

When Better Writing Changes the Order: A Monte Carlo Study of Rank Reversals and Selection Instability 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: L2 writing assessment, AI-assisted revision, rank reversals, pairwise classification, selection accuracy, Monte Carlo simulation, score validity

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11309235/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.

WHEN BETTER WRITING CHANGES THE ORDER: A MONTE CARLO STUDY OF RANK REVERSALS AND SELECTION INSTABILITY IN AI-ASSISTED L2 WRITING ASSESSMENT

Nguyen Huynh Si Tan | Independent Researcher

ORCID: <https://orcid.org/0009-0000-9801-2760>

Abstract

Improvement in AI-assisted writing products does not, by itself, establish preservation of the relative ordering of writers by independent proficiency. This methodological paper separates rank reversals between observed products from newly introduced ordering errors relative to known latent proficiency, and examines how both influence competitive selection and the effectiveness of score calibration. Under an ADEMP-aligned synthetic Monte Carlo design, 1,000 independent primary cohorts of 1,000 learners were generated with four correlated proficiency domains, paired unaided and assisted scores, and 25,000 uniformly sampled unordered writer pairs per cohort. Ten assistance and control mechanisms were assessed, followed by a separately seeded 500-cohort confirmation, 160-cohort recovery benchmarks, 120-replication robustness per condition, and intensity and pair-sampling validation. Uniform gains and strictly monotone compression preserved every observed pair ordering despite changing score levels and dispersion. In the heterogeneous weakness-targeted condition, the newly incorrect pair fraction averaged 16.6%, compared with 2.6% resolved errors, and top-decile oracle selection accuracy declined from 74.5% to 50.5%. Higher-proficiency-targeted gains produced frequent ordering changes but almost no net loss in pairwise accuracy. Product-only monotonic calibration did not restore inverted strict rankings, whereas models with additional synthetic metadata partly recovered the latent order. A separate 240-cohort margin and all-cutoff selection audit shows that net ordering loss varies nonmonotonically with true proficiency separation; a 160-cohort rounded-score analysis demonstrates tie-convention sensitivity for near-zero effects. Conclusions concern conditional mathematical and simulated assessment properties, not estimates of real learners or AI products. The study develops ranking and selection validity as distinct evaluation targets alongside text quality, compression, and domain diagnosis.

Keywords: L2 writing assessment; AI-assisted revision; rank reversals; pairwise classification; selection accuracy; Monte Carlo simulation; score validity

Introduction

Writing assessment often serves purposes beyond describing the linguistic quality of a single text. Programs use writing scores to compare candidates, rank applicants, and make selections when places, opportunities, or distinctions are limited. A score may therefore provide useful evidence for one purpose while remaining insufficient for another: the final quality of a collaboratively produced essay is not necessarily a transparent measure of the writer's independent ability. In an argument-based account of validity, the credibility of a score depends on its proposed interpretation and use, not simply on the score's consistency or apparent sophistication (Kane, 2013). The introduction of AI-assisted revision makes that principle particularly consequential because the same final product can reflect varying amounts and types of external contribution.

Research on AI-supported EFL writing has examined revision practices, differences in feedback use, and changes in scored writing quality. For example, Tuong and Tran (2025) examined differential impacts of writing tools among Vietnamese university learners, and Mekheimer (2025) investigated generative-AI feedback alongside revision and writing quality. Such investigations provide reasons to distinguish assistance mechanisms and learner characteristics, but a demonstrated improvement in average writing scores is not automatically evidence that comparative ranking decisions remain equally accurate. The present study isolates that additional measurement question using known synthetic latent proficiency and controlled assistance gains. It treats the preservation of relative order as a property requiring analysis rather than a background assumption.

Four related preprints by the author address error redistribution, product–process ambiguity, diagnostic masking, and compression of between-writer performance (Nguyen, 2026a, 2026b, 2026c, 2026d). Those works provide conceptual continuity, not independent empirical replication. Compression and ranking instability are especially easy to conflate: a strictly increasing transformation can narrow score differences without changing any writer's position, while a permutation can destroy relative ordering without changing the distribution of scores at all. Moreover, a reversal in observed rank is not synonymous with a wrong decision because the unaided ranking may already contradict the intended independent-proficiency target. A fifth study is justified only if these questions are defined and tested separately.

The core contribution is therefore a paired assessment framework distinguishing observable changes in writer order, newly introduced errors relative to latent independent proficiency, errors resolved by an assistance mechanism, and practical changes in top-ranked selection. Rather than assuming assistance must damage validity, the design includes gain structures that preserve order, amplify existing skill differences, improve particular rankings, or create unpredictable variation. A second contribution is explicit evaluation of what score calibration can and cannot recover: a strictly monotone transformation cannot reverse an existing strict

rank inversion, whereas an assistance-aware model may change order if it uses additional information. A third contribution is to report these effects with simulation uncertainty and independently regenerated cohorts, making the scope and instability of the conclusions inspectable.

Three research questions organize the analyses. RQ1 asks under what combinations of assistance intensity, individual gain heterogeneity, initial proficiency, and measurement uncertainty assistance introduces or resolves pairwise ordering errors. RQ2 asks how these ordering changes affect the validity and stability of top-ranked selections under different selection thresholds. RQ3 asks to what extent transformations and assistance-aware prediction models recover independent proficiency ordering from assisted writing evidence. The following sections explain the inferential distinctions required to answer each question, then specify the simulation model, estimands, validation controls, results, and limits of interpretation.

Related Literature and Theoretical Positioning

Writing-score validity depends on the intended decision

Weigle (2002) situates writing assessment within the interaction of writing constructs, tasks, scores, and their interpretations. The same score may be reliable enough to characterize a response to a particular task yet support a weaker generalization across tasks, raters, or independently produced writing. Brennan's (2001) account of generalizability theory and Barkaoui's (2010) study of rating processes illustrate why measurement error has multiple sources beyond the writer's ability. Relative ordering is especially sensitive when proficiency differences are small compared with the variability introduced by scoring, prompt features, or task constraints. Accordingly, an observed reversal after revision cannot be attributed to AI unless the relevant non-AI sources of measurement variation are controlled or explicitly modeled.

Kane (2013) emphasizes the inferential steps connecting observations to decisions. A decision to rank writers uses a stronger interpretation than a claim that an assisted essay is more polished, because ranking presupposes that observed differences track the target construct in a comparable way for different writers. Competitive selection adds a further consequence: errors concentrated near a cutoff may matter more than numerous reversals among middle-ranked writers. This is conceptually aligned with comparative-judgment research, where relative assessments are an explicit method of ordering performance rather than a hidden consequence of a total score (Whitehouse & Pollitt, 2012). The present simulation does not apply adaptive comparative judgment; it evaluates the meaning of a rank order implied by scalar writing scores.

Score gains and relative comparisons are distinct

A positive gain in an observed score is a within-writer statement, whereas a change in rank is a between-writer comparison. If two writers receive the same additive gain and the error component remains fixed, their difference is unchanged, so their strict order must be preserved. Even unequal gains need not reverse the order when their difference is smaller than the prior score margin. At the other extreme, a lower-scoring writer may receive sufficiently greater assistance to overtake another while both improve in absolute terms. This possibility is a mathematical property of heterogeneous gains, not evidence that actual AI tools necessarily produce such gains.

Compression is another separate property. If assisted scores are obtained through an increasing transformation with derivative between zero and one over an interval, some score margins shrink, yet all strictly ordered pairs remain in the same order. Conversely, permuting a score vector among writers leaves its mean and variance unchanged while generally scrambling rankings. These controls provide decisive falsification tests: a simulation claiming that compression logically entails inversions or that variance preservation guarantees comparability would fail. The current design retains them as complete conditions rather than merely discussing them theoretically.

AI-supported writing as a heterogeneous production condition

Technology-mediated writing may vary by learner, task, revision strategy, and access to particular feedback. Work on AI writing feedback describes changes in revision frequency and perceived quality, while studies comparing proficiency groups emphasize that effects should not automatically be modeled as identical for all writers (Mekheimer, 2025; Tuong & Tran, 2025). The simulation uses these publications only to motivate the need for heterogeneity; their observed effects do not supply the coefficients used here. A targeted gain function and random gain component are deliberate hypothetical mechanisms whose implications can be checked because the latent proficiency profile is known. Any attempt to transport those numerical effects to real classrooms would require independently collected paired writing evidence and measured assistance processes.

The four predecessor SSRN manuscripts examine related features of assisted products but ask different inferential questions. Error migration asks about redistribution of errors (Nguyen, 2026a), non-identifiability asks what can be recovered from products and processes (Nguyen, 2026b), diagnostic masking asks whether latent weak domains remain detectable (Nguyen, 2026c), and proficiency compression asks about narrowing between-writer differences (Nguyen, 2026d). None of these constructs logically determines whether a pair of writers is ordered correctly. The new study isolates ordering and selection as decision-related properties and evaluates them against an explicit latent-proficiency oracle. Self-citation is used to locate the boundary of the research program rather than to establish external credibility.

Rankings and selection validity

Rank correlation summarizes broad order agreement but can conceal where mistakes occur. Davis and Goadrich (2006) distinguish precision–recall from ROC performance in binary classification; the present study computes fixed-top-k overlap rather than full precision–recall or ROC curves. Here, selection precision and recall are meaningful only after defining a positive class, a selection decision, and an oracle target. Kendall-type discordance concerns the ordering of pairs; Spearman correlation considers the monotonic relation between ranked variables; and agreement between two observed rankings concerns stability, not necessarily agreement with an independent criterion. For a fixed-size top-k selection, selection precision and recall coincide numerically when the selected and oracle sets contain the same number of writers, but their interpretation still requires the target set to be defined. An AI-assisted selection may agree poorly with an unaided selection yet agree better with the oracle, or the reverse. These distinctions motivate parallel reporting of observed reversals, newly incorrect pairwise comparisons, net error, and top-k oracle overlap.

The rank-comparison tradition provides a complementary methodological backdrop. Thurstone (1927) formalized comparative judgment, and Bradley and Terry (1952) developed a model for paired-comparison outcomes; neither study concerns AI-assisted L2 revision. Kendall (1938) provides a foundational rank-correlation measure, while the current study separately records transitions in oracle-correct pairwise decisions. These concepts explain why ordinal agreement, comparison accuracy, and fixed-cutoff selection are related but not equivalent properties. Unlike direct comparative-judgment studies, the present simulation creates rankings from numeric writing scores and does not estimate judge-specific preference models.

A final complication is dependence. A cohort of 1,000 writers contains 499,500 unordered pairs, but those pairs share writers and cannot be treated as independent statistical units. Simulation uncertainty must therefore be summarized across independently regenerated cohorts, not by interpreting millions of pair outcomes as millions of independent participants. Morris et al. (2019) recommend prespecifying aims, data generation, estimands, methods, and performance measures (ADEMP) and explicitly characterizing Monte Carlo error. This study applies that logic throughout primary and secondary analyses, including a separate test of the sampled-pair estimator against complete enumeration. The estimator validation is important because computational shortcuts can otherwise silently change the meaning of the reported quantities.

Conceptual and Mathematical Framework

Latent target and observed evidence

Let θ_{ik} denote the independent proficiency of writer i in domain k , where four domains represent grammar, lexical resource, organization, and mechanics. The global target θ_i is

defined as their equally weighted mean. Equal weights are selected for transparent demonstration rather than on the basis of an empirically validated analytic rubric. A multivariate distribution with correlated components generates these proficiency coordinates, after which scoring error is added to produce the unaided score. Latent θ is invariant across conditions: simulated assistance changes products, not independent competence during the paired comparison.

$$\theta_i = \frac{1}{K} \sum_{k=1}^K \theta_{ik} \quad (1)$$

The unaided score is constructed from the latent target and an observation error component. Because this component is independent noise, the unaided ordering is itself imperfect; its errors form the relevant diagnostic baseline. Using unaided ranks as the target would conflate comparison stability with validity. The distinction becomes explicit when the same writers are assigned revised scores under controlled gains. The core analysis holds the original proficiency and unaided error draws fixed, so all changes in observed ordering within a paired comparison originate from the specified assisted-score transformation.

$$U_i = 65 + 12\theta_i + \varepsilon_i \quad (2)$$

The assisted product score adds the gain associated with a given condition. In the primary gain scenarios ξ_i is zero and any person-specific stochasticity enters through the explicitly generated gain; in the measurement-noise control, the added perturbation is treated separately as ξ_i rather than beneficial assistance. The generator does not simulate essay text, actual grammar corrections, or the learning processes that might cause independent proficiency to change over longer periods. Every reported result is therefore a statement about score interpretation in a stipulated measurement system. These boundaries are necessary to avoid attributing changes in a simulated scalar to genuine learner development.

$$A_i = U_i + G_i + \xi_i \quad (3)$$

The reversal condition and its limits

For writers i and j , assume the unaided scores satisfy $U_i > U_j$. A strict observed reversal occurs if the aided scores satisfy $A_i < A_j$. The corresponding algebraic condition is that the difference in gains and additional errors favoring j exceeds i 's original score advantage. This reveals why gain *differences*, rather than mean gains alone, govern pairwise reversal. Uniform additive gains cannot satisfy the inequality for originally ordered scores when ξ is unchanged, whereas heterogeneous gains can. A reversal says only that the observed order changed; its diagnostic status depends separately on the latent ordering of the two writers.

$$U_i > U_j, \quad A_i < A_j \Leftrightarrow (G_j + \xi_j) - (G_i + \xi_i) > U_i - U_j \quad (4)$$

The sampled-pair estimands compare the sign of an observed score difference with the sign of the latent proficiency difference. A pair is newly incorrect if the unaided comparison matches the latent ordering but the assisted comparison does not; a pair is resolved when the reverse occurs. Persistently incorrect pairs are wrong under both conditions, and persistently correct pairs agree with the oracle throughout. Exact or rounded ties in an observed score are treated as failing a strict pairwise ordering whenever the latent target is distinct, rather than silently discarded. This convention is immaterial in the unbounded continuous primary data but becomes important in the bounded-and-rounded robustness analysis.

$$N_{\text{net}} = N_{\text{new}} - N_{\text{resolved}} = P(E_A) - P(E_U) \quad (5)$$

Selection as a distinct assessment target

The decision use examined in RQ2 is selection of a fixed number k of candidates. For each condition, the k writers with the highest observed scores form its selected set; the corresponding oracle set contains the k writers with the highest latent independent proficiency. Selecting k writers makes the share of selected individuals who genuinely belong to the target group numerically equal to the recall of the oracle top- k set. Agreement between assisted and unaided selected sets is a separate statistic: it measures membership stability rather than correctness. Newly excluded and rescued oracle-qualified writers are paired across conditions so that net exclusion can be decomposed rather than simply described as selection churn.

$$Q_k(S) = \arg \max_{|T|=k} \sum_{i \in T} S_i \quad (6)$$

Ranking errors and selection errors may differ sharply. A model could reverse numerous pairs far below a competitive selection cutoff while leaving the selected group unchanged, or produce only a few errors precisely at the margin between successful and unsuccessful candidates. Therefore, RQ2 examines top 5%, 10%, 20%, and 30% rather than assuming one threshold is representative. The simulation also records how many oracle-qualified writers were selected before assistance but excluded afterward and how many were newly included correctly. This addresses whether observed change amounts to an error introduced by assistance or the resolution of an earlier error.

