Investigating Factors Associated With Gains in TOEFL ITP® Scores at a University in Japan

Siwon Park and Tetsuya Fukuda

Faculty of International Liberal Arts, Juntendo University

Abstract

This report presents the outcomes of a project funded by the TOEFL® English-Language Researcher/Practitioner Grant Program. The project examined how much first-year students in a Japanese university English program gained on the TOEFL ITP® test over the course of one academic year and which learner characteristics were associated with those gains. A cohort of 239 students in an intensive liberal arts English curriculum took the TOEFL ITP test three times. The students also completed measures of motivation, resilience, self-regulated learning (SRL), engagement (measured as peer support), and critical thinking, together with an end-of-year survey that asked whether the test reflected their English proficiency. Mean total scores increased from 426.2 to 455.2. The mean paired gain from Time 1 to Time 3 was 27.9 points (d_z = 0.85). However, the design included no comparison group, so the observed gain cannot be attributed to instruction alone. Critical-thinking performance was represented by a linked Rasch person measure that was constructed across three test forms. Critical thinking was the only focal learner variable that had a statistically significant unique association with end-of-year proficiency in regressions that controlled for Time 1 proficiency and form group (β = .358; the equivalent coefficient on the gain metric was .418). It was also associated with all three section scores (β = .239 to .352). No psychosocial composite had a significant unique association with end-of-year performance, although motivation and resilience were correlated with SRL (r = .40 and .42). Most learner measures were administered after the final test; therefore, the associations are concurrent or retrospective, and no causal interpretation is warranted. In the perception survey, 54% of

students rated TOEFL ITP as “more or less” reflective of their English proficiency. The open-ended comments centered on the absence of speaking and writing from the regular score. The report discusses the implications for score interpretation, program evaluation, and future TOEFL ITP research.

Keywords: TOEFL ITP®, score gains, longitudinal assessment, critical thinking, self-regulated learning, motivation, resilience, engagement, learner perceptions, Japanese university students

Introduction

Japanese universities administer the TOEFL ITP® tests on a large scale, and repeated administration is a routine part of program evaluation: A cohort is tested at entry, taught for a year, and tested again. The program is considered successful when mean scores increase. However, average scores obscure the question that matters most for instruction and score interpretation. Some students gain more than 60 points, and others lose points under the same curriculum, teachers, and contact hours. This report presents findings from a project funded through the TOEFL® English-Language Researcher/Practitioner Grant Program. The project followed one first-year cohort at a Japanese university through three administrations of the TOEFL ITP. Its aims were to document the size and timing of score changes and to identify the learner characteristics that were associated with larger and smaller gains.

Our approach treats score change as more than a by-product of instruction. Learners bring different resources to the same classroom, and these differences may help explain why measured proficiency grows more for some students than for others. We examined five such learner resources. Motivation, self-regulated learning (SRL), engagement, and resilience represent the psychosocial side of that account. That is, they refer to the orientation to invest in English, the capacity to manage one’s own study behavior, peer-supported participation in classroom learning, and the ability to keep functioning when the work becomes difficult. Critical thinking represents the cognitive side, that is, the inferential and analytical operations that TOEFL ITP items may also demand. A sixth component examined students’ perceptions of the test at the end of the year. A score that institutions rely on can be interpreted more easily when the students who produce it explain what they think it does and does not reflect.

Background

TOEFL ITP as an Institutional Measure of English Development

The TOEFL ITP tests have a specific role in institutional language assessment. The Educational Testing Service (ETS) positions the test series as suitable for placement, progress monitoring, and exit decisions within English language programs. The Level 1 test that was used in this project contains three multiple-choice sections: listening comprehension (50 items, 35 minutes), structure and written expression (40 items, 25 minutes), and reading comprehension (50 items, 55 minutes). Each section yields a scaled score (31–68 for listening and for structure and written expression, 31–67 for reading). The total score is the average of the three section scores multiplied by 10, and it ranges from 310 to 677. Speaking is assessed by the separately administered TOEFL ITP speaking test and is reported outside of the regular score (ETS, 2025). Two implications follow for the present project. First, the repeated-measures design depends on the progress-monitoring use, which ETS itself has examined using institutional data (Choi & Papageorgiou, 2014). Second, the three sections set the interpretive boundary of the total score. TOEFL ITP total scores reflect receptive and structure-related proficiency, not four-skill communicative ability. This distinction is particularly important in a curriculum that also includes speaking and writing.

Research on the use of TOEFL-family scores supports treating this construct boundary as central to score interpretation. Kokhan and Lin (2014) showed that repeated TOEFL score reports can inform English as a Second Language (ESL) placement decisions. Harsch et al. (2017) evaluated the use of TOEFL iBT® scores in a British university through both predictive relationships and stakeholders' accounts of preparedness. Although TOEFL ITP and TOEFL iBT are not interchangeable, these two studies converge on the same principle: The usefulness of a score depends on how closely its interpretation is aligned with the construct that the test is intended to measure.

Evidence specific to TOEFL ITP leads to the same conclusion. Golubovich et al. (2018) surveyed and interviewed institutional users in Japan, Mexico, and Indonesia. These users described the test as practical and affordable for placement and exit decisions, but they were also aware of its narrower skill coverage. Collins and Miller (2018) found more reserved views in a survey of teachers at a Japanese university: Respondents accepted the institutional role of the test but rated it less favorably as a measure of communicative competence. A test can be a workable institutional benchmark and still raise construct-

related reservations when the surrounding curriculum has a broader aim than the three sections of the test.

Motivation

Motivation remains the default explanation for differential language achievement because it addresses why learners initiate, sustain, and regulate effort over long periods. Dörnyei's (2005) synthesis of learner psychology and the self-determination account of Noels et al. (2003) both emphasize the quality of motivation over its quantity. Enjoyment, integrative interest, instrumental goals, effortful commitment, and an ideal L2 self are related but not interchangeable, and they need not have the same relation to a given outcome. Meta-analytic work indicates the magnitudes that such variables can be expected to show. Masgoret and Gardner (2003) found that motivation correlated with L2 achievement at approximately r = .37 across the 75 studies from the Gardner tradition that they synthesized. Al-Hoorie (2018) found in a meta-analysis of the L2 motivational self system that the ideal L2 self correlated .61 with self-reported intended effort but .20 with measured achievement. Under Plonsky and Oswald's (2014) field-specific benchmark, correlations near .25 and .40 are small and medium, respectively. By that benchmark, motivational variables typically show small-to-medium associations with test performance, and the associations are larger for self-reported effort than for objective scores. In this study, motivation refers to the orientation that a 29-item questionnaire measured, which spanned effort, enjoyment, integrativeness, instrumentality, and ideal L2 self. In the analyses, motivation was represented by the overall composite of those items (see Measures).

Recent work in Japanese and comparable EFL settings fits this description. Ueno and Park (2025) found that motivational and self-regulatory variables were related more strongly to each other than to test scores among Japanese university students who were assessed with TOEFL-based measures. Derakhshan and Fathi (2024) reported that, in their structural model, the associations of growth mindset and self-efficacy with L2 speaking performance were mediated by self-regulation. Both results suggest that motivational resources matter most once they are converted into regulated study behavior. This expectation shaped our own selection of measures.

Self-Regulation and Engagement

Self-regulation is the construct in this project that most closely reflects students' day-to-day learning behavior. Zimmerman (2002) described SRL as a cycle of forethought, performance control, and self-reflection. Tseng et al. (2006) argued that what should be measured is not an inventory of strategies but the capacity of the learner to control strategic behavior, and they distinguished five facets of that control: commitment, metacognitive, satiation, emotion, and environmental control. Their self-regulating capacity questionnaire for vocabulary learning has been the reference instrument for measurement in this tradition. From this perspective, successful learners do more than hold favorable attitudes. They manage attention, mood, boredom, and study conditions in the service of their goals. Meta-analytic estimates show that self-regulatory processes have small average associations with achievement in school-age samples, near $r = .20$ for metacognitive processes and $r = .11$ for cognitive strategy use (Dent & Koenka, 2016). In this study, SRL refers to self-reported control capacity across the five facets of Tseng et al. (2006). It was measured by a 20-item questionnaire and analyzed as an overall composite (see Measures).

Empirical studies have related self-regulation to language outcomes directly and through its links with other learner characteristics. Jackson and Park (2020) found that self-regulation in L2 writing operates jointly with personality and situational states, not as an isolated trait. Yang et al. (2025) showed that an SRL intervention improved motivation, strategic competence, and English proficiency among Chinese EFL learners. Engagement is closely related to self-regulation but is not synonymous with it. Engagement refers to the enacted, observable side of participation in learning. Meta-analytic estimates place its association with achievement near $r = .27$ overall and at .35 for behavioral engagement (Lei et al., 2018). Engagement in this project was measured narrowly as prosocial peer learning support, that is, students' perceptions of giving and receiving help with English among classmates. The construct that we examined is therefore peer-support engagement, which is closest to the behavioral component of the broader construct, instead of engagement in its full multidimensional sense (see Measures). Accordingly, we kept self-regulation and engagement separate throughout and interpreted the engagement findings within that narrower scope.

