

Impact of Online Task-Based Language Teaching on Criterion-Referenced Scores of Thai EFL Learners

WATCHARAPHONG SOONGPANKHAO

SUWANNA KEMDANG*

GARY EDWARD GRAY

Faculty of Liberal Arts, Rajamangala University of Technology Phra Nakhon, Thailand

*Corresponding author email: suwanna.k@rmutp.ac.th

Article information	Abstract
Article history: Received: 28 Nov 2025 Last revised: 22 Sep 2026 Accepted: 25 Sep 2026 Available online: 5 Oct 2026 Keywords: Task-based language teaching Criterion-referenced assessment Online learning EFL Thailand	<i>This mixed-methods study compared online and on-site task-based language teaching (TBLT) for an occupationally relevant giving-directions task. Eighty second-year English for International Communication majors (CEFR A2-B1) at a university in Bangkok were randomly assigned to online (n = 40) or on-site (n = 40) conditions. The task and eight criterion-referenced performance indicators were grounded in a prior task-based needs analysis that identified giving directions as a critical workplace task for Thai university graduates (Soongpankhao & Lambert, 2026). Both groups completed a pretest, a 90-minute TBLT module, and a posttest. Online participants used Google Meet and digital maps, whereas on-site participants interacted face to face with printed maps; consequently, the study compares the two implemented instructional packages rather than isolating communication channel from material format. Both groups improved substantially (online $d = 1.52$; on-site $d = 1.89$). On-site participants obtained higher posttest scores by 0.88 points on the eight-point checklist, $t(78) = 3.45$, $p = .001$, $d = 0.69$. Criterion-level contrasts favored the on-site condition for acknowledging the customer problem, clarifying transportation mode, explaining routes through landmarks, and confirming understanding; the online condition was stronger for quantifying distance. Questionnaire results indicated higher ratings for social interaction, immediate feedback, and technical ease on site, while online learners valued flexibility but frequently reported connectivity problems. The findings show that overall scores can conceal criterion-specific differences and support modality-sensitive task design, interactional scaffolding, and equitable technical provision.</i>

INTRODUCTION

Technology-mediated instruction is now a routine component of language education, and task-based language teaching (TBLT) is increasingly implemented through synchronous video platforms, shared documents, mobile applications, and other digital tools. In TBLT, however, the central question is not simply whether learners acquire decontextualized language knowledge. It is whether they become better able to complete meaningful target tasks

through the second language. Comparisons of online and face-to-face instruction therefore require performance measures that are directly aligned with the task rather than broad measures of general achievement.

Criterion-referenced assessment is well suited to this purpose because performance is judged against predefined indicators of successful task completion rather than against the performance of other learners (Giraldo, 2020; Long, 2015; Norris, 2009). Such indicators can show not only whether overall performance improves but also which components of a task are supported or constrained by a particular instructional environment. A single total score may obscure these differences. For example, an online map may support precise distance estimates while synchronous video may provide fewer readily available embodied cues than physical co-presence. A criterion-level analysis can reveal this uneven profile and provide more useful information for subsequent task design.

This issue is particularly important for complex service encounters because the constituent criteria place different demands on learners. A standard greeting can be prepared and retrieved as a short conventional sequence, whereas a route explanation requires ongoing spatial reasoning, monitoring of the interlocutor, selection of salient landmarks, sequencing, and repair. A comprehension check depends on the learner noticing whether the customer has followed the explanation. Treating these actions as interchangeable indicators of one underlying score can conceal the relationship between task demands and the resources available in each condition. A criterion profile, by contrast, can show whether an instructional package supports routine discourse, information precision, interaction management, or some combination of these.

The English as a foreign language (EFL) context of Thailand offers an important setting for examining this issue. A recent task-based needs analysis (TBNA) of English for International Communication (EIC) graduates working in Thai service industries identified eight recurring workplace task types. Giving directions was ranked highest: 68% of in-service graduates rated it extremely or very important. The TBNA also identified eight sub-task steps and converted them into a criterion-based assessment checklist for workplace-relevant performance (Soongpankhao & Lambert, 2026). This empirical link between workplace needs, task design, and assessment provides a stronger basis for comparing instructional conditions than a task selected only for classroom convenience.