Recovery and calibration

Score linking and equating must also be distinguished from restoration of individual ordering. Kolen and Brennan (2014) discuss equating as a problem involving score scales, designs, assumptions, and equating error; that literature does not support the idea that a strictly increasing score transformation can reorder test takers. This study therefore describes linear and isotonic procedures as recalibration benchmarks, rather than as interchangeable proofs of validity or empirically justified cross-condition equating. When score-only recalibration preserves order, an improvement in level prediction would not by itself correct a pairwise rank

inversion. Any nonmonotonic recovery requires additional predictors, alternative models, or information constraints that must be disclosed.

RQ3 evaluates what can be recovered from the revised score under different information constraints. Linear score calibration is a strictly monotone transformation when its fitted slope is positive, and such a transformation cannot correct a strict rank inversion already present in the score vector. Isotonic calibration is nondecreasing rather than strictly increasing: it can create ties and modify rank-based correlation statistics, but it cannot strictly reorder two cases whose original score ordering is reversed. A fitted multivariate or metadata-aware model can change order when it accesses information beyond a single score, such as an assistance proxy. Any apparent improvement must be interpreted within that information regime, not attributed to a score-only transformation.

$$S_i > S_j \Rightarrow f(S_i) > f(S_j) \quad \text{for any strictly increasing } f \quad (7)$$

The experiment additionally includes a true-gain subtraction comparator that recovers the unaided score by construction. This is not a theoretical upper bound on all possible prediction models, because gain itself may encode information about proficiency and because the unaided score remains noisy. Its role is a transparent information-rich reference: a model that knows the exact gain can remove the assisted product transformation, an ability unavailable to an ordinary assessor. All fitted models are trained on separate learners and then evaluated on held-out learners. This prevents the latent criterion used to construct performance metrics from leaking into test-time estimation.

Methods: ADEMP-Aligned Synthetic Design

Aims, data-generating distributions, and replicates

The primary aims are to estimate newly introduced and resolved pairwise ranking errors (RQ1), changes in rank-based selection accuracy (RQ2), and recoverability under different evidence regimes (RQ3). The main generator samples 1,000 synthetic writers per cohort, each with four standardized latent domain coordinates and equicorrelation .45. Independent baseline score noise has SD 3.5; the observed unaided score has an intercept of 65 and a proficiency slope of 12. Stochastic gains depend on a standard-normal person-specific driver, while an independent beta-distributed ancillary variable is recorded for a placebo-like covariate check; assistance magnitude is varied explicitly in the separate intensity-multiplier experiment. These parameter values are designed for stress testing and are not calibrated to a particular scoring rubric or observed AI system.

The primary study regenerates 1,000 statistically independent cohorts under seed 20261009, and a separately seeded confirmation generates 500 more under seed 20261019. Within each cohort all scenarios share the same latent writer profiles and original scores, so contrasts are

paired. The total population of unordered writer pairs is 499,500, but a reproducible uniform sample of 25,000 without replacement is used for paired transition statistics. Kendall and Spearman rank correlations are computed on the full 1,000-writer cohort. Across-cohort means, empirical SDs, Monte Carlo standard errors, and simulation percentile intervals are retained at replication level.

Table 1
Primary data-generating specification

Element	Primary specification
Synthetic learners per cohort	1,000
Latent domains	4; equicorrelation .45
Latent global target	Unweighted mean of four latent domains
Unaided score	$65 + 12 \times \text{latent global proficiency} + \varepsilon$
Unaided score noise	Independent Gaussian; SD = 3.5
Primary ranking scale	Unbounded continuous scores; no ties by construction
Pair sample	25,000 distinct unordered pairs from 499,500 possible
Primary replication count	1,000 cohorts; root seed 20261009
Independent confirmation	500 cohorts; root seed 20261019
Selection thresholds	Top 5%, 10%, 20%, 30%
Recovery benchmark	160 independent cohorts; 700 training and 300 testing learners
Robustness checks	120 cohorts per condition; alternate covariance, noise, distribution, and bounds

Note. These are synthetic analysis settings chosen for numerical study and must not be interpreted as properties of an observed learner population or production AI tool.

Assistance mechanisms and negative controls

Ten primary conditions are prespecified. Exact zero assistance and uniform additive gain form rank-preservation controls; strictly monotone score compression demonstrates that smaller score variance does not imply an inversion. Deterministic weakness targeting supplies larger gains to lower-proficiency learners without added gain noise, and heterogeneous weakness targeting adds positive person-specific variation. A mirror-like strength-targeted condition favors higher-proficiency writers, while random positive gains and nonlinear weakness targeting test whether the patterns depend on targeting and functional form. An added-noise condition separates nonbeneficial scoring perturbation from positive assistance, and an equal-variance permutation leaves the score distribution unchanged while breaking its alignment with

individual writers. The permutation is an intentionally unrealistic mathematical falsification control, not a behavioral AI model.

Table 2

Full primary scenarios and their mathematical roles

Scenario	Gain/score mechanism	Primary purpose
Zero gain	$A = U$	Exact null for changes
Uniform additive	$A = U + 10$	Common gain cannot change order
Monotonic compression	$A = U + 12 - 2.5 \tanh[(U - 65)/10]$	Positive gain; order preserved; lower dispersion
Weak targeting, deterministic	$A = U + \text{clip}(8 - 6\theta, 0, 28)$	Targeting without gain noise
Weak targeting, varied	$A = U + \text{clip}(8 - 6\theta + 4z, 0, 28)$	Targeting plus person heterogeneity
Strong targeting, varied	$A = U + \text{clip}(8 + 6\theta + 4z, 0, 28)$	Higher proficiency preferential benefit
Random positive gain	$A = U + \text{clip}(8 + 4z, 0, 28)$	Non-targeted gain heterogeneity
Nonlinear targeting	$A = U + \text{clip}(8 - 7 \tanh \theta + 4z, 0, 28)$	Alternative targeted gain functional form
Added scoring noise	$A = U + \eta; \eta \text{ Gaussian, SD } 4$	Non-assistance measurement perturbation
Equal-variance permutation	$A = \text{permutation}(U)$	Rank disruption with exactly preserved dispersion

Note. θ denotes latent global proficiency; z denotes an independent standard-normal gain driver. The final two controls are not modeled as universally positive writing improvements. “Clip” is lower–upper truncation of the gain.

Pairwise estimator and numerical validation

Pair samples are drawn uniformly without replacement from all 499,500 possible unordered writer pairs, using integer triangular-index inversion to avoid allocating the full pair matrix in every replication. The pairwise baseline and assisted outcomes are then compared against the same latent criterion and the same sampled pairs. This design provides unbiased Monte Carlo sampling of the cohort’s pairwise error proportion under the specified sampling scheme, although individual pair observations remain dependent because they share writers. To validate the implementation, 24 independently generated cohorts and five mechanisms are recomputed by exact enumeration of every unordered pair, yielding 120 sample-versus-full comparisons. The difference is assessed descriptively, with cohort-level replication retained as the unit of uncertainty.

The exact-pair audit produced a mean absolute difference of 0.0016 and a maximum difference of 0.0057 between estimated and enumerated net pairwise error proportions. The zero-assistance and uniform-shift controls are additionally checked to produce no strict rank reversals, and the monotone-compression function is verified to have positive derivative and strictly positive gain over the analyzed range. Every primary run checks the identity between new and resolved pairwise errors and the net difference in misordered-pair prevalence. These automated properties prevent a plausible-looking chart from substituting for a correctly implemented estimator.

Table 3

Exact enumeration audit of the pair-subsampling estimator

Metric	Validation result
Validation cohorts and scenarios	24 cohorts $\times$ 5 scenarios = 120 comparisons
Possible unordered pairs per cohort	499,500
Uniform sampled pairs per cohort	25,000
Mean absolute sampled–exact net-error difference	0.0016
Maximum absolute difference	0.0057
Zero-gain and uniform-gain inversion tests	Passed
New – resolved = net identity	Validated per primary replication

Note. All pair-count and identity checks were executed in Python before drafting the Results. The sampled-pair metrics are conditional on an independently regenerated cohort and should not be treated as individual independent Bernoulli trials.

Additional independent validation and recoverability

A separate 160-cohort recovery experiment uses disjoint samples of 700 training and 300 test writers, assessing raw assisted scores, linear score calibration, nondecreasing isotonic calibration, a regression that also uses synthetic metadata, and subtraction of the true simulated gain. The regression models are fitted to the latent target only on training observations. Reported test metrics include Kendall rank correlation against independent proficiency and top-decile oracle overlap, which evaluates selection accuracy rather than agreement with unaided scores. The included synthetic gain driver is a privileged information condition, whereas the independently generated beta covariate is not part of the gain equation; any extra prediction benefit cannot be attributed to that ancillary covariate. The true-gain subtraction benchmark is likewise explanatory, not an implementable classroom procedure.

Robustness analyses independently vary domain correlation, baseline measurement noise, the latent proficiency distribution, and whether scores are bounded and rounded. The crossed sensitivity to gain intensity is evaluated in an additional 160-cohort ensemble with scaling

factors zero, 0.5, 1.0, and 1.5. The bounded/rounded design creates genuine observed ties, which are counted as failures of strict pairwise ordering rather than dropped from denominators. Additional variability in group-specific error remains outside this set of scenarios, and the study does not claim exhaustiveness over all possible AI–learner interaction processes. The aim is to find boundary conditions that contradict overgeneralized claims about ranking deterioration.

Uncertainty, verification, and reporting

All primary and sensitivity statistics are summarized as means over independent synthetic cohorts rather than treating writer pairs as independent replicates. The Monte Carlo standard error of an estimated mean is the between-cohort SD divided by the square root of the number of cohorts. Percentiles of the cohort-level statistics are also included in the files to show variation that is not captured by uncertainty of the mean. In this study the difference matters especially when a mean net error is small: an almost zero mean can conceal considerable rates of newly incorrect and resolved pairwise comparisons that offset one another. The Abstract and Discussion therefore report mechanism-specific mean patterns but avoid asserting that an effect is always present in every synthetic cohort.

$$\widehat{\text{MCSE}}(\bar{T}) = \frac{s_T}{\sqrt{R}} \tag{8}$$

Results — RQ1: Ranking Instability and Error Transitions

Total reversals are not new ranking errors

Table 4 reports pairwise outcomes under every primary scenario, with the unaided scoring process held fixed within each cohort. Rank reversal is the proportion of sampled pairs whose observed ordering changes; newly incorrect and resolved proportions are referenced instead to known latent independent proficiency. Thus the net quantity answers a different question from simple order instability. The columns for newly incorrect and resolved pairs also reveal whether the same net change could conceal considerable movement in both directions. This distinction is the central diagnostic requirement for interpreting RQ1 without mistaking an unstable ranking for an increasingly inaccurate one.

