CELES 2026 · Nara

AI-Based Speaking Assessment of a Short-Term Study Abroad Program

Spoken English Proficiency Before and After a Three-Week Overseas Program

Ken Urano
Hokkai-Gakuen University, Sapporo, Japan

study abroad | spoken English proficiency | AI testing

Slides & references

Background & Motivation

Program Context

3-week overseas program (Univ. of Hawai‘i at Mānoa) for business students in Hokkaido

Output-oriented design: presentations, homestay, English-only classrooms

Part of a larger mixed-methods evaluation project

AI-Based Speaking Test

AISATS (AI Skill Assessment & Training System) by Potential Plus K.K.

Powered by Speechace LLC (US); via EdulinX (JP)

Correlation with human raters: $r = 0.80$, within ±0.5 IELTS

Instant, objective, scalable

Research Questions

RQ1: Did the SA group improve more than the CG on AISATS overall?

RQ2: Were gains component-specific? (esp. fluency — closely tied to output-oriented program design)

Prior Research: Objective Speaking Tests

Hirai (2018)

Meta-analysis of 31 Japanese university study-abroad studies

Short-term (1 month or less) gains are real but modest — much smaller than for longer programs

Sets realistic expectations for a three-week program

Saito (2025)

Versant speaking test; pre/post, one-month program in Australia (*N* = 14)

Significant gain in overall spoken proficiency, plus reduced anxiety

Single group — no comparison

Sekiya et al. (2018)

Rated group-oral test; 64 learners, four one-month programs in the U.S.

Large gains: *d* = 0.61–0.98 but a delayed posttest: most faded after 10 months

Single group — no comparison

The catch: these are mostly single-group — without a comparison group, improvement ≠ program effect (maturation, self-study, test familiarity). → So this study adds a comparison group.

Instrument: AISATS (AI Skill Assessment & Training System)

Test Overview

Questions	5 items (10–15 min)
Format	Read-aloud ×2 · Photo ×1 · Free ×2
Scoring	9-point scale (1–9), + 5 component sub-scores
Predicted	CEFR, IELTS, TOEIC, TOEFL, GTEC, EIKEN
Feedback	Instant — AI avatar interviewer
Engine	Speechace LLC (US); *r* = 0.80 with human raters

Sample Tasks

Read-aloud	Read the on-screen text aloud (×2)
Photo	Tell the story shown in the image (×1)
Free	Answer the interviewer's spoken questions (×2)
How it runs	Browser-based; AI avatar interviewer; scored in seconds
Report	5 skill scores + predicted tests

AISATS in Action: A Concrete Example

5-Component Scoring

Pronunciation
Phoneme-level AI analysis

Fluency
Words/min + pause count

Vocabulary
Range & appropriateness of lexis

Grammar
Structural accuracy & complexity

Task Achievement
Relevance, completeness, creativity

Sample Score Report

テスト結果とフィードバック

全体 (7.3)
ある程度のアクセントはありますが、かなり良い発音です。 流暢さと話す内容の一貫性に優れていますが、たまに話が途中で止まってしまうことがあるようです。 洗練された語彙と慣用的な表現を使用することに優れています。 さまざまな文法を使用して複雑な考えなどを表現することに優れています。 重要なポイントをきちんと押さえており、課題もしっかりと完了できています。

発音 (8.2)
幅広く様々な発音を使つことができています。

流暢さ (6.6)
話すことに詰まることなく、話の内容に一貫性を保ったまま長く話すことができています。

Method

Participants

Study Abroad (SA): *n* = 12

(Univ. of Hawai‘i at Mānoa, 3 weeks)

Comparison (CG): *n* = 9

(remained in Japan, same period)