Resilience

Language learning is a long and effortful process that often involves frustration. This is why resilience has become an important topic in L2 individual-difference research. The construct refers to the capacity to maintain or recover adaptive functioning under difficulty. Recent studies have related it to persistence and strategy maintenance more than to positive affect alone. Park et al. (2025) found that self-regulation was the strongest predictor of L2 proficiency among ESL learners and that resilience was related to proficiency mainly through its association with self-regulation. Park (2025) reported that resilience was associated with the development of oral proficiency in short-term study abroad, which is a setting with frequent challenges. Wang and Patterson (2025) added that academic resilience in EFL settings is shaped by learner beliefs and, to a lesser extent, by classroom climate. Their finding cautions against treating resilience as purely intrapersonal. Meta-analytic benchmarks comparable to those cited for the other constructs are not yet available for trait resilience and language achievement, so expectations about magnitude depend on primary studies such as these. In this study, resilience refers to trait resilience as measured by the 25-item Connor–Davidson Resilience Scale (CD-RISC-25; Connor & Davidson, 2003). The working hypothesis was that resilience relates to sustained study behavior more than to score change directly.

Critical Thinking

Critical thinking has a plausible relation to TOEFL ITP performance because some test items may require cognitive processes that overlap with those included in definitions of the construct. In the Delphi report, an expert panel defined critical thinking in terms of interpretation, analysis, inference, and evaluation (Facione, 1990). Halpern (2014) framed it as the deliberate use of such skills to reach warranted conclusions. Fong et al. (2017) reported in a meta-analysis of community college studies that the overall association between critical thinking and achievement was approximately $r = .28$, which is a small-to-medium value by Plonsky and Oswald's (2014) benchmarks. TOEFL ITP listening and reading items may involve inference from spoken or written input, discrimination among plausible alternatives, and consistency checking under time pressure. The structure and written expression section may reward analytical attention to grammatical relations and rapid elimination of distractors. These correspondences are hypotheses about item demands, not

established features of TOEFL ITP response processes. The present study collected no item-level response-process evidence, and successful performance could reflect English proficiency, grammatical knowledge, processing speed, or test-taking skill instead of critical-thinking ability. In this study, critical thinking refers to the consistency, analysis, and inference skills that are sampled by an English-medium instrument line that was developed for Japanese EFL learners (see below and Measures).

A practical obstacle to examining that relation in Japan has been the availability of appropriate instruments. Recent work has reduced this obstacle, although important measurement issues remain. Hirai et al. (2020) piloted an English-medium critical-thinking test for Japanese EFL learners, and Kubo (2022) reported factor-analytic support for a related pilot instrument. In the fullest validation study, Hirai et al. (2022) found that the intended three subskills (consistency, analysis, and inference) did not emerge as three separate factors. Instead, a two-factor model fit better, with consistency as one factor and analysis and inference combined as the other. Scores correlated $r = .64$ with an English proficiency measure and $r = .55$ with a Japanese-language critical-thinking test. This line of research therefore provides a locally developed measure with emerging validity evidence, not a decontextualized proxy for general ability. At the same time, its scores appear to reflect both English-language ability *and* critical thinking. This is an important consideration when the outcome measure is itself an English proficiency test. Substantive evidence has followed. Park (2022) found that the critical-thinking and self-regulating capacities of L2 learners were associated with their English grammar and reading abilities. The present project extends this cross-sectional evidence to a record of TOEFL ITP scores at three time points. It treats critical thinking as a candidate correlate of score development, not as an enrichment construct that was appended to the study design.

Learner Perceptions and Score Users

An outcomes report should also describe how learners themselves view the assessment that is used to measure their growth. In Bachman and Palmer's (2010) framework, the justification of score use requires that the interpretation be aligned with the intended use and with the construct that the test represents. Learner perceptions do not establish validity. They are stakeholder evidence about perceived relevance, washback, and the consequences of test use, and they cannot by themselves show that an interpretation is

sound or that consequences are beneficial. The TOEFL-related record indicates what to expect. However, it spans different tests and stakeholder groups, so it should be read as context, not as a single accumulating estimate. TOEFL iBT assesses all four language skills, unlike the three-section TOEFL ITP. Students who were admitted to a British university on TOEFL iBT scores generally felt prepared for academic study (Harsch et al., 2017). For TOEFL ITP itself, institutional users in Japan, Mexico, and Indonesia described the test as practical but were aware of its narrower skill coverage (Golubovich et al., 2018). Teachers at a Japanese university accepted its institutional role but rated it less favorably as a measure of communicative competence (Collins & Miller, 2018).

Direct evidence from students is more limited. Graduates of an English department at an Indonesian university judged the TOEFL requirement that they had met (the report does not specify which TOEFL test) as more relevant to entering professional life than to pursuing further study. They accepted the use of the score but distinguished the purposes that it serves (Karjo & Ronaldo, 2019). TOEFL preparation classes have produced favorable attitudes toward the test together with concerns about narrow, strategy-focused teaching (Wang & Huang, 2020). Collectively, these studies imply that a simple positive–negative polarity is the wrong frame. Learners can accept a test as useful for placement or progress monitoring and still withhold endorsement if they hear the question as one about four-skill development. We therefore interpreted the perception data with that possibility in mind.

The Current Study

The constructs described above suggest a layered account of the learner characteristics that may be associated with the development of English proficiency as measured by the TOEFL ITP. Motivation provides the orientation to invest effort in learning. Self-regulation and engagement describe how that orientation is enacted through study behavior. Resilience concerns maintaining that behavior under difficulty. Critical thinking represents the analytic processes that some TOEFL ITP item types may require. We refer to motivation, resilience, SRL, and engagement as psychosocial variables and to critical thinking as a cognitive variable. Learner perceptions complement the score record by providing evidence on how students understand and evaluate the assessment and the scores it produces. One design feature is central to the framing of the study throughout: The learner measures were administered late in the academic year, and most of them were

administered after the final test administration (see Procedure). The project therefore examined these variables as correlates of TOEFL ITP performance and score gains, not as determinants of growth. Within that frame, the project addressed five research questions (RQs):

- RQ 1: To what extent is motivation associated with students' end-of-year TOEFL ITP performance and their Time 1–3 score gain, once initial proficiency is taken into account?
- RQ 2: To what extent is critical-thinking ability associated with students' end-of-year TOEFL ITP performance and their Time 1–3 score gain, once initial proficiency is taken into account?
- RQ 3: To what extent are self-regulated learning, engagement, and resilience associated with students' end-of-year TOEFL ITP performance and their Time 1–3 score gain, once initial proficiency is taken into account?
- RQ 4: How are the psychosocial and cognitive learner variables related to one another and to TOEFL ITP scores?
- RQ 5: To what extent do students perceive TOEFL ITP scores as reflecting their English proficiency?

Method

Design and Setting

We used a longitudinal institutional design with three TOEFL ITP Level 1 administrations across one academic year. The first administration was on March 27, 2025, before program instruction began; the second on July 7, 2025, at the end of the spring semester; and the third on January 20, 2026, near the end of the fall semester. The setting was the intensive English program of the Faculty of International Liberal Arts at a private university in Tokyo, Japan. Students attended six 90-minute English lessons per week across two 15-week semesters. This schedule totaled roughly 270 hours of instruction for the year, of which approximately 225 hours had been delivered between the first and third administrations. Four of the six weekly lessons belonged to English for Liberal Arts, which is a four-skills integrated strand, and the other two to Academic English, which is an academic-skills strand. Courses were taught in English by native-English-speaking and Japanese

teachers. Instruction used in-house liberal arts materials, content-and-language-integrated learning, and explicit grammar teaching, and it addressed reading, listening, grammar, vocabulary, writing, and speaking. Typical class size was 15 to 17 students. Students were placed into one of three proficiency-based tiers on the basis of the first TOEFL ITP administration in late March, and classes were formed within tiers. TOEFL ITP preparation was not addressed directly in lessons, although the academic orientation of the instruction overlaps with the abilities that the test samples.

Participants

The analytic dataset consisted of 239 first-year students who were majoring in International Liberal Arts (135 female, 104 male). Most students (227) were first-language speakers of Japanese. The remaining 12 were international students from China (10) and Sudan (2). Time 1 TOEFL ITP totals were available for all 239 students, Time 2 totals for 238, and Time 3 totals for 230. Questionnaire coverage varied because not all students completed all project components. The motivation composite was available for 204 students, resilience for 223, SRL for 224, engagement for 211, and a linked critical-thinking person measure for 192 unique cohort members. The end-of-year perception survey received 228 responses. Multivariable models that required all predictors simultaneously were estimated on the complete-case subsample of 136 students. Table 1 summarizes variable availability.