Table 4
Paired rank changes and newly introduced versus resolved ordering errors

Condition	Reversals %	New errors %	Resolved %	Net error pp	Kendall τ
No assistance	0.0	0.0	0.0	+0.0	0.768
Uniform gain	0.0	0.0	0.0	+0.0	0.768

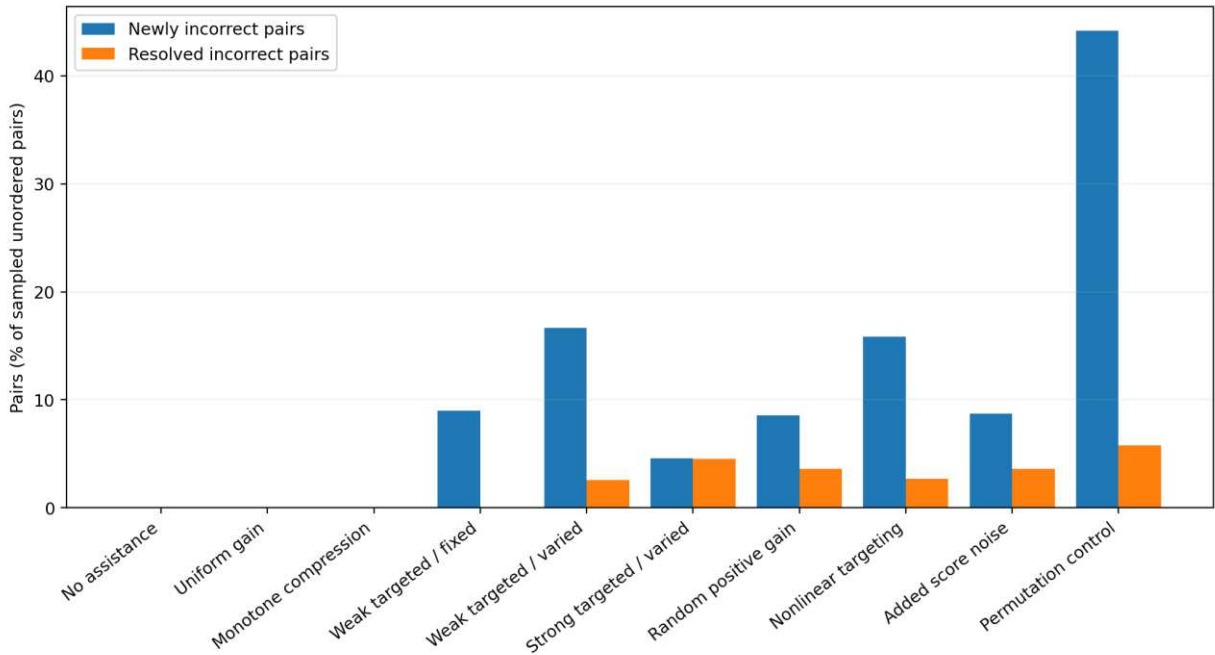
Condition	Reversals %	New errors %	Resolved %	Net error pp	Kendall τ
Monotone compression	0.0	0.0	0.0	+0.0	0.768
Weak-targeted deterministic	9.0	9.0	0.0	+9.0	0.589
Weak-targeted heterogeneous	19.2	16.6	2.6	+14.0	0.488
Strong-targeted heterogeneous	9.1	4.6	4.5	+0.0	0.767
Random positive gain	12.1	8.6	3.6	+5.0	0.669
Nonlinear weak targeting	18.5	15.8	2.7	+13.1	0.506
Added measurement noise	12.3	8.7	3.6	+5.1	0.666
Equal-variance permutation	50.0	44.2	5.8	+38.4	0.000

Note. Percentages are of uniformly sampled unordered pairs; “net error pp” is the percentage-point difference between newly incorrect and resolved pairs. τ compares assisted observed scores with latent proficiency across all 1,000 writers in a cohort. Values are means over 1,000 independent synthetic cohorts.

All three rank-preserving controls behave as required. Exact zero assistance and uniform additive improvement yield no reversals, while monotonic compression reduces score dispersion without creating a single changed ordering. In contrast, heterogeneous weakness targeting reverses approximately 19.2% of sampled pair comparisons, with 16.6% newly incorrect and 2.6% newly corrected. The net deterioration is therefore substantially smaller than the reversal fraction but remains clearly positive. Strength-targeted assistance shows the converse interpretive risk: many pairs reverse, yet newly incorrect and resolved rates almost cancel, leaving little net accuracy change. The full paired decomposition must be retained to avoid labeling all movement as deterioration.

Figure 1 displays the two components of net error separately. The primary gain mechanisms are presented alongside the rank-preservation and permutation controls so that positive and near-zero results are visible rather than only the scenarios producing the largest deterioration. The mirrored rates under strength targeting help explain why the count of reversed comparisons cannot serve as an error rate.

Figure 1
Newly incorrect and resolved pairwise decisions across primary mechanisms

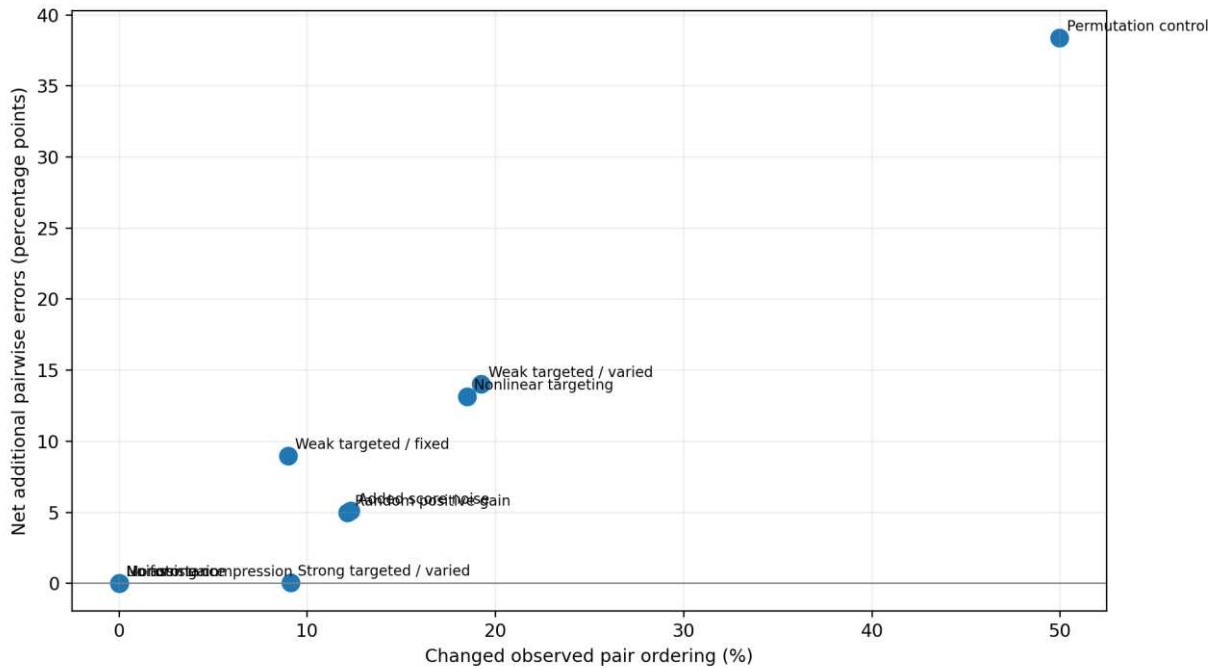


Note. The plotted statistics are synthetic Monte Carlo summaries under the stated generating mechanisms; graphical trends do not establish empirical learner-level effects.

Figure 2 juxtaposes the proportion of writer pairs whose observed order changes with the net change in accuracy against the latent oracle. The deliberately artificial permutation case provides an upper-range illustration: the score distribution can remain unchanged while half the observed pairs change order. Other scenarios occupy more informative regions where a substantial reversal fraction can coexist with little net error.

Figure 2

Observed ranking changes compared with net additional ordering error



Note. The plotted statistics are synthetic Monte Carlo summaries under the stated generating mechanisms; graphical trends do not establish empirical learner-level effects.

Ranking accuracy under targeted and non-targeted gains

The deterministic weakness-targeted condition supplies positive gains that are strongest for lower latent proficiency, without additional gain noise. Its newly incorrect comparisons are not offset by resolved comparisons to the same extent as the stochastic case, indicating that a smooth gain gradient can shift margins in a systematic direction. Adding heterogeneous gain noise both introduces and resolves comparisons, raising total churn while weakening the inferential simplicity of its net effect. This distinction illustrates why a sample-level average gain cannot predict relative-order validity without specifying the gain–proficiency relation and the dispersion of individual gains. The analysis does not claim that a real AI tool follows the linear law encoded by these conditions.

Strength-targeted heterogeneous assistance produces a different pattern, even though all generated gain amounts are nonnegative. Its effects enlarge the differences favoring higher latent proficiency on average, but the stochastic driver still creates pairwise movements, some favorable and some unfavorable. The near-zero mean net error in this condition should not be interpreted as stability of every pair: the newly incorrect and resolved fractions are both nontrivial. This mechanism establishes a boundary against the proposition that any heterogeneous gain must reduce ranking validity. The separate score-noise control shows that a

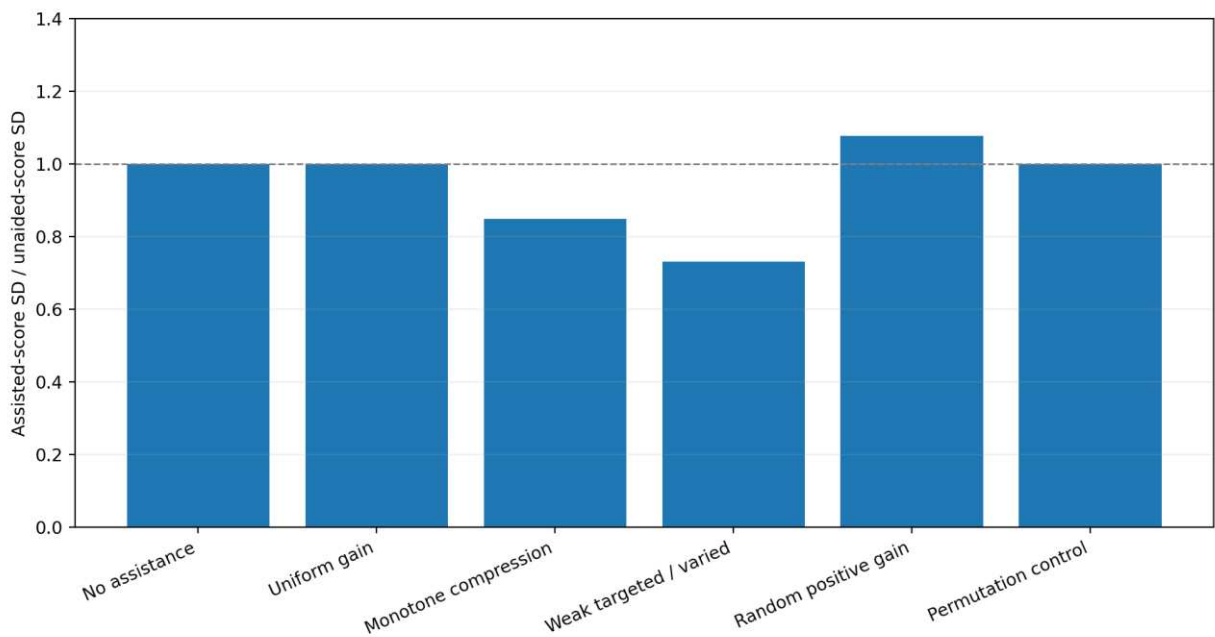
similar amount of instability can emerge without beneficial assistance at all; it is therefore inappropriate to attribute all model-based net error to an improvement process unless gain and measurement perturbation are distinguished.

Compression and variance preservation do not determine order

The compression-only condition is useful precisely because it disproves an appealing but invalid shortcut. Assisted-score SD is about 0.848 times unaided-score SD, yet its monotonic transformation preserves every strict ordering and every fixed-size selected set. In the equal-variance permutation control, the corresponding SD ratio is exactly one, but the placement of scores among writers is randomized, destroying much of the original agreement with latent proficiency. Neither score dispersion nor average improvement therefore uniquely identifies rank preservation or validity. The two controls should be understood as mathematical demonstrations that accompany, rather than replace, the substantive gain scenarios.

Figure 5 presents the assisted-to-unaided SD ratio for controls and selected gain mechanisms. The data should be interpreted jointly with pairwise error results in Table 4; the SD ratio alone is not a measure of how accurately a writer is ranked. It is possible to obtain unchanged variance with profound reordering and smaller variance with exact order preservation.