AISATS Pre & Post

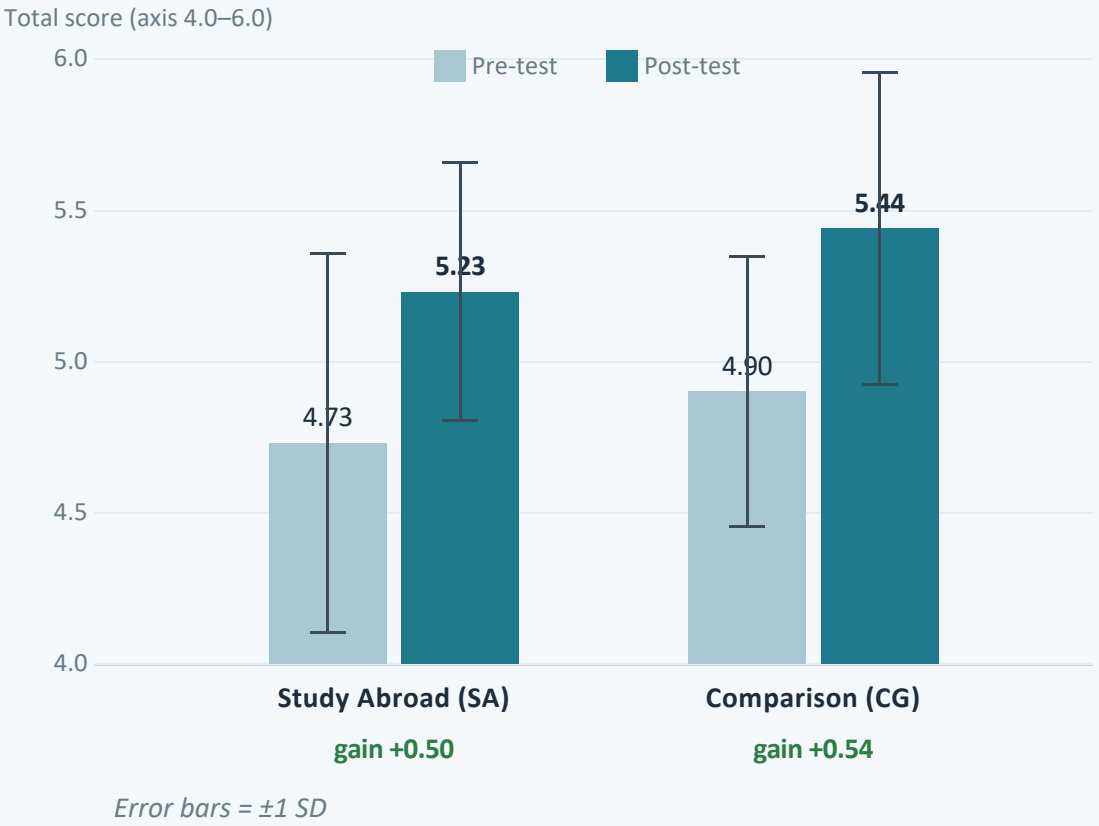
Same test administered at both time points

5 components scored 1–9:

Pronunciation · Fluency · Vocabulary · Grammar · Task Achievement

Study Design

Raw Scores: Total — Both Groups Improve



🔍 Key Findings

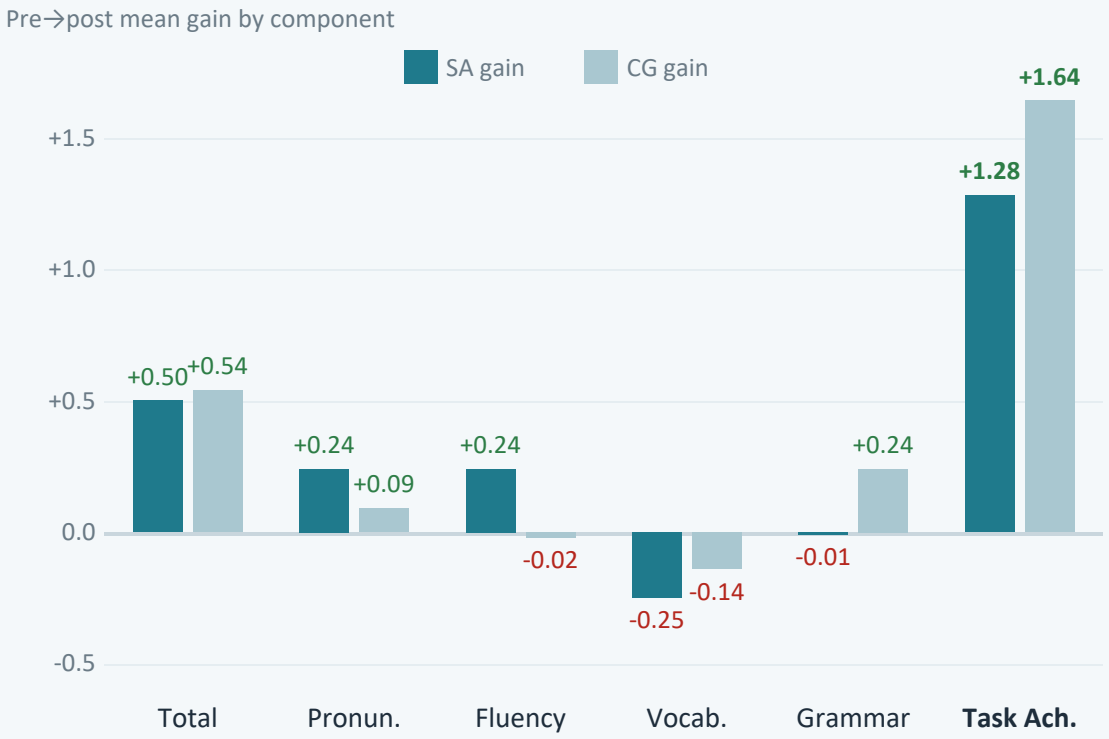
Both groups gained:
SA +0.50, CG +0.54
(both significant within-group)

