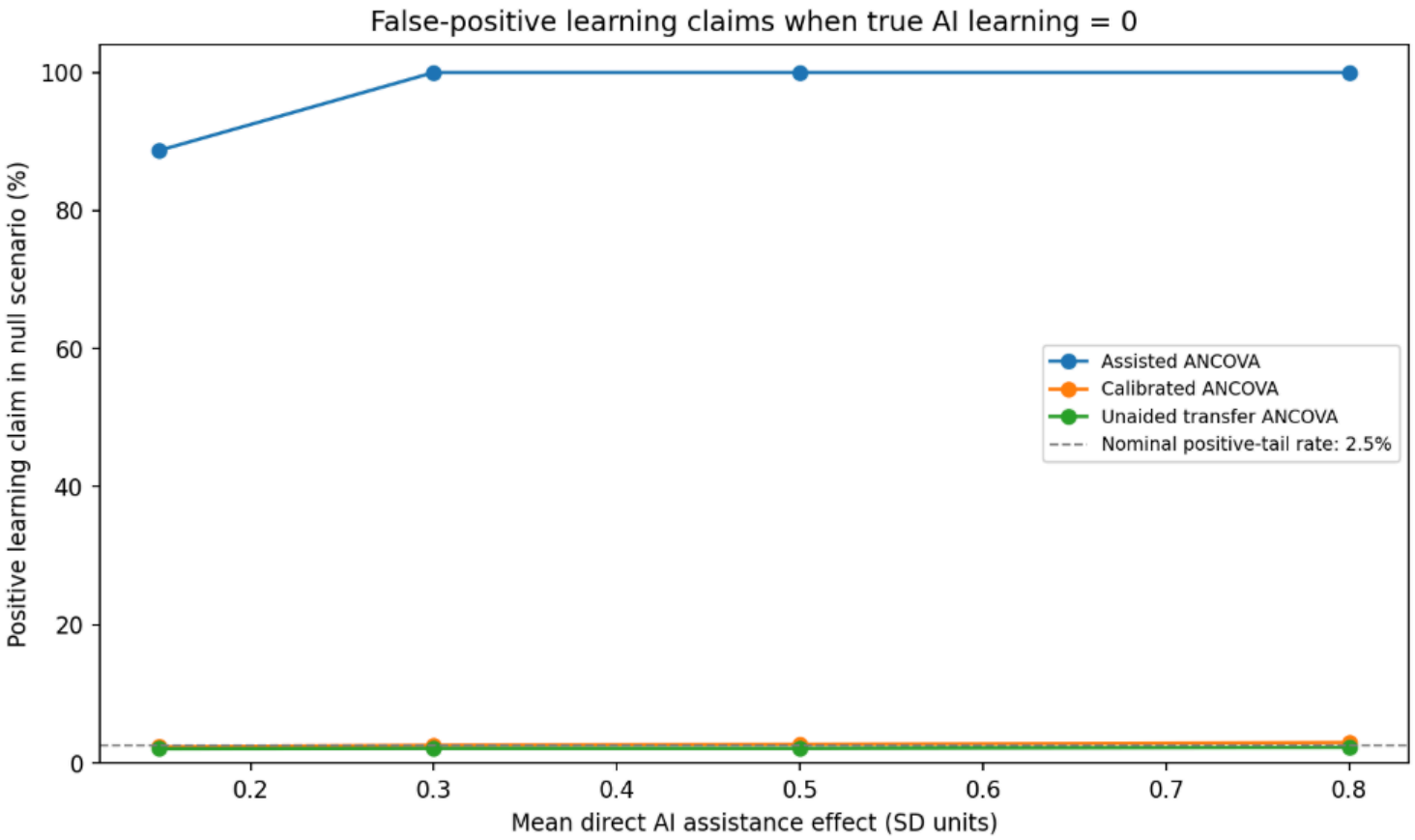


Figure 3

Rate of statistically positive learning claims under zero true learning, by direct assistance magnitude. The dashed reference line is the nominal 2.5% positive-tail rate associated with a two-sided 95% interval.