Figure 5
Score dispersion can change independently of ranking validity

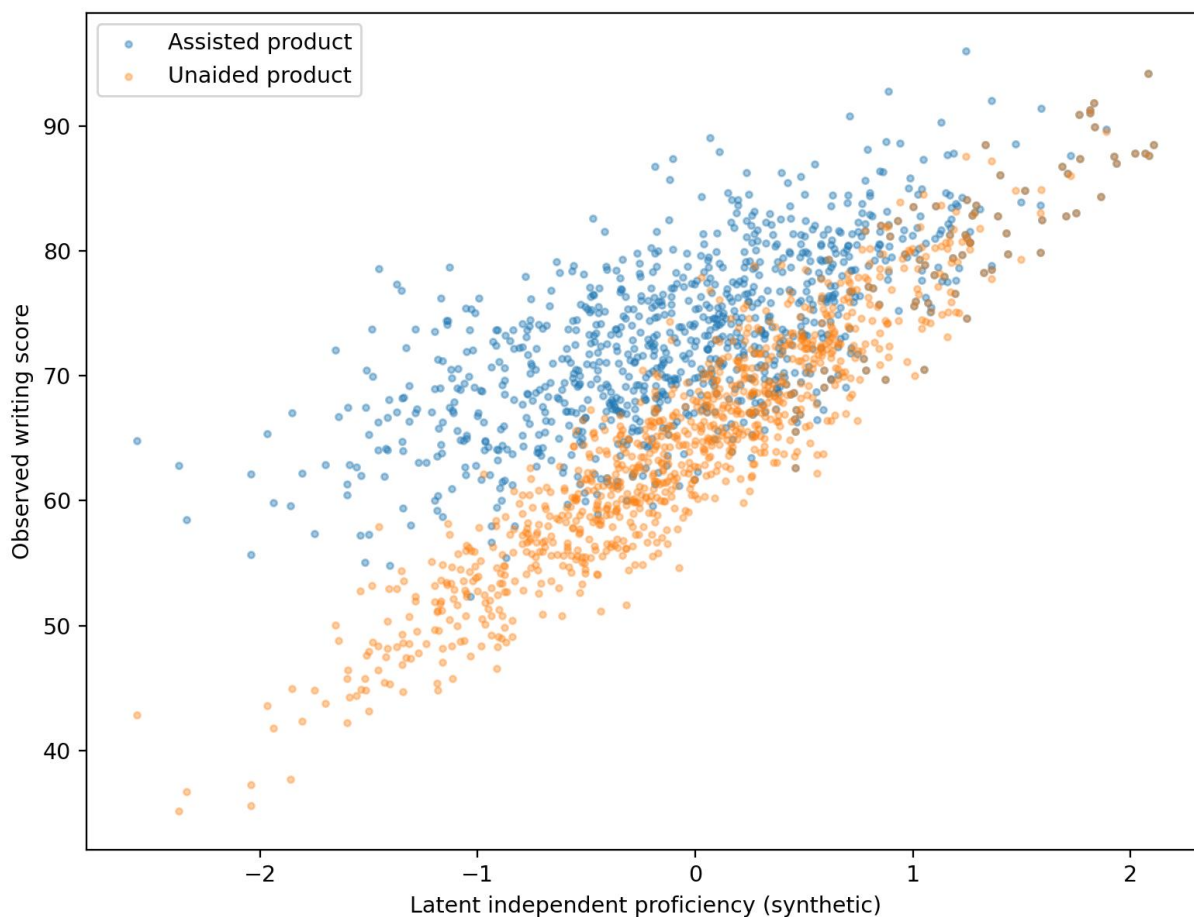


Note. The plotted statistics are synthetic Monte Carlo summaries under the stated generating mechanisms; graphical trends do not establish empirical learner-level effects.

Figure 8 uses one synthetic cohort to show the dispersion of unaided and weakness-targeted assisted products against the same latent proficiency values. The figure is intentionally illustrative rather than a summary of the Monte Carlo evidence; its value is to make clear that product-score gains and proficiency-specific ordering are different axes. The group-level summaries in Tables 4 and 5 carry the replicated quantitative conclusions.

Figure 8

Illustrative latent proficiency and observed scores in one paired cohort



Note. The plotted statistics are synthetic Monte Carlo summaries under the stated generating mechanisms; graphical trends do not establish empirical learner-level effects.

Results — RQ2: Accuracy of Rank-Based Selection

Oracle selection and membership churn

The selected set is fixed in size and evaluated against the writers with the highest latent proficiency, not against the original unaided ranking. Table 5 compares oracle overlap for four selection proportions across the full primary mechanism set. An overlap of 0.75 at top 10% means that approximately 75 of the 100 selected synthetic writers also belong to the latent top

100; it does not imply an observed population probability of selecting qualified candidates. Because both sets contain exactly the same number, selection precision and recall coincide numerically. The paired new-exclusion and rescue components are therefore needed to determine whether a decrease from the baseline reflects errors introduced after assistance or merely turnover among equally uncertain candidates.

Table 5

Oracle selection overlap across thresholds and assistance conditions

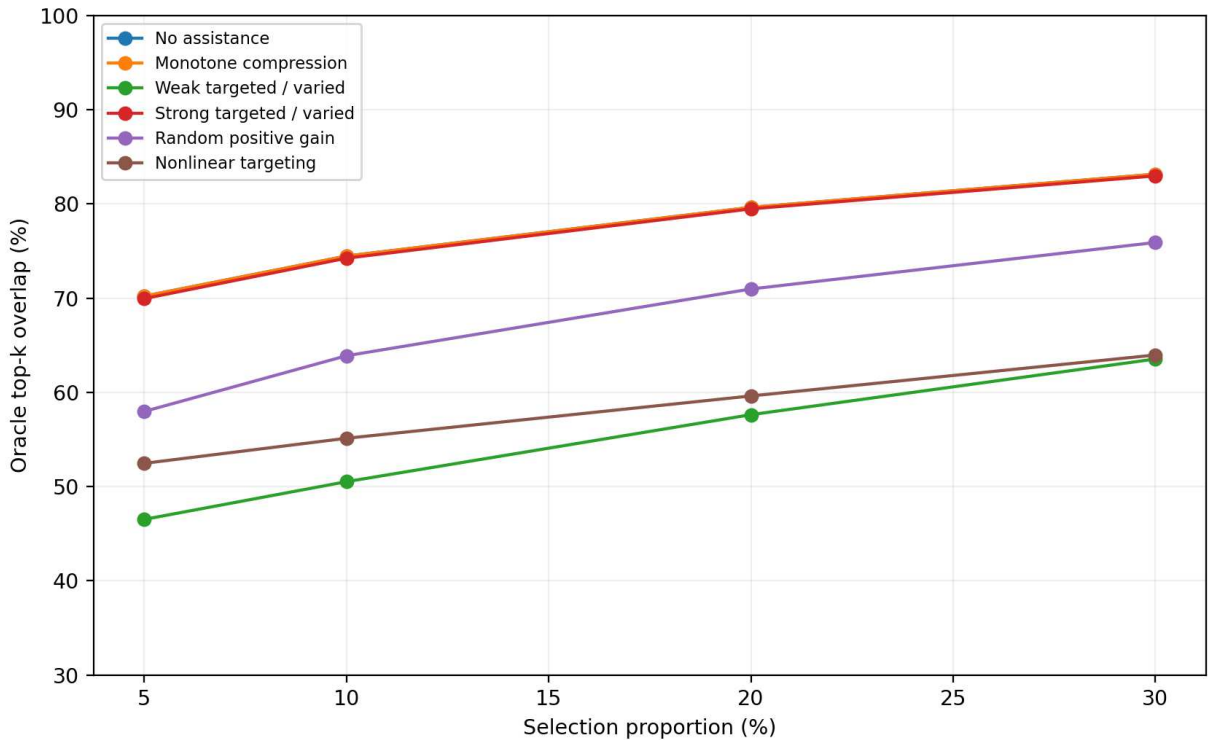
Condition	Top 5%	Top 10%	Top 20%	Top 30%	New false exclusions at 10%	Resolved exclusions at 10%
No assistance	70.2%	74.5%	79.6%	83.2%	0.0%	0.0%
Uniform gain	70.2%	74.5%	79.6%	83.2%	0.0%	0.0%
Monotone compression	70.2%	74.5%	79.6%	83.2%	0.0%	0.0%
Weak- targeted deterministic	57.0%	58.0%	64.4%	70.1%	16.4%	0.0%
Weak- targeted heterogeneous	46.5%	50.5%	57.6%	63.5%	26.0%	2.1%
Strong- targeted heterogeneous	70.0%	74.2%	79.5%	83.0%	8.6%	8.4%
Random positive gain	58.0%	63.9%	71.0%	75.9%	16.1%	5.5%
Nonlinear weak targeting	52.5%	55.1%	59.6%	64.0%	22.2%	2.9%
Added measurement noise	58.0%	63.9%	71.0%	75.8%	16.1%	5.5%
Equal- variance permutation	5.0%	10.1%	20.0%	30.1%	67.0%	2.6%

Note. Selection percentages are averaged across 1,000 synthetic cohorts. Each selection set and the latent oracle have the same size. Newly excluded and resolved members are expressed as fractions of the oracle top-10% group and compared against the same cohort’s unaided selection.

At the top-decile boundary, the unaided oracle overlap is 74.5%, whereas heterogeneous weakness-targeted gains reduce it to 50.5%. This is a substantial change in a concrete decision rather than simply an abstract loss of rank correlation. However, it is not equivalent to the percentage of rank reversals: only some inversions move candidates across the selected-set boundary. The strength-targeted scenario produces frequent pairwise changes while preserving almost all of the baseline top-decile accuracy, showing why RQ2 cannot be inferred directly from RQ1. Conversely, the equal-variance permutation control approaches the expected overlap between unrelated sets of the same size, demonstrating that an unchanged score distribution does not protect competitive selection.

Figure 3 compares oracle selection overlap across the prespecified thresholds for the baseline and five assistance conditions. The size of the selected set changes the expected effect because additional writers farther from the top boundary enter the candidate pool. The figure therefore evaluates the stability of the interpretation across thresholds without assuming that the 10% result generalizes unchanged to other decisions.

Figure 3
Selection accuracy under changing top-k thresholds



Note. The plotted statistics are synthetic Monte Carlo summaries under the stated generating mechanisms; graphical trends do not establish empirical learner-level effects.

Newly excluded versus recovered oracle-qualified writers

Selection instability is not synonymous with wrong selection. The assisted top-k set can differ from the unaided set because it newly excludes oracle-qualified candidates or because it corrects earlier exclusions. Table 5 reports both directions for the top-decile comparison and thereby allows a net accuracy change to be decomposed. This is the selection counterpart of the pairwise new-versus-resolved identity: assisted recall minus baseline recall is the negative of newly excluded minus newly rescued oracle-qualified candidates, divided by the target set size. Paired analysis is essential because two conditions may have the same overall overlap while selecting different specific writers, a distinction that an aggregate percentage alone would obscure.

The interpretation is conditional on the oracle itself. In this simulation the oracle is the equally weighted average of four latent domains, so selected membership represents success relative to that constructed target and nothing else. A real selection system might wish to rank collaborative writing quality, independent language control, disciplinary argumentation, or task-specific competencies, and these would be different targets. The fact that the top-k simulation is operationally interpretable does not justify recommending a particular high-stakes institutional cutoff. External application would require an explicit validity argument concerning the target, tasks, scoring quality, and unequal costs of false inclusion and false exclusion.

Results — RQ3: Recovering Rankings Under Different Evidence Regimes

Strictly monotonic equating and its theoretical limit

A strictly increasing calibration function preserves the sign of all strict observed score differences. Consequently, monotonic equating applied only to assisted scores cannot reverse an ordering that conflicts with the latent target. This identity is not a simulation discovery; the recovery experiment checks whether common fitted procedures behave consistently with it on held-out writers. Linear calibration produces the same pair ordering as the uncalibrated assisted score when its fitted slope is positive, as observed throughout the selected scenarios. A nondecreasing isotonic fit may create ties and can alter Kendall tau-b through the treatment of those ties, but it does not supply genuine strict inversions that correct an erroneous order. For selection, tie-breaking can even reduce oracle overlap slightly; that outcome must not be described as recovery simply because a correlation coefficient increases.

Figure 6 and Table 6 evaluate the substantive comparison in the heterogeneous weakness-targeted scenario. They report top-decile overlap against the latent oracle for score-only and information-enhanced conditions, rather than improvement in agreement with unaided scores. The recovery models use independent training and held-out learners and are not tuned after observing the test result. Their differences therefore reveal what additional evidence can

contribute within this generator. A product-only normalization that appears to improve level prediction still should not be credited with fixing an inverted relative ranking unless it changes the actual ordering of cases.

Table 6

Held-out ranking recovery under explicit information regimes

Scenario	Model/evidence condition	Kendall τ	Top-10 overlap	MCSE overlap
Uniform gain	Uncorrected assisted score	0.768	0.755	0.0048
Uniform gain	Linear score calibration	0.768	0.755	0.0048
Uniform gain	Isotonic score calibration	0.779	0.748	0.0050
Uniform gain	Metadata-aware regression	0.767	0.752	0.0048
Uniform gain	Exact gain subtraction	0.768	0.755	0.0048
Monotone compression	Uncorrected assisted score	0.768	0.755	0.0048
Monotone compression	Linear score calibration	0.768	0.755	0.0048
Monotone compression	Isotonic score calibration	0.779	0.748	0.0050
Monotone compression	Metadata-aware regression	0.767	0.753	0.0048
Monotone compression	Exact gain subtraction	0.768	0.755	0.0048
Weak-targeted heterogeneous	Uncorrected assisted score	0.487	0.512	0.0060
Weak-targeted heterogeneous	Linear score calibration	0.487	0.512	0.0060
Weak-targeted heterogeneous	Isotonic score calibration	0.500	0.511	0.0059
Weak-targeted heterogeneous	Metadata-aware regression	0.596	0.617	0.0056
Weak-targeted heterogeneous	Exact gain subtraction	0.768	0.755	0.0048

Scenario	Model/evidence condition	Kendall τ	Top-10 overlap	MCSE overlap
Strong-targeted heterogeneous	Uncorrected assisted score	0.766	0.750	0.0052
Strong-targeted heterogeneous	Linear score calibration	0.766	0.750	0.0052
Strong-targeted heterogeneous	Isotonic score calibration	0.777	0.746	0.0051
Strong-targeted heterogeneous	Metadata-aware regression	0.835	0.821	0.0045
Strong-targeted heterogeneous	Exact gain subtraction	0.768	0.755	0.0048
Random positive gain	Uncorrected assisted score	0.668	0.650	0.0058
Random positive gain	Linear score calibration	0.668	0.650	0.0058
Random positive gain	Isotonic score calibration	0.680	0.646	0.0057
Random positive gain	Metadata-aware regression	0.767	0.753	0.0049
Random positive gain	Exact gain subtraction	0.768	0.755	0.0048
Nonlinear weak targeting	Uncorrected assisted score	0.504	0.564	0.0060
Nonlinear weak targeting	Linear score calibration	0.504	0.564	0.0060
Nonlinear weak targeting	Isotonic score calibration	0.517	0.555	0.0063
Nonlinear weak targeting	Metadata-aware regression	0.612	0.671	0.0054
Nonlinear weak targeting	Exact gain subtraction	0.768	0.755	0.0048
Added measurement noise	Uncorrected assisted score	0.667	0.649	0.0055
Added measurement noise	Linear score calibration	0.667	0.649	0.0055