At the Total level the two groups look the same

A composite test alone would suggest no program effect — but that is misleading

The Total hides the component-level story →

Raw Scores: Components — Task Achievement Dominates



Key Findings

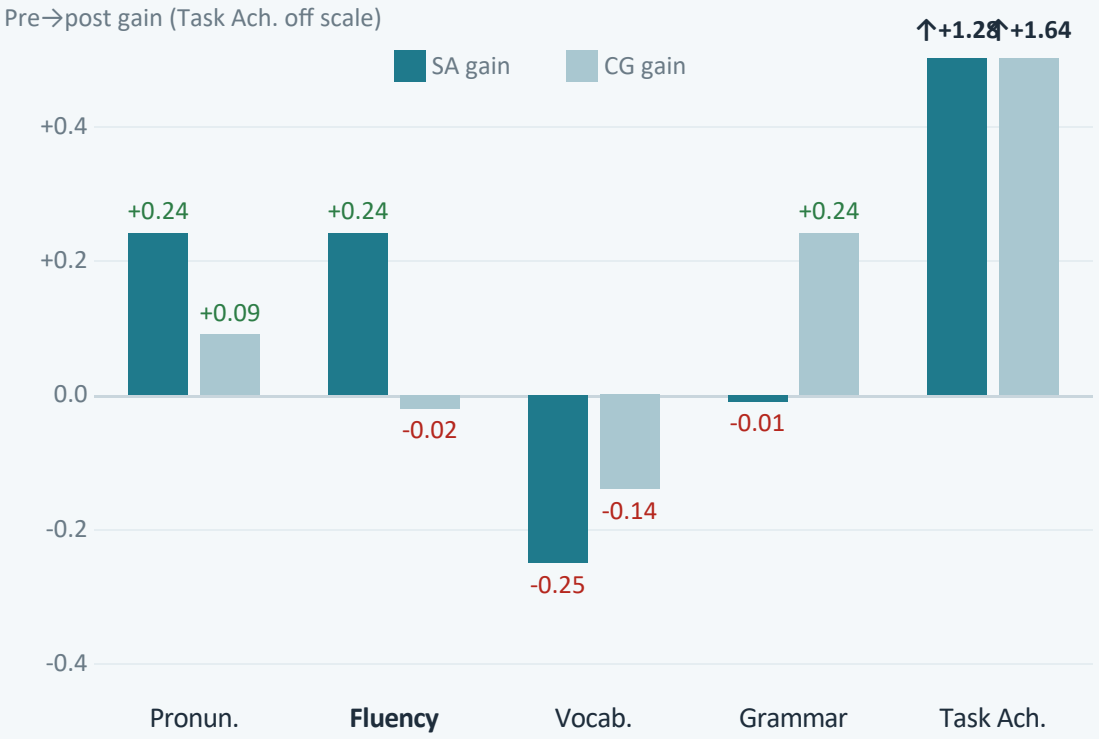
Task Achievement dwarfs everything

It drives the composite — likely test familiarity

On this scale the other five components are barely visible

So we rescale and view them on their own →

Raw Scores: The Five Components Up Close



Key Findings

Task Achievement is off the chart (SA +1.28, CG +1.64)

Remove it, and the other components become visible

Fluency stands out: SA rises, CG slips

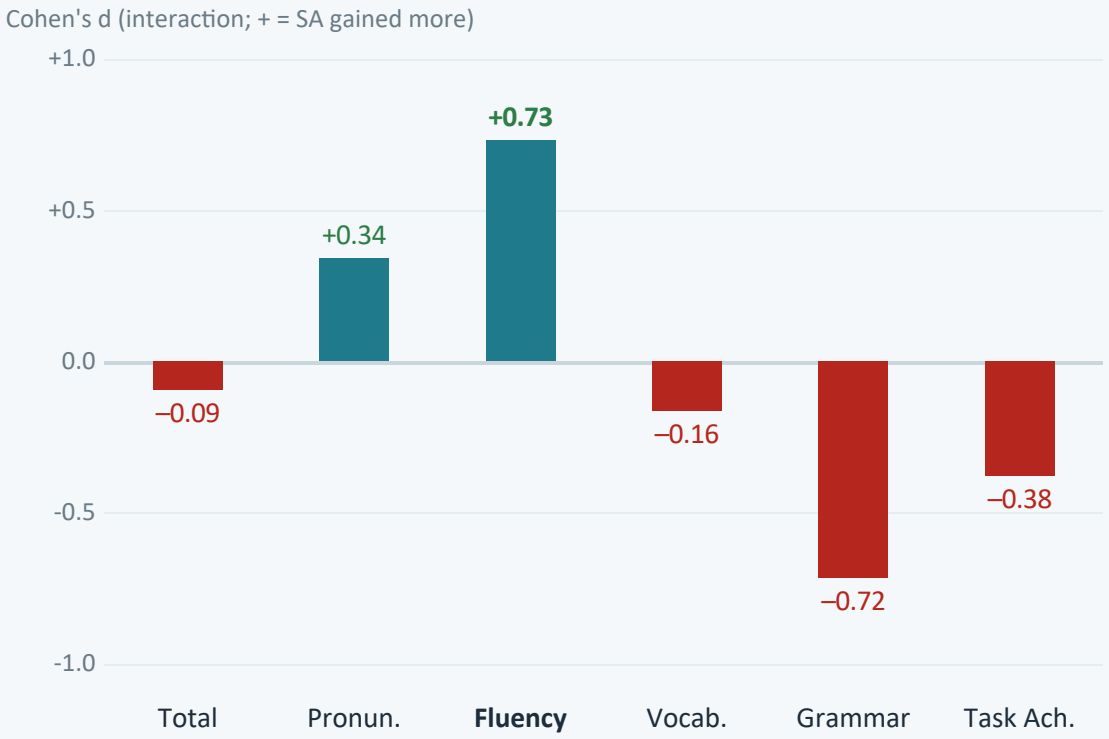
Now we quantify these gaps → the interaction

Results: Focus on the Group × Time Interaction

Component	SA Pre → Post	CG Pre → Post	Interaction (SA – CG)	Cohen's *d*
Total Score	4.73 → 5.23 (0.63) (0.43)	4.90 → 5.44 (0.45) (0.52)	**–0.04**	**–0.09**
Pronunciation	5.29 → 5.53 (0.58) (0.32)	5.50 → 5.59 (0.43) (0.29)	**+0.15**	**+0.34**
Fluency	5.13 → 5.37 (0.48) (0.23)	5.33 → 5.31 (0.35) (0.23)	**+0.26**	**+0.73**
Vocabulary	5.46 → 5.21 (0.66) (0.33)	5.39 → 5.24 (0.49) (0.40)	**–0.11**	**–0.16**
Grammar	5.58 → 5.58 (0.36) (0.33)	5.44 → 5.69 (0.30) (0.31)	**–0.25**	**–0.72**
Task Achievement	2.71 → 3.99 (0.81) (0.89)	2.83 → 4.48 (1.00) (1.09)	**–0.36**	**–0.38**