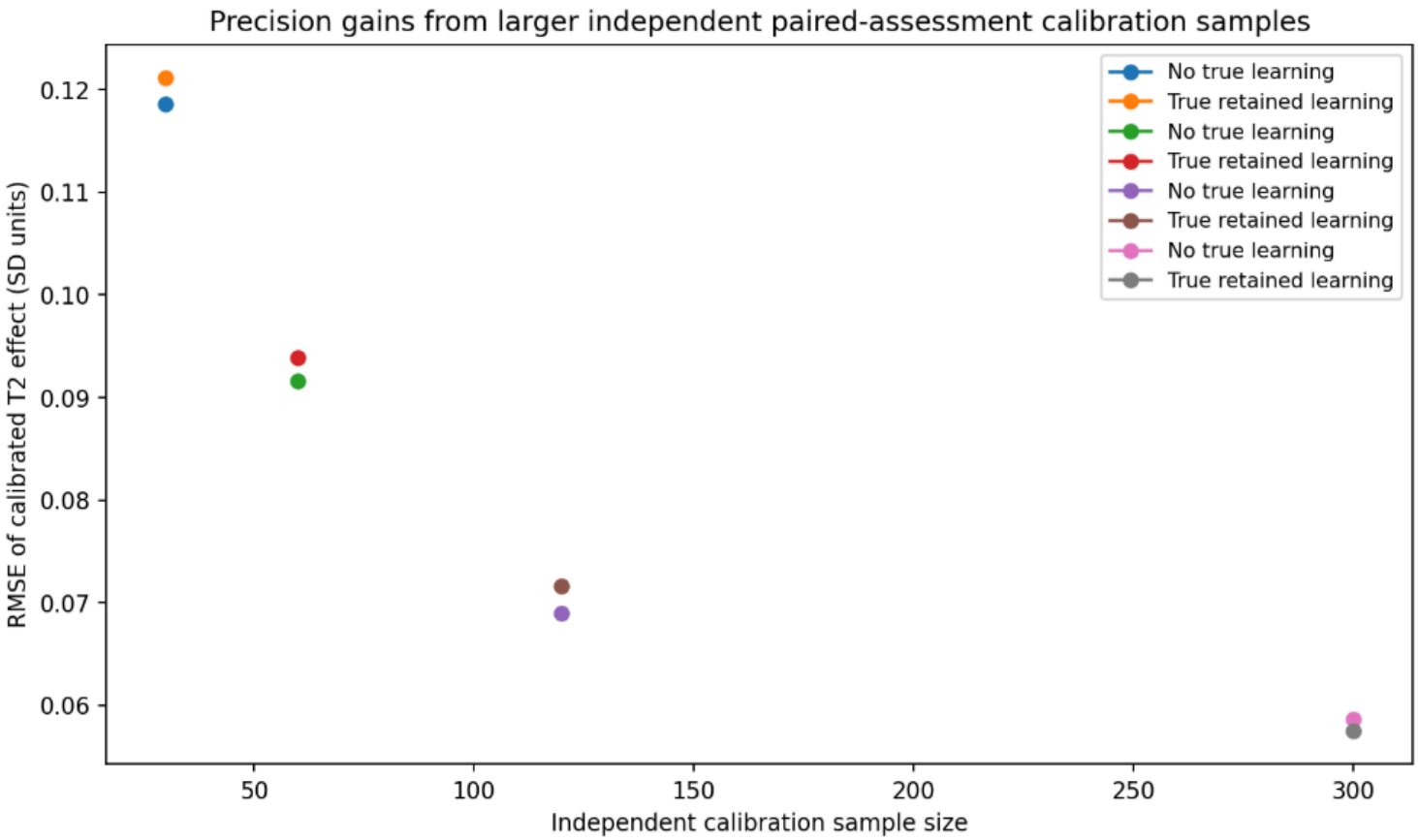


Figure 4

Calibration-sample size and root mean squared error of the corrected endpoint estimator, under the no-learning and retained-learning scenarios.

