順天堂グローバル教養論集　第十一巻（2026）3–18 頁

Original Articles

English Proficiency Gains in Four Weeks Abroad: Evidence from Cebu

Siwon PARK[1)]*, Tetsuya FUKUDA[2)], Hiroaki UMEHARA[3)]

【Abstract】

This study examined short-term changes in English proficiency during a four-week Cebu program and whether perceived communicative gains and baseline traits aligned with those changes. We analyzed TOEFL ITP pre–post scores, CEFR-aligned self-assessments of listening and speaking, and a fully linked subset for trait models. In the TOEFL paired sample, paired *t*-tests showed reliable gains in Listening, Reading, and the Total score, while Structure and Written Expression did not change. Self-assessed Listening and Speaking both rose, indicating perceived growth in communicative ability. Gains in self-assessment were not reliably correlated with TOEFL gain scores. In gain-score models, pre-departure baseline Collaboration negatively predicted Reading gain, while SRL and Collaboration did not predict gains in Listening or Structure and Written Expression in the two-predictor models; effects were interpreted cautiously given the sample size. A delayed posttest and qualitative interviews are planned to examine durability and mechanisms.

Key words: study abroad, TOEFL ITP, CEFR self-assessment, collaboration, self-regulated learning

研究論文

海外の 4 週間での英語力の伸び：セブ島研修の実証研究

朴　シウォン[1)]*, 福田　哲哉[2)], 梅原　洋陽[3)]

【要　旨】

本研究は、セブ島での 4 週間の英語研修プログラムにおける英語力の変化、及びコミュニケーション能力と学習者の特性がその変化と関係するかどうかを検証した。分析には TOEFL ITP の事前・事後スコア、リスニングとスピーキングの自己評価、さらに特性モデル分析のデータを使用した。対応のある *t* 検定の結果、得点向上が確認されたが、リスニング、読解、総合スコアは上昇した一方、文法・語法には変化が見られなかった。自己評価ではリスニングおよびスピーキングの両方が上昇し、主観的にはコミュニケーション能力の向上が示唆されたが、自己評価の伸びは TOEFL ITP の伸びと有意に相関しなかった。伸び（事後－事前）を目的変数としたモデルでは、協働性は読解の伸びを負に予測し、自己調整学習と協働性はいずれもリスニングおよび文法の伸びを予測しなかった（サンプル数を踏まえ慎重に解釈すべきである）。今後は遅延事後テストおよび質的インタビューを実施し、学習効果の持続性と変化メカニズムを検証する予定である。

キーワード：留学、TOEFL ITP、CEFR 自己評価、協働性、自己調整学習

[1)] Faculty of International Liberal Arts, Juntendo University（Email: s.park.ll@juntendo.ac.jp）
[2)] Faculty of International Liberal Arts, Juntendo University（E-mail: t.fukuda.wv@juntendo.ac.jp）
[3)] Faculty of International Liberal Arts, Juntendo University（E-mail: h.umehara.yz@juntendo.ac.jp）
* Corresponding author: Siwon PARK
〔Received on November 7, 2025〕〔Accepted on February 20, 2026〕

1. Introduction

Short-term study abroad has expanded in Japanese higher education as a practical pathway to internationalization within tight academic calendars. A central empirical question is not whether four weeks abroad can matter, but which dimensions of English proficiency are sensitive to that interval, for whom, and under what learning conditions. Prior work suggests that well-structured short stays can foster measurable development, particularly in receptive skills and aspects of oral fluency, while the magnitude and locus of gains vary with program design and learner engagement. These patterns are visible across syntheses and multi-site studies as well as language-specific investigations.

The present study examines a four-week, faculty-guided program in Cebu, the Philippines, estimating short-interval changes using two complementary sources: conservative standardized indices from the TOEFL ITP (section scores and Total) and contextualized CEFR-aligned self-assessments of listening and speaking. We draw on interaction- and mediation-focused perspectives to frame what might change over four weeks in intensive settings, while recognizing that some subskills typically require longer or more targeted instruction. We treat Collaboration as a pre-departure baseline perception and self-regulated learning (SRL) as an individual characteristic that may condition short-interval gains. The study's primary purpose is to estimate how much English proficiency develops over four weeks and to situate observed pre–post differences against expectations from prior research at this timescale. A delayed posttest and qualitative interviews are planned to examine durability and mechanisms. For terminological consistency, we use the official TOEFL ITP section name, *Structure and Written Expression* and reference the CEFR Companion Volume for self-assessment alignment.

2. Literature review

2.1 Design over exposure

Across the recent study-abroad (SA) literature, program architecture matters more than exposure alone. Multi-site work shows that guided mediation, structured tasks, and accountable products predict larger gains than unsupervised immersion of comparable length (Park, 2025; Vande Berg et al., 2009). New syntheses reach a similar conclusion for short programs: effects are small to moderate when interaction is dense and reflective activities are embedded by design (Kinginger, 2009; Llanes & Muñoz, 2009; Park, 2025; Park & Sugita, 2024; Sekiya & Park, 2006; Sekiya et al., 2018). More recent overviews of SA assessment argue that credible short-interval claims require alignment among tasks, instruments, and hypothesized mechanisms of change, not just time in country (Bradly & Iskhakova, 2023; Goldstein, 2022). These points motivate our focus on a four-week program that engineers one-to-one lessons and small-group work.

