Studies in Second Language Acquisition (2026), 1–33
doi:10.1017/S0272263126102034

RESEARCH ARTICLE

Task-based or task-supported language education?

A longitudinal comparative effectiveness study of two pedagogical approaches

Koen Van Gorp[1], Matthew D. Coss[2], Xinyi Guan[1], John Dylan Burton[3] and Ayşen Tuzcu[1]

[1]Linguistics, Languages, and Cultures, Michigan State University, East Lansing, MI, USA; [2]Modern Languages and Classics, The University of Alabama, Tuscaloosa, AL, USA and [3]Applied Linguistics and ESL, Georgia State University, Atlanta, GA, USA
Corresponding author: Koen Van Gorp; Email: vangorpk@msu.edu

(Received 14 April 2026 UTC; Revised 22 August 2026 UTC; Accepted 31 August 2026 UTC)

Abstract

Task-Based Language Teaching (TBLT) has gained prominence in additional language education for its focus on fostering communicative competence through meaningful, real-world tasks. However, there are few longitudinal studies comparing the outcomes of TBLT with those of form-focused instruction (Boers & Faez, 2026). It is not yet clear to what extent TBLT supports the development of explicit linguistic knowledge alongside task-specific communicative proficiency. This issue is especially relevant in educational contexts where students' progress is evaluated through high-stakes (explicit) language knowledge assessments. This quasi-experimental study reports on a yearlong comparative evaluation of a TBLT Business English course in a Chinese college and examines its effectiveness compared to a task-supported course. Student ($N = 80$) learning outcomes were assessed through both task-based and discrete lexicogrammatical measures. Findings show that TBLT was at least equally effective in building students' lexical and grammatical knowledge, and more effective in developing task-specific speaking proficiency.

Keywords: longitudinal study; second language learning; task-based language assessment; task-based language teaching

Introduction

Task-Based Language Teaching (TBLT) has emerged as a prominent approach in the field of additional language (L2) education for its potential to enhance learners' communicative competence through meaningful language use. In TBLT, learners complete authentic, real-world tasks which encourage them to negotiate for meaning and develop communication skills in context. While TBLT has been shown to have

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

advantages over other approaches (e.g., González-Lloret & Nielson, 2015), the widespread adoption of TBLT has encountered challenges in large classes and exam-oriented education systems (e.g., East, 2021; Newton, 2025). Especially in contexts driven by high-stakes language assessments, stakeholders may doubt whether TBLT-based curricula can develop students' language skills effectively for them to earn top marks (Butler, 2011).

A strong task-based approach (Long, 2015) is distinct from structure-oriented L2 teaching approaches. TBLT prioritizes meaning-focused, goal-oriented communication rather than explicit form-focused instruction and decontextualized practice (East, 2021). Despite decades of research on TBLT design, the comparative effectiveness of strong versions of TBLT with structural approaches, including task-supported approaches, for proficiency development remains under-researched, especially longitudinally (Boers, Bryfonski, Faez, & McKay, 2021; Boers & Faez, 2026). Thus, the aim of this study was to compare language learning outcomes between a TBLT- and a language-focused, task-supported syllabus in a two-semester English for Business Purposes course. To capture a comprehensive picture of L2 development, outcomes were measured using two language assessments: a task-based language assessment (TBLA) to evaluate communicative performance and a lexicogrammatical test to evaluate explicit linguistic knowledge. This longitudinal comparative study contributes to calls for sorely needed research on TBLT's effectiveness by examining these variables together, establishing whether TBLT can satisfy the demands of both real-world task performance and high-stakes language testing environments.

Background

Task-based language teaching

TBLT is an approach to language pedagogy which emphasizes language development via the completion of real-world tasks, fostering a learning environment where language is used receptively and productively to achieve specific goals aimed at building students' language proficiency to function in relevant real-world contexts (Van den Branden, 2022). TBLT has been promoted as a strong form of communicative language teaching with a rich research tradition based on cognitive-interactionist, psycholinguistic, and sociocultural perspectives (Ellis, Skehan, Li, Shintani, & Lambert, 2020). Relatedly, TBLT has been argued to be the pedagogical approach most grounded in instructed second language acquisition (ISLA) theory and research to date (Loewen & Sato, 2021; Long, 2015).

Strong-form TBLT uses tasks as the unit of analysis in syllabus (i.e., curriculum) design (Long & Crookes, 1993). Meaningful and relevant real-world tasks are identified in a needs analysis and used to design a language syllabus and assessment plan (Long, 2015). In this approach, language learning occurs as learners engage in tasks that reflect authentic language use, with form-focused instruction provided as needed. However, there is "no *a priori* explicit teaching of language" (Ellis, 2024, p. 11). In contrast, weaker forms of TBLT use communicative tasks to supplement a form-focused curriculum (East, 2021). Rather than driving learning, tasks may be used primarily as pedagogical tools for the practice of pre-taught forms (Samuda & Bygate, 2008), often replicating a Presentation-Practice-Production (PPP) model, with tasks contributing to the practice (Ellis, 2024) and/or production phases (East, 2021). "Weaker" forms of TBLT are often referred to as task-supported language teaching (TSLT) and constitute a fundamentally different approach from TBLT (East, 2021; Ellis, 2024), as they are grounded in Skill Acquisition Theory (e.g., Dekeyser, 2017). While both approaches incorporate tasks,

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

strong TBLT places tasks at the core of the curriculum, whereas weaker forms of TBLT (i.e., TSLT) maintain a primary focus on the teaching and learning of grammatical structures and vocabulary, making "TSLT effectively no different from weak CLT [Communicative Language Teaching]" (East, 2021, p. 75). Ellis (2019, p. 456) described TSLT as "the mainstream methodology in structural language teaching today."

Essential and still unanswered questions for TBLT are 1) the extent to which TBLT facilitates second or additional language (L2) development and 2) the extent to which TBLT outperforms non-TBLT approaches (Norris & Davis, 2021). In a meta-analysis of 52 studies, Bryfonski and McKay (2019) concluded that TBLT programs had a strong positive effect on L2 learning outcomes. However, the authors noted the inconsistent use of instruments for these outcomes and a lack of reporting of how outcomes were tracked. In addition, few programs collected information beyond one semester (15–16 weeks), making it difficult to evaluate the longitudinal impact of TBLT. In a critical reanalysis of this dataset, however, Boers et al. (2021) and Boers and Faez (2026) reported that only one of the 52 studies used strong-form TBLT. Furthermore, a prevalent lack of methodological detail (e.g., amount of instructional time, learning opportunities outside the classroom, etc.) casts doubt on the validity of the claims supporting TBLT, leading the authors to conclude that "much more (and more solid and replicable) empirical work on the comparative effectiveness of TBLT needs to be done" (Boers et al., 2021, p. 15).

There are few longitudinal classroom-based studies comparing a strong TBLT approach with other approaches, and those that exist show mixed results. Ganjouee, Ghonsooly, and Fatemi (2018) and Pomarolli, Artoni, Boschiero, and Piccinin (2023), for example, demonstrated that TBLT can lead to stronger outcomes than alternative approaches, but whereas Ganjouee et al. (2018) found significantly higher speaking test scores in their TBLT condition than their PPP condition, in the Pomarolli et al. (2023) study, the TBLT group outperformed the control PPP group in writing, but not on listening, reading, or speaking. Others found that TBLT and PPP (Harris & Leeming, 2022) or TSLT (Qi, 2023) approaches were equally effective in building L2 speaking proficiency, although Harris and Leeming (2022) found slightly different developmental trajectories for each group. Whereas these studies adopted pre- and post-test designs, an important limitation is that the studies did not compare the development of both task-based performance ability *and* linguistic knowledge.

To our knowledge, González-Lloret and Nielson (2015) is the only study with a strong-form TBLT operationalization that evaluated students' task-based L2 performance and linguistic knowledge and compared a TBLT to a non-TBLT approach. Seven target tasks, selected on the basis of a needs analysis, were implemented in an eight-week strong-form TBLT program that included task-based and non-task-based assessments. The students in the TBLT course significantly outperformed the students in the grammar-based comparison course in terms of oral fluency and syntactic complexity on the task-based assessments, with no differences in lexical complexity or grammatical accuracy between groups. On a standardized language proficiency test, the TBLT students improved over the course of the program on discrete components of sentence mastery and vocabulary, with smaller gains in fluency and pronunciation. Comparison group proficiency data was not available due to contextual constraints.

The capacity for TBLT to enhance discrete linguistic knowledge alongside communicative performance, as observed by González-Lloret and Nielson (2015), is supported by a broader body of empirical research. Regarding grammatical development, Cobb's (2010) meta-analysis demonstrated that engaging adult learners in focused oral communication tasks yields durable, medium-to-large effect sizes for the acquisition of target grammatical structures, indicating that meaning-focused interaction effectively

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

drives morphosyntactic development. Similarly, Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka's (2006) meta-analysis found that task-based interaction in focused tasks (i.e., with predetermined grammatical and lexical features) produced large gains in both grammatical and lexical acquisition, with effects persisting over delayed posttests and being particularly pronounced when tasks elicited pushed output. Together, these findings align with González-Lloret and Nielson's (2015) results, supporting the premise that TBLT effectively supports the acquisition of discrete vocabulary and grammar knowledge typically targeted by TSLT without an overt pre-planned language syllabus (see also Long, 2016). However, while meta-analyses by Cobb (2010) and Keck et al. (2006) demonstrate that task-based interaction effectively fosters grammar and vocabulary acquisition, these reviews heavily emphasize dyadic task interaction rather than monologic oral performance. Furthermore, echoing the methodological limitations of González-Lloret and Nielson (2015), among others, there remains a notable lack of longitudinal studies comparing the outcomes of TBLT with those of form-focused instruction. Therefore, a clear gap remains for longitudinal research examining the effects of TBLT on both explicit linguistic knowledge and monologic task-based speaking performance among adult English for Specific Purposes (ESP) learners.

Task-based language assessment

Task-based language assessment (TBLA) is a form of evaluation that simulates a target task performance in a real-world (or simulated) language use context and assesses that performance "according to highly specific task-criteria like those used in non-assessment settings" (Norris & East, 2021, p. 512). TBLA generally prioritizes learners' ability to successfully complete a task; in other words, task completion is often the primary or only criterion for success or failure (Long, 2015; Norris & East, 2021). TBLA thus aims to make meaningful interpretations about what learners can *do* in a target language use domain (Bachman & Palmer, 2010), contrasting with general language proficiency or achievement tests which aim to measure L2 ability more broadly. Although TBLA may also include evaluations of L2 ability (e.g., fluency, vocabulary, grammar) alongside task-based criteria, Long (2015) questioned this practice, arguing that the "introduction of a linguistic caboose could eventually lead to a reorientation of a task-based course, as a result of negative washback, to one which devotes progressively larger segments of class time to work on language as an object" (p. 333). He instead argued for a holistic assessment of a learner's language performance, focused specifically on task completion, rather than evaluating a learner's language performance at the level of specific linguistic features.

A challenge that TBLT evaluations face is thus selecting the most appropriate type of assessment and outcome measures to assess their effectiveness (Vandommele, Van Gorp, Van den Branden, & De Maeyer, 2022). While many TBLT studies logically opt for TBLA to document gains, TBLT programs may also face pressure from stakeholders to use standardized (often discrete and/or performance-based) language proficiency tests (e.g., González-Lloret & Nielson, 2015). The extent to which TBLA and L2 proficiency tests, especially those proficiency tests that include a language use component assessing grammatical and lexical knowledge through discrete-point formats, show convergent or divergent evidence of L2 development is unclear, much like the enduring question of the L2 developmental affordances of TBLT versus other pedagogical approaches. Indeed, Bachman (2002) argued that the extent to which a TBLA or a standardized proficiency test alone can fully capture L2 development is doubtful given the complexity of the underlying construct. To date, few studies have compared L2

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

development outcomes across different assessment types, and none have done so using an experimental design. These questions are relevant both at the theoretical level, to examine TBLT's efficacy, and at the practical level, especially if TBLT is to show evidence of viability in contexts where L2 abilities are tested using discrete linguistic items in high-stakes tests.

The present study

To date, the effectiveness of TBLT programs, especially of strong-form TBLT programs (Long, 2015), has not been established in a robust way, with a notable lack of longitudinal studies comparing the outcomes of TBLT with those of form-focused instruction (Boers et al., 2021; Boers & Faez, 2026). Furthermore, while studies have evaluated TBLT using either TBLA or general language ability tests, there are significant gaps in evaluating both domains simultaneously. It is not yet clear to what extent TBLT supports the development of explicit linguistic knowledge (i.e., grammatical and lexical knowledge measured through discrete-point assessments) alongside task-specific communicative proficiency. To address these gaps—specifically, the need for longitudinal comparisons of TBLT and TSLT outcomes, and the need to examine these outcomes using both TBLA and linguistic knowledge assessments together—we investigated the following two research questions:

RQ1: To what extent does a strong TBLT course improve students' performance on a task-based speaking assessment and a linguistic knowledge test compared to a task-supported language course?