Measures

TOEFL ITP Proficiency

The outcome measures were the TOEFL ITP Level 1 section scores for listening comprehension, structure and written expression, and reading comprehension, together with the scaled total score, at each of the three time points. For brevity, tables and figures label the structure and written expression section as “grammar.”

Motivation

Twenty-nine 4-point Likert-type items measured five motivational indicators: effort, enjoyment, integrativeness, instrumentality, and ideal L2 self. The items were adapted from Fukuda’s (2020) L2 motivation questionnaire for first-year students at Japanese universities and were administered in English. The composite that was used in the models was the

overall mean across the 29 items (α = .96). The dimensional structure of the battery was not tested in this sample. The composite was therefore interpreted as a summary of broad motivational orientation, not as evidence that the five indicators formed one factor (see Limitations). Sample items appear in the appendix.

Resilience

Resilience was measured using the authorized Japanese version of the CD-RISC-25 (Connor & Davidson, 2003; Ito et al., 2009). The CD-RISC-25 asks respondents to indicate how well each statement describes them with reference to the previous month, using five response categories ranging from "not true at all" to "true nearly all the time." The authorized Japanese version was purchased directly from the copyright holder and used in accordance with the applicable terms of use. In the analytic dataset, responses were coded 1 to 5, and the overall item mean was used. Possible mean scores thus ranged from 1 to 5, and higher scores indicated greater resilience. Internal consistency in the present sample was high (α = .94).

Self-Regulated Learning

A 20-item questionnaire with a 6-point response scale measured SRL across the five control facets that Tseng et al. (2006) distinguished: commitment, metacognitive, satiation, emotion, and environmental control. The items were adapted from Tseng et al.'s (2006) self-regulating capacity in vocabulary learning questionnaire. The survey was administered bilingually in Japanese and English (see the appendix for the first page). The overall scale mean was used in the models (α = .96).

Engagement

Engagement was measured with the 4-item Prosocial Peer Learning Support (PPLS) Engagement Questionnaire, which was administered in English with a 6-point response scale. The construct refers to engagement in which students' perceptions of their mutually constructive participation matter to them in their academic achievement (Fukuda et al., 2021). The items address peer support in English learning. Two items ask about receiving help from classmates in learning and enjoying English, and two mirror items ask about providing that help (see the appendix for all four). The scale thus operationalizes engagement narrowly as peer-support engagement, which is closest to the behavioral

component of the broader multidimensional construct. The overall mean was used in the models (α = .85). The two-item received and provided subsets were not analyzed separately.

Critical Thinking

Critical thinking was measured with an English-medium test for Japanese EFL learners that Hirai et al. (2020, 2022) developed. The test was used with the permission of the developers, on the condition that its items not be reproduced (see the appendix). Each student completed one of three forms that were connected by common items. Forms were assigned at the class level. The allocation of forms to classes was decided arbitrarily for administrative convenience, not by students' scores. However, classes were formed within proficiency tiers, so form group overlapped with class placement and prior proficiency. A post hoc, form-specific screen excluded items with negative corrected item–total correlations. The screen reduced Forms A, B, and C from 49, 50, and 51 items to 33, 38, and 39 items, respectively. This screen was conducted after data collection and was not preregistered, and TOEFL outcomes were not used in it. A total of 10 usable common-item families connected 110 administered item occurrences as 96 linked item columns. All 202 response records entered a dichotomous Rasch calibration. Joint maximum-likelihood estimation was performed in Winsteps 3.65.0 (Linacre, 2008), and the linked person measure was used in all focal critical-thinking analyses. Person measures were transformed as $T = 50 + 10[(\theta - .1363678)/.6608436]$, where .1363678 and .6608436 were, respectively, the mean and standard deviation of the linked logit person measures across all 202 calibration records. The transformation therefore set the full calibration sample to a mean of 50 and a standard deviation of 10. We did not establish common-item invariance, so we describe the result as a conditional link and do not claim operational interchangeability across forms (Kolen & Brennan, 2014).

Perceptions of TOEFL ITP

A follow-up survey that was administered after the third TOEFL ITP administration asked one closed-ended question: whether the student thought that TOEFL ITP reflected their overall English proficiency. The response options were "yes," "more or less," and "not really." An optional open-ended prompt then invited students to explain their view. One of the authors read the open-ended responses in full and grouped them inductively into themes. A second coder was not used (see Limitations). Some students raised multiple

points in a single response, so a response could contribute to more than one theme. The themes that are reported below are the ones that were raised most often, not an exhaustive partition of the comments.

Procedure

The measures were administered across the academic year in the order that the program could accommodate. The resilience scale was administered at the end of the spring semester, after the second TOEFL ITP administration. The critical-thinking test was given in December 2025, roughly one month before the final TOEFL ITP administration. The motivation, SRL, engagement, and perception measures were administered immediately after the third TOEFL ITP administration in January 2026. This sequence has important implications for interpretation. Because students completed the January questionnaires after all three TOEFL ITP administrations, their responses may have been influenced by their results and their experiences of the tests. The critical-thinking test preceded the final time point, but it was not a baseline measure and may itself have reflected learning that occurred during the year. The resilience measure can relate prospectively only to the second-semester interval. Accordingly, we treat the learner variables as correlates of TOEFL ITP performance and score gains, not as temporally prior predictors (see Limitations).

Ethical Considerations

The study was reviewed and approved by the research ethics committee of the Faculty of International Liberal Arts at the participating university (Protocol No. 2024007; approval notification dated May 2, 2025), under the title of this report and with the first author as principal investigator. The TOEFL ITP administrations were part of the regular assessment schedule of the program. The questionnaires and the critical-thinking test were administered for the research. Questionnaire administrations included an explanation that responses would be processed statistically and could not be used to identify any individual, and students indicated their agreement to participate before responding. Participation in the project components was voluntary, as the varying response counts in Table 1 reflect. No personally identifying information entered the analytic dataset. Responses were linked across instruments and score records through pseudonymous student identifiers, and all results are reported in aggregate.

Data Preparation and Reliability

Data preparation proceeded in five steps. First, we merged the score and questionnaire files by student identifier. Second, we recoded placeholder zeros as missing wherever the source files showed that zero marked an absent response. Third, we resolved duplicated critical-thinking identifiers by checking the records against the source files and retaining the verified entry. Fourth, we formed the motivation, resilience, SRL, and engagement composites. Fifth, we computed the Time 1–2, 2–3, and 1–3 gains. All 202 critical-thinking response records (A = 71, B = 59, C = 72) entered the calibration. Three source identifiers appeared twice in the critical-thinking file, apparently through data-entry error. The duplicated records were double-checked, and only the correct entry was retained in each case. The linked measures were then matched to the longitudinal file, which yielded 192 unique cohort members with a linked person measure. Coefficient alpha was computed from the students with complete item responses to each questionnaire (*n*s as in Table 2). The coefficients were .96 for motivation, .94 for resilience, .96 for SRL, and .85 for engagement. KR-20 coefficients for the retained form-specific critical-thinking item sets were .63, .66, and .71. Real person reliability was .64, .67, and .68 in the three complete-form runs and .67 in the linked run. Linked item reliability was .90, with item infit mean squares from .87 to 1.13 and outfit mean squares from .66 to 1.28. First residual-contrast eigenvalues were 2.8, 3.2, and 3.0, and maximum mean-adjusted Q3 values were .580, .521, and .482 for Forms A, B, and C, respectively. These residual diagnostics and the modest person reliability limit strong interpretations of individual scores (Christensen et al., 2017).

Analytic Approach

The analyses were sequenced from descriptive summaries to conditional associations. Descriptive statistics and reliability estimates established the distributional profile of the variables. Paired-change tests and piecewise mixed-effects models tested whether scores changed and, if so, when the change was concentrated. The mixed-effects models included random intercepts for students and separate fixed effects for the two semester intervals. Correlations then described the associations among baseline proficiency, later proficiency, score gains, and the five focal constructs (RQ 4).

A baseline- and form-adjusted regression model addressed the conditional associations in RQs 1–3. Time 3 total proficiency was regressed on Time 1 total, motivation,

resilience, SRL, engagement, the linked critical-thinking T score, and Form B and C indicators simultaneously, with Form A as the reference group. Form indicators were retained because the critical-thinking form group explained 82.4% of the variance in Time 1 total scores in the matched data. We defined gain as the Time 3 score minus the Time 1 score, so the same model reparameterized with gain as the outcome yielded identical unstandardized coefficients and significance tests for each predictor except for the baseline proficiency. Consequently, we report the gain results alongside the Time 3 model, not as a separate analysis.