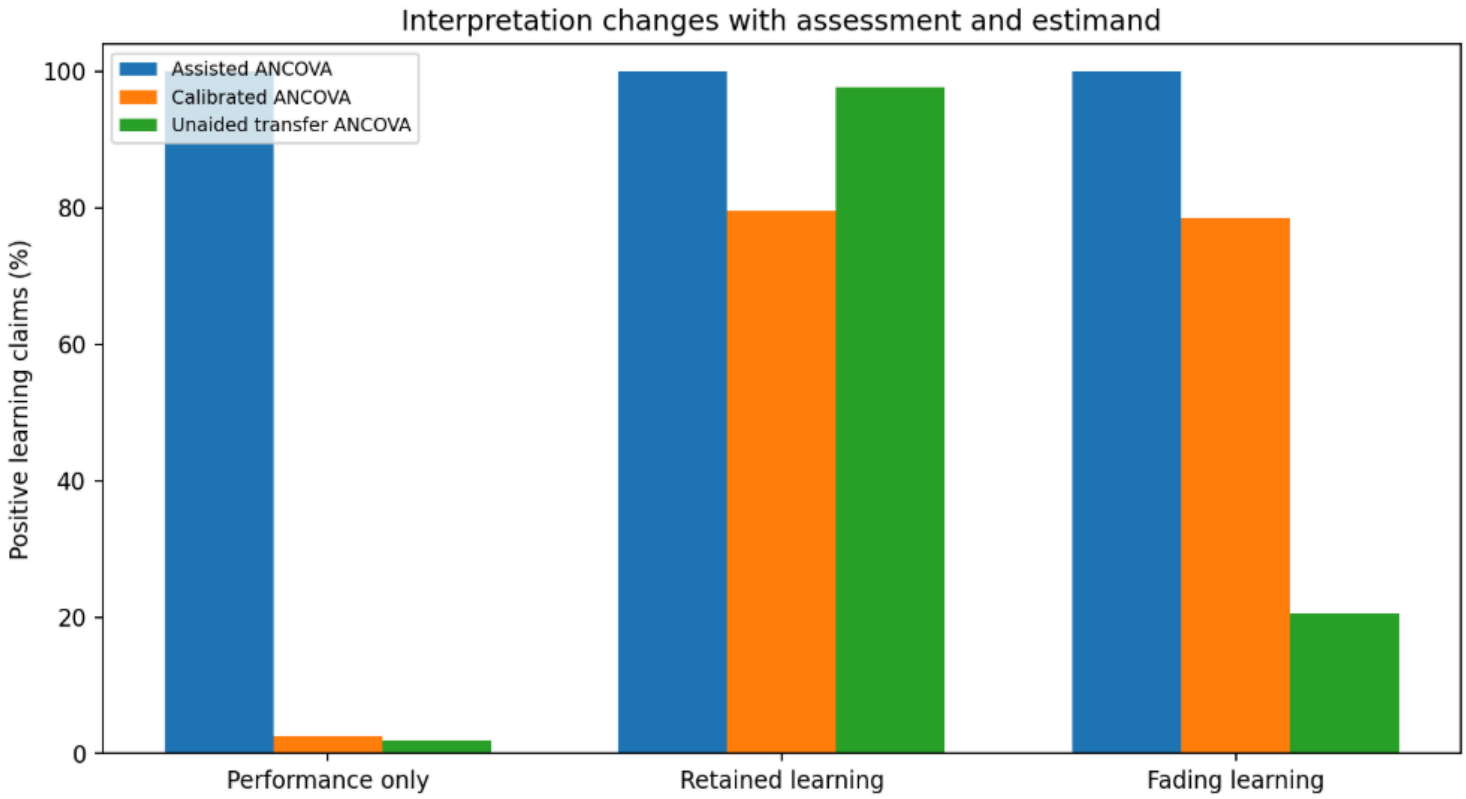


Figure 5

Proportion of statistically positive learning claims across principal scenarios and assessment methods. Rates for true positive effects represent detection probabilities, not false-positive rates.