The same TBNA also showed that workplace informants prioritized effectiveness, clarity, politeness, and natural responsiveness above sophisticated grammar and vocabulary. This ordering supports a functional interpretation of success: learners must complete the service encounter in a way that the customer can follow. It does not imply that linguistic quality is unimportant, but it explains why the primary assessment in the present study focuses on whether the task steps are accomplished. The resulting checklist therefore represents a deliberate link between occupational evidence and classroom evaluation rather than a general-purpose speaking rubric.

Existing research shows that technology-mediated tasks can support language development and engagement when technological resources are integrated purposefully (González-Lloret, 2020; Lai & Li, 2011; Soongpankhao et al., 2023). At the same time, face-to-face and synchronous video interaction can differ in affective engagement, social presence, access to nonverbal cues, and the effort required to coordinate talk and shared resources (Aubrey & Philpott, 2023; Stickler, 2022). These studies establish that both modes can support task work, but they do not fully explain whether specific components of an occupational task respond differently to the two instructional packages.

The present study addresses this problem by comparing online and on-site versions of the same needs-based TBLT module. It examines total pretest-posttest gains, criterion-specific posttest differences, and learners' reported experiences. The intended contribution is therefore not the broad claim that one mode is universally superior. Rather, the study tests whether the resources and constraints associated with each implemented condition are reflected in a differentiated profile of task performance. This distinction is important for teachers designing online, on-site, or blended task sequences because the instructional environment may align more closely with some performance criteria than with others.

LITERATURE REVIEW

Needs-based tasks and criterion-referenced assessment

A defining principle of needs-based TBLT is that target tasks provide the unit of analysis for syllabus design, instruction, and assessment (Long, 2015). Target tasks are activities that learners need to perform outside the classroom, while pedagogic tasks are sequenced classroom approximations designed to prepare learners for those activities. A TBNA identifies the relevant target tasks, the steps involved in completing them, the conditions under which they occur, and the criteria by which successful performance is judged. When these elements are translated into assessment indicators, the resulting test can make a direct inference about learners' ability to complete a meaningful task.

This sequence creates a validity chain from the target domain to the classroom task and finally to the score interpretation. The target task must first be shown to matter in the learners' future context; the pedagogic sequence must then preserve its central communicative demands; and the assessment criteria must represent the actions that knowledgeable domain participants recognize as successful. If any link is weak, a score may be reliable yet have limited occupational meaning. Needs analysis is therefore not merely a preliminary survey. It determines what is taught, how performance is elicited, and what evidence counts as improvement. Ellis (2019) similarly argues for modular task-based curricula in which coherent task units connect pedagogic work with program-level design.

Soongpankhao and Lambert (2026) provide the empirical foundation for the task used in the present study. Their six-cycle TBNA drew on job-placement records, interviews with graduates and workplace managers, a means analysis, a confirmatory survey of 70 in-service graduates,

and an analysis of target discourse based on authentic workplace calls. The analysis identified giving directions as the most critical of eight task types and specified eight recurring sub-task steps: greeting the customer, acknowledging the problem, identifying the mode of transportation, explaining the route through visible landmarks, clarifying distance, offering additional support, confirming understanding, and closing the conversation. These steps were converted into a pass-fail criterion-based assessment. The present study evaluates how performance on this same task and its constituent criteria responds to two instructional conditions.

The analysis of target discourse in that study is also important for instructional authenticity. Recorded workplace calls showed that direction giving was not a simple list of imperatives. Employees moved through an opening, acknowledgement of the caller's need, identification of transport mode, repeated route segments, distance estimates, optional support, comprehension checks, and a closing. Some steps recurred within a single call as the employee monitored the customer and elaborated the route. The module used in the present study drew on this discourse structure and on the earlier development work reported by Soongpankhao (2023), allowing input and role-play tasks to model the functional sequence rather than isolated phrases.

Criterion-referenced task assessment differs from norm-referenced testing because it interprets performance in relation to explicit requirements rather than rank order (McNamara, 1996). In task-based assessment, the criteria should represent successful task completion and should be sufficiently transparent to support instruction and feedback (Giraldo, 2020; Norris, 2009). A task-specific checklist also limits the tendency to append broad linguistic traits that may not be central to the target activity. Its disadvantage is that a binary criterion captures whether a functional step is completed but not the full quality of the language used. For this reason, the checklist in the present study is interpreted as a measure of task completion, not as a comprehensive measure of linguistic proficiency.