2.2 Duration and intensity

Short stays can work when intensity is high. A comprehensive 2023 systematic review of intensive and short-term mobility reports positive but heterogeneous outcomes, with duration interacting with contact quality and structure (Bradly & Iskhakova, 2023). Classic comparative work already established that oral fluency benefits when learners accrue meaningful interactional contact, a point directly relevant to compressed timelines that maximize one-to-one and small-group interaction (Segalowitz & Freed, 2004; Sekiya & Park, 2006). In L2 development specifically, recent cohort studies during 3–6 week terms show selective growth, often strongest where practice is repeated and task-linked (Trentman, 2021). Our program's design follows these intensity principles.

2.3 Skill trajectories and assessment

Short-interval growth is asymmetric. Theory and meta-analytic evidence indicate that repeated comprehensible input and interaction tend to move listening, lexical access, and some oral measures more quickly, while discrete grammatical accuracy often requires explicit focus on form to shift detectably (Long, 1996; Spada & Tomita, 2010). Recent work confirms that even during short SA, grammatical and lexical development can occur, but effects concentrate where tasks push noticing and production (Serrano et al., 2016). In Japanese university settings, TOEFL ITP remains common for monitoring receptive skills because it offers conservative section indices and a stable Total score. Official guidance cautions against over-interpreting very small differences over short windows and recommends reading section movement alongside other evidence (Educational Testing Service, 2025). These strands justify our choice of TOEFL ITP sections and Total for objective indices, paired with a communicative lens.

2.4 Collaboration as mechanism

We treat collaboration as structured cooperative work rather than unstructured grouping. Foundational cooperative learning research identifies five design elements that underpin productive collaboration: positive interdependence, individual accountability, promotive interaction, appropriate social skills, and group processing (Johnson et al., 1999; Johnson & Johnson, 2009).

In L2 contexts, collaborative work increases opportunities for negotiation of meaning and feedback, which facilitate comprehension, noticing, and uptake (Mackey, 1999; Pica, 1994). When mutuality and balanced control are present, learners engage in language-related episodes that link form and function, yielding richer learning opportunities than dominant-passive patterns (Storch, 2002, 2013). Recent quantitative syntheses reinforce these mechanisms. Meta-analyses of collaborative writing and peer-interaction tasks report positive effects on accuracy and complexity, particularly when tasks include planning, role rotation, and explicit criteria (Elabdali, 2021; Fan, 2024; Zou et al., 2016). Our design incorporates these elements, so collaboration is both a program lever and a baseline perception to examine as a predictor.

2.5 Self-regulated learning as condition

Self-regulated learning (SRL) refers to proactive regulation of cognition, motivation or affect, behavior, and context across planning, monitoring, and reflective phases (Pintrich, 2004). Recent L2-focused reviews emphasize metacognitive monitoring, strategy control, and effort regulation as processes that help learners convert opportunities into measurable outcomes, although effects can be modest over short spans when pretest proficiency dominates variance (Hiver et al., 2021; Teng & Zhang, 2022). Broader education meta-analyses also confirm small-to-moderate links between SRL and achievement, with stronger effects when instruction explicitly supports strategy use (Broadbent & Poon, 2015; Richardson et al., 2012). In a four-week window, SRL is plausibly stable at the trait level, yet variation in deployment may still condition individual differences in gains. This motivates testing SRL as a person-level predictor after accounting for pretest.

2.6 Self-assessment and alignment

For communicative outcomes, CEFR-aligned self-assessment can complement standardized testing when items are localized and midpoints are avoided. The CEFR Companion Volume updates descriptors for interaction and mediation and remains the key reference for can-do interpretation (Council of Europe, 2020). A recent meta-analysis finds moderate convergence between self-ratings and

external measures overall, with variation by skill and level, which is consistent with using self-assessment as a complementary lens in short-stay evaluation (Li & Zhang, 2020). Short-interval studies also note calibration shifts. After intensive experiences, learners may apply stricter internal standards to the same descriptors, producing flat or lower self-ratings even when performance improves. This phenomenon is discussed in the broader measurement literature as response shift and has been observed in post-program self-reports where expectations change with exposure (Georgeson et al., 2021; Gogol et al., 2014; Kidd, 2004). These considerations inform our alignment analyses and our interpretation of self-assessment trajectories over four weeks.

Building on these themes, we estimate short-interval change with TOEFL ITP sections and Total, and we use a contextualized CEFR-aligned self-assessment of listening and speaking to capture perceived communicative ability. We treat Collaboration as a pre-departure baseline perception and SRL as a trait-like characteristic that may condition gains. The design choices follow current guidance on short-term program evaluation and assessment alignment, and they set up the analysis plan for Research Questions (RQs) 1–4.

2.7 Research questions

Building on evidence that short-term study abroad yields skill-specific gains when interaction is dense and instruction is structured, the present study asks how much English proficiency changes over a four-week program and whether learner factors help explain individual differences. Proficiency is indexed by TOEFL ITP section and Total scores and perceived communicative ability by a contextualized CEFR-aligned self-assessment of listening and speaking. Collaboration is treated as a baseline perception measured at pre-departure and SRL as a relatively stable individual difference over a four-week window.

- RQ1. To what extent do students improve English proficiency over four weeks, as indexed by TOEFL ITP section scores (Listening, Structure and Written Expression, Reading) and Total?
- RQ2. Do students show pre–post change in perceived communicative ability, as captured by CEFR-aligned self-assessments of Listening and Speaking?
- RQ3. To what degree do within-person gains in CEFR-aligned Listening and Speaking align with gains in TOEFL ITP (with particular attention to Listening)?
- RQ4. Do baseline Collaboration and baseline SRL account for short-interval gains in (a) TOEFL ITP outcomes and (b) CEFR-aligned self-assessments, after adjusting for pretest proficiency?