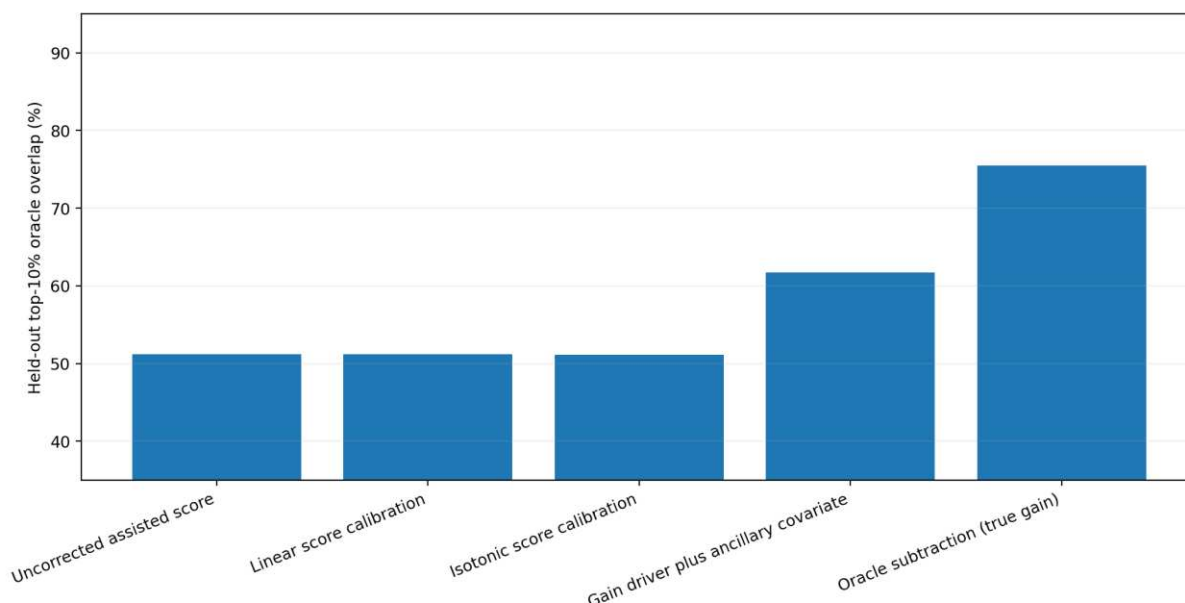
Scenario	Model/evidence condition	Kendall τ	Top-10 overlap	MCSE overlap
Added measurement noise	Isotonic score calibration	0.679	0.644	0.0056
Added measurement noise	Metadata-aware regression	0.666	0.646	0.0053
Added measurement noise	Exact gain subtraction	0.768	0.755	0.0048

Note. Recovery results use 160 independently generated cohorts, with 700 training and 300 test writers per cohort. True gain subtraction uses unavailable simulator truth, not a feasible assessor’s model or an upper bound on all statistical procedures. Isotonic calibration may induce ties, complicating direct interpretation of tau-b changes.

Figure 6 isolates the most diagnostically relevant weak-targeted scenario and contrasts score-only transformations with models given privileged information. Improvements under metadata-aware regression are conditional on the simulator providing an accurate gain driver and unrelated ancillary beta covariate, so their magnitude is not evidence of a real-world recovery tool.

Figure 6

Top-decile ranking recovery under heterogeneous weakness targeting



Note. The plotted statistics are synthetic Monte Carlo summaries under the stated generating mechanisms; graphical trends do not establish empirical learner-level effects.

Metadata and true-gain benchmarks

The metadata-aware regression uses the observed assisted score, the synthetic gain driver, and an ancillary beta-distributed variable that does not enter the gain generator. In the weak-targeted scenario, its mean top-decile overlap is 61.7%, versus 51.2% for the product alone. This improvement arises principally from access to the true gain driver and is meaningful only within the stipulated generator; the independently sampled ancillary variable contributes no designed gain information. The true-gain subtraction reference restores the original noisy unaided score before regression and therefore recovers approximately the unaided ranking. Its performance is not uniformly the best across all scenarios because assistance gains themselves can be informative about proficiency and a flexible metadata model can exploit that signal.

This contrast explains the practical boundary of RQ3. Ranking loss cannot be repaired by transforming one scalar monotonically, but some of the lost rank information can be recovered by adding information about the gain mechanism. The stronger model should therefore be read as a conditional information benchmark, not evidence that a specific real AI platform records enough accurate metadata to recover independent competence. A valid empirical application would require measured assistance use, evidence of its relation to independent proficiency, and a separated validation population. None of those requirements is satisfied by synthetic gain variables alone.

Independent Confirmation and Extended Robustness

Replication under an independent seed family

To test whether the primary effects depend on one particular Monte Carlo random stream, 500 new cohorts were generated using a distinct seed family without reusing primary observations. Table 7 compares the resulting net pair error and top-decile oracle overlap with the primary estimates for each prespecified mechanism. Agreement in direction and similar magnitude supports numerical reproducibility under the same score-generating assumptions, but it is not an alternative model of the world. The separate simulation family does not supply empirical validation, because it shares the same latent structure and gain equations. Its purpose is to distinguish stable conditional Monte Carlo patterns from results driven by an unusually favorable random draw.

Table 7

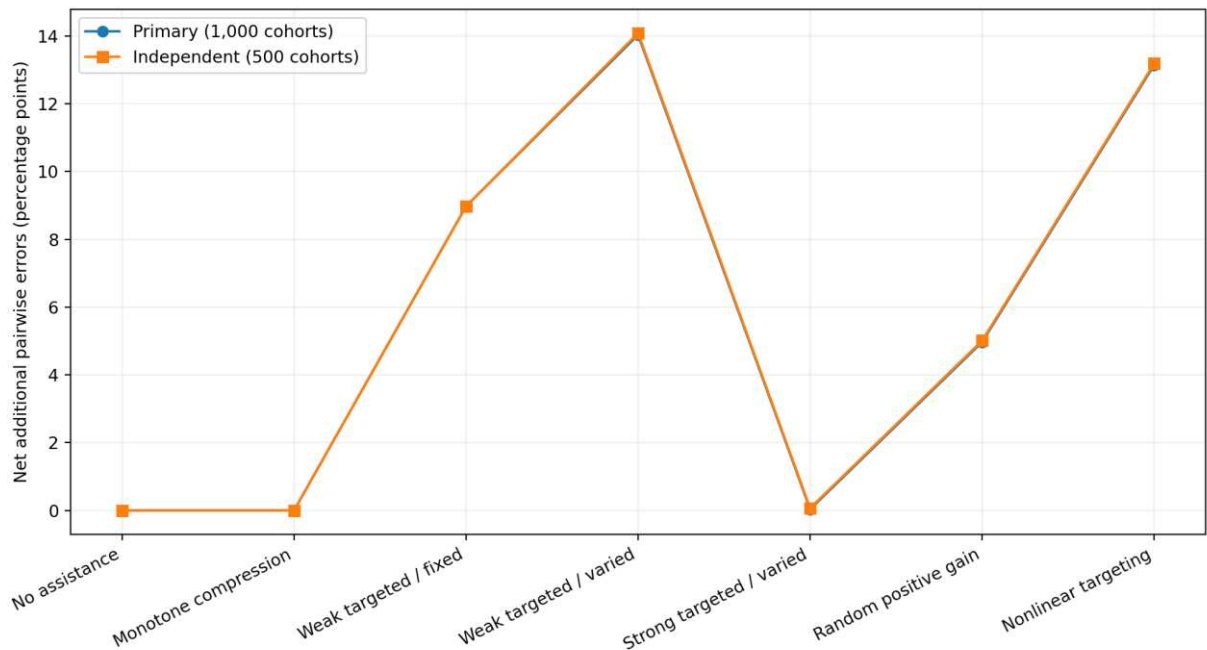
Primary and independently regenerated confirmation estimates

Condition	Primary net error pp	Confirmation net error pp	Primary top-10 overlap	Confirmation top-10 overlap
No assistance	+0.00	+0.00	74.5%	74.6%
Uniform gain	+0.00	+0.00	74.5%	74.6%
Monotone compression	+0.00	+0.00	74.5%	74.6%
Weak-targeted deterministic	+8.98	+8.97	58.0%	58.1%
Weak-targeted heterogeneous	+14.03	+14.08	50.5%	50.9%
Strong-targeted heterogeneous	+0.04	+0.07	74.2%	74.4%
Random positive gain	+4.97	+5.01	63.9%	64.0%
Nonlinear weak targeting	+13.14	+13.19	55.1%	55.3%
Added measurement noise	+5.12	+5.14	63.9%	63.9%
Equal-variance permutation	+38.38	+38.45	10.1%	9.9%

Note. Primary means are from 1,000 synthetic cohorts; confirmation means are from 500 independently regenerated cohorts using seed 20261019. Net errors are percentages of sampled pairs and changes relative to the paired unaided baseline.

Figure 4 displays the primary and confirmation net error means for a representative subset of conditions. The concordant trajectories support computational stability within the specified simulation assumptions; they should not be interpreted as reproducibility across independent empirical populations or different AI systems.

Figure 4
Independent confirmation of net pairwise error



Note. The plotted statistics are synthetic Monte Carlo summaries under the stated generating mechanisms; graphical trends do not establish empirical learner-level effects.

Changes in domain dependence, score error, latent distribution, and score bounds

The robustness ensemble changes features of the latent proficiency distribution and observation process rather than merely increasing the primary replication count. Low and high domain correlations alter the reliability of the aggregate target; low and high baseline score error alter the proportion of errors already present before assistance. Heavy-tailed and skewed proficiency distributions challenge normality, while clipped-and-rounded outcomes create observed ties absent from the primary specification. Table 8 retains each factor and compares targeted, strength-targeted, and random gain cases against the zero condition. Every row represents 120 independently regenerated cohorts, and its uncertainty is available in the replication-level file.

Table 8*Full sensitivity grid across latent distributions and score properties*

Robustness condition	Assistance	Net error pp	Reversals %	Top-10 accuracy %	Tie pairs %
Low correlation	No assistance	+0.0	0.0	69.3	0.0
Low correlation	Weak-heterogeneous	+15.4	21.9	41.0	0.0
Low correlation	Strong-heterogeneous	+0.2	11.3	69.1	0.0
Low correlation	Random positive gain	+5.8	14.5	58.0	0.0
Low correlation	Nonlinear weak targeting	+15.2	21.7	44.4	0.0
High correlation	No assistance	+0.0	0.0	78.4	0.0
High correlation	Weak-heterogeneous	+12.8	17.2	59.1	0.0
High correlation	Strong-heterogeneous	+0.1	7.8	78.0	0.0
High correlation	Random positive gain	+4.5	10.7	68.5	0.0
High correlation	Nonlinear weak targeting	+11.6	16.2	63.7	0.0
Low score error	No assistance	+0.0	0.0	84.9	0.0
Low score error	Weak-heterogeneous	+15.9	20.3	56.6	0.0
Low score error	Strong-heterogeneous	+3.0	8.9	78.0	0.0
Low score error	Random positive gain	+7.5	12.7	68.7	0.0

Robustness condition	Assistance	Net error pp	Reversals %	Top-10 accuracy %	Tie pairs %
Low score error	Nonlinear weak targeting	+15.2	19.6	61.1	0.0
High score error	No assistance	+0.0	0.0	59.9	0.0
High score error	Weak-targeted heterogeneous	+12.2	17.1	40.5	0.0
High score error	Strong-targeted heterogeneous	-3.0	9.5	66.9	0.0
High score error	Random positive gain	+2.7	11.1	54.9	0.0
High score error	Nonlinear weak targeting	+11.2	16.3	44.5	0.0
Heavy-tailed proficiency	No assistance	+0.0	0.0	75.9	0.0
Heavy-tailed proficiency	Weak-targeted heterogeneous	+14.3	20.2	50.4	0.0
Heavy-tailed proficiency	Strong-targeted heterogeneous	+0.0	10.2	75.7	0.0
Heavy-tailed proficiency	Random positive gain	+5.2	13.2	65.0	0.0
Heavy-tailed proficiency	Nonlinear weak targeting	+13.8	19.7	54.4	0.0
Skewed proficiency	No assistance	+0.0	0.0	81.1	0.0
Skewed proficiency	Weak-targeted heterogeneous	+14.3	20.2	60.9	0.0
Skewed proficiency	Strong-targeted heterogeneous	+0.1	10.1	81.1	0.0