A criterion checklist also changes the role of reliability. Raters must decide whether observable evidence of a task step is present, which can be more transparent than assigning a global impression score, but ambiguity can still arise. An acknowledgement may be explicit or implicit; a landmark may be mentioned without being integrated into a usable route; and a comprehension check may be formulaic without genuinely inviting a response. Clear descriptors, shared examples, rater familiarization, and double scoring are therefore necessary. Internal consistency is relevant to the total score, but it should not be taken to mean that all criteria measure an identical trait. In a task-based interpretation, heterogeneous criteria can still belong together because they jointly define successful completion of the target activity.

Criterion-level analysis and the modality question

Most comparisons of instructional modality emphasize an overall achievement score, a general proficiency measure, or learner attitudes. Such analyses answer whether conditions differ on average, but they provide limited guidance when the task itself combines qualitatively

different operations. A criterion-level analysis treats the total as a profile. It asks whether an observed overall difference is concentrated in interaction-dependent components, information-precision components, routine discourse, or another subset of the task.

This approach does not assume that a single feature of the learning environment directly causes performance on a particular criterion. Instead, criterion-level results can help identify possible relationships between learners' performance and their experiences in each condition. For example, if online learners perform less well on comprehension checks and also report difficulty receiving immediate feedback, these findings together suggest that limited interactional feedback may have contributed to their performance, although this cannot be confirmed from the present study. Similarly, if online learners perform better on distance estimates and report that screen-shared maps were helpful, the digital map may have supported this aspect of the task. Such patterns can therefore be used to develop specific hypotheses for future research rather than to make broad claims that one instructional mode is generally better than another.

Technology-mediated and face-to-face TBLT

Technology-mediated TBLT retains the central characteristics of a task—a primary focus on meaning, a communicative gap, learner use of available linguistic resources, and a nonlinguistic outcome—while redistributing interaction across digital tools and channels (González-Lloret & Ortega, 2014; Lai & Li, 2011). Digital platforms can provide shared visual resources, opportunities for repeated input, text chat, recordings, and flexible access. These affordances can support task preparation and performance, but they also create additional requirements, including interface management, device access, connectivity, and the coordination of talk with on-screen materials (Stickler, 2022; Winke & Goertler, 2008).

Synchronous video interaction is especially complex because it combines spoken communication with a technologically framed visual field. Participants may see only the head and shoulders of a partner, experience delay or audio overlap, and divide attention among the video window, task instructions, chat, and a shared map. At the same time, digital tools can externalize information that would otherwise need to be held in memory. A learner can enlarge a map, inspect station names, or compare route segments while speaking. The relevant question is therefore not whether technology adds or removes support in general, but which support becomes available for which part of the task and what new coordination cost accompanies it.

Empirical work suggests that online and face-to-face task performances should not be treated as interchangeable experiences. Aubrey and Philpott (2023), for example, found differences in learner engagement when spoken decision-making tasks were completed face to face and through synchronous video, even where both conditions supported task performance. González-Lloret (2020) similarly argues that online collaborative tasks require deliberate orchestration of interaction and resources rather than a direct transfer of classroom procedures to a video platform. These findings indicate that the effect of a delivery condition may depend on the component of performance being examined.

The distinction between task equivalence and procedural identity is useful here. Two conditions can share the same goal, information gap, time allocation, and assessment while differing in the means through which learners access information and coordinate with a partner. Exact procedural identity may be impossible because tools are part of the environment being studied. What researchers can do is specify the common task core, document all condition-specific resources, and limit causal claims accordingly. This transparency is particularly important when a material such as a map is central to performance.

Research in Thai EFL settings also demonstrates the potential of combining tasks and technology. Durongbhandhu and Suwanasilp (2023) reported vocabulary gains from computer-assisted, task-based lessons, while Soongpankhao et al. (2023) found that goal tracking influenced engagement in language use within a fully online TBLT module. Such studies support the viability of online task work, but they also point to the importance of implementation features. An online condition is not a single variable: it may include a video platform, digital worksheets, shared maps, messaging applications, and particular forms of teacher support. Likewise, an on-site condition combines co-presence with physical materials and locally available interactional resources.

Earlier Thai classroom research also found that teachers and learners could respond positively to a task-based course when tasks were localized and supported appropriately (McDonough & Chaikitmongkol, 2007). Together, this work suggests that the relevant contrast is not between a culturally suitable physical classroom and an inherently unsuitable online environment. It is between concrete designs that vary in task sequencing, technological demands, teacher mediation, and opportunities for interaction. The present study narrows the comparison by holding the task objective and pedagogic sequence constant while documenting the resources that could not be held constant.