3. Method

3.1 Context, participants, and timing

The Summer 2025 cohort completed a four-week intensive English program at two partner institutes located on Cebu Island and Mactan Island, the Philippines, from 24 August to 21 September 2025. Students were enrolled in the Faculty of International Liberal Arts at a private university in the Tokyo metropolitan area, Japan. The cohort comprised 57 undergraduates (49 first-year; 8 second-year). Analyses for this study use listwise complete cases from the first-year group only: 24 students provided complete pre–post TOEFL ITP scores, CEFR-aligned self-assessments, and baseline questionnaires, and therefore constitute the analytic sample (female = 16; male = 8). Pretest profiles indicated that most students were in the A2–B1 range on the CEFR based on the project's self-assessment instrument.

3.2 Program

Students studied at one of two long-standing

partner institutes in Cebu. Both programs prioritized TOEFL ITP preparation. Daily lessons targeted Listening Comprehension, Structure and Written Expression, and Reading Comprehension, with regular practice tests and explicit instruction in test-taking strategies aligned to the ITP format. Alongside test-oriented work, each program scheduled meaning-focused one-to-one and small-group classes and guided task-based activities designed to increase interaction, negotiation of meaning, and feedback opportunities.

Learning extended beyond formal class time. Typical days comprised 8 to 10 lessons followed by optional evening self-study sessions. Students lived in shared dormitories, ate together in on-site cafeterias, and had access to facilities such as gyms, cafés, and study rooms, which created additional contexts for informal English use. Weekend schedules encouraged local excursions and recreational activities, including visits to beaches, shopping areas, and community spaces. The multicultural environment, with international peers and instructors, afforded authentic social interaction and supported the development of both linguistic proficiency and intercultural awareness.

3.3 Measures and procedure

TOEFL ITP Level 1. Listening Comprehension, Structure and Written Expression, and Reading Comprehension were administered under proctored conditions on campus. The pretest took place on 7 July 2025 and the posttest on 24–25 September 2025. Section scores are reported on Educational Testing Service (ETS) scaled metrics, and the Total score ranges from 310 to 677. We follow ETS section labels in all tables and figures. Regarding internal consistency, because TOEFL ITP scoring is conducted at the local ETS office, raw responses are not available for reliability estimation of the three sections. Nonetheless, ETS has published presumed reliability coefficients based on large-scale calibration samples, indicating consistently high internal consistency across all sections.

CEFR-aligned self-assessment (Listening and Speaking). A 20-item instrument developed for this project, with 10 Listening and 10 Speaking items, used a six-point scale anchored to contextualized CEFR can-do descriptors for Japanese undergraduates in short-stay contexts. The self-assessment was delivered via Google Forms one week before departure and within one week after return. Item content and anchor phrasing were aligned to the CEFR Companion Volume. Internal consistency in the present sample was excellent. For the pre-administration, Listening showed $\alpha = .97$, and Speaking showed $\alpha = .96$. For the post administration, Listening showed $\alpha = .93$, and Speaking showed $\alpha = .94$.

Collaboration (baseline). A nine-item composite administered one week before departure measured perceived interdependence, balanced participation, and joint problem solving in lessons and field tasks. Collaboration is treated as a pre-departure baseline perception in this wave. Internal consistency was good ($\alpha = .84$).

Self-regulated learning (SRL). A 19-item baseline instrument assessed five dimensions on the same six-point agreement scale: emotion and motivation control, goal setting and persistence, attention and concentration, task management and time use, and environment or resource management with reflection. SRL is treated as relatively stable over four weeks. Items were adapted from commonly used tertiary SRL questionnaires ($\alpha = .94$).

Background and exposure (post). A brief post-program questionnaire recorded prior overseas experience and English use during the program in and out of class.

3.4 Ethics, consent, and data governance

Before accessing any items, students opened an online information sheet embedded at the beginning of the Google Forms. The sheet explained the study purpose, voluntary participation, the right to decline or withdraw without penalty, confidentiality, and anonymity. Students then recorded consent by selecting an explicit consent checkbox. Only records from students who consented were retained for analysis, and only those who consented had their TOEFL ITP scores linked to survey responses using anonymous identifiers. Participation had no consequences for grades or eligibility for mobility programs. For data governance, analysis files were de-identified; the linkage key was stored separately on a restricted-access drive available only to the research team. This e-consent procedure follows widely accepted guidance allowing electronic informed consent when the information sheet and documentation of consent are securely delivered and retained.

3.5 Analytic strategy and data handling

Records from testing and survey platforms were merged using anonymous identifiers. Obvious data entry errors were screened and corrected when verifiable; duplicate records were resolved conservatively. Analyses relied on listwise deletion within the model, yielding $n = 25$ complete cases.

Pre–post change was evaluated with paired-sample *t*-tests, reporting mean difference, Cohen's d_z, and 95 percent confidence intervals. To examine baseline predictors net of initial level, we estimated linear models that predicted each posttest outcome from its corresponding pretest score, baseline SRL, and baseline Collaboration, using HC3 robust standard errors. To probe individual differences in change, we calculated residualized gains by adjusting posttest scores for pretest, and correlated these adjusted scores with baseline traits and changes in self-assessment. All tests were two-tailed with $\alpha = .05$. We emphasize effect sizes and confidence intervals over null-hypothesis tests alone.