Parallel models were estimated for each TOEFL ITP section with the corresponding Time 1 section score as the baseline. Finally, a longitudinal covariance-structure model summarized the measurement structure of TOEFL ITP proficiency across the three time points. This model involves only the nine section scores and includes no learner variables, and it is reported as measurement evidence, not as a structural claim about learning. Data screening followed standard multivariate practice (Hair et al., 2018; Tabachnick & Fidell, 2019). The learner-variable models used complete cases (n = 136, roughly 17 cases per predictor). Students who were missing questionnaire components may differ from completers (see Limitations). Regression models were estimated by ordinary least squares, piecewise growth models using mixed-effects maximum likelihood, and the longitudinal covariance-structure model using maximum likelihood.

Results

Descriptive Statistics

Table 1 shows the cohort profile and the availability of variables. Table 2 reports descriptive statistics and internal consistency.

TOEFL ITP coverage remained nearly complete across the year, with attrition of nine students by Time 3. Cohort mean scores increased at each time point, and the questionnaire composite scores were above their respective scale midpoints. The linked critical-thinking T score had a mean of 49.75 and a standard deviation of 9.92 among the 192 matched cohort members. These values are close to the 50/10 scaling that was defined across all 202 calibration records.

Table 1. Project Cohort and Variable Availability

Project variable	Count
Unique students in the analytic data file	239
Female students	135
Male students	104
TOEFL ITP total available, Time 1	239
TOEFL ITP total available, Time 2	238
TOEFL ITP total available, Time 3	230
Motivation composite available	204
Resilience composite available	223
Self-regulated learning composite available	224
Engagement composite available	211
Critical thinking (unique cohort members with linked person measure)	192
Follow-up perception survey responses	228

Note. Counts reflect variable availability in the analytic dataset used for the reported analyses.

Table 2. Descriptive Statistics and Internal Consistency of the Main Project Variables

Measure	*n*	M	*SD*	α/KR-20
TOEFL ITP total, Time 1	239	426.22	41.78	—
TOEFL ITP total, Time 2	238	445.54	44.49	—
TOEFL ITP total, Time 3	230	455.23	42.32	—
Pre–post gain (Time 3 − Time 1)	230	27.94	32.97	—
Motivation (29 items)	204	3.42	0.50	.96
Resilience (25-item mean)	223	3.68	0.70	.94
Self-regulated learning (20 items)	224	4.06	0.95	.96
Engagement (4 items)	211	4.85	0.91	.85
Critical thinking (linked Rasch T score)	192	49.75	9.92	.63/.66/.71

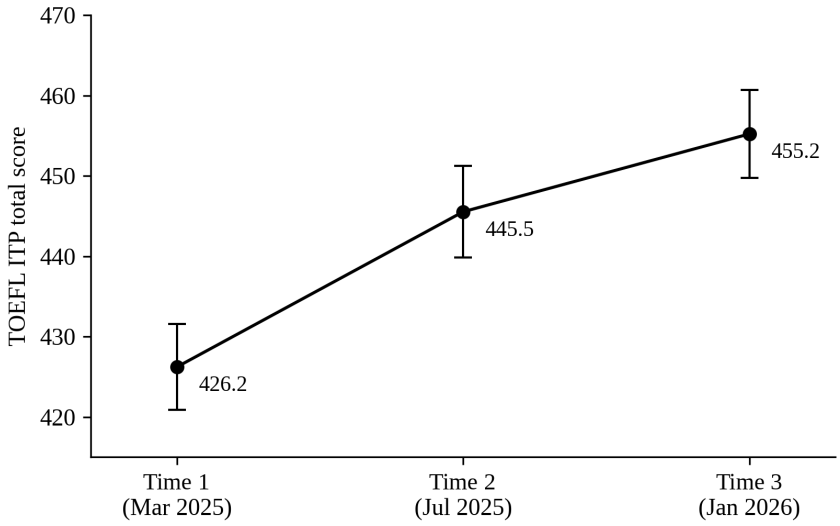
Note. Internal consistency values for the questionnaire measures are coefficient alpha estimates. Resilience descriptives are reported on the 1–5 item-mean metric used in the analyses. For critical thinking, the three values are form-specific KR-20 coefficients for the retained item sets. The critical-thinking mean and standard deviation are based on the 192 matched cohort members; the T scale was centered at 50 with a standard deviation of 10 across all 202 calibration records. Real person reliability in the linked calibration was .67.

Growth in TOEFL ITP Scores

Mean total scores rose from 426.2 at Time 1 to 445.5 at Time 2 and 455.2 at Time 3 (Figure 1). Mean scores also increased in all three sections (Figure 2). Table 3 reports the paired changes. The total gain from Time 1 to Time 3 averaged 27.9 points, 95% confidence interval (CI) [23.7, 32.2]. This gain is a large within-student effect (d_z = 0.85). Change was

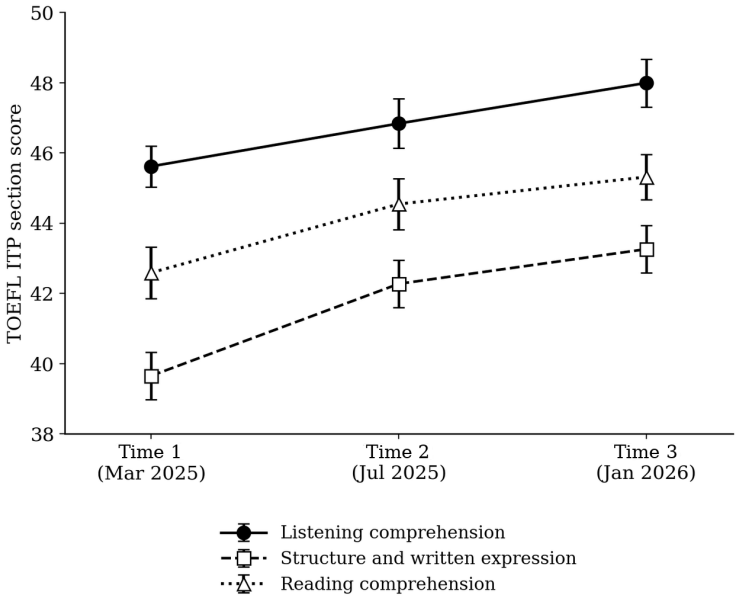
front-loaded: The first-semester gain (19.3 points) was greater than the second-semester gain (9.2 points). However, the second-semester change remained statistically significant for the total score and for each section (Table 3). At the section level, grammar scores increased most in the first semester. In contrast, listening and reading scores improved more gradually across the year.

Figure 1. TOEFL ITP Total-Score Trajectory Across the Academic Year



Note. Error bars are 95% confidence intervals around the cohort mean at each time point (ns = 239, 238, and 230).

Figure 2. Section-Score Trajectories Across the Academic Year



Note. Error bars are 95% confidence intervals. Grammar = structure and written expression.

Table 3. Repeated-Measures Change in TOEFL ITP Total and Section Scores

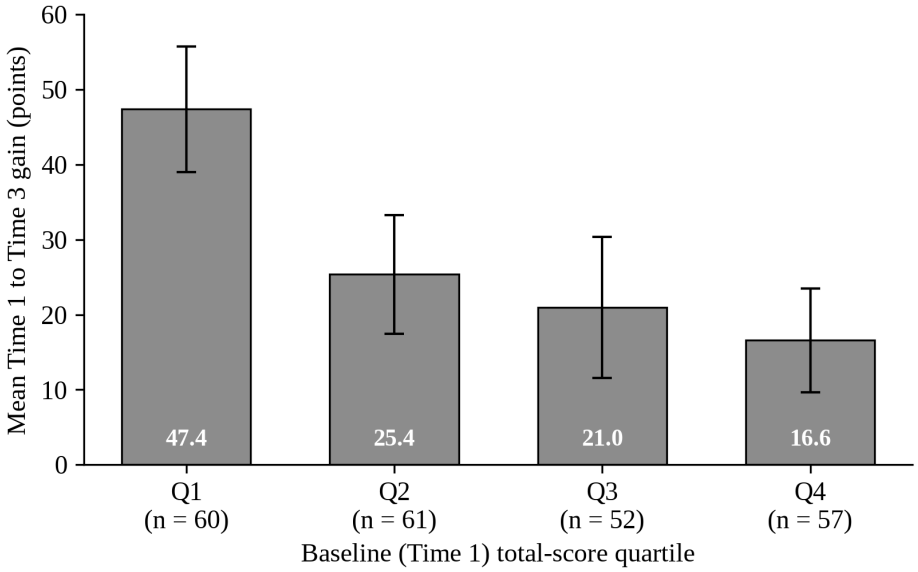
Outcome	Comparison	*n*	M_{change}	95% CI	t	*p*	d_z
Total	Time 1 to Time 2	238	19.31	[15.22, 23.39]	9.31	< .001	0.60
Total	Time 2 to Time 3	229	9.17	[5.30, 13.03]	4.67	< .001	0.31
Total	Time 1 to Time 3	230	27.94	[23.66, 32.22]	12.85	< .001	0.85
Listening	Time 1 to Time 2	238	1.21	[0.70, 1.72]	4.70	< .001	0.30
Listening	Time 2 to Time 3	229	1.10	[0.60, 1.61]	4.34	< .001	0.29
Listening	Time 1 to Time 3	230	2.33	[1.86, 2.79]	9.79	< .001	0.65
Grammar	Time 1 to Time 2	238	2.61	[1.95, 3.28]	7.71	< .001	0.50
Grammar	Time 2 to Time 3	229	0.92	[0.27, 1.56]	2.81	.005	0.19
Grammar	Time 1 to Time 3	230	3.57	[2.85, 4.28]	9.87	< .001	0.65
Reading	Time 1 to Time 2	238	1.97	[1.23, 2.71]	5.24	< .001	0.34
Reading	Time 2 to Time 3	229	0.73	[0.05, 1.42]	2.11	.036	0.14
Reading	Time 1 to Time 3	230	2.50	[1.78, 3.22]	6.81	< .001	0.45

Note. Each comparison uses the students with paired observations. CI = confidence interval. Grammar = structure and written expression. d_z is the mean change divided by the standard deviation of the change scores.