Learner experience and modality-specific constraints

Learner experience is relevant because engagement, access, and coordination shape the conditions under which task performance develops. Face-to-face interaction gives participants immediate access to gaze, gesture, body orientation, and rapid repair, whereas synchronous video can reduce or rearrange these cues. Online environments may nevertheless offer flexibility, private preparation space, and convenient access to digital information. Their benefits depend on stable connectivity, usable devices, and familiarity with the platform (Stickler, 2022; Winke & Goertler, 2008). Learner perceptions do not by themselves demonstrate learning, but they can help explain why particular criteria improve more readily in one condition than another.

Experience data are most informative when tied to identifiable aspects of task performance. A general statement that online learning is convenient does not explain a route score, but a report that screen sharing made station names and distances easier to inspect may be relevant to precision criteria. Likewise, a preference for face-to-face interaction becomes analytically useful when learners describe immediate signals of confusion, gesture, or rapid clarification and those same functions appear in the criterion profile. The present study therefore uses

perceptions as explanatory evidence while avoiding the assumption that satisfaction and learning are equivalent.

The present study therefore integrates criterion-referenced scores with questionnaire data. Quantitative performance results indicate what learners accomplished; Likert-scale and open-ended responses provide evidence about how they experienced the two instructional packages. The integration of these sources is intended to produce a more defensible interpretation than either data type could provide alone (Creswell & Creswell, 2023).

Research questions

The study addressed the following research questions:

1. To what extent do conditions (online vs. on-site) affect learners' total criterion-referenced posttest scores after accounting for pretest performance?
2. Which individual task-performance criteria differ between the online and on-site conditions?
3. How do learners in the two conditions evaluate their learning experience?

METHOD

Research design

The study used a convergent mixed-methods, randomized pretest-posttest design. Participants were randomly assigned to an online or on-site instructional condition, and both groups completed the same target task before and after a parallel 90-minute TBLT module. Quantitative data consisted of total and criterion-level task scores and Likert-scale questionnaire responses. Open-ended questionnaire responses were analyzed qualitatively and integrated with the quantitative results in the discussion. Because communication mode and material format covaried—digital maps were used online and printed maps on site—the study estimates the effects of the two implemented instructional packages rather than a pure effect of communication mode.

The quantitative component was primary because the first two research questions concerned change in observable task performance. The qualitative component was explanatory: it was used to identify how learners understood the interactional, technical, and organizational features of each condition. The two strands were collected within the same intervention, analyzed separately, and brought together after the statistical results were established. This sequencing prevented open-ended comments from being used to redefine the performance criteria after the fact.

Participants and context

Participants were 80 second-year EIC majors at a public university in Bangkok, Thailand. Their English proficiency ranged from CEFR A2 to B1. A priori power analysis in G*Power 3.1, based

on a medium effect size ($d = 0.50$), $\alpha = .05$, and power = .80, indicated a minimum sample of 64 participants (Faul et al., 2009). The available cohort exceeded this requirement. Participants were randomly allocated in equal numbers to the online ($n = 40$) and on-site ($n = 40$) conditions.

The EIC program prepares students for communication-intensive employment in hospitality, retail, recreation, and related service sectors. Students typically enter the program after accuracy-oriented secondary or vocational English instruction and have limited experience completing extended oral service tasks. Most classroom communication in the program is in Thai, although English is used in language-focused courses. These contextual features made a tightly scaffolded, repeated task sequence appropriate for the participants and also justified allowing Thai responses in the experience questionnaire.

Table 1
Demographic characteristics of participants

Characteristic	Online (n = 40)	On-site (n = 40)
Age, M (SD)	20.5 (1.9)	20.1 (1.7)
Gender (female/male)	25/15	24/16
Prior TBLT experience, n (%)	11 (27.5%)	12 (30.0%)

Note. M = mean; SD = standard deviation; TBLT = task-based language teaching.

The cohort was broadly comparable across conditions in age, gender distribution, and prior exposure to TBLT (Table 1). Participation was voluntary, written informed consent was obtained, and participants could withdraw without penalty. The study was approved by the university's institutional review board (IRB-COE-052-2025).