4. Results

Analyses used listwise-matched pairs for each outcome ($n = 36$). Two-tailed paired-sample t-tests were evaluated at $\alpha = .05$. For each section of TOEFL ITP Level 1 (Listening Comprehension, Structure and Written Expression, and Reading Comprehension) along with the Total score, Table 1 reports the pretest and posttest means and standard deviations, the mean gain (Post − Pre) with its standard deviation, *t*, degrees of freedom, *p*, within-person effect size d_z, and a 95% confidence interval (CI) for the mean gain.

4.1 TOEFL ITP pre–post change

Table 1 summarizes Pre (early July) and Post (late September) means and standard deviations for each TOEFL section and Total ($n = 36$).

Table 1 Descriptive Statistics and t-test Results for TOEFL ITP Scores ($n = 36$)

Outcome	Pre *M* (*SD*)	Post *M* (*SD*)	Gain *M* (*SD*)	*t*	*df*	*p*	*dz*	95% CI (low)	95% CI (high)
Listening	44.00 (4.90)	45.75 (3.65)	1.75 (3.94)	2.65	35	.012	0.44	0.41	3.09
Structure & Written Expression	40.00 (4.78)	39.19 (6.20)	−0.81 (5.36)	−0.63	35	.535	−0.11	−3.36	1.75
Reading	43.03 (5.52)	45.14 (5.29)	2.11 (4.59)	2.75	35	.009	0.46	0.55	3.67
Total	423.94 (35.15)	434.17 (41.48)	10.22 (27.83)	2.21	35	.034	0.37	0.84	19.61

Across the four-week interval, students improved significantly in Reading and in the Total score. Reading increased by 2.11 scaled points, 95% CI [0.55, 3.67], with a small-to-moderate within-person effect, $t(35) = 2.75$, $p = .009$, $d_z = 0.46$. The Total score rose by 10.22 points, 95% CI [0.84, 19.61], representing a small-to-moderate effect, $t(35) = 2.21$, $p = .034$, $d_z = 0.37$. Listening also showed a statistically reliable gain of 1.75 points, 95% CI [0.41, 3.09], $t(35) = 2.75$, $p = .012$, $d_z = 0.44$. In contrast, Structure and Written Expression was unchanged over the interval; the mean difference was −0.81, 95% CI [−3.36, 1.75], $t(35) = -0.63$, $p = .535$, $d_z = -0.11$.

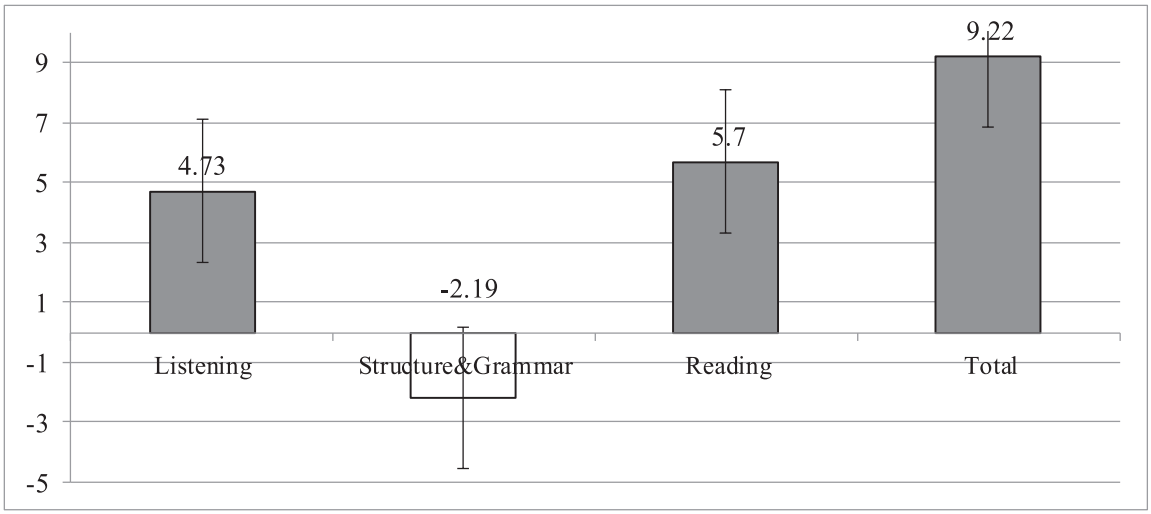
Figure 1 visualizes these differences on a comparable scale. Because the composite Total is reported on a different metric (average of the three section scores × 100), we normalized all outcomes to the percentage of the section scale range (31–68; range = 37). The normalized results indicate percentage gains of +4.73% in Listening, −2.19% in Structure and Written Grammar, +5.70% in Reading, and +9.22% in the composite Total score derived from the average section change. The normalized display clarifies that movement is most visible for Reading and Listening, while Structure and Written Grammar remains flat to slightly negative.

Taken together, the section-level profile shows detectable movement on conservative receptive indices most sensitive to frequent, meaning-focused exposure (Reading and Listening), with corresponding growth in the composite Total score. The discrete-point grammar section remained comparatively stable during this short window, a pattern that will be revisited in the Discussion when considering instructional mechanisms and assessment sensitivity.