RQ2: How do task completion scores compare to other linguistic criteria (e.g., organization, comprehensibility, language use) on a monologic speaking task across time and instructional conditions?

Based on González-Lloret and Nielson (2015) and the argument that TBLT is more effective than form-focused approaches in facilitating language acquisition (Ellis et al., 2020; Long, 2015), we hypothesized that students in the TBLT condition would outperform students in the TSLT comparison condition on a TBLA measure, and that they would perform at least equally well on a linguistic knowledge test. Given the exploratory nature of the second research question, we did not make an *a priori* hypothesis for RQ2.

Method

We collected data for the present quasi-experimental study from two intact Business English classes at a public three-year polytechnical vocational college in Guangzhou, China. Both classes were taught by the same instructor (the third author). As a student in a master's program in TESOL in China and the U.S., the third author familiarized herself extensively with TBLT at both theoretical and classroom-practice levels. Instruction and data collection took place during student participants' second and third semesters of college, from March 2020 to December 2020. English was a required subject for all students, and all students were required to take a high-stakes government-mandated English proficiency test (the College English Test Band-4 [CET-4]) during the year of the quasi-experimental intervention.

Participants

Participants in the study were L1 speakers of one or more varieties of Chinese. Learners enrolled in one of two intact L2 English classes prior to the outset of the study

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

(i.e., learners were not assigned to classes by the research team). The two classes were randomly assigned to a comparison condition (n = 40) or an experimental TBLT intervention (n = 40). The sample size in this study was determined primarily by practical constraints, as only two intact classes, each consisting of 40 students, were available at the research site during the intervention period. As a result, an *a priori* power analysis was not feasible for determining the optimal sample size. The learners were between 19 and 21 years old (experimental group, M = 20, SD = .69; comparison group, M = 20.02, SD = .77), and 12% (n = 9) identified as male. The L2 learners had been studying English for an average of 11.5 years (TBLT, M = 11.33, SD = 1.24; Comparison, M = 12.33, SD = 2.07) in formal primary and secondary schools. Their language proficiency at the outset of the treatment ranged from roughly A2 to B1 on the Common European Framework of Reference (CEFR) scale as measured by a version of the Quick Placement Test (UCLES, 2001), with no statistically significant differences in proficiency between the groups (TBLT, M = 29.85, 95% CI [28.44, 31.26], SD = 4.42; Comparison, M = 27.50, 95% CI [26.00, 29.00], SD = 4.71).

Instructional conditions: curricula, methodology, and implementation

The two groups took part in a two-semester (one academic year, 34 weeks, 128 lessons of 40 minutes) Business English program, which began with 15 weeks of online instruction during the COVID-19 pandemic before transitioning to face-to-face teaching. The shared institutional goal for both groups was to develop the functional English abilities needed for business-related communication. However, the two groups followed distinct curricula and pedagogical approaches. In the TBLT condition, this broad goal was operationalized through a task-based needs analysis (reported in Van Gorp, Guan, & Coss, 2026b), which identified 41 target tasks representing the personal and occupational situations in which students were expected to use English. These target tasks provided the basis for the design and sequencing of the TBLT syllabus, pedagogic tasks, and task-based assessment. The comparison group pursued the same broad institutional goal through the institution's pre-existing textbook-based curriculum via TSLT methodology. Although both groups addressed broadly similar domains of workplace and business communication, their underlying curricular design, lesson procedures, and orientation toward language learning differed substantially. Both syllabi, one sample speaking lesson, and one writing lesson for both conditions are available on the Open Science Framework (Van Gorp, Coss, Guan, Burton, & Tuzcu, 2026a) (see S*yllabi and sample tasks* subfolder). We selected speaking and writing lesson plans that were comparable across conditions and that included the same task in both conditions to highlight the different position and role the speaking and writing tasks occupied in the student learning activities for the two conditions. These tasks are referenced in the following sections.

Comparison group (textbook-based TSLT). The comparison group followed the already existing local curriculum built around two textbooks: *Century Business English Integrated Course: Foundation II* (Dai & Zheng, 2014) during the first semester and *Intercultural Communication for Business* (Ma & Pan, 2011) during the second semester. These textbooks emphasize practical workplace communication (semester one) and cross-cultural business (semester two) topics. Semester one was intended to strengthen general listening, speaking, reading, and writing skills with introductory business content, while semester two built on these skills by emphasizing students' ability to use English for oral communication and pragmatic action in professional

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

contexts, with attention to intercultural business communication. As required by the institutional English program administrators, the instructor focused on vocabulary and grammar instruction, reading comprehension, listening, speaking (role plays, group discussions), short essay writing, and translation in this course. In practice, lessons followed a task-supported PPP sequence in which tasks were used in the practice and production phases of lessons. The teacher first presented new grammar and vocabulary (e.g., inversion structures for formality, contract terminology, polite requests), then guided students through controlled practice activities (multiple-choice or true-false comprehension exercises, sentence transformations, gap-fill exercises, translations, structured dialogues), and finally engaged the students in short role-play tasks or writing tasks in which the taught forms were applied. The textbook for the second semester focused on cross-cultural communication, cultural business customs, and intercultural business negotiation, addressing topics like etiquette, verbal and non-verbal communication, values, gender differences, etc. The textbook put a strong focus on translation and analysis of textbook passages which the teacher supplemented with videos and student whole-class demonstrations. While lesson topics often referenced workplace or intercultural scenarios, they did not center tasks in the TBLT sense. Instead, tasks were used as practice of pre-taught forms rather than as the basis of curriculum, instruction, and assessment.

For example, in writing lessons, students learned grammatical structures (e.g., inversion for emphasis) and email etiquette before composing a short business email that had to include one inversion sentence and adhere to etiquette conventions. In speaking lessons, students listened to and read passages on contract negotiations, answered comprehension questions, practiced target vocabulary, translated key sentences into Chinese, and only in the final stage performed short role plays simulating a buyer–seller exchange. Tasks served primarily as practice vehicles to demonstrate mastery of taught forms, as is the case typically in TSLT pedagogy. Thus, although the syllabus included workplace-relevant content and tasks, the methodology remained form-focused and accuracy-oriented, distinguishing it from task-based instruction.

TBLT group (task-based syllabus). In contrast, the intervention group followed a newly developed TBLT syllabus. Following internationally recognized scholarship in TBLT, the curriculum was designed by the third author in collaboration with the first author through a task-based needs analysis (Long, 2005, 2015) (reported in Van Gorp et al., 2026b). The needs analysis utilized questionnaires, interviews, and focus groups with students, teachers, and business-domain experts to identify 41 authentic target tasks relevant to the students' personal and occupational domains. These target tasks were then used as the basis for pedagogic tasks adapted to learners' proficiency and the classroom context. The pedagogic tasks met Ellis and Shintani's (2014) four task criteria —focus on meaning, an information or opinion gap, reliance on learners' own resources, and a clearly defined communicative outcome—while also providing opportunities for feedback and focus on form.

To form the yearlong syllabus, these tasks were sequenced by thematic coherence and task complexity, informed by Robinson's (2011) Triadic Componential Framework. The first semester targeted personal-domain tasks (e.g., introducing oneself, making a hotel reservation, seeing a doctor), while the second focused on occupational tasks (e.g., developing business relations, promoting products, providing customer service, conducting a sales pitch). Within each lesson, instruction adopted the three-stage task cycle (pre-task, task, post-task) described by Willis and Willis (2007). Students first engaged in input and planning activities (e.g., analyzing model texts, watching authentic product-pitch videos, brainstorming persuasive strategies), then

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

performed the main task collaboratively (e.g., drafting an email to a potential client, presenting a product to foreign buyers at a simulated trade fair), and concluded with post-task reflection and form-focused feedback. Unlike the TSLT condition, attention to form occurred reactively or incidentally during task performance and was reinforced afterward, often in response to teacher-observed errors or gaps during students' task performance.

Curricular vs. methodological distinction. Both conditions addressed similar business English content domains but differed fundamentally in the organization of instruction. To illustrate, consider the business email writing unit. In the TBLT group, students analyzed authentic samples, brainstormed expressions for polite requests, and then drafted and revised emails to potential clients through multiple cycles of peer and teacher feedback, with the communicative purpose of securing future business at the center of the lesson. In the TSLT comparison group, students first practiced inversion grammar and textbook email conventions, then wrote a short email that demonstrated their use of these forms, with peer review guided by a checklist (which included the use of the inversion structure). Similarly, in speaking-focused lesson sequences, the TBLT group engaged in extended simulations of a trade fair where they presented products and negotiated with rotating roles, while the TSLT comparison group read textbook passages on business deals, practiced vocabulary, and only then performed brief role plays with scripted prompts. These contrasts exemplify how the distinction between conditions was methodological as well as curricular, with a task-based approach in the intervention condition versus a task-supported approach in the comparison condition.

Implementation fidelity. Implementation fidelity was closely attended to in both course planning and delivery. During the intervention, only limited syllabus adaptations were made reflecting responsive implementation rather than a departure from the task-based syllabus. For example, in the first online semester, Tasks 2 and 3 were repeated after Task 20 because the instructor observed that students had not fully mastered these earlier tasks, yet their content was needed for subsequent task performance. In addition, as the instructor noticed that initially students showed resistance to engaging in online group work, an interaction-heavy task (Task 4) scheduled early in Semester 1 was moved to the beginning of Semester 2, when face-to-face teaching resumed.

The same instructor taught both groups, ensuring comparability of teaching style and classroom management. Teacher logs documented some implementation and especially classroom management challenges in the first semester of the TBLT group (e.g., not always enough time for post-task phase, slow transitions from whole class to group work). These issues occurred primarily during the initial quarter of the semester (March 2020); from Lesson 10 onward (early April 2020), lessons proceeded as intended, and in the second semester both conditions were implemented fully according to plan. Fidelity was documented through additional sources beyond teacher reflection logs, including email correspondence with the first author (who guided the curriculum design), lesson recordings, and learners' post-task surveys. The email correspondence captured ongoing guidance on pedagogical dilemmas, teaching decisions, and the extent to which lessons preserved authentic task-based practice. Although reflection logs, lesson recordings, emails, and post-task surveys were not analyzed as data sources in the present study, they served as supplementary documentation of implementation fidelity.

Instructor-as-researcher role. The third author served as the instructor for both groups and, in her researcher role, contributed to the needs analysis, curriculum design, and documentation of the two conditions. This dual role reflected the practitioner-led

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

nature of this study and the limited institutional and research resources available. The project emerged from her interest in exploring whether TBLT could address Chinese EFL learners' persistent difficulties in developing communicative competence despite years of formal English instruction (Hu, 2005). Because this study was designed with the hypothesis that the TBLT condition would yield stronger outcomes, we acknowledge that this dual role may have introduced potential expectancy effects as a limitation; it is possible that the instructor's enthusiasm or commitment to the intervention could have unintentionally influenced instructional delivery. However, having the same instructor teach both groups helped control instructor-related differences in teaching style, classroom management, pacing, feedback, rapport with students, contextual familiarity, and responsiveness to learner difficulties. Furthermore, avoiding variance introduced by an instructor-pedagogy confound has been highlighted in TBLT scholarship as a core design feature of high-quality comparative studies (Shintani, 2020). The instructor sought to provide comparable guidance and support across conditions. As a novice teacher, she was invested in supporting students in both groups, including the comparison group, where students' silence and passive learning prompted reflection on adaptive strategies to motivate and engage them, such as diversifying classroom activities, adding occasional competitive elements, and using short theme-related videos as lesson warm-ups, without altering the textbook-based TSLT curriculum. This dual role and its potential limitations are therefore considered when interpreting the findings.

Assessment instruments

All students completed three complementary measures of language development: (a) a freely available version of the Quick Placement Test (QPT) (UCLES, 2001), a discrete-point multiple-choice test of English lexicogrammar knowledge, (b) a monologic task-based speaking assessment (TBLA), and (c) the official government-mandated Chinese English Test Band Four (CET-4), which did not include a speaking section.

Language Proficiency Assessment. For reasons of practicality, we used the freely available pen-and-paper *Quick Placement Test Version 1* (UCLES, 2001)[1] to measure linguistic knowledge. The Quick Placement Test (QPT) is "a flexible test of English language proficiency developed by Oxford University Press and Cambridge ESOL to give teachers a reliable and time-saving method of finding a student's level of English" (Geranpayeh, 2003, p. 8). The QPT is a 60-item multiple-choice test, using gap-filling cloze and sentence-level exercises, which focuses on lexis and grammar, with test-score interpretations ranging from below A1 to C2 on the CEFR (Geranpayeh, 2003). The QPT is a curriculum-independent placement test and screening examination, the use of which provides a standardized external benchmark that is not tied to the specific instructional materials used in either group. The same version of the QPT was administered as a pre-test before the beginning of semester one (January 2020) and as a post-test at the end of semester two (December 2020). All QPTs were administered face-to-face, with QPT1 completed before the COVID-19 lockdown.