Task rationale and instructional module

The target task required a service employee to give a customer directions to a Bangkok destination. Its occupational relevance and discourse structure were established in the TBNA and analysis of target discourse reported by Soongpankhao and Lambert (2026). The module used the eight sub-task steps identified in that study as its organizing sequence. Input included custom listening materials featuring Thai speakers giving directions in English, reading materials, maps, and repeated role-play performances. The instructional sequence alternated input-based work with interactive task performance so that learners encountered models of the target discourse and then applied the relevant information in new locations.

The module began with an initial task performance so that learners attempted the complete encounter before receiving targeted input. Listening and reading activities then directed attention to how workplace speakers opened the call, established the customer's location and transport mode, organized route segments, referred to landmarks, estimated distances, checked comprehension, and offered further support. Learners subsequently repeated the task with a new partner and destination. Repetition preserved the functional demands while changing the route information, reducing the likelihood that improvement reflected memorization of one map or one dialogue.

Three performance cycles were used to create repeated opportunities for planning, production, feedback, and adjustment within the 90-minute lesson. Partner changes required learners to coordinate with different interlocutors rather than settle into one pair routine. The customer and employee roles were alternated so that every learner experienced the information demands of both sides of the exchange, although only the employee role was scored. The pedagogic sequence followed the needs-analysis-to-evaluation development process described in Soongpankhao (2023).

The same task goals, discourse content, lesson sequence, planning time, and number of performances were used in both conditions. The online group met synchronously through Google Meet and used PDF maps, LiveWorksheets, screen-shared materials, and LINE for distributing files and messages. The on-site group met in physical classrooms and used printed versions of the same maps and worksheets. Both interlocutors in both conditions had access to a map. To reduce an obvious advantage from physical pointing, on-site participants were instructed not to trace the route for their partner on the paper map. Nevertheless, digital and printed maps offered different affordances; this limitation is considered explicitly in the interpretation.

Criterion-referenced task assessment

The pretest and posttest used the same eight functional criteria derived from the workplace-informed sub-task steps in Soongpankhao and Lambert (2026). Each criterion was coded pass (1) or fail (0), producing a total score from 0 to 8. Only the employee role was assessed because the customer role served to elicit the direction-giving performance. Three experienced EFL instructors reviewed the checklist for alignment with the task objectives and instructional content. Two independent raters double-scored 20% of the performances. Inter-rater agreement was acceptable (Cohen's kappa = .85), and internal consistency for the checklist was Cronbach's alpha = .78.

A pass indicated that the learner produced sufficient observable evidence of the function during the exchange. The checklist did not require a predetermined sentence or grammatical form. For example, transportation mode could be clarified through an open question, a choice question, or a confirmation, provided the learner established whether the customer was walking, driving, or using public transport. This functional scoring was consistent with the TBNA finding that workplace success was defined primarily through effectiveness and clarity rather than linguistic sophistication.

Table 2
Criterion-referenced assessment checklist

No.	Performance criterion	Score
1	Greets the customer with a standard greeting.	Pass / Fail
2	Acknowledges the customer's problem.	Pass / Fail
3	Clarifies the customer's mode of transportation.	Pass / Fail
4	Explains the route using visible landmarks.	Pass / Fail
5	Clarifies distances in minutes, metres, stops, streets, or other relevant units.	Pass / Fail

No.	Performance criterion	Score
6	Offers additional support.	Pass / Fail
7	Confirms that the customer understood key information.	Pass / Fail
8	Closes the conversation with a standard closing and thanks the customer.	Pass / Fail

Note. Pass = 1; Fail = 0. Total possible score = 8.

Exit task and experience questionnaire

The exit task was a dialogic role-play situated in a realistic Bangkok travel scenario. Online pairs completed the task through Google Meet with digital maps and worksheets; on-site pairs completed it face to face with printed materials. The task objectives and the employee-role assessment criteria were identical across conditions.

The post-module questionnaire contained parallel five-point Likert items (1 = strongly disagree to 5 = strongly agree) covering four dimensions: social interaction, immediate feedback, technical ease, and flexibility. Condition-appropriate wording was used while retaining the same underlying dimensions. Open-ended prompts asked learners to identify an effective element of the module and suggest improvements. Participants were permitted to respond in Thai so that limited English proficiency would not constrain the detail of their comments. The questionnaire was piloted with 10 students from the same program, and wording was revised for clarity.