4.2 CEFR-aligned self-assessments of listening and speaking

Students completed a six-point CEFR-aligned self-assessment one week before departure and within one week of return. As shown in Table 2, self-assessed Listening increased from 4.13 ($SD = 1.01$) to 4.72 ($SD = 0.71$), a mean gain of 0.59 points, $t(25) = 4.15$, $p < .01$, $d_z = 0.81$, 95% CI [0.30, 0.89]. Self-assessed Speaking rose from 3.91 ($SD = 1.18$) to 4.64 ($SD = 0.87$), a mean gain of 0.73 points, $t(25) = 4.56$, $p < .01$, $d_z = 0.89$, 95% CI [0.40, 1.07]. Posttest means clustered above 4.6/6, indicating meaningful increases in perceived functional ability over four weeks.

These results establish that students perceived



Note. Total reflects the composite average expressed on the same percentage scale (error bars show standard errors).

Fig. 1. Normalized TOEFL ITP Section and Total Gains (%) (n = 36)

Table 2 Self-assessed Listening and Speaking: Pre–post Comparisons (n = 26)

Outcome	Pre M (SD)	Post M (SD)	Gain M	95% CI [low, high]	t(df)	p	dz
Listening (SA, 6-pt)	4.13 (1.01)	4.72 (0.71)	0.59	[0.30, 0.89]	4.15 (25)	<.01	0.81
Speaking (SA, 6-pt)	3.91 (1.18)	4.64 (0.87)	0.73	[0.40, 1.07]	4.56 (25)	<.01	0.89

Note. SA = self-assessment. Gains are Post – Pre.

sizable growth in listening and speaking facility over the program interval, which will be contrasted with receptive test movement in the Discussion.

4.3 Alignment between self-assessed gains and TOEFL ITP gains

To examine whether perceived growth tracks change on conservative receptive measures, we correlated self-assessment gains with TOEFL ITP gains using complete cases that linked all variables (n = 24). Table 3 shows that self-assessed Listening gain did not correlate with TOEFL Listening gain, r = .04, p = .849, nor with gains in Grammar, Reading, or Total ($|r| \leq .17$, all $p \geq .433$). Self-assessed Speaking gain likewise showed no reliable association with gains on any TOEFL outcome ($|r| \leq .10$, all $p \geq .648$).

The absence of strong cross-measure alignment in this wave will be revisited in the Discussion in light of instrument focus (receptive vs. communicative), timing, and potential response-shift effects.

4.4 Baseline traits and short-interval outcomes

We next tested whether baseline self-regulated learning (SRL) and baseline Collaboration (both collected pre-departure) were associated with short-interval gains on TOEFL ITP. Using the same complete-case sample (n = 24), we modeled each gain score (Post–Pre) as a function of SRL and Collaboration with ordinary least squares and heteroskedasticity-consistent (HC3) standard errors.

As summarized in Table 4, Collaboration was a significant negative predictor of Reading gain in the two-predictor model (b = −3.70, SE = 1.50, t = −2.47, p = .013), with model R^2 = .375. Neither SRL nor Collaboration predicted gains in Listening or Grammar in the multiple model (both p > .46), and the Total-score model did not yield significant predictors at α = .05. Given the moderate correlation between SRL and Collaboration (r = .64), we also estimated simple regressions with a single predictor. In those models, higher baseline Collaboration predicted smaller gains in Reading (b = −4.83, p = .01) and Total (b = −18.04, p = .031), while higher baseline SRL was negatively related to Reading gain (b = −2.99, p = .006). These single-predictor patterns attenuated once both traits were entered together, consistent

Table 3 Correlations between Self-assessed Gains and TOEFL ITP Gains (n = 24)

Self-assessment gain	TOEFL gain	r	p
SA Listening gain	Δ TOEFL Listening	0.04	.849
SA Listening gain	Δ TOEFL Grammar	0.02	.916
SA Listening gain	Δ TOEFL Reading	−0.17	.433
SA Listening gain	Δ TOEFL Total	−0.06	.778
SA Speaking gain	Δ TOEFL Listening	0.04	.857
SA Speaking gain	Δ TOEFL Grammar	0.10	.648
SA Speaking gain	Δ TOEFL Reading	−0.10	.661
SA Speaking gain	Δ TOEFL Total	0.03	.873

Note. SA = self-assessment. Gains are Post – Pre.

Table 4 Baseline SRL and Collaboration Predicting TOEFL ITP Gains (OLS with HC3 SEs; $n = 24$)

Outcome (Δ)	Predictor	*b*	SE (HC3)	*t*	*p*	Model R^2
	Intercept	9.78	5.39	1.81	.070	.065
TOEFL Listening	SRL	−0.67	1.18	−0.57	.568	
	Collaboration	−1.28	1.82	−0.70	.482	
	Intercept	−7.16	11.74	−0.61	.542	.003
TOEFL Grammar	SRL	0.52	2.09	0.25	.803	
	Collaboration	0.82	3.65	0.22	.822	
	Intercept	21.37	5.93	3.60	<.001	.375
TOEFL Reading	SRL	−0.23	0.83	−0.28	.779	
	Collaboration	−3.70	1.50	−2.47	.013	
	Intercept	80.96	37.96	2.13	.033	.176
TOEFL Total	SRL	−4.11	7.16	−0.57	.572	
	Collaboration	−15.51	10.34	−1.50	.146	

Note. Δ = Post – Pre. Coefficients are unstandardized. In simple (one-predictor) models, Collaboration negatively predicted Reading and Total gains; SRL negatively predicted Reading gain. In the multiple model, only Collaboration remained a significant predictor of Reading gain, likely due to shared variance between SRL and Collaboration ($r = .64$).

with shared variance between SRL and Collaboration.