Task-Based Assessments. The task-based assessment was a monologic speaking task with a clear communicative purpose. Presenting products and services through a sales pitch was one of the occupational target tasks identified in the needs analysis. We

[1]The pen-and-paper version was downloaded from https://www.vhs-aschaffenburg.de/fileadmin/vhs-aschaffenburg/PDF/OoxfordTest.pdf

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

selected it because it was broadly relevant to the students' anticipated business communication contexts and elicited an extended individual performance that could be administered and scored consistently across groups and measurement occasions. Students were provided with a task sheet in English (see Appendix A) and given one minute to prepare before delivering a two-minute presentation, which was video-recorded. This standardized format supported longitudinal comparison; however, we acknowledge that a single monologic task captures only part of learners' communicative ability.

The same sales-pitch assessment task was administered three times: at the beginning of semester one (March 2020), the end of semester one (June 2020), and the end of semester two (December 2020). The pre-test was conducted online due to COVID-19 restrictions, whereas Post-test 1 and Post-test 2 were conducted face-to-face. In both groups, students encountered a product-pitching task that was similar to the assessment task (though with different content) once in their respective syllabi. In the TSLT condition, this occurred before the first post-test (at the end of semester 1), whereas in the TBLT condition it occurred in the second semester.

To analyze student performance on this assessment task, we collaboratively developed a 7-point analytic rubric with four criteria: task completion, comprehensibility, organization, and language use (see Appendix B for the full rubric). Seven-point scales have been shown to have higher psychometric precision than five or fewer points (Simms, Zelazny, Williams, & Bernstein, 2019). The empty categories for 2, 4, and 6 served as "halfway" points when performances demonstrated some but not all aspects of two adjacent categories with descriptors (in line with operational practice in Cambridge Assessment scales and IELTS). Both test-internal (test purpose and score use) and test-external (existing speaking rubrics and theory/literature) sources informed our scale construction (Knoch, Deygers, & Khamboonruang, 2021). Considering test purpose and score use, we included a task-specific task completion criterion in addition to three task-independent criteria which were modeled after standardized speaking task rubrics (IELTS, TOEFL, Advanced Placement) as well as recent rubrics from functional adequacy research (e.g., Kuiken & Vedder, 2022). It is important to note that performing well on the task completion criterion required more than simply mentioning the prompts provided in the task instructions. As detailed in our analytic rubric, top scores were awarded only to students who successfully "elaborated beyond the content provided on the task sheet." After drafting the rubric, we completed four rounds of rater training and consensus building using randomly selected samples from the full dataset, each of which resulted in revisions to the rubric and the scales. Three raters (authors 2, 3, and 5), with varied language backgrounds (English, Chinese, and Turkish as L1s, respectively) and varied experience with teaching L2 learners of English with an L1 Chinese background[2] (from extensive to none), each independently rated all 240 student performances on the TBLA (40 students × 2 classes × 3 test sessions). The order of the 240 test samples was randomized for each rater.

The CET-4. As described in Jin (2019), "the CET is an exit test to assess the English language proficiency of tertiary-level learners upon their completion of the College English program" (p. 101). This test is generally administered to students in their fourth semester of college, after they have completed the first four levels of the College English curriculum. The CET-4 is made up of four parts, taking "a componential view of the

[2]This balancing of rater backgrounds ensured that rater-background-induced variance (see Winke & Gass, 2013) was minimized.

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

construct of language proficiency, assessing language skills independently" (Jin, 2019, p. 104). The test contains sections for listening, reading, translation, and writing. Each section is given a score, and the composite score on the whole test ranges from 290 to 710. The passing score is 425. We included CET scores in this analysis because this was the test with the most published validity evidence available in our dataset, all test takers were required to take it during this intervention, and it is a high-stakes, locally meaningful test, and therefore consequential when evaluating the viability of TBLT in this context. Furthermore, it served as a covariate in our models to account for variance attributable to underlying proficiency; that is, to more appropriately model the trajectories of the outcome variables.

Because the CET-4 was administered just two weeks into the second semester (September 16, 2020),[3] the scores primarily reflect students' underlying English proficiency developed prior to the intervention and during the first semester. This timing was methodologically appropriate for our study, as the CET-4 was utilized as a covariate in our statistical models to control for baseline proficiency rather than as a summative outcome measure. However, we acknowledge that this early administration timing limits the extent to which the CET-4 scores reflect the full impact of the two-semester intervention. It is entirely possible that a later administration at the end of the second semester might have yielded different results, potentially capturing further divergence in language ability between the instructional conditions.

Data analysis

To answer how instructional groups compared on a linguistic knowledge test (RQ1), we analyzed QPT scores prior to starting the course (pre-test) and at the end of the course (post-test). Internal consistency was below operational range for the two administrations ($\alpha_{\text{Pretest}} = .54$, $\alpha_{\text{Posttest}} = .56$), likely because the restricted range of ability among the learners attenuated this estimate. To ensure our test scores were as meaningful as possible, we ran additional quality assurance checks on the test items, reported in Appendix C. We were satisfied with the quality of the items, and because our focus was on total scores (an aggregated sum of all 60 items), rather than item-level inferences, we used the QPT for analysis.

We calculated the dependent variable as total test scores out of 60 possible correct responses with no weighted items. We conducted a 2 × 2 factorial ANOVA to investigate the effects of the task intervention on language ability as measured with the QPT. We used the *aov* function in R (v.4.3.2) to account for both within- and between-person effects. We specified repeated measurements in the model using the syntax *Score~Class*Time+Error(ID)*. The outcome measure was the test scores, and the independent variables were group membership (Comparison vs. Experimental) and time. We calculated effect sizes using eta squared rather than partial eta squared (Norouzian & Plonsky, 2018) using the *eta_squared* function (effectsizes package v.0.8.9). Despite the presence of one outlier, the assumptions for the statistical model were met. Test scores for each group at each time were distributed normally with non-significant Shapiro-Wilk tests (*p*s = .29, .29, .27, .12). Levene's test for the homogeneity of variance held across both instructional groups, $F(1, 158) = 0.02$, $p = .89$, and test time, $F(1, 158) = 0.01$, $p = .94$. Sphericity was not applicable in this design since the study involved only two time points, and observations were independent across participants.

[3]As this is an official examination, we had no control over test administration timing.

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

To answer how the instructional groups compared on the task-based speaking assessment, we followed two main steps. First, we confirmed that the scoring mechanism for the task-based language assessment (TBLA) was psychometrically robust prior to using any of the data in any inferential models. Robustness, reported in Appendix C, was important to confirm for the four scales used in the study as well as rater and test-taker behavior. After confirming that the scales functioned as intended, we ran inferential statistical analyses to determine the relationships between group, time, and the four rated outcomes. Because this dataset consisted of repeated measurements and ordinal values on seven-point scales, we analyzed the outcomes using ordinal mixed-effects regression using *clmm* in R (v.2023.12-4), with a logit link and flexible thresholds. The first model (RQ1) included all rated outcomes together (an omnibus model), with raters, test takers, and differences in criteria modeled as random effects. This model is roughly equivalent to an explanatory rating scale Many Facet Rasch model. To answer RQ2, we also modeled each rated outcome separately. For the omnibus model and the four separate models, we modeled the interaction between group membership (Comparison × Experimental) and time (three time points). The comparison group and the first administration of the TBLA served as reference values in the model. We included random effects of test taker and rater in the model (and criteria for the omnibus model). We also controlled for the test takers' language proficiency using their official CET scores as a covariate predictor. We controlled for CET given that mean scores differed slightly between the comparison group, M = 429.58, and the experimental group, M = 436.17, although this difference was not significant in a Welch's independent samples t-test, $t(76.01) = -0.65$, $p = .52$. We calculated model fit based on comparisons of the full model with the null model with no predictors and the model with random effects removed. Effect sizes are reported as odds ratios. Conditional and marginal R^2 is reported using Nakagawa's R^2 (Nakagawa & Schielzeth, 2013), computed in *sjPlot* (v.2.8.17). We adjusted alpha to 0.01 to account for the five repeated statistical tests (Bonferroni correction).

Assumptions of ordinality and a lack of multicollinearity were met. The assumption of proportional odds was generally met in the models of individual criteria using Brant's tests (Brant, 1990) using the *brant* package (v.0.3-0) (on models with random effects removed [*polr* models], as there are no methods to check this assumption for *clmm* models with random effects). Nonetheless, for our purposes, we were interested in the *average* odds ratios rather than predicted discrete outcomes, which Harrell (2020) stated was a less problematic violation. R code for the descriptive and inferential data is available in Van Gorp et al. (2026a).

Results

QPT results

RQ1 investigated whether learning gains on a discrete language test would be present in both the TBLT and comparison group. Table 1 displays the mean scores for each group at the two testing times. Figure 1 shows the distribution of these trends in both violin and boxplots. Both groups demonstrated a normal distribution around a mean that was roughly 50% of the total test score (30/60) at both time periods, with small differences between groups by experimental condition and at each time point.

A 2 × 2 factorial ANOVA confirmed that there was a significant main effect of time, $F(1, 78) = 26.62$, $p < .001$, with $\eta_p^2 = .25$, $\eta^2 = .07$, which is a medium effect size (Cohen, 1969). Test scores, regardless of the grouping variable, grew from a mean of 28.7 [27.60,

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Table 1. Descriptive Statistics for TBLT and Comparison Groups at Times 1 and 2 (QPT)

	Time 1: Pre-test			Time 2: Post-test		
	Mean	95% CI	SD	Mean	95% CI	SD
Comparison	27.50	[26.00, 29.00]	4.71	30.73	[29.36, 32.09]	4.26
Experimental	29.85	[28.44, 31.26]	4.42	31.70	[30.03, 33.37]	5.23

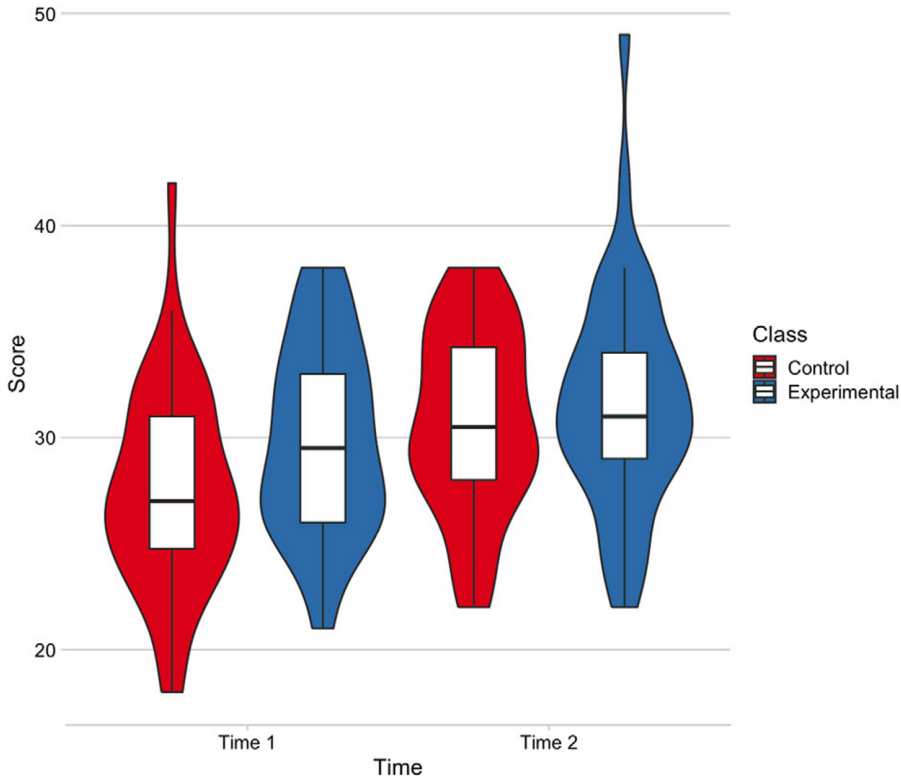


Figure 1. Violin plots and boxplots of QPT.

29.70] in the pre-test to a mean of 31.20 [30.20, 32.20] in the post-test. However, this relationship was not significant for the grouping variable, $F(1, 78) = 3.25$, $p = .08$, $\eta_p^2 = .04$, $\eta^2 = .03$, or the interaction between group and time, $F(1, 78) = 1.96$, $p = .17$, $\eta_p^2 = .02$, $\eta^2 < .001$, both with small effect sizes.