The Likert items were intended to compare dimensions common to both conditions rather than to create a single satisfaction score. Social interaction addressed access to and engagement with the partner; immediate feedback addressed the speed with which misunderstanding could be noticed and resolved; technical ease addressed the effort required to use the available tools and materials; and flexibility addressed convenience of access and participation. The open-ended prompts allowed learners to qualify these ratings and identify condition-specific issues not anticipated in the closed items.

Procedure

Both conditions followed a 3.5-hour schedule consisting of a 60-minute pretest block, a 90-minute TBLT module, and a 60-minute posttest and questionnaire block (Table 3). The groups participated on separate days to reduce cross-condition contamination. Five experienced Thai EFL instructors used a common lesson plan and implementation briefing. Participants worked in groups of 8–10 and in changing pairs across the task cycles.

The common protocol specified the sequence of instructions, preparation time, task cycles, input activities, partner changes, and assessment procedures. The research team met before implementation to review the protocol and again after each session to record deviations or technical problems. Instructors could clarify procedures but were not to provide route language during an assessed performance. This standardization was intended to reduce teacher-level variation, although teacher effects were not modeled statistically.

Table 3
Pedagogic structure and timing

Pretest block	TBLT module	Posttest and questionnaire block
60 minutes Instructions, 5-minute planning, paired role-play, role change, transitions	90 minutes Three task-performance cycles, listening and reading input, repeated tasks with new destinations	60 minutes Instructions, 5-minute planning, paired role-play, role change, post-module questionnaire

Note. The 60-minute assessment blocks included instructions, preparation, both role turns, and transitions. Individual speaking time was not separately logged.

Before each performance, pairs received five minutes to review the instructions and map. They then completed both role turns, with a new destination used when roles changed so that a learner who had first heard the route as the customer did not reproduce the same information as the employee. During the module, learners completed three iterative task cycles with different partners and destinations. The study recorded the duration of the overall blocks but did not time each individual role-play; individual task duration therefore could not be examined as a covariate.

Data analysis

For Research Question 1, paired-samples *t* tests examined pretest-posttest change within each condition, and an independent-samples *t* test compared posttest totals. ANCOVA was used as a sensitivity analysis with posttest score as the dependent variable, condition as the fixed factor, and pretest score as the covariate. Cohen's *d* and 95% confidence intervals were reported to indicate practical magnitude. Normality, homogeneity of variance, and independence were checked before the analyses.

Effect sizes were interpreted alongside statistical significance because the sample was large enough for small differences to become detectable and because the educational relevance of a difference depends on its magnitude (Cohen, 1988). The ANCOVA did not replace the pretest-posttest comparisons; it provided a check on whether the posttest contrast remained after adjustment for initial score. Prior TBLT experience was examined as a potential moderator, and no condition-by-experience interaction was detected ($p = .312$).

For Research Question 2, each binary criterion was compared between conditions using the original independent-samples test plan. A Bonferroni-adjusted alpha of .006 (.05/8) controlled familywise Type I error across the eight criteria. Because the criteria were pass-fail indicators, criterion-level results are presented as the direction of the difference, test statistic, adjusted significance decision, and effect size rather than as if they were continuous proficiency scales. For Research Question 3, independent-samples *t* tests compared the four Likert dimensions. Open-ended Thai responses were coded from the Thai originals in NVivo 12 following Braun and Clarke's (2006) six-phase thematic procedure. English was introduced only when theme labels and aggregate summaries were prepared for reporting; no translated participant quotations are used in the article. Two coders independently coded a subset of responses, discussed discrepancies, and achieved intercoder agreement of Cohen's kappa = .82. Quantitative and qualitative findings were integrated at the interpretation stage.

RESULTS

Overall criterion-referenced performance

Both conditions showed substantial improvement from pretest to posttest (Table 4). The on-site group increased from $M = 3.82$ ($SD = 1.12$) to $M = 6.71$ ($SD = 0.89$), $t(39) = 12.87$, $p < .001$, $d = 1.89$. The online group increased from $M = 3.65$ ($SD = 1.08$) to $M = 5.83$ ($SD = 0.95$), $t(39) = 10.42$, $p < .001$, $d = 1.52$. The mean gain was 2.89 points on site and 2.18 points online.