5. Discussion

This study asked how much English proficiency changes over a four-week, faculty-guided Cebu program and whether perceived communicative development and baseline traits align with that change. With paired TOEFL data ($n = 36$), students showed a moderate gain in Reading and a smaller gain in Total across the program window, while the Listening section and the Structure and Written Expression section were statistically stable. In the questionnaire sample ($n = 26$), self-assessed Listening and Speaking both increased significantly, indicating perceived communicative growth over the same interval. In the complete-case trait sample ($n = 24$), baseline SRL did not predict short-interval gains once pretest level was controlled, and higher baseline Collaboration was associated with smaller adjusted gains in Reading and Total. Below, we interpret these patterns in relation to the four research questions and the literature reviewed above.

5.1 Objective proficiency gains (RQ1)

The section-level profile matches prior evidence that short windows yield selective, not uniform, growth when programs intensify contact and structure tasks. Multi-site and review work shows that well-designed short stays produce small-to-moderate effects, especially when guided reflection, mediation, and accountability are built into activities rather than relying on immersion alone (Kinginger, 2009; Llanes & Muñoz, 2009; Park, 2025; Vande Berg et al., 2009). The Georgetown Consortium Project remains the canonical example and documents stronger outcomes when programs add structured interventions to exposure, a design principle directly relevant to four-week formats.

A Reading lift of moderate magnitude alongside a modest Total increase is also plausible on a conservative, reliable instrument, given a curriculum that includes repeated timed practice and test-relevant strategies for the ITP sections. ETS guidance cautions that small, short-interval changes should be interpreted in light of score precision and the standard error of measurement, and recommends using multiple indicators (e.g., section trends plus other evidence)

when making inferences about learning outcomes (Educational Testing Service, 2022). This aligns with our interpretation of section-specific movement rather than across-the-board change.

By contrast, discrete grammar often requires either longer exposure or explicit focus on form to shift detectably in four weeks, and scripted listening formats tend to be less sensitive to short-interval change unless the training directly targets their specific demands. Meta-analytic and theoretical work on form-focused instruction and output-driven learning explains this asymmetry: explicit attention to form benefits accuracy, and pushed production helps consolidate form-function links that do not arise automatically from input (Long, 1996; Spada & Tomita, 2010; Swain, 2005).

5.2 Perceived communicative ability (RQ2)

Self-assessed Listening and Speaking both increased significantly over the four-week program, with mean gains of 0.59 and 0.73 points respectively on a six-point CEFR-aligned scale. These results suggest that students perceived meaningful improvements in their functional communicative abilities during the program. This upward movement is consistent with the structured immersion and daily opportunities for interaction built into the curriculum, particularly in one-to-one and small-group settings.

At the same time, it remains important to interpret these gains cautiously. The CEFR Companion Volume notes that self-assessment is inherently subjective and sensitive to internal calibration. In some short-term programs, perceived gains may under- or overestimate actual development due to evolving self-awareness or comparison standards. This is often referred to as response-shift calibration. Although the current data indicate increases rather than declines, the possibility remains that learners' internal benchmarks also shifted alongside ability.

Finally, the test-preparatory emphasis of the program, which focused on receptive skills and strategy, may have amplified students' awareness of their communicative limits in more spontaneous or unstructured contexts. This supports the rationale for using both standardized test indices and contextualized self-assessments together. Taken as a whole, the observed increases in self-ratings strengthen the case that students perceived real gains in communicative competence over the four-week interval.

5.3 Alignment of perceived and objective change (RQ3)

Objective and perceived indicators did not move in parallel in this wave. That pattern fits prior reports that self-ratings align only moderately with external measures and can de-align under response-shift conditions or when the objective dashboard omits a speaking task (Council of Europe, 2020; Oscarson, 1997; Ross, 1998). As the project adds a short rubric-based speaking performance and fully links samples, within-person gain correlations should yield clearer alignment estimates, especially for Listening, which theory and meta-review suggest is most sensitive to comprehensible input plus interaction in short windows (Long, 1996; Swain, 2005).

5.4 Baseline traits as predictors (RQ4)

Treating Collaboration and SRL as baseline traits, neither predicted Listening or Speaking changes, and SRL did not explain short-interval variance in TOEFL outcomes after accounting for pretest. Over a one-month interval on bounded scales, pretest dominates posttest, and SRL effects typically accumulate over longer spans through planning, monitoring, and adaptive strategy use (Cronbach & Furby, 1970; Panadero, 2017; Pintrich, 2004; Rogosa, 1995). The negative association between baseline Collaboration and adjusted gains in Reading and Total admits two nonexclusive explanations. First, there may have been a headroom effect. Learners

reporting higher pre-departure baseline perceived collaboration may have had less measurable room to improve on conservative receptive outcomes (Cronbach & Furby, 1970; Rogosa, 1995). Second, a time-allocation trade-off may have occurred. Students who devoted more effort to collaborative oral work may have reduced the time available for silent reading practice that most directly boosts TOEFL reading performance, whereas peers who balanced collaboration with individual reading tasks realized larger receptive gains. Because Collaboration here is self-perceived and the sample is modest, this finding is provisional; future waves should incorporate exposure metrics that differentiate oral interaction, individual reading, and test-strategy practice (Johnson & Johnson, 2009; Storch, 2002).