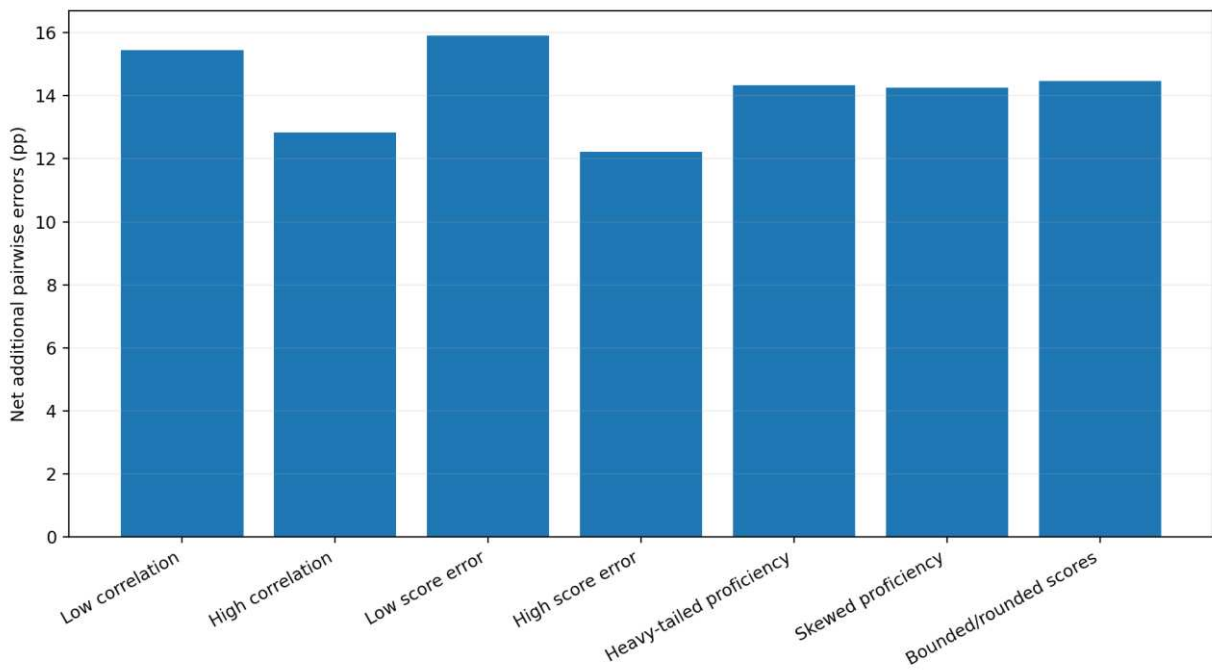
Robustness condition	Assistance	Net error pp	Reversals %	Top-10 accuracy %	Tie pairs %
Skewed proficiency	Random positive gain	+5.2	13.2	72.5	0.0
Skewed proficiency	Nonlinear weak targeting	+13.8	19.9	64.7	0.0
Bounded/rounded scores	No assistance	+0.0	0.0	74.7	2.8
Bounded/rounded scores	Weak-targeted heterogeneous	+14.5	22.3	50.8	3.9
Bounded/rounded scores	Strong-targeted heterogeneous	-0.4	11.4	74.1	2.1
Bounded/rounded scores	Random positive gain	+4.8	14.6	64.0	2.7
Bounded/rounded scores	Nonlinear weak targeting	+13.5	21.4	55.2	3.8

Note. Each condition×scenario cell contains 120 synthetic cohorts. Net errors and reversals derive from 18,000 uniformly sampled unordered pairs per cohort. Bounded/rounded outcomes treat observed ties as failures of strict oracle ordering; the primary unbounded scenarios have negligible ties.

Weakness-targeted gains produce positive net additional pairwise error under all seven tested score structures, though the size varies with domain correlation and observation noise. Strength-targeted gains are not uniformly neutral: in the high-noise condition they produce negative mean net error, indicating a simulated *improvement* in ordering accuracy relative to the noisy baseline. This is a scientifically important counterexample to the claim that heterogeneity necessarily worsens ranking. The bounded/rounded scenario introduces tie pairs, showing why tie handling must be specified rather than silently changing the denominator. Taken together, the grid supports a mechanism-contingent conclusion with identifiable boundary conditions instead of a universal claim about AI-assisted revision.

Figure 7 isolates the weakness-targeted scenario to make its robustness pattern readable. The variation across the conditions emphasizes that the numerical effect depends on the signal-to-noise relation and latent distribution even when the direction is relatively stable. The table provides the complementary results for other gains, including cases where the net error is close to or below zero.

Figure 7
Robustness of weakness-targeted net ranking error to changed score assumptions



Note. The plotted statistics are synthetic Monte Carlo summaries under the stated generating mechanisms; graphical trends do not establish empirical learner-level effects.

Assistance intensity as a graded mechanism

A separate 160-cohort experiment rescales each realized gain by a factor of zero, one-half, one, or one-and-one-half, holding the latent profiles and unaided score fixed. This is more informative than comparing only a weak-targeted and a strong-targeted functional form because it demonstrates whether the same gain-allocation rule changes ranking accuracy as its magnitude increases. Table 9 reports the full three-by-four intensity grid for weakness-targeted, nonlinear, and untargeted stochastic positive gains. The zero scale is an exact null and is checked to produce no net newly introduced error. The remaining values should be read as conditional dose-response curves, not estimates of what would happen if an actual student accepted a particular percentage of AI feedback.

Table 9
Full gain-intensity sweep across three heterogeneous mechanisms

Gain mechanism	Multiplier	Net error pp	Top-10 overlap %	New pair errors %
Nonlinear weak targeting	0.0	+0.0	74.6	0.0

Gain mechanism	Multiplier	Net error pp	Top-10 overlap %	New pair errors %
Nonlinear weak targeting	0.5	+4.5	69.0	6.2
Nonlinear weak targeting	1.0	+13.1	55.7	15.8
Nonlinear weak targeting	1.5	+23.2	36.9	26.5
Random positive gain	0.0	+0.0	74.6	0.0
Random positive gain	0.5	+1.5	71.5	3.9
Random positive gain	1.0	+4.9	64.2	8.5
Random positive gain	1.5	+8.8	56.3	13.0
Weak-targeted heterogeneous	0.0	+0.0	74.6	0.0
Weak-targeted heterogeneous	0.5	+4.7	67.1	6.4
Weak-targeted heterogeneous	1.0	+14.0	50.9	16.6
Weak-targeted heterogeneous	1.5	+25.4	29.0	28.6

Note. Each of the 12 cells summarizes 160 independently regenerated synthetic cohorts. The multiplier is applied to the realized gain with latent profiles and baseline scores held fixed within a cohort. Scale zero must preserve all original pairwise and selection outcomes.

Net pairwise error rises with gain scale under all three mechanisms in this design, but the slopes differ markedly. Weakness-targeted gains combine score shifts systematically related to independent proficiency with a random component, so a sufficiently large magnitude erodes the original ordering more sharply than a smaller shift. Untargeted stochastic gains also cause deterioration by introducing unpredictable variation, yet their net effect grows differently from the targeted mechanism. The experiment does not establish a universal dose-response law because clipping, heterogeneity, and domain scores differ across plausible settings. Its contribution is to show that intensity must be distinguished from gain direction and distribution when evaluating comparative validity.

Monte Carlo uncertainty and inferential boundaries

Precision of the conditional Monte Carlo mean is not the same as variability among synthetic cohorts. Table 10 therefore provides both MCSE and the 2.5th–97.5th percentile interval of independently generated cohort-level net error for selected scenarios. With 1,000 primary replications the mean may be estimated very precisely even though cohort-level outcomes vary materially. This difference matters particularly for strength targeting, where a near-zero mean arises from positive and negative changes that offset one another, not necessarily from a scenario in which every cohort has no ranking changes. The manuscript does not use significance testing against real-world population hypotheses because the simulated mechanisms themselves have been chosen rather than estimated from a sampled educational population.

Table 10

Mean net ordering error and variation among independent synthetic cohorts

Condition	Mean net error pp	MCSE (pp)	2.5th percentile pp	97.5th percentile pp
No assistance	+0.00	0.000	+0.00	+0.00
Monotone compression	+0.00	0.000	+0.00	+0.00
Weak-targeted heterogeneous	+14.03	0.020	+12.83	+15.25
Strong-targeted heterogeneous	+0.04	0.012	-0.66	+0.77
Random positive gain	+4.97	0.014	+4.14	+5.89
Nonlinear weak targeting	+13.14	0.020	+11.94	+14.36

Note. MCSE is the estimated SD of cohort-level net error divided by $\sqrt{1,000}$ and measures numerical uncertainty in the simulated mean. Percentiles describe distribution across synthetic cohorts, not confidence intervals for effects in an observed population.

The distinction also limits what independent confirmation means. Similar primary and confirmation means support that the reported conditional comparisons are not driven by one seed family, but they cannot resolve misspecification of the data-generating model. Robustness across alternate distributions, observation error, score bounds, and gain functions addresses some model uncertainty without eliminating it. An authentic evaluation of relative proficiency ranking after AI assistance would require independent writing samples, defensible proficiency anchors, comparable tasks, measured revision opportunities, and possibly repeated decisions across occasions. The present design is useful because these requirements and decision

consequences are explicit, not because synthetic outcomes automatically transport to policy settings.

Pairwise ranking errors conditional on true proficiency separation

The aggregate pairwise estimand weights all unordered writer pairs equally, regardless of whether two writers have nearly identical latent proficiency or substantially different values. This may conceal an educationally important boundary condition because pairs separated by a large proficiency margin need larger differential assistance gains to reverse their order. A new 240-cohort experiment therefore conditions sampled pairs on four prespecified bands of absolute latent-proficiency difference (at most .25; .25–.50; .50–1.00; and above 1.00). Each independent cohort comprises 1,000 synthetic writers with a common set of 25,000 uniformly sampled pairs across six paired gain conditions, including an exact no-assistance control. The experiment retains the original scoring model but also adds an explicitly new alternative in which the beta-distributed individual assistance variable actually multiplies the intended positive gain. Specifically, the new gain is $\text{clip}\{\text{clip}(8 - 6\theta + 4z, 0, 28) \times (0.25 + 1.5a), 0, 28\}$, with a distributed $\text{Beta}(2.2, 2.2)$, z standard normal, and θ denoting the same independent global proficiency as in the primary generator. The 240-cohort ensemble uses root seed 2026100991, and the separately regenerated 160-cohort tie-policy analysis uses root seed 2026100992. Unlike the earlier ancillary variable in the recovery benchmark, this intensity factor enters the new gain equation and is fully disclosed as a separate sensitivity mechanism.

Table 11 reports the complete six-condition by four-margin-band grid. For each scenario and band, it decomposes net additional ordering error into newly incorrect comparisons and errors resolved relative to latent independent proficiency. Denominators are all sampled pairs within the relevant proficiency band, not all possible writer pairs in the cohort; therefore, the rows should not be added without their corresponding band weights. Replication-level data also retain the number and prevalence of sampled pairs in each band, baseline and assisted error, reversal rate, cohort-seeded estimates, MCSE, and cohort quantiles. This conditional design can expose a nonmonotone risk pattern that aggregate Kendall correlation cannot describe.

Table 11*Newly incorrect, resolved, and net pairwise ordering errors by proficiency separation*

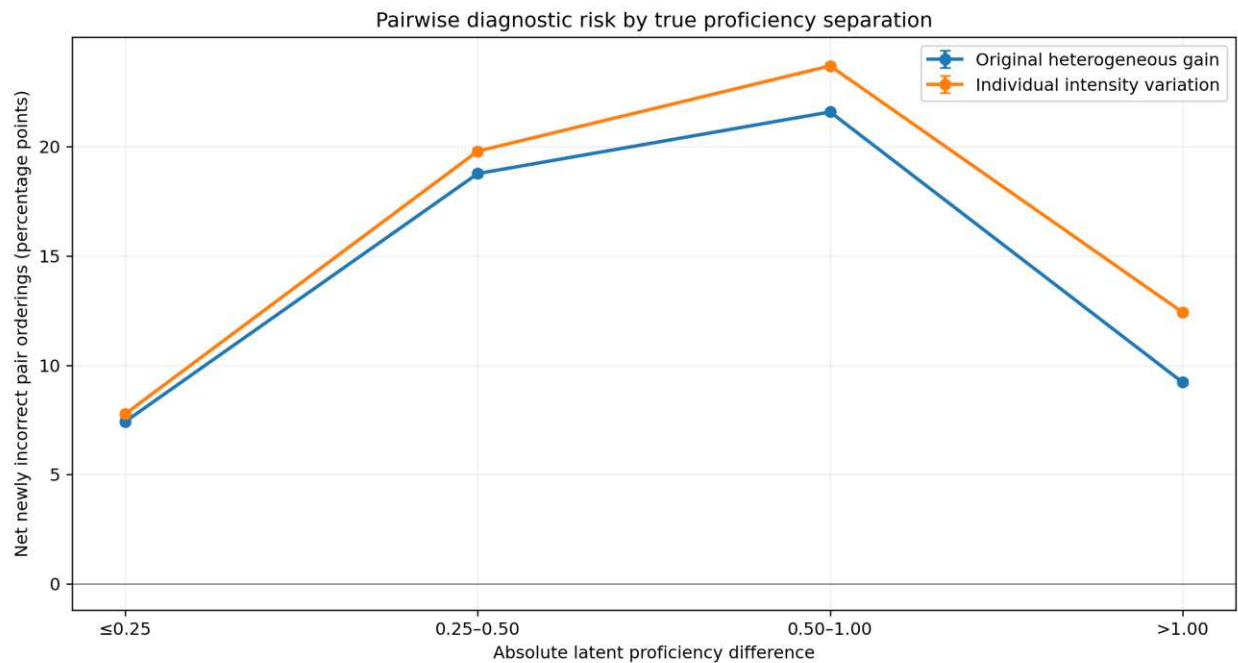
Mechanism	$\Delta\theta$ band	New %	Resolved %	Net pp
Z	≤ 0.25	0.0	0.0	+0.0
Z	0.25–0.50	0.0	0.0	+0.0
Z	0.50–1.00	0.0	0.0	+0.0
Z	> 1.00	0.0	0.0	+0.0
W	≤ 0.25	16.8	9.4	+7.4
W	0.25–0.50	22.6	3.8	+18.8
W	0.50–1.00	22.3	0.7	+21.6
W	> 1.00	9.2	0.0	+9.2
WI	≤ 0.25	18.4	10.6	+7.8
WI	0.25–0.50	24.3	4.5	+19.8
WI	0.50–1.00	24.6	0.9	+23.7
WI	> 1.00	12.4	0.0	+12.4
H	≤ 0.25	12.3	12.2	+0.1
H	0.25–0.50	8.5	8.4	+0.1
H	0.50–1.00	2.9	2.7	+0.1
H	> 1.00	0.1	0.1	+0.0
R	≤ 0.25	14.9	11.1	+3.8
R	0.25–0.50	15.1	6.1	+9.0
R	0.50–1.00	9.5	1.6	+7.9
R	> 1.00	1.4	0.0	+1.4
N	≤ 0.25	16.8	9.5	+7.2
N	0.25–0.50	22.2	4.0	+18.3
N	0.50–1.00	21.4	0.8	+20.6
N	> 1.00	7.8	0.0	+7.8

Note. Mechanism codes: Z = no assistance; W = heterogeneous weakness targeting; WI = intensity-dependent weakness targeting; H = higher-proficiency targeting; R = untargeted positive gain; N = nonlinear weakness targeting. Six conditions $\times$ four margin bands; 240 synthetic cohorts per condition (1,000 learners and 25,000 sampled pairs per cohort). Each percentage is conditional on the indicated margin band; net pp equals newly incorrect minus resolved. For all conditions, the same latent target and unaided scores are used within a paired cohort. Intensity-dependent weak targeting is a newly specified robustness mechanism, not a replacement for the primary analysis.