TBLA results

Descriptive statistics for the TBLA data are shown in Table 2. Median score values were centered around a score of 4 on the seven-point scale for task completion, comprehensibility, and organization, with interquartile ranges one point above and below each scale and no differences between the comparison and TBLT groups. Language use was

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Table 2. Descriptive Statistics of Rating Scale Data

Group	Criterion	Q1	Median	Q3	SE
Comparison	Task Completion	3	4	5	.09
	Comprehensibility	3	4	5	.08
	Organization	3	4	5	.07
	Language Use	2	3	4	.06
Experimental	Task Completion	3	4	5	.09
	Comprehensibility	3	4	5	.07
	Organization	3	4	5	.07
	Language Use	3	3	4	.06

Note: Median and first (Q1) and third (Q3) interquartile ranges are reported for ordinal variables.

rated somewhat more severely, with medians in each group at 3. The comparison group exhibited more variance in language use scores, with interquartile ranges from 2 to 4, while the TBLT group showed more centrality with the ranges from 3 to 4.

In terms of rater reliability (further discussed in Appendix C), the rater separation statistic indicated that raters differed in their severity but in a very narrow range. We inspected these differences further by calculating Krippendorff's Alpha (a measure of *agreement* among raters on ordinal scores while taking into account chance agreements), which was .69. Both statistics suggest that exact agreement was only medium. However, we estimated the effect that these differences might bias any findings by decomposing the intraclass correlation coefficient (ICC) by raters. ICC indicated that rater effects accounted for only 1% of total score variance ($ICC_{RATER} = .01$), suggesting that while raters could be statistically distinguished, systematic severity differences were negligible in magnitude.

Table 3 shows the polychoric correlations among the rated scales. The lowest correlation was between task completion and comprehensibility (.43), while language use and organization correlated the strongest (.78). Overall, language-related elements tended to correlate more strongly together than with task completion.

Four Outcome (Omnibus) Model. The omnibus model of all four outcomes (rubric criteria) modeled together is shown in Table 4. Both the interaction between group and time and their main effects were significant predictors in this model, with odds ratios ranging from 2.81 to 4.80. The strongest effect of the model was the interaction between the TBLT group and the final test, in which being in the TBLT group led to scores being 4.80 times more likely to be higher than the comparison group. The covariate of CET score had little predictive effect on the model, odds ratio = 1.04. Nakagawa's marginal R^2 was .22 and conditional R^2 was .53, indicating that the predictors and random effects explained 53% of the variance and the predictors alone explained 22% of the variance on the latent scale. The intraclass correlation coefficient (ICC) was .41, indicating that approximately 41% of the total variance in scores was attributable to group-level clustering. Variance in the random effects was small for raters, indicating only small differences in scoring patterns. Criterion variance showed that the four outcomes did not behave uniformly. The interaction between group and time is visualized in Figure 2.

Table 3. Polychoric Scale Correlations

	Task Completion	Comprehensibility	Organization
Comprehensibility	.43		
Organization	.70	.70	
Language Use	.58	.65	.78

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Table 4. Model of All Outcomes

Coefficients	*b*	95% CI	SE	*z* value	*p*	OR	95% CI
TBLT	−0.29	[−0.93, 0.36]	.33	−0.86	.39	0.75	[0.39, 1.44]
Time 2	1.03	[0.80, 1.27]	.12	8.54	< .001*	2.81	[2.22, 3.56]
Time 3	1.31	[1.07, 1.55]	.12	10.65	< .001*	3.71	[2.92, 4.73]
CET	0.04	[0.02, 0.06]	.01	4.51	< .001*	1.04	[1.02, 1.06]
TBLT:Time 2	1.00	[0.67, 1.33]	.17	5.89	< .001*	2.72	[1.95, 3.80]
TBLT:Time 3	1.57	[1.22, 1.91]	.17	9.04	< .001*	4.80	[3.41, 6.74]
	Random effects						
	Groups	Variance	SD				
	Raters	.07	.27				
	Test takers	1.87	1.37				
	Criteria	.31	.56				

Note: TBLT = experimental group. *b* = standardized regression coefficient, 95% CI = 95% confidence interval, SE = standard error, *z* value = z-statistic, OR = odds ratio, *p < .01.

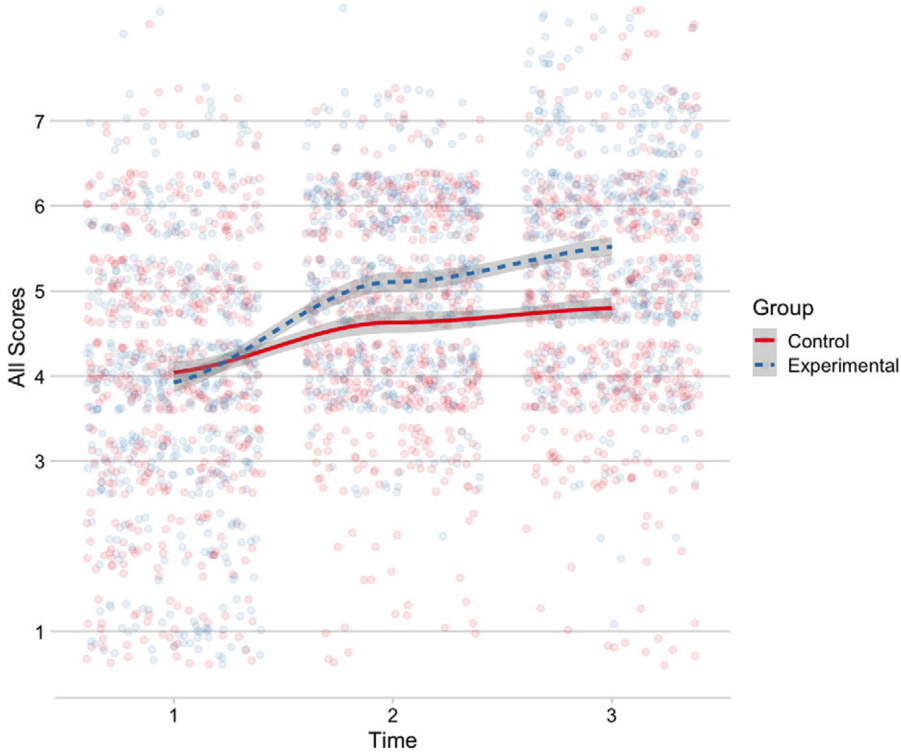


Figure 2. Model of All Scores (Omnibus Model).

Task Completion. The final model of task completion is shown in Table 5. Both the interaction between group and time and their main effects were significant predictors of this outcome, with odds ratios ranging from 5.30 to 15.51. The strongest effect of the model was the interaction between the TBLT group and the final test, in which being in the TBLT group led to scores being 15.51 times more likely to be higher than the

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Table 5. Model of Task Completion

Coefficients	*b*	95% CI	SE	*z* value	*p*	OR	95% CI
TBLT	−0.85	[−1.82, 0.12]	.50	−1.71	.09	.43	[0.16, 1.13]
Time 2	1.93	[1.42, 2.45]	.26	7.34	< .001*	6.90	[4.12, 11.55]
Time 3	1.67	[1.15, 2.17]	.26	6.49	< .001*	5.30	[3.20, 8.78]
CET	0.04	[0.02, 0.07]	.01	3.23	.001*	1.04	[1.02, 1.07]
TBLT:Time 2	1.62	[0.92, 2.32]	.36	4.54	< .001*	5.05	[2.51, 10.16]
TBLT:Time 3	2.74	[2.02, 3.46]	.37	7.47	< .001*	15.51	[7.56, 31.85]
	Random effects						
	Groups	Variance	SD				
	Raters	.99	.99				
	Test takers	3.67	1.91				

Note: *b* = standardized regression coefficient, 95% CI = 95% confidence interval, SE = standard error, *z* value = *z*-statistic, OR = odds ratio, *p < .01.

comparison group. The covariate of CET score had little predictive effect on the model, odds ratio = 1.04. Nakagawa's marginal R^2 was .27 and conditional R^2 was 0.70, indicating that the fixed effects explained a reasonable amount of variance in task completion, with the random effects accounting for most of the variance (ICC = .59). Variance in the random effects was greater for test takers than raters, and of the four outcomes, residual variance in raters was highest for task completion, indicating larger between-rater differences than other outcomes. The interaction between group and time is modeled in Figure 3.

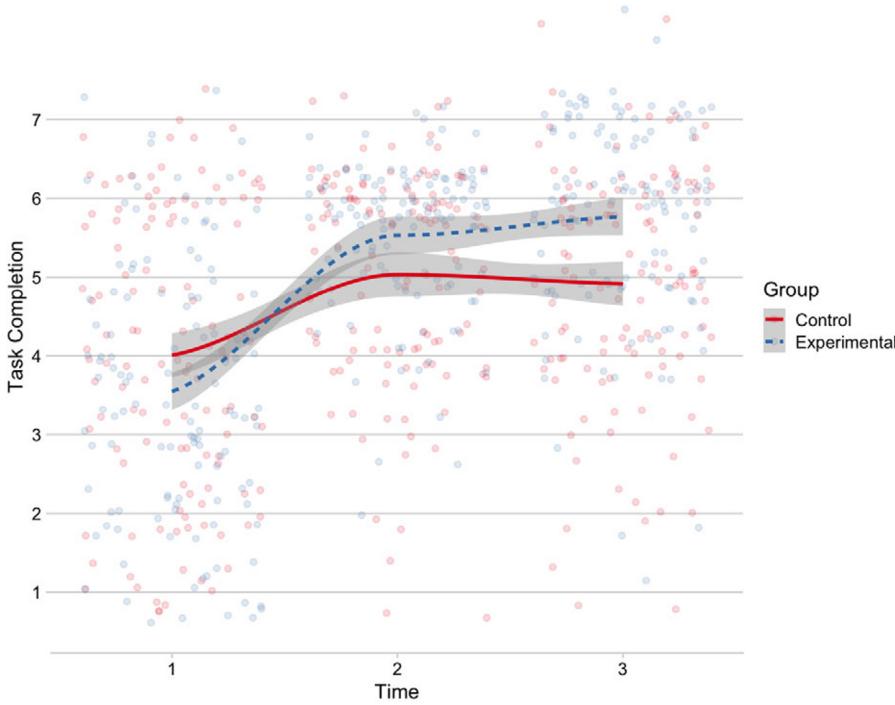


Figure 3. Model of Task Completion.

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Table 6. Model of Comprehensibility

Coefficients	*b*	95% CI	SE	z value	*p*	OR	95% CI
TBLT	.40	[−0.39, 1.94]	.40	1.00	.32	1.50	[0.68, 3.30]
Time 2	.76	[0.30, 1.25]	.24	3.21	.001*	2.17	[1.35, 3.49]
Time 3	1.38	[0.90, 1.87]	.25	5.63	< .001*	3.99	[2.46, 6.46]
CET	.03	[0.01, 0.05]	.01	2.92	.004*	1.03	[1.01, 1.05]
TBLT:Time 2	.12	[−0.53, 0.78]	.34	.37	.71	1.13	[0.59, 2.19]
TBLT:Time 3	.53	[−0.14, 1.20]	.34	1.56	.12	1.70	[0.87, 3.32]
	Random effects						
	Groups	Variance	SD				
	Raters	.05	.23				
	Test takers	2.06	1.43				

Note: *b* = standardized regression coefficient, 95% CI = 95% confidence interval, SE = standard error, *z* value = *z*-statistic, OR = odds ratio, *p < .01.

Comprehensibility. The final model of comprehensibility is shown in Table 6. Only the main effects of time and CET were significant predictors in this model, with odds ratios ranging from 1.03 to 3.99. The strongest effect of the model was the effect of time, showing that compared to time 1, all test takers were 4 times more likely to improve on comprehensibility regardless of their instructional condition. The covariate of CET score had little predictive effect on the model, odds ratio = 1.03. Nakagawa's marginal R^2 was .14 and conditional R^2 was .47, indicating that the fixed effects explained a smaller amount of variance for comprehensibility than task completion on the latent scale, with the random effects of the grouping variables accounting for substantial variance (ICC = .39). Variance in the random effects of raters was negligible (.05), showing high consistency. The model is visualized in Figure 4.

Organization. The final model of organization is shown in Table 7. Both the interaction between group and time and their main effects were significant predictors of this outcome, with odds ratios ranging from 2.34 to 17.41. The strongest effect of the model was the interaction between the TBLT group and the last test, in which being in the TBLT group led to scores being 17.41 times more likely to be higher than the comparison group. The covariate of CET score had little predictive effect on the model, odds ratio = 1.05. To calculate R^2, the near 0 residual variance of rater after accounting for fixed effects caused a singularity issue; thus, variance was estimated with the random effect removed: Nakagawa's marginal R^2 was .35 and conditional R^2 was .65, indicating that the fixed effects explained a large amount of variance for organization on the latent scale, with the random effects of the grouping variable accounting for substantial variance (ICC = .46). The interaction between group and time is modeled in Figure 5.