5.5 Measurement and design implications

Short-window inference benefits from a mixed evidence base and section-sensitive interpretation. TOEFL ITP offers stable receptive indices for institutional monitoring, but small differences should be read against score precision and potential retest effects; confidence intervals and effect sizes are essential (Cronbach & Furby, 1970; Educational Testing Service, 2025; Rogosa, 1995). For CEFR-aligned self-assessment, localization to context and midpoint-free scales aid interpretation, but calibration shifts after intensive experiences are common; analytic approaches such as residualized change or (when sample size permits) latent change models can help (Council of Europe, 2020; Panadero, 2017; Ross, 1998). Programmatically, three priorities follow from the present pattern and prior syntheses:

1) Preserve interactional density through daily one-to-one/small-group work and structured, accountable collaboration (Sekiya & Park, 2006; Sekiya et al., 2018; Vande Berg et al., 2009).
2) Insert targeted focus-on-form touchpoints to address persistent grammatical bottlenecks that rarely move without explicit attention in four weeks (Long, 1996; Spada & Tomita, 2010; Swain, 2005).
3) Make individual reading practice visible and accountable, complementing collaboration so that reading gains continue to surface on section indices (Llanes & Muñoz, 2009; Segalowitz & Freed, 2004).

6. Conclusion

This study estimated short-interval development in English proficiency during a four-week, faculty-guided program in Cebu and examined whether perceived communicative gains and baseline learner traits aligned with observed change. Using matched pairs, students showed a moderate increase in TOEFL ITP Reading and a smaller increase in the Total score, while the Listening section and the Structure and Written Expression section were stable over the same interval. CEFR-aligned self-ratings of Listening and Speaking both increased significantly, indicating that students perceived growth in communicative ability during the program. These perceived gains complement the observed movement in receptive test scores and suggest that students experienced progress on multiple dimensions. Guidance from test documentation supports cautious interpretation of small pre–post differences and highlights the value of combining standardized testing with self-report and performance-based evidence.

The pattern aligns with a central claim in the study-abroad literature: program design, not immersion alone, drives learning. Multi-site and review evidence shows that short stays yield small to moderate effects when programs deliberately engineer interactional density, guided reflection, and accountable tasks. In that context, a reading-focused lift is plausible, especially in designs that include test-relevant practice. At the same time, movement in

discrete grammatical accuracy and test-specific listening is less likely without sustained focus on form.

Although self-assessed Listening and Speaking increased overall, gains did not correlate with objective test changes. This is compatible with known properties of self-assessment. Learners' internal standards may shift after intensive exposure, leading to changes in how they interpret can-do descriptors. The CEFR Companion Volume remains the reference for interpreting these measures and supports their use as part of a broader assessment strategy. As the project incorporates a brief, rubric-based speaking performance and additional linked data, stronger alignment estimates may emerge.

Program implications follow directly from the evidence. First, preserve interactional density through one-to-one and small-group instruction and make participation accountable through structured collaboration. Second, add targeted focus-on-form touchpoints to address grammatical bottlenecks least likely to change over short intervals. Third, maintain a mixed assessment dashboard: use TOEFL ITP sections and Total for conservative monitoring, and pair them with calibrated CEFR-aligned self-assessment and a brief, rubric-based speaking performance so that communicative development is visible even when receptive test scores move slowly. Cooperative-learning work also suggests that collaboration is most productive when key design elements are present, including positive interdependence and individual accountability, which can guide task scripting in future iterations.

Several boundaries of inference shape what we can conclude. The analytic sample was modest; only two time points were available for each measure in this wave; the testing and questionnaire datasets only partially overlapped; and Collaboration was indexed as a self-perceived baseline trait. Analyses used listwise deletion, which can introduce bias if missingness is not random. The larger project will address durability with a delayed post-TOEFL and add qualitative interviews with students, teachers, and administrators to document how collaboration and strategy support are enacted. These steps align with recommendations to combine conservative receptive testing with complementary indicators and to interpret small score changes against measurement precision.

Within these bounds, the study contributes short-window evidence that a carefully engineered four-week program can yield measurable gains in both perceived and tested English proficiency. For universities operating under tight academic calendars, the practical message is straightforward: design collaboration rather than assume it, preserve intensive interaction, and pair meaning-focused pedagogy with targeted form-focused support and balanced individual practice. This combination is consistent with programs that report the strongest short-stay outcomes in the literature and offers a credible path to near-term development while maintaining attention to student experience and equity of access.

References

Bradly, A., & Iskhakova, M. (2023). Systematic review of short-term study abroad outcomes and an agenda for future research. *Journal of International Education in Business, 16*(1), 70-90. https://doi.org/10.1108/JIEB-02-2022-0012

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education, 27*, 1-13. https://doi.org/10.1016/j.iheduc.2015.04.007

Council of Europe. (2020). *Common European Framework of Reference for Languages: Companion volume.* Council of Europe.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—Or should we? *Psycholog-*

ical Bulletin, 74(1), 68-80. https://doi.org/10.1037/h0029382

Educational Testing Service. (2022). *Major Field Tests: Guide to score interpretation.* ETS.

Educational Testing Service. (2025). *TOEFL ITP test taker handbook.* ETS.