The conditional results reveal why the statement that ranking errors mainly concern near-identical writers would be incomplete. Under heterogeneous weakness targeting, the net increase in wrong pair orderings is approximately 7.4 percentage points for latent differences

no greater than .25, rises to 18.8 points between .25 and .50, and peaks at about 21.6 points between .50 and 1.00 before declining to 9.2 points for differences above 1.00. At the smallest differences, many comparisons are already unreliable in the unaided condition, so assistance-induced changes can both introduce and correct errors. For intermediate differences the initial ordering is more often accurate, yet targeted assistance can still close and reverse the margin. At the greatest separations larger relative gains are needed to overturn the original order. This is a conditional property of the specified gain scale, not a general population law.

Figure 9
Change in pairwise ranking error across absolute latent-proficiency margins



Note. Points are independently regenerated Monte Carlo means over 240 cohorts, with error bars showing ± 1.96 Monte Carlo standard errors, not real-world confidence intervals. The intensity-dependent mechanism adds explicit learner-level variation in gain magnitude. See Table 11 for newly incorrect and resolved components.

The monel in which individual beta-distributed intensity affects the gain produces the same nonmonotone profile, with somewhat larger deterioration than the original stochastic weak-targeting mechanism across the four bands. The comparison strengthens mechanistic robustness without validating the assumed beta distribution empirically. It also clarifies what the earlier model does not show: the original beta covariate was generated but not used in the main gain function, so the new case, rather than the original analysis, is the valid basis for discussing explicitly heterogeneous individual intensity. Figure 9 isolates the two weakness-targeted structures to make their conditional patterns visible. More flexible interactions,

correlation between engagement and baseline scoring error, and longitudinal competence change remain untested.

Selection transitions at all cutoffs and tie-policy sensitivity

The primary selection table reports top-k oracle overlap for all four cutoffs but decomposes newly excluded and rescued oracle-qualified writers only for the top-decile case. A second new analysis reports that decomposition for the 5%, 10%, 20%, and 30% selection boundaries under all six paired mechanisms. Each fraction uses the fixed-size latent oracle set at that cutoff as its denominator, not the total cohort size. The paired net-exclusion identity is checked independently for every simulated cohort and decision threshold. This presentation permits comparison of selection churn and selection accuracy without assuming that errors near one cutoff generalize to another.

Table 12*Oracle-based selection transitions for four cutoffs across alternative gain mechanisms*

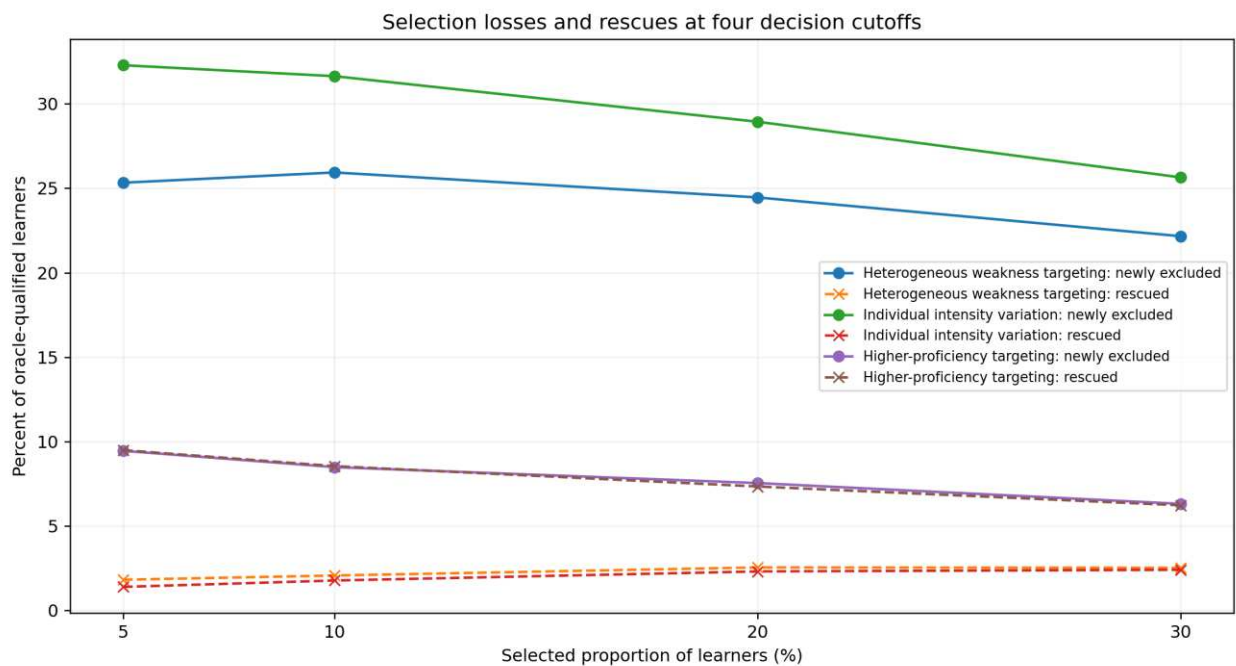
Mechanism	Cutoff	Baseline %	Assisted %	New exclusions %	Rescued %
Z	5%	70.3	70.3	0.0	0.0
Z	10%	74.7	74.7	0.0	0.0
Z	20%	79.7	79.7	0.0	0.0
Z	30%	83.2	83.2	0.0	0.0
W	5%	70.3	46.8	25.3	1.8
W	10%	74.7	50.8	25.9	2.1
W	20%	79.7	57.8	24.5	2.5
W	30%	83.2	63.5	22.2	2.5
WI	5%	70.3	39.5	32.3	1.4
WI	10%	74.7	44.8	31.6	1.8
WI	20%	79.7	53.1	28.9	2.3
WI	30%	83.2	59.9	25.6	2.4
H	5%	70.3	70.4	9.4	9.5
H	10%	74.7	74.7	8.5	8.5
H	20%	79.7	79.5	7.5	7.3
H	30%	83.2	83.1	6.3	6.2
R	5%	70.3	58.3	18.0	5.9
R	10%	74.7	64.4	15.8	5.5
R	20%	79.7	71.1	13.8	5.2
R	30%	83.2	76.1	11.6	4.5
N	5%	70.3	52.9	21.1	3.6
N	10%	74.7	55.4	22.2	2.9
N	20%	79.7	59.8	22.6	2.6
N	30%	83.2	63.9	21.6	2.4

Note. Mechanism codes: Z = no assistance; W = heterogeneous weakness targeting; WI = intensity-dependent weakness targeting; H = higher-proficiency targeting; R = untargeted positive gain; N = nonlinear weakness targeting. 240 synthetic cohorts of 1,000 writers for each scenario, with identical latent and baseline scores across paired comparisons. Baseline and assisted columns are top-k precision/recall against the latent oracle; new exclusions and rescued entries are proportions of the corresponding oracle top-k group. Net exclusion equals new exclusions minus rescued entries and equals baseline oracle overlap minus assisted oracle overlap. All values are conditional simulation averages.

Figure 10 displays the countervailing newly excluded and rescued proportions across decision cutoffs for heterogeneous weakness targeting, the new intensity-dependent mechanism, and higher-proficiency targeting. The two rates often differ substantially even when assisted and baseline oracle overlap are close. Consequently, stable selection accuracy can coexist with membership replacement, while a large decline in overlap indicates newly rejected oracle-qualified writers not offset by comparable rescues. The complete Table 12 identifies the same components for neutral, nonlinear, and no-assistance conditions rather than selecting only the most adverse case. The comparison does not assign educational utilities to different mistakes; it quantifies allocation validity against one explicitly defined simulated target.

Figure 10

New exclusion and rescue of oracle-qualified learners across competitive cutoffs



Note. All plotted proportions refer to the latent oracle-qualified group at the indicated selection cutoff. Results are simulated means over 240 paired cohorts, not observations of educational admissions or scholarship selection.

A further sensitivity check isolates the consequence of the tie convention. In 160 separately regenerated cohorts, unaided and assisted scores are bounded to 25–100 and rounded to integers. The original primary convention scores every observed tie as an incorrect strict comparison. Alternatives assign one-half error credit to ties, or compare only pairs that remain untied in both conditions. The last rule changes the denominator and accordingly answers a narrower question; it is shown as a robustness contrast rather than claimed to be directly equivalent. Table 13 reports the resulting paired net-error changes and baseline/assisted tie prevalence under the no-assistance, weak-targeted, strong-targeted, and nonlinear scenarios.

Table 13*Sensitivity of net ranking error to rounded-score tie policies*

Mechanism	Unaided ties %	Assisted ties %	Ties wrong, net pp	Half credit, net pp	Both untied, net pp
No assistance	2.9	2.9	+0.00	+0.00	+0.00
Weak- targeted varying	2.9	3.9	+14.59	+14.06	+13.22
Strong- targeted varying	2.9	2.1	-0.32	+0.07	+0.16
Nonlinear weak- targeted	2.9	3.8	+13.60	+13.15	+12.34

Note. 160 independently regenerated synthetic cohorts per mechanism, with paired samples of 25,000 unordered writer pairs each. “Ties wrong” assigns error 1 to any observed tie, “Half credit” assigns error .5, and “Both untied” restricts analysis to pairs without either unaided or assisted tie; the latter changes the estimand. Negative net error represents an improvement relative to unaided scores.

Strong-targeted gains are especially sensitive because the estimated net effect is close to zero. Under the ties-as-wrong convention, their mean paired net change is approximately -0.32 percentage points, whereas half-credit produces about $+0.07$ and the common untied-pair comparison about $+0.16$ points. These are small effects, but the sign is not stable to scoring conventions, so the correct conclusion is near-neutrality under this bounded setting, not a universal improvement. In contrast, heterogeneous weakness targeting yields a substantial positive net error under all three conventions, although its exact magnitude changes. This check makes the bounded-score robustness claim more defensible while revealing a genuine interpretive boundary of near-zero results.

Discussion

RQ1: improvement, reordering, and relative validity

The primary result for RQ1 is that assistance-induced improvement and the correctness of a comparative judgment are separable quantities. Uniform gains and increasing compression leave rank ordering unchanged, whereas heterogeneous gains can alter ordering even if every assisted score is at least as large as its unaided counterpart. The direction and balance of those changes depend on whether gains are related to proficiency or distributed independently of it. Under heterogeneous weakness targeting, many more pairs become incorrect than are resolved,

producing a clear net deterioration. Under higher-proficiency targeting, newly introduced and corrected pairwise errors nearly cancel despite nontrivial reversal rates. This is not a contradiction: moving an observed score relative to other noisy scores can sometimes worsen and sometimes correct the initial approximation to the latent target.

The zero, uniform, and compression controls serve an important scientific function beyond convenience. They establish that a monotonic score transformation and an increase in score level should not, on their own, be interpreted as an impairment of rank validity. The equal-variance permutation provides the complementary counterexample: identical mean and variance reveal nothing about how scores are assigned to individual writers, so they cannot guarantee accuracy of comparative interpretation. The score-noise condition adds a further distinction by demonstrating ranking deterioration in the absence of a positive assistance gain. In combination, these controls prevent the simulation from building the conclusion into its outcome labels and show why rank correctness requires an oracle-referenced error estimand.

The between-cohort analyses also calibrate the strength of the inference. Mean results under weakness targeting remain positive across the tested settings, but their sizes change with latent proficiency correlation, measurement precision, and gain magnitude. The negative mean net error under strength targeting and high score noise reveals a condition in which preferential gains can improve comparative accuracy, while the same mechanism under other settings is approximately neutral. These boundary findings should be foregrounded rather than treated as exceptions to be explained away. They reinforce the argument that AI assistance cannot be declared ranking-neutral or ranking-damaging without specifying its mechanism and the assessment target.