Language Use. The final model of language use is shown in Table 8. Only the interaction between instructional group and Time 3 was significant, as well as the main effects of time. Odds ratios ranged from 2.82 to 3.93. The strongest effect of the model was the interaction between the TBLT group and the final test, in which being in the TBLT group led to scores being 3.93 times more likely to be higher than the comparison group. The covariate of CET score had little predictive effect on the model, odds ratio = 1.06. Nakagawa's marginal R^2 was .26 and conditional R^2 was .55, indicating that the fixed effects explained a sizeable amount of variance for language use on the latent scale, with the random effects of test takers accounting for substantial variance (ICC = .39). The interaction between group and time is modeled in Figure 6.

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

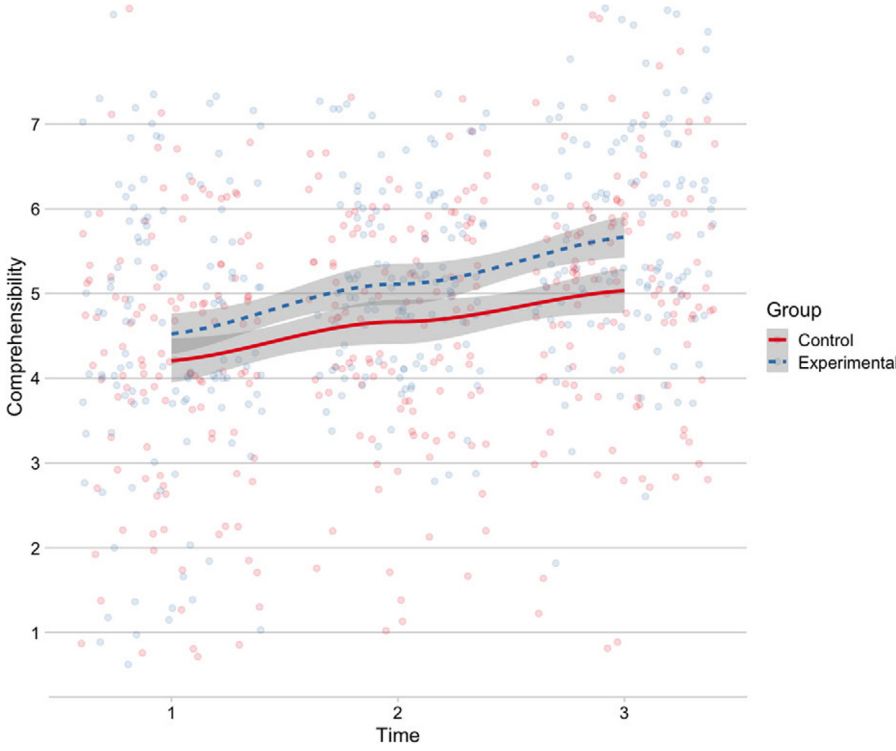


Figure 4. Model of comprehensibility.

Table 7. Model of Organization

Coefficients	*b*	95% CI	SE	z value	*p*	OR	95% CI
TBLT	−0.92	[−1.81, −0.03]	.46	−2.02	.04	.40	[0.16, 0.97]
Time 2	.85	[0.35, 1.35]	.25	3.36	< .001*	2.34	[1.43, 3.85]
Time 3	1.52	[1.02, 2.04]	.26	5.86	< .001*	4.60	[2.76, 7.66]
CET	.06	[0.04, 0.08]	.01	5.14	< .001*	1.05	[1.04, 1.09]
TBLT:Time 2	2.23	[1.51, 2.94]	.36	6.11	< .001*	9.26	[4.54, 18.91]
TBLT:Time 3	2.85	[2.12, 3.59]	.37	7.65	< .001*	17.41	[8.37, 36.19]
	Random effects						
	Groups	Variance	SD				
	Raters	< .01	< .01				
	Test takers	2.82	1.68				

Note: *b* = standardized regression coefficient, 95% CI = 95% confidence interval, SE = standard error, *z* value = z-statistic, OR = odds ratio, *p < .01.

Discussion

In this study, we examined the extent to which a strong-form TBLT approach would lead to stronger task-based speaking performance and equivalent or better linguistic gains compared to a task-supported language course focused on vocabulary and grammar. Given that past research has generally either measured outcomes on a

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

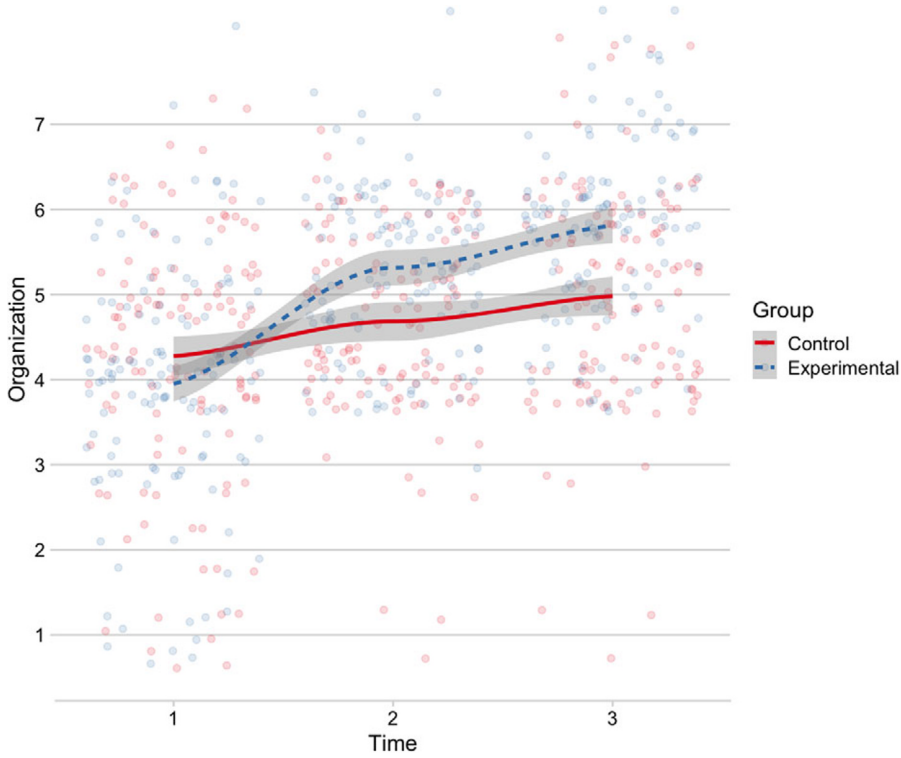


Figure 5. Model of organization.

Table 8. Model of Language Use

Coefficients	*b*	95% CI	SE	z value	*p*	OR	95% CI
TBLT	−0.09	[−0.89, 0.72]	.41	−0.21	.83	.92	[0.41, 2.04]
Time 2	1.04	[0.54, 1.53]	.25	4.09	< .001*	2.82	[1.72, 4.64]
Time 3	1.31	[0.80, 1.81]	.26	5.06	< .001*	3.70	[2.23, 6.14]
CET	.06	[0.04, 0.08]	.01	5.34	< .001*	1.06	[1.04, 1.08]
TBLT:Time 2	.73	[0.04, 1.43]	.35	2.07	.04	2.08	[1.04, 4.15]
TBLT:Time 3	1.37	[0.66, 2.07]	.36	3.81	< .001*	3.93	[1.94, 7.96]
	Random effects						
	Groups	Variance	SD				
	Raters	.05	.22				
	Test takers	2.04	1.43				

Note: *b* = standardized regression coefficient, 95% CI = 95% confidence interval, SE = standard error, *z* value = z-statistic, OR = odds ratio, *p < .01.

linguistic knowledge assessment *or* a task-based assessment but not both (Boers & Faez, 2026), we were interested in how the two measures would compare, as there is a risk in assuming that different assessment methods provide equivalent evidence of language development (Bachman, 2002). In addition, by using two complementary outcome measures, we aimed to evaluate TBLT's role in supporting both communicative

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

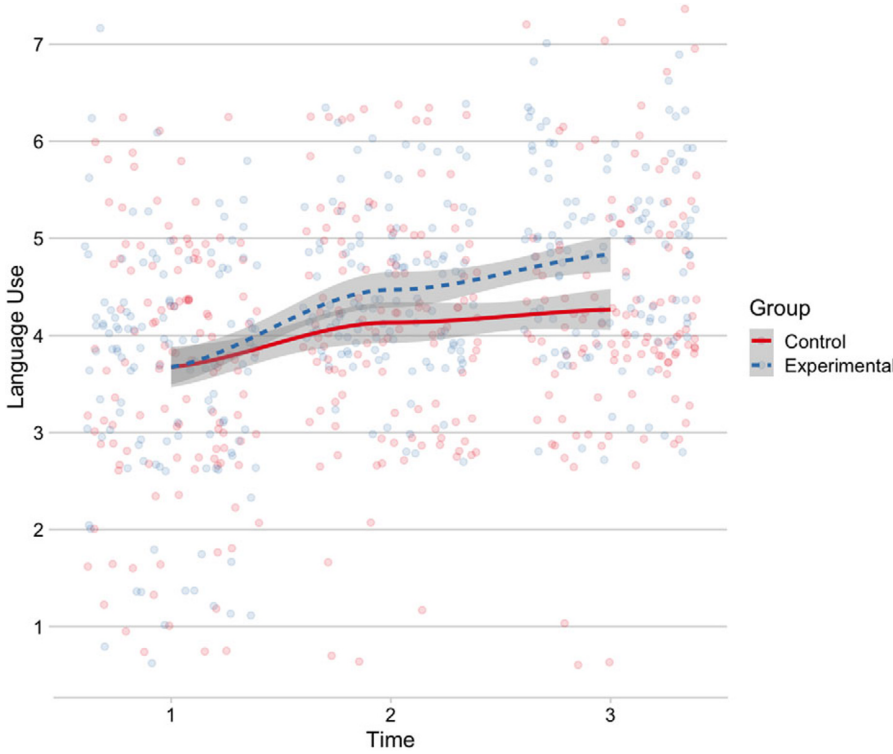


Figure 6. Model of language use.

competence and explicit linguistic knowledge development. Furthermore, we explored how changes in students' task completion ability—a central criterion in strong-form TBLT/A—compared to growth on linguistic and textual dimensions of language performance. In doing so, the study contributes to a better understanding of the robustness and instructional relevance of different performance criteria in task-based assessment. This study is novel in its use of a longer-term longitudinal design, which lasted the course of an academic year, and its use of both task-based and explicit linguistic knowledge assessments. Crucially, this investigation was carried out in a high-stakes-test-oriented East Asian tertiary educational setting, an environment often seen as an impediment to the successful implementation of TBLT.

To answer our first research question, we examined students' performance on a task-based speaking assessment and on the Quick Placement Test (QPT), a test of linguistic knowledge, to address how TBLT students perform compared to TSLT students. Our findings indicate that the TBLT group significantly outperformed the comparison group on the speaking task at the end of both first and second semesters. This result aligns with expectations (Ellis et al., 2020; Long, 2015) and prior research by González-Lloret and Nielson (2015) (see also Ganjouee et al., 2018), as the TBLT course emphasized speaking and communicative competence in real-world workplace-related tasks. The work-related speaking task as a focal outcome measure might explain the differences between our findings and González-Lloret and Nielson's (2015) findings with Harris and Leeming's (2022) and Pomarolli et al. (2023) findings, who reported

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

that both TBLT and non-TBLT (i.e., PPP in their two studies) were equally effective in supporting speaking development. The Harris and Leeming (2022) study used a speaking assessment procedure that was almost identical to our procedure, but, like Pomarolli et al. (2023), focused on everyday-life content and used typical classroom speaking assignments with no stated audience or communicative purpose (cf. Norris, 2016). The ability of TBLT students in our study to successfully complete real-world communicative tasks supports TBLT's pedagogical philosophy, which prioritizes authentic task performance as the driver of language development (Long, 2015).

Interestingly, on the vocabulary and grammar test (the QPT), the TBLT group performed comparably to the comparison group, despite the latter receiving two semesters of explicit grammar and vocabulary instruction, while the TBLT group's language instruction was limited to reactive focus on form (Long, 2000). This suggests that although the TBLT program did not explicitly focus on lexical and grammatical knowledge, TBLT still facilitated the acquisition of these domains of L2 knowledge to a similar degree as a structural syllabus. This finding is also in line with González-Lloret and Nielson's (2015) findings. Although the scores only represented an overall 3–4 point gain on the QPT (out of 60 total possible points), this represents progress of approximately half a skill level between B1 and B2 on the CEFR (Geranpayeh, 2003), which is in line with language development trends in similar classroom-based contexts.

