STUDY RESULTS REPORT

Making Learning Visible

A retrospective cohort study of AI-supported English practice

Jinso analyzed historical learning activity from partner institutions to examine whether learners who participated more in recorded AI-supported English practice showed greater estimated change in modeled language performance.

740
LEARNERS
Included in the pooled first-to-last change analysis.

82
CLASSROOMS
Represented across five analytic cohorts.

5
ANALYTIC COHORTS
From four anonymized partner institutions.

Primary finding

Across the pooled analysis, greater recorded participation had a positive association with estimated change. The pooled estimate was +0.055 baseline SD per five recorded assignments, with an 80% confidence interval from -0.005 to +0.115.

INTERPRETATION

The findings are early and promising, but they are not causal evidence. The study is observational, relies on model-derived outcomes, and includes heterogeneous implementations and follow-up periods.

Confidentiality

Partner institutions and binational centers are anonymized throughout this report. Institution labels are used only to distinguish study cohorts and do not identify participating organizations.

EXECUTIVE SUMMARY

Greater participation was positively associated with estimated change

The analysis compared recorded participation with first-to-last modeled ability change while adjusting for starting ability, observation length, and classroom.

+0.055 SD
POOLED ESTIMATE
Per five recorded assignments.

4 of 5
POSITIVE COHORT ESTIMATES
One cohort's 80% interval excluded zero.

80% CI
-0.005 TO +0.115
Pooled primary interval.

95% CI
-0.039 TO +0.149
Wider interval for the same pooled estimate.

What this means

- The evidence is consistent with a positive relationship between more regular recorded participation and greater modeled improvement.
- The design does not establish that additional Jinso assignments caused the observed changes.
- A stronger next study would use prospective curricular integration, common external assessments, and predefined exposure rules.

STUDY TAKEAWAY

The strongest conclusion is not that the effect size is settled. It is that the observed signal is worth testing more rigorously in prospective implementations.

STUDY CONTEXT

Why measure learning while it is happening?

Traditional assessment often reveals a problem only after substantial instructional time has passed. Jinso's working premise is that institutions can make better instructional decisions when practice itself produces usable evidence of learning.

A closed-loop learning model

01
Institutional objectives
Define the skills and outcomes that matter.

02
Practice
Complete curriculum-aligned practice.

03
Measure
Create evidence from learner interactions.

04
Learn and adapt
Use evidence to adjust instruction.

05
New evidence
Test whether the next cycle improves.

The present study evaluates one part of that loop: whether recorded participation in AI-supported practice is associated with estimated change in learner performance.

WHY THIS MATTERS

If a learner is struggling today, evidence from current practice can potentially support action before the next high-stakes assessment.

MEASUREMENT

From conversation to inspectable evidence

The historical data contained rich learner conversations and lesson context, but there was no randomized experiment and no common external assessment administered at both the beginning and end of the observation period.

Conversation + lesson context – **CEFR-tagged performance evidence** – **Task-level observations** – **Estimated ability**

How observations were classified

- Successful independent performance.
- Unsuccessful clearly elicited attempts.
- Insufficient or unreliable evidence was not scored.

Two related analyses

Measurement

Estimate learner performance at different points in time.

Comparison

Estimate how recorded participation relates to change, adjusting for starting ability, observation length, and classroom.

Validation and limits

Checks included computational reproducibility, comparisons between observed and predicted response patterns, and sensitivity analyses across coverage rules, observation windows, and specifications. These checks support interpretation of fitted outputs; they do not establish external assessment accuracy.

STUDY DESIGN

An observational comparison within institutions and classrooms

For each learner, recorded participation was defined as the count of qualifying assignments during the observed window. Estimated change was defined as the difference between the first and last supplied modeled ability estimates in that window.

Main model

RECORDED PARTICIPATION

Count qualifying assignments during each learner's observed window.

ESTIMATED CHANGE

Compare the first and last supplied ability estimates in that window.

ADJUSTED FOR

Starting ability
Observation length
Classroom

Working definition of good practice

- Connect assignments to the skills and learning objectives currently taught in class.
- Make practice regular by placing Jinso lessons inside a clear learning sequence.
- Use the evidence by reviewing participation and recurring learner difficulties.

For the participation count used in the analysis, a recorded assignment required at least two learner messages during the observation window.

What this design cannot establish

The design cannot rule out unmeasured differences between learners, classrooms, or implementations. It estimates association, not causal impact.

STUDY COHORT

Who was included

The source data came from four partner institutions using Jinso as part of instruction. To protect partner confidentiality, names are removed. One institution contributed two analytic level cohorts, producing five cohorts in the pooled analysis.