RQ2: selection accuracy is not captured by rank correlations alone

The RQ2 analysis shows that rank errors have decision consequences, but the mapping from pairwise instability to selection harm is not one-to-one. Weakness-targeted heterogeneous gains reduce top-decile overlap with the latent oracle substantially, while strength-targeted gains create some pairwise reversals without materially reducing the fraction of selected writers who truly belong to the latent top group. This difference follows because rank changes far from the selection boundary can have little effect on the selected set. At the same time, a small number of reversals near that boundary may exchange an oracle-qualified writer for an unqualified one. Reporting four selection thresholds and both newly excluded and rescued candidates makes the assessment use visible rather than relying on a single global rank coefficient.

The correct target of the selection inference is especially important. Comparing AI-assisted selections against unaided selections measures stability, but unaided scores also contain error. Some AI-related changes might improve oracle agreement by correcting unaided mistakes, a possibility visible in the strength-targeted condition. Consequently, a policymaker cannot

equate low agreement with an unauthorized advantage or high agreement with valid ranking. The simulation does not propose a particular admissions, scholarship, or placement rule; it illustrates the information required to evaluate such uses. Real selection thresholds would additionally need evidence about decision costs, task sampling, score reliability, and fairness across populations rather than relying on one synthetic top-k percentage.

RQ3: why normalization is not recovery of ordering

The mathematical limit of score-only monotone equating is straightforward but often overlooked in discussions of AI-adjusted assessment scales. A strictly increasing function may repair location, spread, or the numerical interpretation of score levels while preserving all strict pairwise relationships, including incorrect ones. The simulated linear calibration reproduces that rank-invariance property on held-out cases. Nondecreasing isotonic calibration requires extra care because it can group scores into ties, changing certain rank statistics or arbitrary tie-break positions without restoring the latent relative order. This is why a change in Kendall tau-b after isotonic fitting should not be described as a recovered strict rank when top-decile oracle overlap does not improve. The decisive question is whether additional evidence supports a justified reordering of writers.

The metadata-aware model illustrates the role of such evidence while exposing an equally important limitation. Its inputs include synthetic variables directly linked to the gain generator, allowing it to distinguish some product-score changes associated with assistance from variation in proficiency. Its improved top-decile overlap is therefore an information-advantage result, not an indication that a common AI-use disclosure would achieve the same accuracy in a classroom. Similarly, subtracting the true gain returns the unaided score by construction but is not automatically the optimal possible predictor; gain may contain its own information about proficiency. Honest labeling of these regimes prevents a benchmark from being mistaken for a practical correction procedure and links RQ3 to the wider validity requirement that the evidence used by an assessor be observable and defensible.

Implications for educational measurement and future research

The methodological implication is that evaluation of AI-assisted writing scores should make its comparison target explicit. If an assessment is intended to measure the quality of a supported text, a change in relative ranking may be appropriate or even desirable. If it is intended to rank writers by independently deployable writing proficiency, the same change raises a different validity question requiring external evidence. A comparative validity evaluation should therefore track latent or independently anchored proficiency, produce paired unaided and assisted scores under comparable tasks, and separate changes in order from changes in correctness. Such evidence could be collected through repeated unaided anchors, observed revision traces, and carefully documented scoring procedures. None of these steps can be bypassed by strong association between assisted and unaided total scores.

The paper also offers a design principle for statistical research on assisted assessments. The unit of inferential precision should be independent cohorts or appropriately clustered learners, not the enormous number of overlapping pairwise comparisons. Selection analyses should report errors at actual decision margins, and model comparisons should specify the information available at training and at test time. Synthetic gain controls should include positive, negative, and null conditions to prevent an assumed harm narrative from becoming self-fulfilling. These principles follow from the study's falsification results and from established argument-based approaches to score validity (Kane, 2013; Morris et al., 2019). They are methodological recommendations, not evidence that any particular school or test currently violates those principles.

Proficiency margins, selection cutoffs, and sensitivity to tie conventions

The extended pair-margin audit adds an important qualification to the global RQ1 inference. Very similar writers are more likely to start with an imperfect unaided ordering, so raw reversal probability or newly incorrect pairs alone may exaggerate the severity of net degradation without accounting for errors simultaneously corrected. In the simulated weakness-targeted process, the largest net deterioration instead occurs at intermediate latent distances, where the unaided ranking carries useful information but assistance can still close the performance margin. This pattern cautions against presuming either that all pair inversions are trivial near-ties or that they are all severe contradictions of a stable proficiency hierarchy. An empirical study should define acceptable skill-separation thresholds before analyzing the consequences of changed pairwise ordering.

For RQ2, the complete selection-transition grid reinforces the distinction between changes in membership and wrong selections against the specified target. Newly excluded and rescued oracle-qualified writers must be shown together; a small mean difference can mask considerable offsetting turnover. A bounded rating scale introduces a separate interpretive problem because ties are not genuine evidence of a strict ordering. The rounded-score sensitivity analyses show that the sign of a very small strong-targeted net effect is not invariant to the chosen tie convention, even when major weakness-targeted deterioration persists. Such results do not invalidate ranking analysis; they define the decision rule and reporting standard necessary before translating an ordinal statistic into a claim about assessment accuracy.

Methodologically, the new intensity-dependent scenario also corrects a potential misinterpretation of earlier ancillary covariates. The original gain equation did not multiply gains by the beta-distributed variable, whereas the extension now tests a prespecified alternative that does. Its similar conditional shape supports the relevance of heterogeneous gains but does not make the beta distribution empirically grounded. The most important remaining boundaries are the absence of genuinely observed assistance behavior, repeated-task scoring, differential rater or prompt effects, correlated gain and baseline measurement errors,

and an independently validated real-world proficiency criterion. These issues require new empirical evidence or specifically designed simulations; they cannot be resolved through additional Monte Carlo replication under unchanged assumptions.

Limitations and Directions for Empirical Validation

The first limitation is the absence of real learner writing data. The simulation assigns latent proficiency as a known quantity and converts it into a scalar score with selected error and gain functions, whereas actual proficiency is estimated from tasks and raters rather than observed without error. The domain labels are illustrative and do not constitute validated separate linguistic attributes. AI interactions are represented by stipulated score gains and do not model prompted revision decisions, response quality, or learning over time. Although the results demonstrate mathematical and conditional computational possibilities, they do not identify the prevalence, causal magnitude, or demographic distribution of ranking changes in an actual institution. Real-world claims would require prospective paired writing observations and independent target measures.

A second limitation concerns the pairwise sample and metric conventions. The 25,000-pair estimator is uniformly sampled and numerically checked against complete enumeration, but its estimates retain finite pair-sampling variation inside each cohort. Observed score ties are counted as strict-order failures, an explicit decision that may differ from conventions assigning half-credit or excluding ties. Selection uses deterministic fixed-size top-k sets and a latent equally weighted global target, rather than utility-weighted admissions outcomes or observed writing competencies. The robustness analysis includes bounded and rounded scales but does not cover every tie-breaking policy, rubric, or correlated scoring process. Such design choices should be evaluated against the intended score interpretation in any empirical extension.

A third limitation arises from the recovery models. The regression and isotonic comparisons use conventional fitted functions on 700 training learners and are evaluated on 300 independent test learners; they are not exhaustive searches for the best possible ranking procedure. The metadata-aware and true-gain reference conditions use simulator variables that may be unavailable or unreliable outside the model. Even an algorithm that restores ordering in a simulation would need cross-task and cross-population validation, an explicit independent-proficiency criterion, and a documented information regime before deployment. The current conclusions are therefore bounded by the specified predictor families, sample sizes, and gain structure. The value of these comparisons lies in separating mathematical impossibility for monotonic score-only reordering from recoverability when justified additional evidence is present.

Future empirical work could construct repeated unaided writing anchors, independently scored analytic domains, and matched assisted tasks for the same learners while recording the

prompts, feedback, and extent of revision. Such work should test whether revisions change relative ranking beyond task and rater error, and whether any changes introduce or correct mistakes against an independent criterion. It should also examine differential opportunities for assistance and clarify whether the intended construct is independent writing proficiency or human–AI collaborative performance. Where selection is the proposed use, researchers should explicitly prespecify cutoff fractions and the relative consequences of false inclusion and exclusion. These steps would allow the proposed framework to be tested rather than assumed, without transporting the synthetic percentages into policy recommendations.

Additional independently generated synthetic ensembles examine proficiency-gap-conditioned pair accuracy, selection transitions at all four cutoffs, and alternative conventions for scoring tied comparisons on a bounded rating scale. The new findings sharpen rather than replace the principal conclusions: the net error induced by a gain mechanism varies with the latent distance between writers; selection instability must be decomposed against the oracle at the actual cutoff; and near-zero ordering effects may be sensitive to tie handling. The supplementary individual-intensity mechanism is explicitly separate from the original model, whose beta-distributed covariate was ancillary. These results define further conditions under which the methodology can be falsified without making claims about authentic AI revision or real learner rankings.

Conclusion

This study provides a reproducible demonstration that score improvement and relative proficiency ranking are distinct assessment properties. Under the primary synthetic generator, monotone compression and uniform additive gains preserve strict ordering, while heterogeneous weakness-targeted gains create considerably more new ordering errors than they resolve. Other mechanisms can produce many reversals with near-zero net harm, or improve comparative accuracy under unfavorable baseline noise conditions. Consequently, RQ1 is answered conditionally: changes in ranking accuracy depend on the relationship between assistance gains, original proficiency margins, and score error rather than on improvement alone. The result is a statement about the defined data-generating system, not about typical real AI use.

RQ2 shows that selection consequences must be assessed against an explicit oracle rather than inferred from pairwise instability or agreement with unaided scores. The top-decile and alternative cutoff analyses demonstrate that the same gain mechanism can substantially change who is selected, but the impact depends on the selection boundary. RQ3 establishes the mathematical limit of strictly monotonic score-only equating and illustrates partial ranking recovery when the model is allowed to use synthetic assistance metadata. Independent confirmation, alternative distributions, intensity sweeps, and complete-pair estimator checks support the internal stability of these conditional findings. The central methodological

contribution is an auditable separation of observed rank reversals, newly introduced ranking errors, corrected errors, and decision-level selection accuracy.

Data and Reproducibility Statement

All observations and outcomes are synthetic. The companion research package contains 1,000 primary and 500 confirmation replication ensembles, exact independent cohort seeds, the 25,000-pair subsampling code, a representative 10,000-row learner-scenario example, recovery-model and robustness replications, an assistance-intensity experiment, complete-pair numerical verification, figures, and a styled summary workbook. Primary replication output includes 10,000 cohort–mechanism records; confirmation includes 5,000. The provided Python scripts regenerate the tables and validation files, and separate integrity checks verify pair identities, zero-effect controls, selection set sizes, and reproducibility outputs. No real learner scripts, personal data, or actual AI-revision traces were processed. The separately scripted audit extension provides an additional 240-cohort six-mechanism analysis of proficiency-gap-conditional pair outcomes and selection transitions across all four cutoffs, plus 160 cohorts comparing tie-scoring rules. It includes full replication records, summary data, exact root seeds, two figures, and programmatic paired-identity tests. The new model explicitly using learner assistance intensity is clearly distinguished from the main model where the ancillary beta variable does not affect gain.

Ethics, Funding, and Competing Interests

Ethics: This manuscript reports a synthetic methodological simulation; no human subjects or identifiable personal records were involved in generating the reported datasets. Funding: AUTHOR CONFIRMATION REQUIRED. Competing interests: AUTHOR CONFIRMATION REQUIRED. These declarations must be completed truthfully before public preprint posting or journal submission; they are not assumed to be absent.

References

- Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54–74.
<https://doi.org/10.1080/15434300903464418>
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3–4), 324–345.
<https://doi.org/10.1093/biomet/39.3-4.324>
- Brennan, R. L. (2001). *Generalizability theory*. Springer. <https://doi.org/10.1007/978-1-4757-3456-0>
- Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). Association for Computing Machinery. <https://doi.org/10.1145/1143844.1143874>

- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedem.12000>
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1–2), 81–93. <https://doi.org/10.1093/biomet/30.1-2.81>
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer. <https://doi.org/10.1007/978-1-4939-0317-7>
- Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: A study on proficiency, revision frequency and writing quality. *Discover Education*, 4, Article 170. <https://doi.org/10.1007/s44217-025-00602-7>
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. <https://doi.org/10.1002/sim.8086>
- Nguyen, H. S. T. (2026a, October 7). *Correction or redistribution? A simulation-based error-migration framework for AI-assisted EFL writing revision* [Preprint]. SSRN. <https://doi.org/10.2139/ssrn.7575258>
- Nguyen, H. S. T. (2026b, October 7). *Same score, different competence? A study of product–process non-identifiability in AI-assisted L2 writing assessment* [Preprint]. SSRN. <https://doi.org/10.2139/ssrn.7576659>
- Nguyen, H. S. T. (2026c, October 8). *When better writing hides weakness: A Monte Carlo study of diagnostic masking in AI-assisted L2 writing assessment* [Preprint]. SSRN. <https://doi.org/10.2139/ssrn.7583178>
- Nguyen, H. S. T. (2026d, October 7). *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>
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. <https://doi.org/10.1037/h0070288>
- Tuong, T. P. L., & Tran, T.-T. (2025). The differential impact of AI tools among EFL university learners: A process writing approach. *International Journal of Learning, Teaching and Educational Research*, 24(5), 452–471. <https://doi.org/10.26803/ijlter.24.5.24>
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732997>
- Whitehouse, C., & Pollitt, A. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment*. AQA Centre for Education Research and Policy. <https://www.aqa.org.uk/about-us/our-research/research-library/using-adaptive-comparative-judgement-to-obtain-a-highly-reliable-rank-order-in-summative-assessment>