The piecewise mixed-effects models were fitted to the 707 score records contributed by the 239 students and yielded the same result when all available data were used. The fixed effect for the first interval was 19.12 total-score points, and the fixed effect for the second interval was 9.02 points. Section models estimated first-semester increments of 1.20 (listening), 2.60 (grammar), and 1.94 (reading) scaled-score points and second-semester increments of 1.12, 0.95, and 0.68, respectively. The growth continued throughout the year but at a slower rate.

Figure 3 shows the pre–post gain by baseline quartile. Students in the lowest Time 1 quartile gained 47.4 points on average, which is roughly three times the mean gain of 16.6 points in the highest quartile. One part of this gradient reflects genuine room to grow at the bottom of the distribution. Another part reflects regression toward the mean, because students with negative measurement error at Time 1 are overrepresented in the lowest quartile. The pattern should thus be interpreted descriptively, not as evidence that instruction served lower-proficiency students three times as well.

Figure 3. Pre–Post Gain by Baseline TOEFL ITP Quartile



Note. Bars represent mean score gains from Time 1 to Time 3 by baseline total-score quartile, with 95% confidence intervals. The observed gradient partly reflects regression toward the mean.

Correlations Among the Project Variables

Table 4 reports the correlations among the TOEFL ITP scores and the learner variables. Rows 5–9 show the bivariate side of RQs 1–3, that is, the associations of each construct with end-of-year performance and gain. The intercorrelations among the learner variables address RQ 4. The linked critical-thinking score was weakly related to Time 1 proficiency (r = .22), more clearly related to Time 3 proficiency (r = .45), and positively related to gain (r = .26).

However, these bivariate associations do not by themselves identify temporal or causal effects. Motivation and resilience were related more strongly to SRL and engagement than to the TOEFL outcomes. This is an early sign of the pattern that the adjusted model examined. Baseline proficiency correlated negatively with gain (r = −.38). This negative correlation is the usual result of bounded growth combined with measurement error in difference scores.

Table 4. Correlations Among the TOEFL ITP Scores and Learner Variables

Variable								
1. TOEFL ITP total, Time 1	—							
2. TOEFL ITP total, Time 2	.73	—						
3. TOEFL ITP total, Time 3	.69	.76	—					
4. Gain (Time 3 − Time 1)	−.38	.06	.41	—				
5. Motivation	.24	.17	.13	−.12	—			
6. Resilience	.03	−.02	−.01	−.04	.40	—		
7. Self-regulated learning	.18	.19	.21	.05	.40	.42	—	
8. Engagement	.15	.19	.11	−.04	.31	.17	.30	—
9. Critical thinking (linked T)	.22	.33	.45	.26	.12	.14	.21	.18

Note. Pairwise correlations; *n*s range from 164 to 238 for the coefficients shown. Gain = Time 3 total minus Time 1 total; linked T = linked Rasch T score. Dashes mark the diagonal.

Associations With End-of-Year Proficiency and Gain

Table 5 reports the baseline- and form-adjusted model for Time 3 total proficiency. With all focal learner variables and form group entered, Time 1 proficiency remained associated with Time 3 proficiency (β = .486, p = .002). The linked critical-thinking score had a positive unique association (β = .358, p < .001). The model explained 51.0% of the variance. No psychosocial composite or form indicator had a statistically significant unique coefficient.

Because we defined gain as Time 3 minus Time 1, the gain parameterization of the same model is not a second analysis. The unstandardized coefficients and tests for the non-baseline predictors are identical (on the gain metric, the standardized critical-thinking coefficient is β = .418). Only the baseline coefficient changes in sign and size (β = −.530, p = .004), and R^2 = .333. Accordingly, the answers to RQs 1–3 about end-of-year performance apply equally to gain.

The absence of statistically significant unique coefficients for the psychosocial composites should not be interpreted as evidence that these variables are irrelevant. Their correlations with SRL and engagement, and with each other, were substantial (Table 4). The model shows that, in this cohort, the psychosocial composites had no significant unique

relation to end-of-year TOEFL ITP performance after baseline proficiency, form group, and critical thinking were controlled.

Table 5. Baseline- and Form-Adjusted Model of End-of-Year TOEFL ITP Proficiency

Predictor	β	*p*
Time 1 total	.486	.002
Motivation	−.033	.653
Resilience	−.062	.391
Self-regulated learning	.065	.411
Engagement	−.056	.430
Critical thinking (linked)	.358	< .001
Form B	−.055	.614
Form C	.080	.655

Note. Standardized coefficients from the complete-case model (*n* = 136) predicting Time 3 total score. Form A is the reference group for the two form indicators. R^2 = .510. As gain = Time 3 minus Time 1, the gain parameterization yields identical unstandardized coefficients and tests for all predictors except the baseline (critical thinking: β = .418 on the gain metric, *p* < .001); its baseline coefficient is β = −.530 (*p* = .004) and its R^2 = .333.

Section-Level Models

Table 6 reports selected coefficients from the parallel baseline- and form-adjusted models for the three sections.

Table 6. Selected Coefficients from Section-Specific Baseline- and Form-Adjusted Models

Section	Predictor	β	*p*
Listening	Time 1 listening	.640	< .001
Listening	Critical thinking (linked)	.239	< .001
Grammar	Time 1 grammar	.136	.166
Grammar	Critical thinking (linked)	.352	< .001
Grammar	Form C	.348	.003
Reading	Time 1 reading	.347	.003
Reading	Critical thinking (linked)	.243	.002

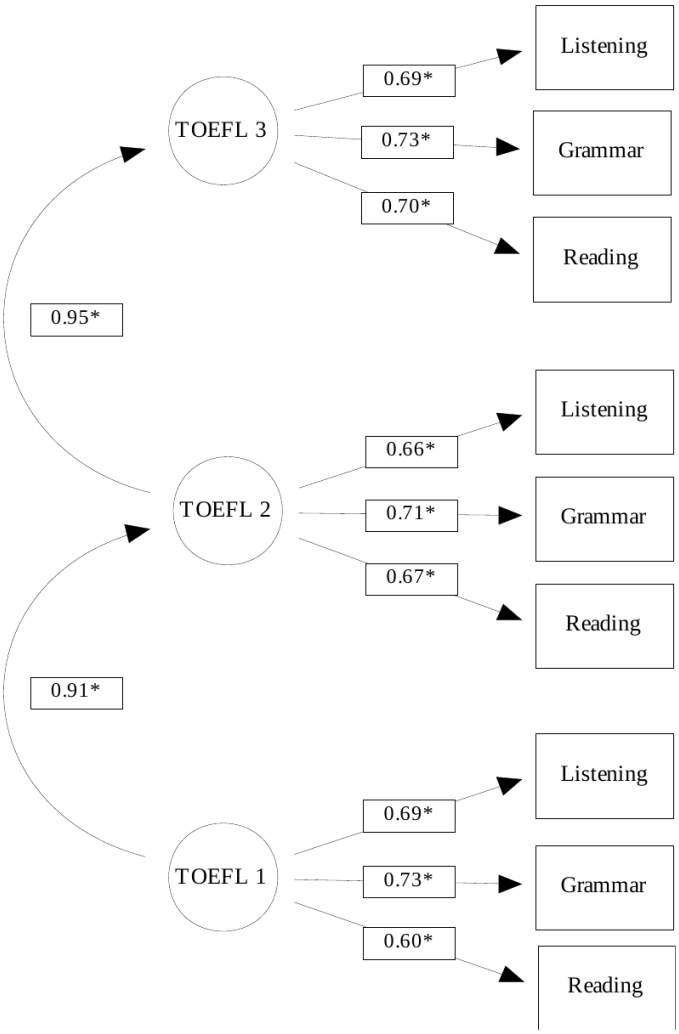
Note. Each model regressed the Time 3 section score on the corresponding Time 1 section score, motivation, resilience, SRL, engagement, the linked critical-thinking score, and Form B and C indicators (*n* = 136). The table shows each baseline and critical-thinking coefficient plus the statistically significant Form C coefficient in the structure-and-written-expression model. R^2 = .582 (listening), .340 (structure and written expression), and .302 (reading). Form A is the reference group; form, class placement, teacher, and prior proficiency were not independently separable.