Anonymized institution	Students with scored activity*	Activity window	Analytic contribution
Partner Institution A	713	Feb-Sep 2026	Cohorts A1 and A2
Partner Institution B	403	Apr-Sep 2026	Cohort B
Partner Institution C	176	From Apr 2026	Cohort D
Partner Institution D	254	Dec 2025-Sep 2026	Cohort C

*Institution-level counts represent distinct students with at least one exchange submitted for scoring, not enrolment totals or analytic sample sizes.

Analytic cohorts

Cohort	Learners	Classrooms
Cohort A1	140	10
Cohort A2	149	9
Cohort B	77	34
Cohort C	233	16
Cohort D	141	13
Pooled	740	82

PRIMARY RESULTS

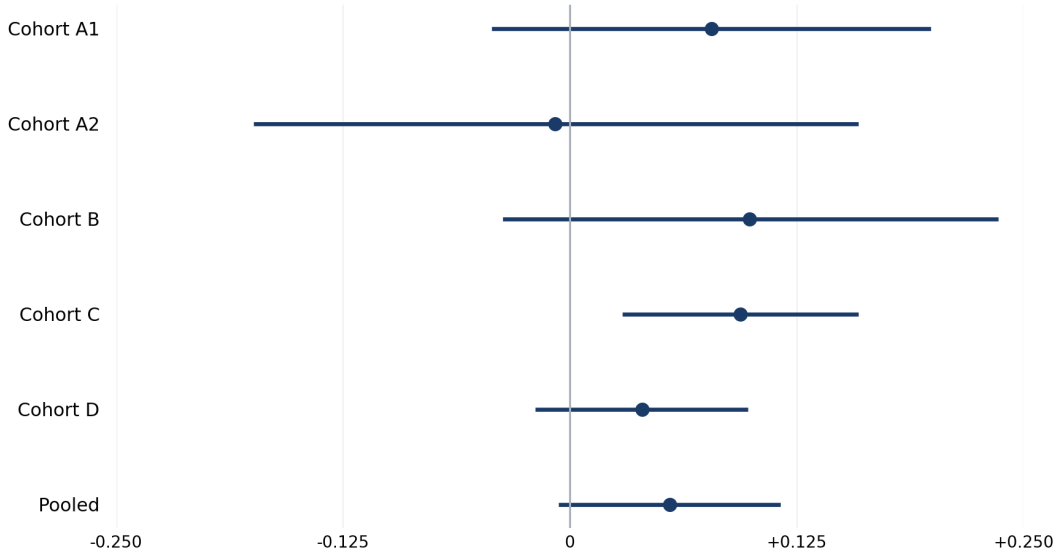
Participation and estimated change across cohorts

Estimated association per five recorded assignments. Bars show 80% confidence intervals.

+0.055 baseline SD

Pooled estimate per five recorded assignments

80% CI: -0.005 to +0.115



Four of five cohort estimates were positive. Cohort C had the clearest positive estimate in the primary cohort comparison, with an 80% interval that did not include zero. The pooled 80% interval narrowly included zero, and the corresponding 95% interval was wider.

COHORT ESTIMATES

Results by anonymized cohort

The table reports model-relative baseline standard deviations of first-to-last estimated change per five recorded assignments.

Cohort	Estimate	80% confidence interval
Cohort A1	+0.078	-0.042 to +0.198
Cohort A2	-0.008	-0.173 to +0.158
Cohort B	+0.099	-0.036 to +0.235
Cohort C	+0.094	+0.030 to +0.158
Cohort D	+0.040	-0.018 to +0.097
Pooled	+0.055	-0.005 to +0.115

HOW TO READ THIS TABLE

Confidence intervals show a range of estimates compatible with the model and available data. An interval including zero means the analysis does not rule out a null association at that interval level.

Important reporting context

- 80% intervals have lower coverage than 95% intervals; they do not add observations or increase certainty.
- Institutional displays used their original classroom and cohort rules. The pooled analysis used a harmonized rule and sometimes different samples.

ILLUSTRATIVE COHORT VIEWS

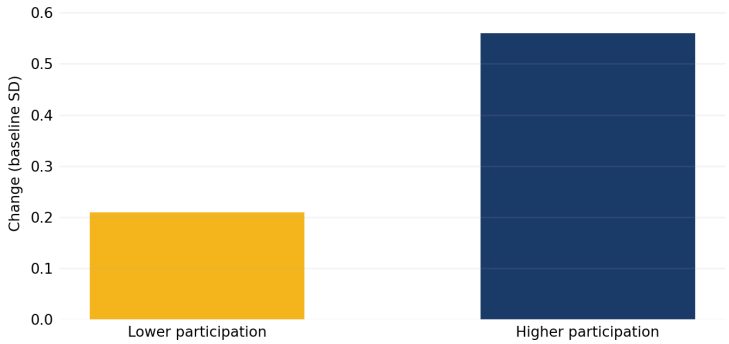
What the cohort-level patterns looked like

Partner Institution A - lower-level cohort

In one lower-level cohort, 89% of learners had a higher final point estimate. Lower participation (2-27 assignments; n=74) had a median increase of +5.99 logits, while higher participation (28-37 assignments; n=67) had a median increase of +6.66 logits. These are descriptive group summaries from the same retrospective model.