Elabdali, R. (2021). Are two heads really better than one? A meta-analysis of the L2 learning benefits of collaborative writing. *Journal of Second Language Writing, 52*, Article 100788. https://doi.org/10.1016/j.jslw.2020.100788

Fan, N. (2024). Effects of collaborative vs. individual pre-task planning on EFL learners' L2 writing: Transferability of writing quality. *Humanities & Social Sciences Communications, 11*, Article 1365. https://doi.org/10.1057/s41599-024-03758-z

Georgeson, A. R., Valente, M. J., & Gonzalez, O. (2021). Evaluating response shift in statistical mediation analysis. *Psychological Assessment, 33*(7), 596-610. https://doi.org/10.1177/25152459211012271

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & Preckel, F. (2014). "My questionnaire is too long!" The assessment of motivational–affective constructs by optimized single-item and multiple-item measures. *Contemporary Educational Psychology, 39*(3), 188-205. https://doi.org/10.1016/j.cedpsych.2014.04.002

Goldstein, S. B. (2022). A systematic review of short-term study abroad research methodology and intercultural competence outcomes. *International Journal of Intercultural Relations, 87*, 26-36. https://doi.org/10.1016/j.ijintrel.2022.01.001

Hiver, P., Al-Hoorie, A. H., Vitta, J. P., & Wu, J. (2021). Engagement in language learning: A systematic review of 20 years of research methods and definitions. *Language Teaching Research, 28*(1), 201-230. https://doi.org/10.1177/13621688211001289

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38*(5), 365-379. https://doi.org/10.3102/0013189X09339057

Johnson, D. W., Johnson, R. T., & Holubec, E. J. (1999). Making cooperative learning work. *Theory Into Practice, 38*(2), 67-73. http://dx.doi.org/10.1080/00405849909543834

Kidd, P., Parshall, M. B., Wojcik, S., & Struttmann, T. (2004). Assessing recalibration as a response-shift phenomenon. *Nursing research, 53*(2), 130-135. https://doi.org/10.1097/00006199-200403000-00009

Kinginger, C. (2009). *Language learning and study abroad: A critical reading of research.* Palgrave Macmillan.

Li, M., & Zhang, X. (2020). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189-218. https://doi.org/10.1177/0265532220932481

Llanes, À., & Muñoz, C. (2009). A short stay abroad: Does it make a difference? *System, 37*(3), 353-365. https://doi.org/10.1016/j.system.2009.03.001

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). Academic Press.

Mackey, A. (1999). Input, interaction, and second language development. *Studies in Second Language Acquisition, 21*(4), 5573-587. https://doi.org/10.4324/9780203053560-8

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language*

and education (Vol. 7: *Language testing and assessment*, pp. 175-187). Kluwer Academic.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Park, S. (2025). The impact of resilience, collaboration, and language learning strategies on L2 oral proficiency development in short-term study-abroad programs. *Studies in English Education, 30*(1), 59–89. https://doi.org/10.22275/SEE.30.1.03

Park, S., & Sugita, M. (2024). The effects of online and onsite study abroad experiences on students' perceptions of global challenges and decisions on their career pathways. *Journal of Kanda University of International Studies, 36*, 163–183.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning, 44*(3), 493-527. https://doi.org/10.1111/j.1467-1770.1994.tb01115.x

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385-407. https://doi.org/10.1007/s10648-004-0006-x

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353-387. https://doi.org/10.1037/a0026838

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research," plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3-66). Erlbaum.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis. *Language Testing, 15*(1), 1-20. https://doi.org/10.1177/026553229801500102

Sanz, C. (2014). Contributions of study abroad research to our understanding of SLA processes and outcomes: The SALA project, an appraisal. In C. Pérez-Vidal (Ed.), *Language acquisition in study abroad and formal instruction contexts* (pp. 1-14). John Benjamins. https://doi.org/10.1075/aals.13.01ch1

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at-home and study-abroad contexts. *Studies in Second Language Acquisition, 26*(2), 173-199. https://doi.org/10.1017/S0272263104262027

Sekiya, Y., & Park, S. (2006). An experimental short-term study-abroad program in the United States: Its design, implementation, and effects on participants' oral proficiency. *Studies in Linguistics and Language Teaching, 17*, 167-193.

Sekiya, Y., Park, S., & Tsuji, R. (2018). Effects of a short-term study-abroad program. *Studies in Linguistics and Language Teaching, 29*, 161-180.

Serrano, R., Llanes, À., & Tragant, E. (2016). Examining L2 development in two short-term intensive programs for teenagers: Study abroad vs. "at home". *System, 57*, 43-54. https://doi.org/10.1016/j.system.2016.01.003

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*(2), 263-308. https://doi.org/10.1111/j.1467-9922.2010.00562.x

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52*(1), 119–158. https://doi.org/10.1111/1467-9922.00179

Storch, N. (2013). *Collaborative writing in L2 classrooms*. Multilingual Matters.

Swain, M. (2005). The output hypothesis: Theory

and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471-483). Lawrence Erlbaum Associates.

Teng, L. S., & Zhang, L. J. (2022). Can self-regulation be transferred to second/foreign language learning and teaching? Current status, controversies, and future directions. *Applied Linguistics, 43*(3), 587-597. https://doi.org/10.1093/applin/amab041

Trentman, E. (2021). Arabic study abroad: Critical contextualization and research-based interventions. In K. Ryding & D. Wilmsen (Eds.), *The Cambridge Handbook of Arabic Linguistics* (pp. 106–126). Cambridge University Press

Vande Berg, M., Connor-Linton, J., & Paige, R. M. (2009). The Georgetown Consortium Project: Interventions for student learning abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad, 18*(1), 1-75. https://doi.org/10.36366/frontiers.v18i1.251

Zou, B., Wang, D., & Xing, M. (2016). Collaborative tasks in wiki-based environments in EFL learning. *Computer Assisted Language Learning, 29*(5), 1000-1018. https://doi.org/10.1080/09588221.2015.1121878