The linked critical-thinking score was positively associated with listening (β = .239, p < .001), structure and written expression (β = .352, p < .001), and reading (β = .243, p = .002). Baseline section performance remained statistically significant for listening (β = .640, p < .001) and reading (β = .347, p = .003). Form C had a statistically significant residual coefficient in the model for structure and written expression (β = .348, p = .003). In that model, the Time 1 section coefficient was not significant (β = .136, p = .166). The models explained 58.2%, 34.0%, and 30.2% of the variance, respectively. Form group coincided with class placement and prior proficiency, so the Form C coefficient should not be interpreted as an isolated form effect.

Longitudinal Measurement Structure of TOEFL ITP Scores

Figure 4 shows the measurement-and-stability model for TOEFL ITP proficiency across the three time points.

Figure 4. Longitudinal Measurement-and-Stability Model for TOEFL ITP Proficiency



Note. Rounded standardized estimates. Complete-case model (n = 229). ML fit: $\chi^2(16) = 13.19$, $p = .659$, CFI = 1.000, RMSEA = .000, SRMR = .018; robust Satorra–Bentler $\chi^2(16) = 12.93$, $p = .678$, CFI = 1.000, RMSEA = .000. LC = listening comprehension; SWE = structure and written expression; RC = reading comprehension.

The model was estimated on the 229 students who had complete listening, structure and written expression, and reading scores at all three time points. Listening, grammar, and reading loaded on a single proficiency factor at each time point, with standardized loadings between .60 and .73. The latent stability paths were high (β = .914 from Time 1 to Time 2; β = .951 from Time 2 to Time 3). The model fit the data closely, $\chi^2(16) = 13.19$, $p = .659$, CFI = 1.000, RMSEA = .000, SRMR = .018, and the robust Satorra–Bentler $\chi^2(16) = 12.93$, $p = .678$. We therefore treated the three section scores as coherent indicators of one developing proficiency dimension across the year. The model contains only the nine TOEFL ITP section scores. It involves no learner variables and does not support claims about why the scores changed.

Student Perceptions of TOEFL ITP

Students were asked whether they thought that TOEFL ITP reflected their overall English proficiency. Of the 228 students who answered the closed-ended item, a majority (123) selected "more or less," and outright rejection was a minority position (Table 7). The middle category therefore dominated the distribution. Most students declined full endorsement but accepted the test as at least somewhat representative of their English development.

Table 7. Follow-Up Survey Responses on Whether TOEFL ITP Reflects Students' English Proficiency

Response category	Count	Proportion
Yes	39	17.1%
More or less	123	53.9%
Not really	66	28.9%

Note. n = 228 respondents to the closed-ended item ("Do you think the TOEFL ITP reflects your overall English proficiency?"), administered after the third TOEFL ITP administration.

Fifty-one respondents wrote an open-ended comment. The comments converged on three recurring themes (Table 8). The most frequent concern was construct coverage. Thirty-one students observed that the regular TOEFL ITP does not directly assess speaking and writing, skills that were emphasized in their classes. A second theme concerned test difficulty and length. A third theme involved topical or cultural distance: Some students found items that were tied to American university life remote from their own experience. A small number of comments addressed the conditions of the administration instead of the test itself, such as room temperature or noise in the examination room. These comments were not among the three themes. The pattern suggests conditional acceptance. Students' comments reflected awareness of the test construct boundaries, particularly the distinction between receptive and productive skills.

Table 8. Themes in the Open-Ended Follow-Up Comments

Theme	Summary	Comments (*n*)
Construct coverage	Students observed that the regular TOEFL ITP does not directly assess speaking and writing, despite these skills being emphasized in class.	31
Difficulty and length	Students described the test as difficult and long.	18

Topical or cultural distance	Some students found items tied to American university life or other culturally specific contexts remote from their experience.	11

Note. Themes were drawn from the optional open-ended responses accompanying the closed-ended item (51 respondents). A response could contribute to more than one theme. Comments concerning the conditions of the test administration rather than the test are not represented in the table.

Discussion

The project yielded a consistent pattern of findings across analyses. TOEFL ITP scores increased across the academic year by an amount that is considered large in within-student terms. The linked critical-thinking score was associated with higher end-of-year TOEFL ITP scores, and equivalently with larger gains, after controlling for baseline proficiency and form group. The psychosocial variables were related more strongly to one another, particularly to SRL, than to score change. Students generally regarded the test as a partial, not a complete, measure of their English development. Because most learner measures were completed at or after the final TOEFL ITP administration, these relationships should be interpreted as concurrent associations. The following sections interpret the findings within this constraint.

TOEFL ITP Score Growth Across an Academic Year

The mean gain of 27.9 points, with statistically significant increases in each section and in both semesters, indicates that TOEFL ITP scores were responsive to change across a full academic year of intensive study. What produced the change is less certain. The design had no comparison group, so the observed gain cannot be partitioned among instruction, out-of-class English exposure, maturation, repeated testing, and growing familiarity with the test format. All of these are plausible contributors. One interpretation of the front-loaded trajectory is that students' early adjustment to a demanding new program produced the largest gains, followed by slower consolidation. However, a retest-familiarity account also predicts front-loading, because the effects of increased familiarity are likely to be strongest in the first retest interval. The second semester nevertheless contributed an additional gain of approximately nine points. This persistence argues against familiarization as a full account, although it does not establish instruction as the cause. These results are consistent with the progress-monitoring use that ETS (2025) describes for the series and with prior ETS work that tracked skill development with TOEFL ITP (Choi & Papageorgiou, 2014). The present study extends that work by using measurement at three time points across a full academic year within a single instructed cohort.

Critical Thinking as the Strongest Non-Baseline Correlate (RQ 2)

The linked critical-thinking measure was the only focal learner variable that showed a consistent unique association with end-of-year proficiency across the main and section regressions. It correlated .22 with Time 1 TOEFL ITP, .45 with Time 3, and .26 with gain. Its baseline- and form-adjusted coefficient was β = .358 for Time 3 performance, which is equivalent to .418 on the gain metric. Three qualifications apply to that coefficient. First, the Time 3 and gain figures are one analysis in two parameterizations, that is, one result and not converging replications. Second, the test was administered in December, shortly before the final time point. It was therefore not a baseline measure, and performance may instead reflect learning during the year. Third, the critical-thinking test and TOEFL ITP are both English-medium, selected-response assessments, and the validation study of the instrument reported *r* = .64 with an English proficiency measure (Hirai et al., 2022). The observed association therefore cannot be interpreted as the contribution of language-free critical thinking. Even so, the association is consistent with a content-based interpretation. TOEFL ITP items may require extracting relevant information, weighing alternatives, and making warranted inferences, which are operations represented by the critical-thinking instrument (Hirai et al., 2020, 2022; Kubo, 2022). However, response-process evidence would be needed to support that interpretation directly. For scale, the adjusted coefficient is at or above the meta-analytic association between critical thinking and achievement of approximately *r* = .28 (Fong et al., 2017), which is medium-sized by Plonsky and Oswald's (2014) benchmarks. The shared language demands may well have inflated the coefficient in the present study.

Critical thinking was associated with all three sections, and the largest coefficient was in structure and written expression (β = .352). The association was therefore not confined to reading. However, the design cannot separate reasoning operations from English-language access or from the shared selected-response method. A residual Form C coefficient also appeared in the model for structure and written expression. Form, class placement, teacher, and prior proficiency were confounded, so this coefficient cannot be interpreted as a form effect. Form-specific KR-20 coefficients ranged from .63 to .71, and linked person reliability was .67, so attenuation and person-score uncertainty remain material. The result extends Park's (2022) cross-sectional associations between critical thinking and grammar and reading abilities to a score record spanning three time points.

For ETS and institutions, the findings identify critical-thinking instruction as a hypothesis for prospective testing, not as an established means of raising TOEFL ITP scores. Intervention and response-process evidence would be needed to determine whether instruction in inference, principled elimination, and consistency checking improves test performance for the proposed reasons.