A key observation from our longitudinal mixed-effects regression analyses is that students showed the most substantial improvement between their first and second speaking assessments (i.e., the first semester), with sustained but more moderate gains in the second semester. This pattern is in line with Harris and Leeming's (2022) findings. They not only found a differential outcome between the TBLT and PPP groups, but also found that the growth curve (i.e., slope) between Time 1 and Time 2 was much steeper than the slope between Time 2 and Time 3 in the TBLT group, with the PPP group showing the opposite pattern. In contrast, our comparison group showed more consistent growth (i.e., a more consistent slope across time). The growth pattern of the TBLT group suggests that the initial exposure to communicative tasks and feedback played a substantial role in immediate language gains, while continued improvements reinforced long-term benefits of TBLT, contrary to Harris and Leeming's (2022) findings where their TBLT and PPP groups turned out to be equally effective over a similar two-semester timeframe. Two additional possible contributors to the relatively stronger growth during semester one versus semester two for the TBLT group could be (a) the fact that the TBLA closely resembled the format of the speaking tasks throughout the TBLT course (and thus possibly contributed to students' confidence, positively impacting performance) and/or (b) the novelty of the TBLT approach enhancing students' motivation to perform well, given that their past L2 learning experience had not included TBLT pedagogy. Overall, the findings for RQ1 aligned with our hypotheses, supporting the argument that task-based instruction enhances both task-specific language performance and other dimensions of linguistic competence.

An interesting question emerges as to why the TSLT group, whose instruction focused primarily on discrete language knowledge, did not substantially *outperform* the TBLT group on measures of explicit linguistic knowledge (the QPT and the CET-4). As our data do not isolate specific mechanisms, we can only speculate about underlying reasons. It may be that the TBLT syllabus, organized around tasks rather than language, provided more robust grounds for interleaved use of language, which has been shown to result in more robust long-term retention than massed, focused "practice" (Rogers, 2017; Suzuki, 2023). Additionally, there may be a motivational component to these

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

longitudinal results: if the TSLT group consistently felt exceedingly challenged or discouraged by the language-based, accuracy-focused curriculum, this might have led to decreased engagement and, over time, less retention of language knowledge, and thereby lower-than-expected test performance.

Task completion and other (analytic) assessment criteria (A "linguistic caboose")

Our second research question examined the relationship between the four rated TBLA outcomes (task completion and other criteria), the instructional group and the interaction between group and time. The analyses showed that at both Times 2 and 3 the TBLT group outperformed the comparison group on task completion and organization; a clear indication that the TBLT group was more adept at expanding upon basic prompts with original information. With respect to the language use criterion, the TBLT group outperformed the comparison group only at Time 3 (i.e., the end of the course). There was no difference between the TBLT group and the comparison group on the comprehensibility of the speaking performances. The TBLT course made an impact on speaking performance (marginal R^2 = .22) overall, but especially in terms of task completion (marginal R^2 = .27) and organization (marginal R^2 = .35), which it impacted the most over time. The TBLT intervention facilitated the development in language use over time, moderately, and it did not differentially improve comprehensibility scores, as everyone improved on this criterion.

These results revealed a pattern consistent with the overall TBLA results: raters' judgment of task completion aligns well with the overall quality of speaking performances, as evidenced by the organization and language use criteria. Because this systematic relationship suggests that growth in task completion mirrors growth in language use, practitioners can make more informed decisions regarding the inclusion and weighting of language-use criteria based on specific assessment purposes. For summative TBLAs (i.e., "target tasks", Long, 2015), practitioners might confidently rely on holistic task completion as the primary, heavily weighted criterion; this reliably reflects overall language development and minimizes the potential negative washback of a "linguistic caboose"—where an over-emphasis on discrete language forms risks shifting attention away from the task as the unit of analysis (Long, 2015). However, because task completion alone does not fully account for the complexity of language development and may obscure key linguistic strengths and weaknesses (Norris, Brown, Hudson, & Yoshioka, 1998), there is likely a time and place for more detailed assessments. To effectively align assessment practices with instructional goals (Norris & East, 2021), practitioners can intentionally include specific language-use criteria alongside task completion to pinpoint areas for targeted diagnostic feedback and to identify students who need individual support when using assessments for formative purposes. Ultimately, determining how and when to balance holistic task completion with analytic language criteria remains a ripe area for further TBLT and TBLA research.

Interestingly, while the TBLT group initially lagged behind in language use, by the third measurement point, they had surpassed their non-TBLT peers. This finding may be an indication that the students were getting used to a new methodology, or that they were differentially impacted by the type of instruction (Spada & Tomita, 2010). In Qi's (2023) study, the TSLT group did make more noticeable attempts to use new grammar structures than the TBLT group. In contrast, the TBLT group showed an advantage, although not a statistically significant advantage, on grammar knowledge tests. As we

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

did not analyze the learners' performances in terms of complexity, accuracy, fluency, or specific linguistic structures, we cannot provide a clearer idea of why the TBLT group lagged behind in the present study. Given that task-based instruction devotes less time to explicit L2 learning, lexicogrammatical development may emerge at a more gradual pace compared to a structure-based course. Nevertheless, our results suggest that students in the task-based course developed the necessary textual and grammatical knowledge (Bachman & Palmer, 2010) to deliver increasingly coherent and structured spoken performances, at least as measured in the present study's TBLA. Although the TBLT group and the comparison group performed similarly on discrete lexicogrammar items in the QPT—that is, the TBLT group did not perform significantly differently than the comparison group—the TBLT students demonstrated a greater ability to integrate this knowledge into spoken communication, demonstrating the dual benefits of TBLT.

Interestingly, despite the overall success of TBLT, comprehensibility showed a different pattern compared to other assessment criteria. Although TBLT students consistently received slightly higher comprehensibility ratings, the difference was not statistically significant. Similar findings have been reported in González-Lloret and Nielson (2015) and studies on L2 pronunciation gains (e.g., Saito, 2021). Given that pronunciation improvements are often linked to explicit phonetic instruction, it is possible that neither instructional approach, as implemented, provided a sufficient pronunciation focus. In addition, it must be acknowledged that a single monologic task might not be the most appropriate speech elicitation instrument to assess comprehensibility, and that task complexity is likely to affect the comprehensibility of speech (Crowther & Isbell, 2024). Therefore, gauging comprehensibility, and especially changes in comprehensibility, may require more numerous and varied (e.g., interactive) tasks. Furthermore, we cannot discount the possibility that the change in comprehensibility over time in our results might be a result of practice effects or task repetition (Lambert, Kormos, & Minn, 2017).

High-stakes language testing and TBLT

An additional indicator of TBLT's effectiveness was students' performance on the CET-4, a high-stakes standardized English test in China. Given the strong washback effect of the CET-4 on Chinese higher education English curricula (Sun, 2016), one might expect the comparison group, whose curriculum was designed to prepare students for this test, to outperform the TBLT group. However, this was not the case. This finding reinforces the argument that proficiency-based, communicative approaches like TBLT can be equally effective, even in test-driven educational environments.

Implications of the study

Theoretical implications. The findings strengthen the theoretical case for tasks as the central unit of curriculum and instruction. Over two semesters, a strong-form TBLT program produced greater development in learners' ability to complete a relevant communicative task while supporting gains in explicit lexicogrammatical knowledge comparable to those achieved through a form-focused, task-supported curriculum. This pattern suggests that sustained meaning-focused task performance, combined with reactive focus on form, can support both functional language use and the development of linguistic knowledge. At the same time, the differential results across task completion, organization, language use, and comprehensibility indicate that

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

communicative development is multidimensional and that gains associated with TBLT should not be assumed to occur uniformly across all aspects of performance.

Methodological implications. The study illustrates the value of evaluating TBLT as a coherent program rather than as an isolated classroom activity. Connecting needs analysis, syllabus design, implementation, and assessment made it possible to examine the effectiveness of a needs-based strong-form TBLT program under ecologically valid conditions. The longitudinal comparison and use of both a task-based performance assessment and a discrete-point measure also yielded complementary evidence that either measure alone would have missed. Future comparative evaluations should similarly document curricular and methodological contrasts, monitor implementation over time, and employ multiple outcome measures aligned with both communicative and institutional goals.

Pedagogical implications. For practitioners in examination-oriented settings, the findings suggest that organizing instruction around needs-based tasks need not come at the expense of performance on lexicogrammatical measures. A task-based curriculum may therefore help learners develop real-world communicative ability while remaining responsive to institutional testing demands. More generally, the assessment findings suggest that task completion can serve as a practical primary criterion in summative TBLA, whereas analytic criteria such as organization and language use remain valuable for formative feedback and diagnosis—which may be a useful takeaway across diverse contexts that employ TBLA. Curriculum designers should maintain clear alignment among target tasks, classroom pedagogic tasks, and assessments while recognizing that additional tasks and more focused support (e.g., pronunciation instruction) may be needed for improvement along dimensions such as comprehensibility.

Limitations and future research directions

This study has several limitations that should be acknowledged when interpreting findings. First, the study was conducted in an English for Specific Purposes (ESP) context at a higher education institution in China. As a result, the findings may not be directly transferable to other educational settings, such as primary or secondary education within China or EFL settings in other countries. However, the strong similarities between our results and those of González-Lloret and Nielson (2015) suggest that, at least in the context of language for specific purposes and adult learning, some generalizability of our findings is plausible.

Second, our design involved a comparison of both curricular and methodological differences (needs-based TBLT vs. textbook-based TSLT), which limits the extent to which instructional methods can be isolated from curriculum. At the same time, this design reflects how these approaches are typically implemented in practice, which enhances the ecological validity of our approach. In addition, a more detailed qualitative analysis of students' language use and the classroom practices in both the TBLT and TSLT groups would allow for a finer-grained interpretation of our findings. Given that both groups were taught by the same instructor, had the same number of contact hours, and were randomly assigned to condition, many variables in the study were controlled. However, to offer a maximally robust account of TBLT compared to other pedagogical approaches, examining both the products (i.e., language outcomes) and the processes (i.e., pedagogy), including classroom interactions and instructional practices, of both TBLT and the comparison condition is necessary. Furthermore, such an analysis could help establish clearer links between teacher actions and student language learning

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

outcomes (Norris & Davis, 2021). Teacher logs revealed some implementation challenges in the first semester of the TBLT condition related to instructor inexperience. Although these issues were resolved from lesson 10 onward and both conditions were implemented as intended in the second semester, future research should include a more in-depth study of implementation fidelity, ideally carried out by an external researcher to provide an independent account of how TBLT and TSLT interventions unfold in classroom practice.

Third, TBLA focuses on "the elicitation and evaluation of language use (across all modalities)" (Norris, 2016, p. 232). Thus, a broader range of task-based assessments, including varied speaking tasks and integrated skills assessments, is needed to capture L2 development more comprehensively. It is well established that multiple tasks per skill are necessary for making generalizable inferences about a learner's language performance, as proficiency in one task does not necessarily transfer to another task with different content or complexity (Norris, 2016). To fully assess what students learn within a task-based syllabus, future research should examine a more extensive repertoire of language use, incorporating pragmatic competence (Timpe-Laughlin, 2024) alongside other linguistic dimensions. To achieve this, future studies should integrate both formative and summative classroom-based assessments throughout the semester, in addition to end-of-semester language tests. Moreover, a stronger alignment between instructional standards, large-scale external exams, and classroom-based task-based assessment practices is essential to accurately describe and promote student learning (Van Gorp, 2022). Strengthening this alignment will contribute to a more robust empirical foundation for TBLT, ensuring its validity and effectiveness in real-world educational settings.

Conclusion

This study is unique in that it demonstrates that TBLT not only fosters students' ability to engage in real-world communicative tasks but also supports the development of linguistic knowledge, even within test-driven educational contexts. The extended duration of the intervention and the use of multiple assessment measures provide robust evidence that TBLT can outperform structural approaches over time. Furthermore, the growth pattern of task completion mirrored the overall growth pattern and, to a large extent, the growth pattern of the individual linguistic criteria. This opens interesting perspectives for teachers to use task completion as a passkey criterion to gauge quickly and efficiently the overall development of students in their classroom.

These results reinforce calls within the TBLT literature to move beyond short-term interventions by demonstrating the longitudinal comparative effectiveness of strong-form TBLT against task-supported instruction. Furthermore, by evaluating outcomes through both task-based language assessment (TBLA) and discrete-point measures of explicit linguistic knowledge, our findings highlight the necessity of refining assessment practices to capture a comprehensive picture of L2 development. Ultimately, this study positions TBLT as a highly viable pedagogical approach capable of satisfying both the pedagogical goal of building real-world communicative competence and the institutional demands of high-stakes testing environments, such as those in China.

Data availability statement. The experiment in this article earned Open Data and Open Materials badges for transparent practices. The data files are available on https://osf.io/t8dbv/overview.