Interaction = SA gain − CG gain (Group × Time contrast; cf. RELC companion study). Cohen's d: + = SA gained more. Cells show M (SD).
No interaction reached significance (all p > .10). Within-group gains in Total Score and Task Achievement were significant in both groups; within-group change alone is not the program effect. Gains are computed from unrounded means; table cells are rounded, so a cell difference may differ by .01–.02.

Results: Between-Group Effect Sizes (Cohen's d on Gains)



Interpretation

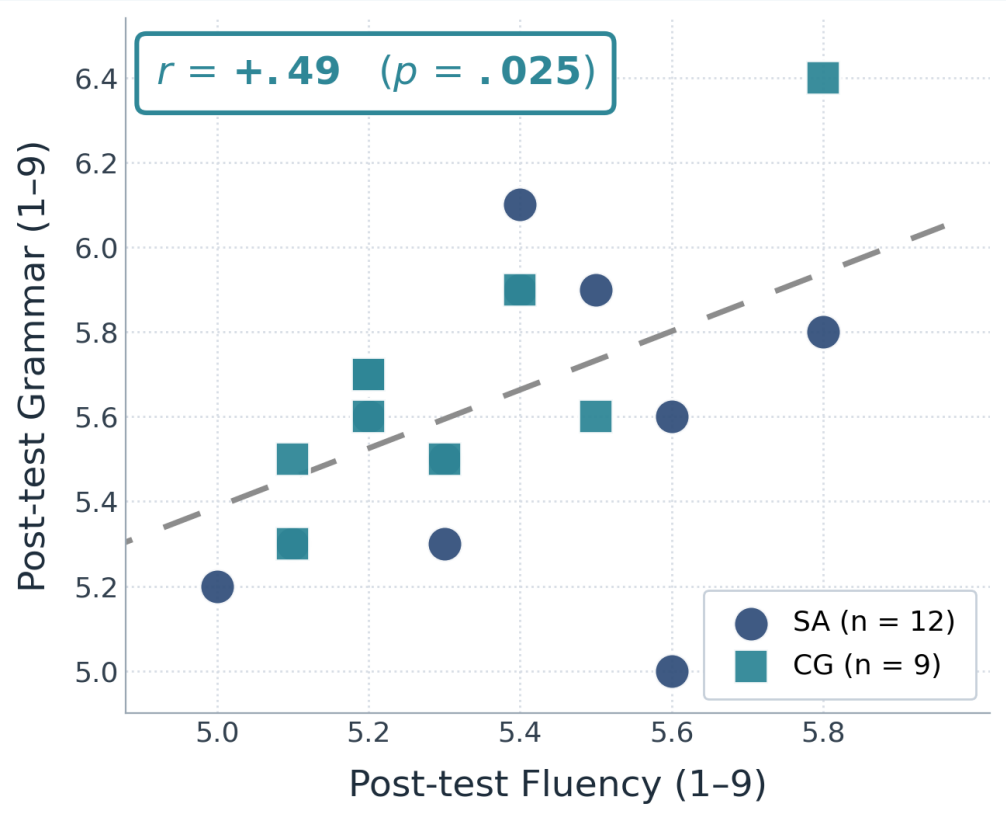
Fluency *d* = +0.73
Medium effect (Plonsky & Oswald 2014)

Pronunciation *d* = +0.34
small, consistent direction

Grammar *d* = –0.72 (favours CG)
CG made a small grammar gain while SA stayed flat

No component significant for interaction (all $p > .10$)

Checking a Fluency–Accuracy Trade-off



The question

Could the opposite signs for Fluency and Grammar reflect a speaking-time trade-off — prioritising fluency at the cost of accuracy?