Motivation, Self-Regulation, Engagement, and Resilience (RQs 1, 3, and 4)

The psychosocial variables largely behaved in line with the recent literature. They were strongly related to each other and to the organization of study, and they were not uniquely associated with end-of-year scores once baseline proficiency was controlled. Motivation and resilience correlated with SRL at .40 and .42, and engagement correlated with SRL at .30. However, none of the four composites had a statistically significant unique coefficient for Time 3 proficiency after baseline, form group, and critical thinking were controlled (RQs 1 and 3). The pattern matches that of Ueno and Park (2025), in which motivational and self-regulatory variables were related to each other more than to TOEFL scores. It is also consistent with accounts in which motivational resources affect outcomes through regulated study behavior (Derakhshan & Fathi, 2024; Tseng et al., 2006). The present data cannot test such an account. The January questionnaires followed the outcome that they would need to precede, and concurrent associations cannot separate influence on learning from reflection on a year of visible progress (RQ 4). The magnitudes are also in line with meta-analytic expectations for objective outcomes. Motivational variables correlate with measured achievement at roughly .20 to .37 (Al-Hoorie, 2018; Masgoret & Gardner, 2003), engagement at about .27 (Lei et al., 2018), and self-regulatory processes at about .11 to .20 in school-age samples (Dent & Koenka, 2016). These magnitudes are small enough that unique coefficients can vanish once the model contains a strong baseline.

Resilience followed the same broad pattern. It correlated with SRL about as strongly as motivation did, but it had no statistically significant unique association with end-of-year TOEFL ITP performance after the other variables were controlled (cf. Park et al., 2025; Wang & Patterson, 2025). A one-year window on a three-section proficiency test may not be the most suitable setting for detecting the effects of resilience.

Measurement Timing and the Interpretation of Learner-Variable Associations

One design feature constrains the interpretation of each learner-variable finding and deserves a separate discussion. The motivation, SRL, engagement, and perception measures were completed immediately after the third TOEFL ITP administration. Resilience was measured at mid-year. The critical-thinking test preceded the final time point by roughly one month. Students thus answered the January questionnaires knowing how the year had gone and, in many cases, how their scores had changed. Their responses may therefore partly reflect test experiences and results, not dispositions that operated during the year. Associations that are estimated under these conditions are concurrent or retrospective. They can identify variables that are worth measuring prospectively, and they can rule out strong concurrent relations. However, they cannot establish temporal precedence, and nothing in this report shows that any learner characteristic produced score gains. This caution applies most forcefully to the critical-thinking association. That association is the most prominent learner-variable finding in the report, and it is also the finding most exposed to shared English-language demands.

Student Perceptions and Qualified Acceptance (RQ 5)

The perception distribution was neither endorsement nor rejection. A majority chose the middle option, and the open-ended comments identified the limits of the test where its construct boundary lies: Speaking and writing are taught in the program but are not part of the regular TOEFL ITP score. In validity terms, this is qualified acceptance under perceived construct underrepresentation. It provides stakeholder evidence about how the alignment between the test and the curriculum is experienced (Bachman & Palmer, 2010). It is not a verdict on the test, and it cannot by itself show that score interpretations or uses are sound.

The pattern is consistent with the existing stakeholder record, namely institutional users who value TOEFL ITP while knowing its coverage (Golubovich et al., 2018) and teachers who reserve judgment on communicative competence (Collins & Miller, 2018). The student evidence adds that learners themselves recognize the boundary between receptive and productive skills. The practical implication is communicative, not psychometric. When a program uses TOEFL ITP as a progress benchmark, clearly explaining to students what the score does and does not represent costs little. On the present evidence, such an explanation would reinforce an accurate understanding that many students already appear to hold.

Implications for Institutional Score Use

Two implications relate to the score itself. First, scores rose steadily during a year of intensive instruction. This rise is consistent with the progress-monitoring use of the test in comparable programs, although the design cannot establish how much of the increase was attributable to instruction. Second, score reports may be better understood and trusted when institutions communicate the construct boundary explicitly to students.

Two further implications concern learners. Critical thinking is a promising variable for TOEFL-related learner research. Whether critical-thinking instruction can improve score development was not tested in the present project and would require intervention evidence. The lack of unique coefficients for the psychosocial variables in a baseline-adjusted model does not indicate that they play no part in learning. In this cohort, they were associated with the organization of study (SRL in particular), not with score change over a relatively short period. Prospective measurement would be needed to test whether they support learning indirectly.

Limitations

The results should be interpreted with reference to the limitations of the study. The design followed a single cohort at a single institution without a comparison group, so the observed score gains cannot be attributed to instruction. Out-of-class English exposure, maturation, repeated testing, and growing familiarity with the format are confounded with teaching effects, even if the persistence of second-semester growth makes familiarization alone an unlikely full account. The size and setting of the cohort also limit generalization to other program types and proficiency ranges.

Measurement timing is the central limitation for the learner-variable findings. Critical thinking was measured in December, before the final time point but not at baseline. Motivation, SRL, and engagement were measured after the third administration, so their coefficients describe concurrent association, not temporal prediction. Resilience was measured at mid-year. Reverse influence, in which a year of visible progress and three score reports shape end-of-year self-reports, cannot be excluded for the January questionnaires. The questionnaire composites were formed without factor-analytic verification of their structure, and engagement was measured narrowly as peer support. Null results for the composites therefore do not exclude associations for other engagement dimensions or for

particular motivation types. The multivariable models were based on 136 complete cases, and students with missing questionnaire components may differ from those with complete data.

The linked critical-thinking score places the retained responses on a conditional cross-form scale, but it does not establish operational equating. Item retention followed a post hoc screening for negative item–total correlations conducted in the same student samples used for the analyses, and the procedure was not cross-validated with independent data. The 10 linking families represented only four of seven blueprint categories (Analysis, Assumptions, Arguments, and Inferences). Form-level reliability was modest, and residual diagnostics indicated localized multidimensionality or dependence. Common-item invariance was not tested directly, and form group overlapped strongly with prior proficiency and class placement. The critical-thinking test was administered in English, so its association with TOEFL ITP may partly reflect English proficiency, processing ability, or selected-response test-taking skill rather than critical thinking as a distinct construct. The validation study of the instrument reported $r = .64$ with an English proficiency measure (Hirai et al., 2022). Regression standard errors treated the linked person measure as observed and did not propagate person-score or linking uncertainty (Kolen & Brennan, 2014). Gain scores and the quartile display are subject to the usual caveats about difference scores, including regression toward the mean. Finally, the perception measure consisted of one closed item with optional comments, and one author coded the comments without an independent second coder. The measure could therefore identify the direction and grounds of students' judgments, but not their depth.

Directions for Future Research

The associations reported here provide a basis for testing the hypothesis identified above, that critical-thinking instruction can raise TOEFL ITP scores, but they do not predict its outcome. A prospective instructional study should therefore test whether explicit instruction on inference, analysis, and consistency checking improves TOEFL ITP development relative to conventional preparation. A larger, independently assigned sample should also be used to replicate the linked calibration, test common-item invariance and differential item functioning, cross-validate the post hoc item screen, and propagate person-score and linking uncertainty into downstream estimates.

Two extensions would address the limitations of this study directly. First, if the psychosocial variables were measured at program entry, and again at mid-year, the concurrent associations would become temporal ones, and mediation through SRL could be tested properly. Larger samples would also permit analysis at the subcomponent level, such as distinguishing received from provided peer support, which the present sample could not support without overfitting. Second, if TOEFL ITP were paired with locally scored speaking and writing measures, the results would show how receptive and structure-related growth relates to productive development, which is where student reservations concentrated. The redesigned TOEFL iBT, which was launched in January 2026, offers a standardized route to the same comparison. A follow-up cohort could take the iBT at the start and end of the program year, or both tests could be administered to the same students. Either design would place speaking and writing on reported score scales instead of local ones. It would allow four-skill development to be compared directly with the receptive and structure-related development that is documented here, and it would show whether the linked critical-thinking measure has comparable prospective associations with the development of productive skills.

Beyond this cohort, multi-site replication would test whether the pattern that was observed here holds across institutions, curricular models, and cultures of score use. That pattern consists of strong growth, an association between the linked critical-thinking measure and later TOEFL ITP performance, and qualified student acceptance. Longer observation periods would show whether the front-loaded trajectory recurs across academic years or study-abroad transitions. Student-perception research that preserves open-ended comments alongside closed responses also deserves a larger place in arguments about TOEFL ITP score use, given the evidence here that students reason about construct coverage in recognizably technical terms.

Conclusion

Students in an intensive Japanese university English program gained substantially on TOEFL ITP scores across two semesters. The test registered this development consistently across sections and time points. Critical thinking was represented by a conditional linked Rasch measure that was administered shortly before the final time point. It was the strongest unique non-baseline correlate of end-of-year performance, and equivalently of

gain, after form group was considered. The association was concurrent and partly confounded with English proficiency. Motivation, resilience, SRL, and engagement were more closely associated with one another than with score change. Students, for their part, accepted the test as a partial mirror of their development: They found it useful and demanding but recognized that it did not cover the productive skills taught in their program.

For ETS, the project offers three usable results. First, it provides longitudinal evidence that TOEFL ITP registered score growth across a year of intensive instruction in a Japanese university program. Second, the linked critical-thinking measure showed a baseline- and form-adjusted association with end-of-year performance that warrants prospective replication. Third, the student-perception evidence supports pairing institutional score use with clear statements of what the score represents.