Acknowledgments. The authors acknowledge the use of ChatGPT-5.5 (OpenAI) during the preparation and revision of this manuscript to improve the clarity and readability of the language. All suggestions

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

generated by the tool were carefully reviewed, edited, and verified by the authors, who take full responsibility for the final content of the manuscript.

Competing interests. The authors declare no competing interests.

References

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476. https://doi.org/10.1191/0265532202lt240oa

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2021). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *43*(1), 2–24. https://doi.org/10.1017/S0272263120000327

Boers, F., & Faez, F. (2026). Meta-analysis to estimate the relative effectiveness of TBLT programs: Are we there yet? *Language Teaching Research*, *30*(3), 1525–1543. https://doi.org/10.1177/13621688231167573

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, *46*(4), 1171–1178. https://doi.org/10.2307/2532457

Bryfonski, L., & McKay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research*, *23*(5), 603–632. https://doi.org/10.1177/1362168817744389

Butler, Y. G. (2011). The implementation of communicative and task-based language teaching in the Asia-Pacific region. *Annual Review of Applied Linguistics*, *31*, 36–57. https://doi.org/10.1017/S0267190511000122

Cobb, M. (2010). *Meta-analysis of the effectiveness of task-based interaction in form-focused instruction of adult learners in foreign and second language teaching* (Unpublished doctoral dissertation). University of San Francisco. https://repository.usfca.edu/diss/389

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic Press.

Crowther, D., & Isbell, D. R. (2024). Second language speech comprehensibility: A research agenda. *Language Teaching*, *57*, 533–549. https://doi.org/10.1017/S026144482300037X

Dai, Y., & Zheng, L. (2014). *Shiji shangwu Yingyu zonghe jiaocheng: Jichu pian II [Century business English integrated course: Foundation II]* (5th ed.). Dalian University of Technology Press.

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15–32). Routledge.

East, M. (2021). *Foundational principles of task-based language teaching*. Taylor & Francis.

Ellis, R. (2019). Towards a modular language curriculum for using tasks. *Language Teaching Research*, *23*(4), 454–475. https://doi.org/10.1177/1362168818765315

Ellis, R. (2024). Task-based and task-supported language teaching. *International Journal of TESOL Studies*, *6*(3), 1–13. https://doi.org/10.58304/ijts.20240401

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. Routledge Taylor & Francis Group.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2020). *Task-based language teaching: Theory and practice*. Cambridge University Press.

Ganjouee, A. A., Ghonsooly, B., & Fatemi, A. H. (2018). The impact of task-based instruction on the enhancement of Iranian intermediate EFL learners' speaking skill and emotional intelligence. *Applied Research on English Language*, *7*(2), 195–214.

Geranpayeh, A. (2003). A quick review of the English Quick Placement Test. *Research Notes*, *12*, 8–10.

González-Lloret, M., & Nielson, K. B. (2015). Evaluating TBLT: The case of a task-based Spanish program. *Language Teaching Research*, *19*(5), 525–549. https://doi.org/10.1177/1362168814541745

Harrell, F. (2020, September 20). Violation of proportional odds is not fatal. *Statistical Thinking*. https://www.fharrell.com/post/po

Harris, J., & Leeming, P. (2022). The impact of teaching approach on growth in L2 proficiency and self-efficacy: A longitudinal classroom-based study of TBLT and PPP. *Journal of Second Language Studies*, *5*(1), 114–143. https://doi.org/10.1075/jsls.20014.har

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Hu, G. (2005). Contextual influences on instructional practices: A Chinese case for an ecological approach to ELT. *TESOL Quarterly*, *39*(4), 635–660. https://doi.org/10.2307/3588525

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jin, Y. (2019). Testing tertiary-level English language learners: The College English Test in China. In I. Lily, W. Su, C. J. Weir, & J. R. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 101–130). Routledge.

Keck, C. M., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). John Benjamins.

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, *38*(4), 602–626. https://doi.org/10.1177/0265532221994052

Kuiken, F., & Vedder, I. (Eds.). (2022). The assessment of functional adequacy in language performance (Special Issue). *TASK: Journal on Task-Based Language Teaching*, *2*(1), 1–158. https://doi.org/10.1075/task.21009.kui

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196. https://doi.org/10.1017/S0272263116000085

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Loewen, S., & Sato, M. (2021). Exploring the relationship between TBLT and ISLA. *TASK: Journal on Task-Based Language Teaching*, *1*(1), 47–70. https://doi.org/10.1075/task.00003.loe

Long, M. H. (2000). Focus on form in task-based language teaching. In R. L. Lambert & E. Shohamy (Eds.), *Language policy and pedagogy* (pp. 179–192). John Benjamins.

Long, M. H. (2005). *Second language needs analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667299

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. John Wiley & Sons.

Long, M. H. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics*, *36*, 5–33.

Long, M. H., & Crookes, G. (1993). Units of analysis in syllabus design: The case for task. In G. Crookes & S. M. Gass (Eds.), *Tasks in a pedagogical context* (pp. 9–54). Multilingual Matters.

Ma, L., & Pan, L. (Eds.). (2011). 商务跨文化交际 [*Intercultural communication for business*]. Xi'an Jiaotong University Press.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Newton, J. (2025). The coursebook in TBLT: Lost cause or launching pad? In M. East (Ed.), *Broadening the horizon of TBLT: Plenary addresses from the second decade of the International Conference on Task-Based Language Teaching* (pp. 60–79). John Benjamins.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, *34*(2), 257–271. https://doi.org/10.1177/0267658316684904

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230–244. https://doi.org/10.1017/S0267190516000027

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (No. *18*). National Foreign Language Resource Center.

Norris, J. M., & Davis, J. M. (2021). Evaluating task-based language programs. In M. Ahmadian & M. Long (Eds.), *The Cambridge handbook of task-based language teaching* (pp. 529–548). Cambridge University Press. https://doi.org/10.1017/9781108868327

Norris, J. M., & East, M. (2021). Task-based language assessment. In M. Ahmadian & M. Long (Eds.), *The Cambridge handbook of task-based language teaching* (pp. 505–528). Cambridge University Press. https://doi.org/10.1017/9781108868327.029

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Pomarolli, G., Artoni, D., Boschiero, M., & Piccinin, S. (2023). Inclusive and accessible teaching of Russian L2 to absolute beginners. *Instructed Second Language Acquisition*, *7*(2), 166–193. https://doi.org/10.1558/isla.26047

Qi, S. (2023). *Task-based vs. task-supported language teaching in a Chinese class for beginners* (Unpublished doctoral dissertation). Columbia University.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 3–38). John Benjamins.

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, *38*(6), 906–911. https://doi.org/10.1093/applin/amw052

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, *55*(3), 866–900. https://doi.org/10.1002/tesq.3027

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Palgrave Macmillan.

Shintani, N. (2020, October). *Comparing task-based language teaching and traditional instruction: Findings and methodological issues.* ESCR-JSLARF Online Symposium, Online. https://www.youtube.com/watch?v=o1G3YWQLCyA

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

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

Sun, Y. (2016). *Context, construct, and consequences: Washback of the College English Test in China*. Doctoral dissertation, Queen's University, Kingston, Canada. Retrieved from https://qspace.library.queensu.ca/handle/1974/13985

Suzuki, Y. (2023). *Practice and automatization in a second language*. Routledge.

Timpe-Laughlin, V. (2024). Pragmatically speaking: Contexts, forms, and functions in L2 speaking assessment. In L. Davis & J. M. Norris (Eds.), *Challenges and innovations in speaking assessment* (pp. 149–170). Routledge. https://doi.org/10.4324/9781315201504-11

UCLES – University of Cambridge Local Examinations Syndicate (2001). *Quick placement test*. Oxford University Press.

Van den Branden, K. (2022). *How to teach an additional language: To task or not to task?* John Benjamins.

Vandommele, G., Van Gorp, K., Van den Branden, K., & De Maeyer, S. (2022). Assessing the development of L2 speaking skills in inside-school and outside-school settings through growth modeling. *Applied Linguistics*, *43*(6), 1094–1115. https://doi.org/10.1093/applin/amac024

Van Gorp, K. (2022). Designing a classroom-based task-based language assessment framework for primary schools: Blurring the lines between teaching, learning and assessment. In M. H. Long & M. J. Ahmadian (Eds.), *The Cambridge handbook of task-based language teaching* (pp. 585–602). Cambridge University Press. https://doi.org/10.1017/9781108868327

Van Gorp, K., Coss, M., Guan, X., Burton, J. D., & Tuzcu, A. (2026a, June 18). Effectiveness of TBLT in higher education in China. https://doi.org/10.17605/OSF.IO/T8DVB

Van Gorp, K., Guan, X., & Coss, M. (2026b). From needs to tasks: A research-informed five-stage framework for task-based syllabus development. In R. Gilabert & A. Malicka (Eds.), *From task-based needs analysis to task and syllabus design: Insights from across educational contexts* (pp. 197–222). John Benjamins.

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford University Press.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, *47*(4), 762–789. https://doi.org/10.1002/tesq.73

Appendix A: task sheet

Speaking Task: A business presentation of a product

Time: 3 minutes

Now, I'm going to give you a business situation and I'd like you to **talk about it for 2 minutes**.

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Table A1. New product presentation sheet.

1. Pros • A foldable phone: Unfolded (8-inch); folded (6.6-inch) • SuperCharge (85% in 30 minutes) • Faster wireless reverse charge	**2. Cons** • Lack Google Play services due to the US ban
3. Specifications • Fingerprint • Supports the 5G, 4G, 3G, and 2G networks • Memory: RAM: 8GB; Storage: 512GB • Price: 2500 EURO	**4. Basic company information** • Huawei was established in 1987(33 years until now) • Supplies products in over 140 countries • The US ban of Huawei products and services

Before your talk, you'll **have one minute to think about** what you are going to say and outline main points you want to make and you can make notes if you wish. Do you understand?

Situation: You are a sales representative at Huawei Technologies Co., Ltd. Recently your company launched a new smartphone—Huawei Mate Xs (see Table A1). You are asked to ***give a presentation about the new product*** **to a business client who is from Europe**. Your task is to sell the new smartphone, that is**, to convince the business client to buy the product for the European market**.

Appendix B: speaking task rubric

Table B1. Speaking task rubric.

		Task Completion	Comprehensibility	Organization	Language Use
General		To what extent was the task completed? (specific task requirements)	How easily is the speaker understood? (How much effort is required on the part of the listener?)	How smoothly is the message delivered? (fluency + organization/ coherence/cohesion)	What quantity and quality of language is used to accomplish the task? Do language-related miscommunications occur?
Task-specific criteria		To what extent (fully or minimally) are all of the following task points mentioned: - **Introduction** (person and product) - **Product strengths** (at least 4: combo of features/specs) - **Product** price (at least the word "price" "cost/s," etc. is sufficient) ***fully addressed** = **elaborated** beyond the content provided on the task sheet	Assuming the audience has only the background knowledge that they are 'attending a sales pitch of a new phone', how much of the presented information is understandable, and with what amount of effort on the part of the listener (minimal or intensive)?	How does the pace and organization of the performance impact the quality—to what extent is the performance smooth (without extensive pauses between ideas and without interruption/re-starting) and easy to follow (progression between topics/sub-topics)?	How varied are the structures being used (single simple sentences, single compound sentences, strings of sentences, extended discourse); questions…? How precise and varied is the vocabulary being used? To what extent, if any, does the quality/quantity of the speaker's language cause miscommunication (obscuring the speaker's intended meaning or communication of inaccurate information)?
Excellent	7	The task was completed very successfully. **All** of the task points were **fully*** addressed.	Completely **comprehensible** performance with minimal (or no) listener effort.	Delivery is **fluent** (minimal or no disfluencies). Ideas are **smoothly** connected by content and linguistic transitional elements.	Demonstrates a **wide range** of language (structures and vocabulary as appropriate). **Minimal imprecise or inaccurate use** of vocabulary or grammatical structures that does not obscure meaning.
	6				
Good	5	The task was completed successfully. **All** of the task points	**Mostly comprehensible**, performance rarely necessitates special listener effort.	Delivery is **mostly fluent**, disfluencies (hesitations, re-starting,	Demonstrates a **range** of language in which **some imprecise or inaccurate use** of vocabulary or

(*Continued*)

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

Table B1. (*Continued*)

		Task Completion	Comprehensibility	Organization	Language Use
		were at least **minimally** addressed.		repetitions) are generally **not distracting from the message**. Relationships between statements (connectedness of one idea to the next) may at times not be immediately clear. Some transitional/ connecting elements are used.	grammatical structures may occur, but they generally do not obscure the speaker's intended meaning.
	4		*Occasionally requires listener effort*		
Adequate	3	The task was completed moderately successfully. **Most** of the task points were addressed.	**Somewhat** comprehensible, performance sometimes necessitates **special listener effort** (e.g., filling in some gaps).	**Disfluencies distract** from the message. The performance is made up of mainly isolated (short) statements with **no clear connection from one to the next** —the sentences could be randomly re-ordered without changing the message. Message is **somewhat disorganized** and **(somewhat) hard to follow.**	**Somewhat limited range and control** of language *(e.g., simple sentences, high-frequency words)*, that sometimes obscures the speaker's intended meaning, and may lead to misinformation (inaccurate information being understood by the listener). May rely heavily on **practiced language or formulaic expressions**.
	2		*Often requires listener effort*		
Weak	1	The task was completed unsuccessfully. Only a **few** of the task points were addressed.	**Minimally comprehensible**, performance necessitates **constant listener effort** (intensive listening and thinking, and filling in the gaps) to achieve minimal comprehensibility (would be incomprehensible to a non-sympathetic interlocutor).	**Highly disorganized and highly disfluent**, both **severely impacting the message** to the point that much of the content is lost in the performance.	**Range and control** of language are **severely limited and frequently obscure the speaker's intended meaning** and lead to inaccurate information. Reads verbatim exclusively, or near-exclusively, from the "New product presentation sheet" (And is not using language not written on the sheet).
	0	The task was not completed at all—the response was entirely **off task.**	**Mostly or completely Incomprehensible** even with **intense listener effort**.		