Partner Institution D

A raw descriptive comparison showed average estimated change of +0.21 baseline SD among lower-participation learners and +0.56 baseline SD among higher-participation learners. After adjustment, the higher-minus-lower difference narrowed to +0.10 SD (80% CI -0.08 to +0.27).



Partner Institution C

Among 141 learners with repeated estimates, 66% had a higher final ability estimate and median estimated change was +0.30 model logits. A cohort-specific participation analysis estimated +0.24 logits per five recorded assignments (80% CI +0.02 to +0.45). This cohort-specific logit scale should not be directly compared with pooled baseline-SD estimates.

FROM EVIDENCE TO INSTRUCTION

Using learner evidence to guide intervention

The practical value of the measurement system is not only to produce a score. It is to surface recurring patterns that academic teams can investigate and address while instruction is still underway.

Repeated difficulty

A learner repeatedly omits required forms in question construction.

Academic review

Across 916 question-asking attempts, 28% failed; 14 of 145 learners failed three or more times across at least two lessons.

Targeted practice

Teachers review the relevant grammar and connect follow-up practice to current instruction.

Later evidence

Subsequent learner work can be inspected for improved performance.

Review what changed

Measure the same task-failure rate again for the same classrooms.

THE INTENDED CLOSED LOOP

Detect a recurring difficulty. Review it at the academic level. Intervene instructionally. Then use later evidence to determine whether the pattern changes.

INTERPRETATION

How to read the findings

Evidence consistent with a positive relationship

The pooled analysis and most cohort estimates point in a positive direction: learners with more recorded participation tended, on average, to show greater modeled change after adjustment for the variables included in the analysis.

Not a causal estimate

Learners who participate more may differ in motivation, attendance, teacher support, prior exposure, or other factors not fully captured in the models. The study supports an association, not a conclusion that assigning additional practice will itself produce a specific amount of improvement.

Model-derived outcomes

The outcomes come from retrospective model fits. Reported intervals do not cover every source of uncertainty or model misspecification. Task alignment to CEFR was modest in several cohorts and absent in one primary analysis; task labels should not be interpreted as direct learner CEFR boundaries.

Different analysis versions answer different questions

The pooled first-to-last analysis estimated +0.055 baseline SD per five assignments in 740 learners across 82 classrooms. A compatible supplied growth-slope analysis estimated +0.0009 baseline SD per 30 days per five assignments in 516 learners across 92 classrooms. Samples and estimands differ.

CONCLUSION

The next study should be more prospective, standardized, and integrated

The strongest conclusion from this retrospective work is that there is a signal worth studying more carefully. Greater AI-supported participation was positively associated with estimated change in the pooled analysis, but the current evidence remains limited by observational design, heterogeneous implementation, short follow-up in some cohorts, and reliance on model-derived outcomes.

Recommended next steps

- Collect more data from implementations where Jinso practice is explicitly integrated into the curriculum.
- Use common external assessments at defined baseline and follow-up points where feasible.
- Predefine assignment exposure, observation windows, and cohort inclusion rules.
- Track implementation variables such as teacher use, scheduling, and instructional follow-up.
- Continue reporting uncertainty and sensitivity results alongside headline estimates.

EVIDENCE PROGRAM

Make learning visible. Then use the evidence to improve the next cycle of learning.

APPENDIX

Statistical context and reporting notes

Analysis	Estimate	Intervals	Sample
Pooled first-to-last change	+0.055 baseline SD per five assignments	80% CI -0.005 to +0.115 95% CI -0.039 to +0.149	740 learners 82 classrooms
Compatible supplied growth slopes	+0.0009 baseline SD per 30 days per five assignments	80% CI -0.0088 to +0.0106 95% CI -0.0145 to +0.0164	516 learners 92 classrooms

Reporting notes

- 80% intervals have lower coverage than 95% intervals; they do not add observations or increase certainty.
- Institutional displays retained their original classroom rules. The pooled analysis used a harmonized rule and sometimes different samples.
- One short-follow-up partner cohort showed weaker learner-level first-to-last change in a separate institutional display: among 191 learners with repeated dates, 45.5% had positive point changes, median follow-up was 10 days, and median change was -0.111 model logits.
- Source analyses were completed and synthesized in September 2026.
- Institution and binational-center names have been removed from this report. Anonymized labels preserve the study structure while protecting partner confidentiality.