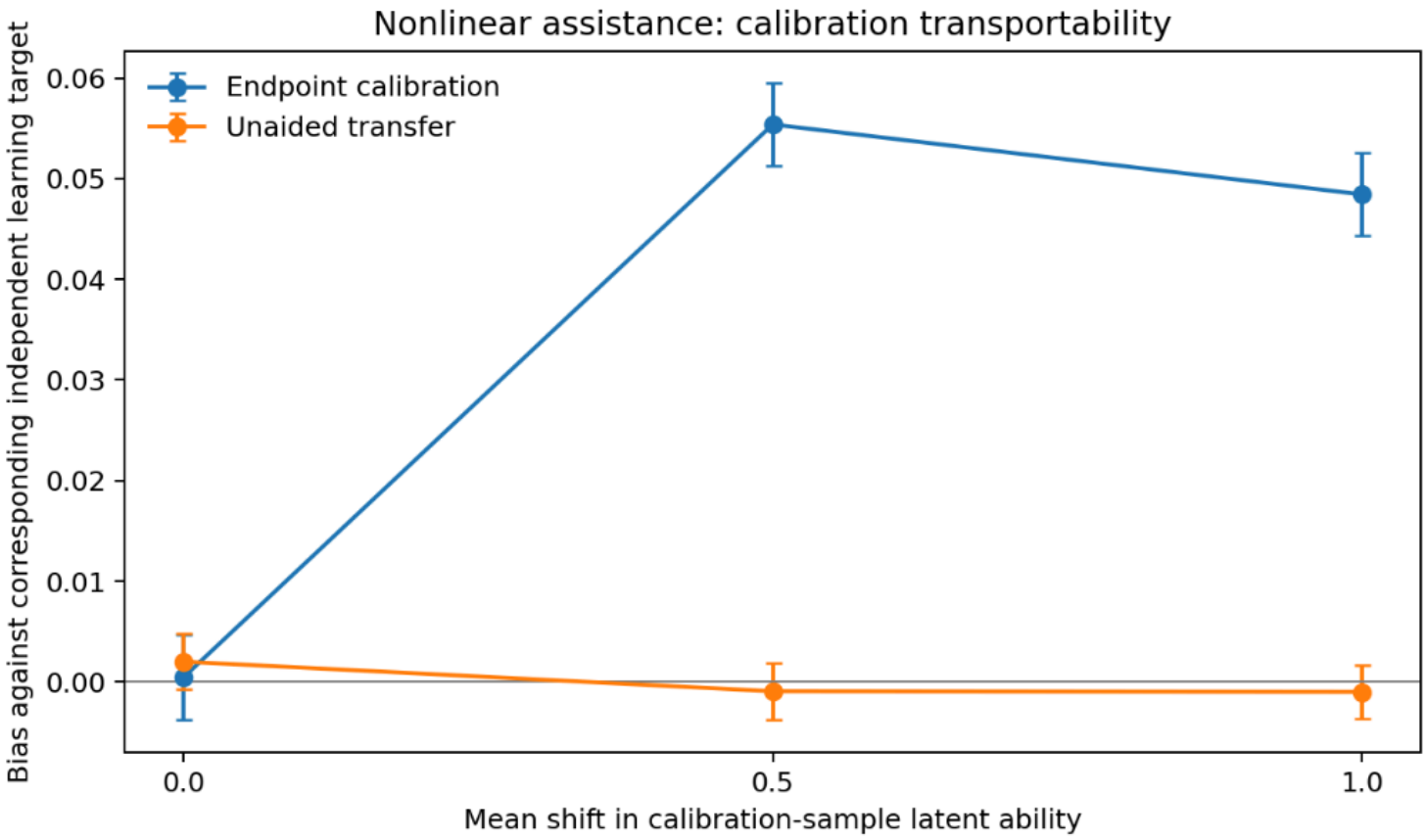


Figure 6

Nonlinear proficiency-dependent assistance: mean calibration bias and Monte Carlo uncertainty across calibrator proficiency shifts. Error bars are ±1.96 Monte Carlo standard errors, not learner-level confidence intervals.