Acknowledgments

This project was funded by the TOEFL English-Language Researcher/Practitioner Grant Program, administered by ETS. We thank the participating students of the Faculty of International Liberal Arts and the colleagues who supported test and questionnaire administration. The views expressed are those of the authors and do not necessarily reflect the views of ETS.

References

Al-Hoorie, A. H. (2018). The L2 motivational self system: A meta-analysis. *Studies in Second Language Learning and Teaching, 8*(4), 721–754. https://doi.org/10.14746/ssllt.2018.8.4.2

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Choi, I., & Papageorgiou, S. (2014). *Monitoring students' progress in English language skills using the TOEFL ITP® assessment series* (Research Memorandum No. RM–14-11). ETS.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Collins, J. B., & Miller, N. H. (2018). The TOEFL® (ITP): A survey of teacher perceptions. *Shiken: JALT Testing & Evaluation SIG Newsletter, 22*(2), 1–13. https://tevalwp.jalt.org/wp-content/uploads/2025/12/22_02_1_Collins_Miller_TOEFL-1.pdf

Connor, K. M., & Davidson, J. R. T. (2003). Development of a new resilience scale: The Connor–Davidson Resilience Scale (CD-RISC). *Depression and Anxiety, 18*(2), 76–82. https://doi.org/10.1002/da.10113

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, 28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

Derakhshan, A., & Fathi, J. (2024). Growth mindset, self-efficacy, and self-regulation: A symphony of success in L2 speaking. *System, 123*, Article 103320. https://doi.org/10.1016/j.system.2024.103320

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* Lawrence Erlbaum Associates.

ETS. (2025). *TOEFL ITP® test taker handbook.* https://www.ets.org/pdfs/toefl-itp-test-taker-handbook.pdf

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. California Academic Press.

Fong, C. J., Kim, Y., Davis, C. W., Hoang, T., & Kim, Y. W. (2017). A meta-analysis on critical thinking and community college student achievement. *Thinking Skills and Creativity, 26*, 71–83. https://doi.org/10.1016/j.tsc.2017.06.002

Fukuda, T. (2020). *School belonging and L2 motivation of first-year students at four Japanese universities* (Publication No. 28027020) [Doctoral dissertation, Temple University]. ProQuest Dissertations and Theses Global.

Fukuda, T., Fukada, Y., Falout, J., & Murphey, T. (2021). How ideal classmates priming increases EFL classroom prosocial engagement. In P. Hiver, A. H. Al-Hoorie, & S. Mercer (Eds.), *Student engagement in the language classroom* (pp. 182–201). Multilingual Matters.

Golubovich, J., Tolentino, F., & Papageorgiou, S. (2018). *Examining the applications and opinions of the TOEFL ITP® assessment series test scores in three countries* (TOEFL Research Report No. RR-84). ETS. https://doi.org/10.1002/ets2.12231

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate data analysis* (8th ed.). Cengage Learning. https://eli.johogo.com/Class/CCU/SEM/_Multivariate%20Data%20Analysis_Hair.pdf

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press.

Harsch, C., Ushioda, E., & Ladroue, C. (2017). *Investigating the predictive validity of TOEFL iBT® test scores and their use in informing policy in a United Kingdom university setting* (TOEFL iBT-30). ETS. https://doi.org/10.1002/ets2.12167

Hirai, A., Maeda, H., Oka, H., Kato, T., & Nakano, E. (2020). The development of an English test for measuring critical thinking skills: A pilot study. *JACET Journal, 64*, 205–225. https://doi.org/10.32234/jacetjournal.64.0_205

Hirai, A., Oka, H., Kato, T., & Maeda, H. (2022). Development and validation of an English test measuring EFL learners' critical thinking skills. *Language Testing in Asia, 12*(1), Article 45. https://doi.org/10.1186/s40468-022-00193-2

Ito, M., Nakajima, S., Shirai, A., & Kim, Y. (2009, November). *Cross-cultural validity of the Connor-Davidson Scale: Data from Japanese population* [Poster presentation]. 25th

Annual Meeting of the International Society for Traumatic Stress Studies, Atlanta, GA, United States.

Jackson, D. O., & Park, S. (2020). Self-regulation and personality among L2 writers: Integrating trait, state, and learner perspectives. *Journal of Second Language Writing, 49*, Article 100731. https://doi.org/10.1016/j.jslw.2020.100731

Karjo, C. H., & Ronaldo, D. (2019). The validity of TOEFL® as entry and exit college requirements: Students' perception. In *Proceedings of the Eleventh Conference on Applied Linguistics (CONAPLIN 2018)* (pp. 326–330). Atlantis Press. https://doi.org/10.2991/conaplin-18.2019.277

Kokhan, K., & Lin, C.-K. (2014). Test of English as a Foreign Language (TOEFL®): Interpretation of multiple score reports for ESL placement. *Papers in Language Testing and Assessment, 3*(1), 1–23. https://doi.org/10.58379/ZJDV7047

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

Kubo, Y. (2022). The development of a pilot test to measure the critical thinking skills of Japanese EFL learners: Validation by exploratory factor analysis. *LET Kanto Journal, 7*, 37–54. https://doi.org/10.24781/letkj.7.0_37

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: An International Journal, 46*(3), 517–528. https://doi.org/10.2224/sbp.7054

Linacre, J. M. (2008). *Winsteps Rasch measurement computer program* (Version 3.65.0) [Computer software]. Winsteps.com.

Masgoret, A.-M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning, 53*(1), 123–163. https://doi.org/10.1111/1467-9922.00212

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2003). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning, 53*(S1), 33–64. https://doi.org/10.1111/1467-9922.53223

Park, S. (2022). Exploring the relationship between L2 learners' critical thinking and self-regulating capacities and English grammar and reading abilities. *Studies in Linguistics and Language Teaching, 32*, 163–175.

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., Ueno, S., & Sugita, M. (2025). The impact of resilience and self-regulation on L2 proficiency among ESL learners. *International Journal of TESOL Studies, 7*(1), 105–125. https://doi.org/10.58304/ijts.20250107

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

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tseng, W.-T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics, 27*(1), 78–102. https://doi.org/10.1093/applin/ami046

Ueno, S., & Park, S. (2025). Exploring the relationship between self-regulation, grit, motivation, and English proficiency among university students in Japan. *Studies in Self-Access Learning Journal, 16*(2), 418–436. https://doi.org/10.37237/160208

Wang, H., & Patterson, M. M. (2025). Academic resilience in learning English as a foreign language: Relations to classroom climate and beliefs about language learning. *Studies in Educational Evaluation, 87*, Article 101517. https://doi.org/10.1016/j.stueduc.2025.101517

Wang, Y., & Huang, B. H. (2020). Washback of TOEFL® preparation courses on students' attitudes and score improvement. *International Journal of Linguistics, 12*(3), 84–104. https://doi.org/10.5296/ijl.v12i3.16940

Yang, T., Wang, Y., & Yang, C. (2025). Unravelling the effectiveness of self-regulated language learning intervention on Chinese EFL students' motivation, strategic competence and English proficiency: A mixed methods study. *British Educational Research Journal, 51*(1), 4–24. https://doi.org/10.1002/berj.4061

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Appendix. Sample Instrument Content

English Critical Thinking Test, Form A. Test items are not reproduced here, in accordance with the developers' restrictions.

Resilience Measure (CD-RISC-25). Scale items are not reproduced here, in accordance with the developers' restrictions.

L2 Motivation Questionnaire (Sample Items)

- Even if English were not a compulsory subject, I would choose to study it.
- I regularly use English with my classmates in class.
- Even if the teacher is not near me or cannot hear me, I still speak English with my classmates in class.
- I make an effort to speak English outside of class.
- I try to use English in daily life.
- I try to learn English from other sources, such as the Internet.
- I am prepared to put a great deal of effort into learning English.
- I enjoy learning English in class.

Survey on English Learning Self-Regulation (Sample Page). The survey was administered bilingually; the first page is reproduced in the following figure.



Perceptions of TOEFL ITP (Sample Items)

- Section A (closed-ended): Do you think the TOEFL ITP reflects your overall English proficiency?
- Section B (optional, open-ended): Why do you think the TOEFL ITP does or does not reflect your English proficiency well?

PPLS Engagement Questionnaire (All Four Items)

- My classmates are helping me to learn English.
- I am helping my classmates to learn English.
- My classmates are helping me to enjoy English.
- I am helping my classmates to enjoy English.

Suggested citation: Park, S., & Fukuda, T. (in press). *Investigating factors associated with gains in TOEFL ITP® scores at a university in Japan* (Research Memorandum). ETS.

Copyright © 2026 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, TOEFL, TOEFL IBT, and TOEFL ITP are registered trademarks of Educational Testing Service (ETS). All other trademarks are the property of their respective owners.