The test

A trade-off predicts that, at post-test, more fluent speakers should be *less* accurate — i.e. a negative correlation.

The result: the opposite

Post-test Fluency and Grammar are **positively** correlated (r = +.49, p = .025). More fluent speakers tend to be more accurate, not less — no sign of a trade-off.

Discussion

📌 Task Achievement Dominated the Picture

Both groups showed the largest gains in Task Achievement (SA +1.28, CG +1.64) — most likely test-format familiarity, as students grew comfortable with the AI interviewer. AI scoring may also reward longer, less hesitant speech, so simply talking more on the retake could inflate this score. This dynamic may have masked subtler between-group differences.

✅ Fluency: Selective Gain Worth Noting

Setting Task Achievement aside, Fluency stands out: SA gained (+0.24), CG slightly declined (−0.02). AISATS Fluency captures words/min and pause count — arguably the component most sensitive to oral automaticity gained through immersive exposure. Between-group d = +0.73 (medium).

⚠️ No Significant Interaction — Power Issue

No component reached significance (all $p > .10$). Small n and the Task Achievement "noise" both reduce the chance of detecting real between-group differences. This study therefore analyses each component separately and re-reads the composite with Task Achievement removed — which is what surfaces the Fluency signal.

🔗 Triangulation with Self-Assessment

A companion CEFR-J self-assessment study (under review) shows the same domain-specific pattern — largest gains in Spoken Production. The constructs are related but not identical: CEFR-J Production is self-rated output; AISATS Fluency is words/min + pause count. Convergence is at the construct-family level.

Limitations

01 Small Sample Size

SA *n* = 12, CG *n* = 9. All CIs are wide. Directional consistency across components provides some support, but magnitude estimates carry substantial uncertainty. Replication with different cohorts is expected.

02 Non-random Group Assignment

Pre-existing differences in motivation and proficiency cannot be excluded, but the pretest scores were similar across groups on AISATS.

03 AI Scoring Validity

The vendor reports validation against human raters (*r* =.80, ±0.5 IELTS; UCR study, *n* = 100). But that figure is vendor-supplied and at the total-score level — component-level validity for this student population and task format has not been independently verified.

04 Single Instrument & Cohort

Results reflect one cohort at one institution using one AI test. The CEFR-J self-assessment and semi-structured interview data (reported separately) will complement and qualify these findings.

Conclusion

- Both groups improved on Total Score and Task Achievement, suggesting general practice or test-familiarity effects.
- Fluency gain was observed only in the SA group (d = +0.73, medium), suggesting selective benefit of immersive oral experience.
- No Group × Time interaction reached significance — small sample size is the primary limitation.
- AISATS Fluency component appears sensitive to the output-oriented program design; findings triangulate with self-assessment data.
- AI-based speaking tests offer scalable, objective pre–post measurement for program evaluation — promising tool for EFL contexts.

Slides & references

References

Hirai, A. (2018). The effects of study abroad duration and predeparture proficiency on the L2 proficiency of Japanese university students: A meta-analysis approach. *JLTA Journal, 21*, 102–123. https://doi.org/10.20622/jltajournal.21.0_102

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In E. D. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks* (pp. 135–163). Cambridge University Press.

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Potential Plus K.K. (2025). *AISATS: AI Skill Assessment & Training System* [Automated speaking test, built on the Speechace scoring engine; distributed in Japan by EdulinX]. https://potentialplus.co/ja/aisats/

Saito, Y. (2025). Effects of a short-term study abroad program on students’ English proficiency and foreign language anxiety. *Eigo Eibei Bungaku* [English Language and Literature], *65*, 133–153. https://chuo-u.repo.nii.ac.jp/records/2002166

Sekiya, Y., Park, S., & Tsuji, R. (2018). Effects of short-term study abroad programs. *Studies in Linguistics and Language Teaching, 29*, 161–180. https://doi.org/10.69236/0000001600

Urano, K. (under review). *Domain-specific effects of short-term study abroad on self-assessed spoken English: A CEFR-J scale-transformation analysis* [Manuscript submitted for publication].