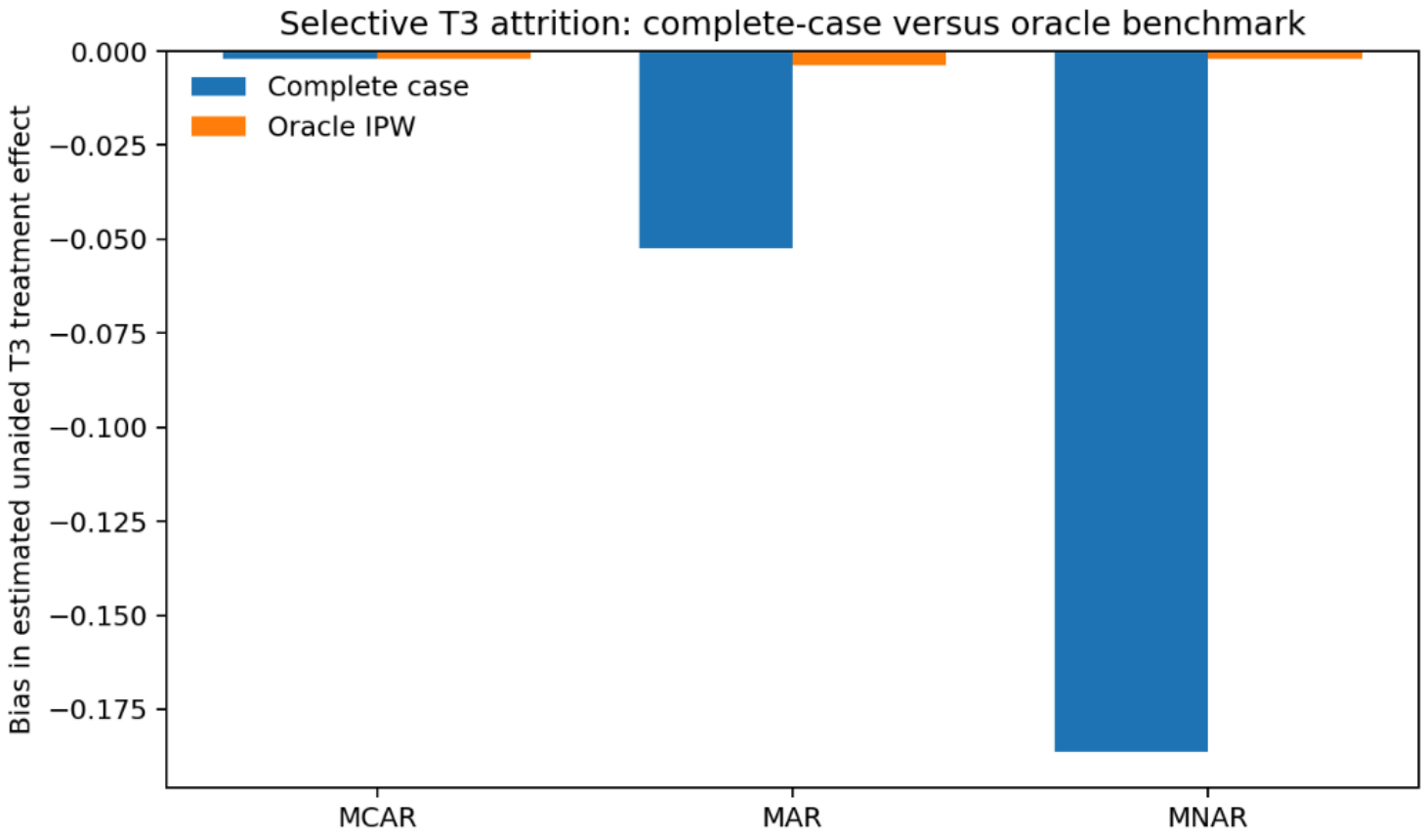


Figure 7

Bias in unaided T3 effects by missingness mechanism for complete-case ANCOVA and the oracle-weighted benchmark. MNAR oracle weights rely on unobserved outcome information and are not operationally identifiable.