Appendix C: quality control analyses

QPT analysis

Although the tests were the same across the two time points, the QPT test results exhibited somewhat low internal consistency ($\alpha_{\text{Pretest}} = .54$, $\alpha_{\text{Posttest}} = .56$). We thus conducted a preliminary item-level analysis to confirm that the test functioned as intended in the two administrations. We checked item-level test statistics for the QPT using a 1PL Rasch model for each administration to confirm that the test functioned properly.

We ran a 1PL Rasch model in R. We ran the model on all 60 items. We did not combine all items into one analysis because the test takers' abilities (thetas) were in flux across the two sessions. The Wright maps are displayed in Figure C1 for Time 1 and Figure C2 for Time 2. R produces these slightly differently from Winsteps. On the left is a histogram of the ability levels of the test takers, and on the right are the difficulty estimates of the items, which are numbered on the x-axis. They roughly get more difficult as the test goes on

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

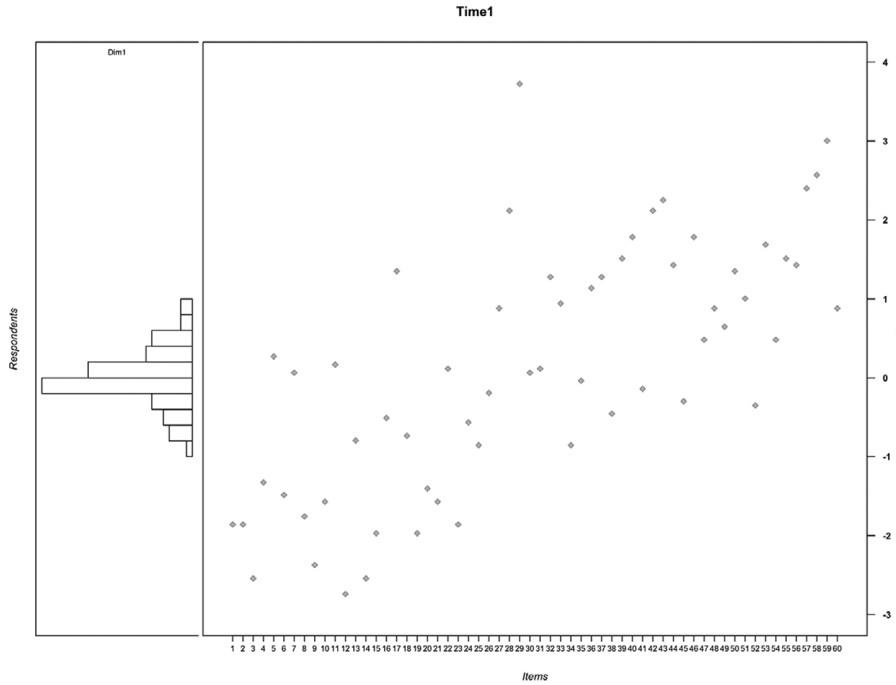


Figure C1. Wright map of the QPT data at Time 1.

for all tests. The item-difficulty patterns correlated at .95 for the two administrations. For the test takers, the histograms show a broader spread of ability estimates in the second test administration compared to the first.

Item infit statistics showed that all items were within Linacre's (2002) acceptable bounds of .50 and 1.50 and even Bond and Fox's (2015) more conservative range of .75 to 1.3. The plots in Figure C3 show the distribution of the infit values across the two tests. The x-axis contains the 60 individual items and the y-axis identifies the infit statistics.

Person infit statistics showed that most participants were also within Linacre's (2002) acceptable bounds of .50 and 1.50. In the first administration, only one test taker (ID = 1) had an outlying fit statistic, and in the second administration three (IDs = 1, 7, and 5). In these four cases, fit was 1.64–1.67, which indicates slightly more unpredictable response patterns (such as answering a question that is calculated as easy for one's ability incorrectly). However, these statistics are all below 2, so this was relatively unproblematic for the model. The plots in Figure C4 show the distribution of the person infit values across the two tests. The x-axis contains the 80 individuals and the y-axis identifies the infit statistics.

TBLA analysis

We checked the quality of the TBLA rater data by running a partial-credit many-facet Rasch model (MFRM) using Facets (v.3.83.6). MFRM assumes unidimensionality across scales (that is to say, that all scales load onto one factor), of which we found evidence by running a one-factor confirmatory factor analysis using the *lavaan* package in R (v.0.6-18). A one-factor model was a reasonable fit for the rating scale data, suggesting that the scales measured the same construct, thus meeting the unidimensionality assumption (CFI > .95, TLI > .95, SRMR < .05; Hu & Bentler, 1999). RMSEA was 0.13, which is greater than the cutoff of .05, but we continued with the analysis given that CFI, TLI, and SRMR indices were within acceptable limits. MFRM showed that the scale categories fit the model with model infit and outfit values ranging from .8 to 1.24, well within Bond and Fox's (2015) conservative range of .75 to 1.3. The separation (3.11) and strata (4.49) values showed that the raters were distinct in their ratings (separation reliability = .91). The three raters showed

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

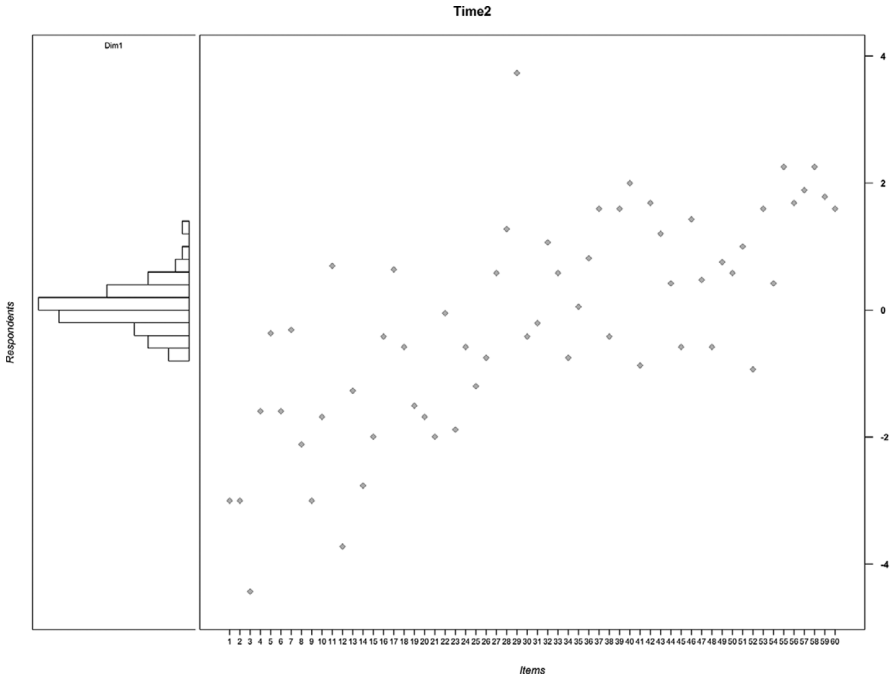


Figure C2. Wright map of the QPT data at Time 2.

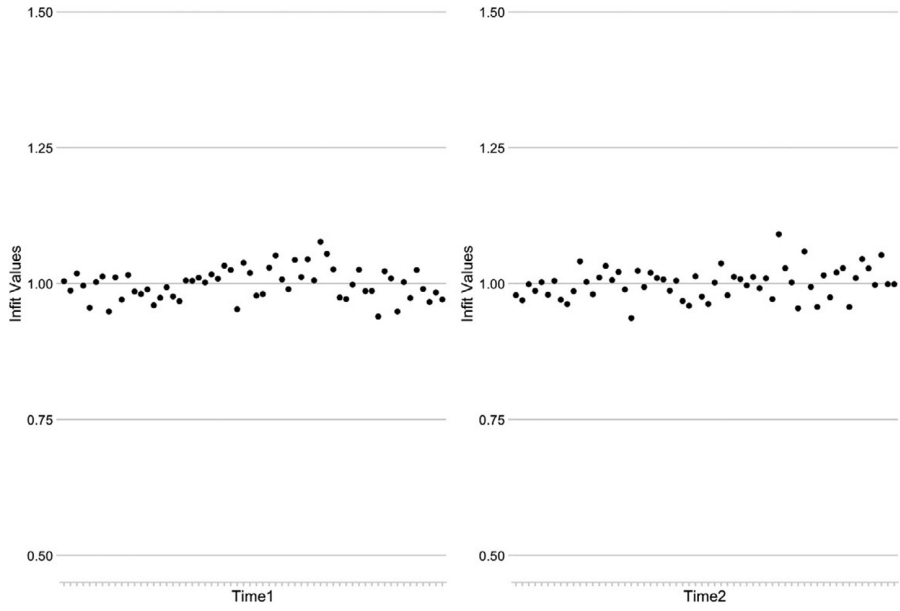


Figure C3. Item fit distributions.

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press

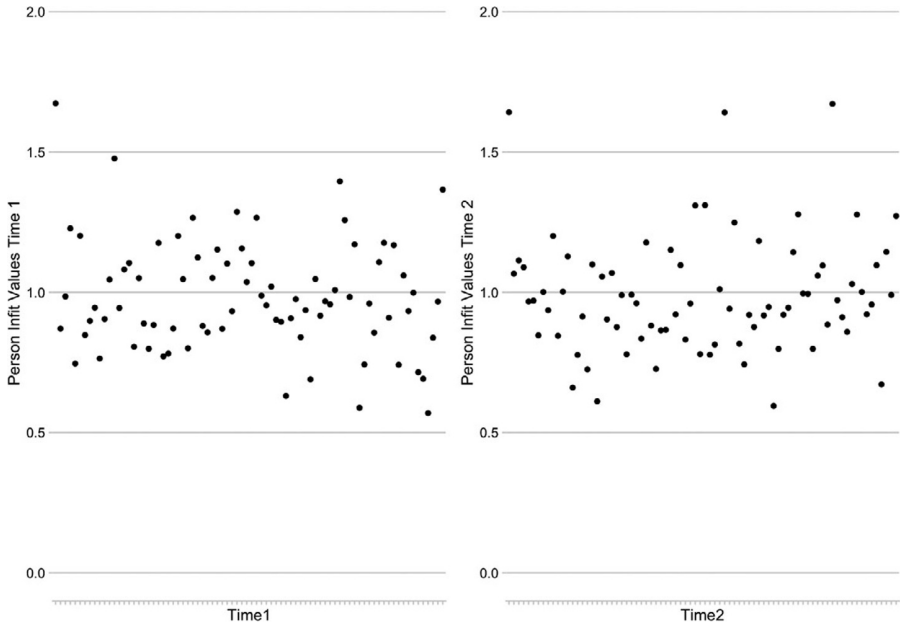


Figure C4. Person fit distributions.

values that fit the model (.84–1.32) using Linacre's (2002) more lenient criterion of model fit (.5–1.5), demonstrating a good level of interrater reliability (exact agreement = 50%). The raters did not, however, demonstrate a high degree of difference in severity, with an absolute difference between the strictest (.11) and most lenient (−0.16) rater being .27 logits. The majority of the test takers also fit the model (85%), demonstrating approximately four levels of ability (separation = 3.77, reliability = .93).

Cite this article: Van Gorp, K., Coss, M. D., Guan, X., Burton, J. D., & Tuzcu, A. (2026). Task-based or task-supported language education?: A longitudinal comparative effectiveness study of two pedagogical approaches. *Studies in Second Language Acquisition,* 1–33. https://doi.org/10.1017/S0272263126102034

https://doi.org/10.1017/S0272263126102034 Published online by Cambridge University Press