Table 4
Pretest and posttest scores by instructional condition

Condition	Pretest M (SD)	Posttest M (SD)	Mean gain	t(39)	p	Cohen's d [95% CI]
On-site	3.82 (1.12)	6.71 (0.89)	2.89	12.87	< .001	1.89 [1.55, 2.23]
Online	3.65 (1.08)	5.83 (0.95)	2.18	10.42	< .001	1.52 [1.20, 1.84]

Note. M = mean; SD = standard deviation; CI = confidence interval. Total possible score = 8.

The between-condition posttest comparison favored the on-site condition. The difference was 0.88 points on the eight-point checklist, 95% CI [0.52, 1.24], $t(78) = 3.45$, $p = .001$, $d = 0.69$. The ANCOVA result remained significant after controlling for pretest score, $F(1, 77) = 8.12$, $p = .006$, partial eta squared = .10.

The confidence interval indicates that the estimated posttest difference was unlikely to be trivial on the eight-point scale, while the large gains in both groups show that the comparison should not be reduced to success versus failure. Online participants completed, on average, more than two additional criteria after the module, and on-site participants completed nearly three additional criteria. The key issue for the subsequent analysis was therefore which criteria accounted for the remaining between-condition difference.

Criterion-specific differences

The criterion-level analysis showed that the overall on-site advantage was not uniform (Table 5). After Bonferroni correction, on-site participants performed better when acknowledging the customer's problem, clarifying the mode of transportation, explaining the route through visible landmarks, and confirming understanding. Online participants performed better when clarifying distance. Greetings, offers of additional support, and closings did not differ significantly. The corrected probability for Criterion 8 was $p = .655$.

Table 5
Criterion-specific posttest contrasts

Criterion	Direction of difference	t(78)	p	d
1. Standard greeting	No significant difference	1.02	.311	0.18
2. Acknowledges problem	On-site higher	3.11	.003	0.62
3. Clarifies transportation mode	On-site higher	2.89	.005	0.51
4. Uses visible landmarks	On-site higher	4.67	< .001	0.93
5. Clarifies distance	Online higher	2.98	.004	0.59
6. Offers additional support	No significant difference	0.87	.387	0.10
7. Confirms understanding	On-site higher	3.89	< .001	0.77
8. Standard closing and thanks	No significant difference	0.45	.655	0.05

Note. N = 80 (40 per condition). Bonferroni-adjusted alpha = .006. Effect sizes are reported in the direction of the higher-performing condition.

The largest reported condition difference occurred for explaining routes through landmarks ($d = 0.93$), followed by confirming understanding ($d = 0.77$). The online advantage for quantifying distance was moderate ($d = 0.59$). Because digital-map format was confounded with the online condition, the distance result should not be attributed solely to online communication. It may instead reflect the combined affordances of the digital map, screen sharing, zoom, and learners' familiarity with digital navigation.

The pattern also shows that the on-site advantage was not limited to explicitly spatial language. Acknowledging the problem and clarifying transportation mode occur before the route itself, while confirming understanding recurs during and after route segments. These criteria share a need to adapt the exchange to the customer. In contrast, distance clarification depends more directly on inspection and quantification of route information. This functional grouping informed the interpretation in the Discussion.

Learner experience

Learners' ratings differed across all four experience dimensions (Table 6). On-site participants reported higher social interaction, immediate feedback, and technical ease. Online participants rated flexibility more highly. The largest differences concerned social interaction and technical ease. Connectivity was the most frequent difficulty in the online open-ended responses, reported by 62% of online participants. Other comments referred to device controls, sound, and coordinating the digital map with the conversation. On-site responses frequently emphasized direct partner interaction, gesture, and the convenience of printed materials. Online responses frequently identified screen-shared maps and access from home as useful.

The dimensions should be interpreted separately. The high online flexibility rating did not eliminate technical difficulty, and the high on-site technical-ease rating did not mean that learners preferred every aspect of physical attendance. Open-ended comments showed that advantages and disadvantages coexisted within each condition. This is consistent with an affordance-and-constraint interpretation rather than a single global preference.

Table 6
Likert-scale ratings for learning experience

Experience dimension	On-site M (SD)	Online M (SD)	t(78)	p
Social interaction	4.4 (0.6)	2.7 (1.0)	6.34	< .001
Immediate feedback	4.2 (0.7)	3.1 (1.0)	3.89	< .001
Technical ease	4.6 (0.4)	2.9 (1.1)	5.95	< .001
Flexibility	2.8 (0.9)	4.5 (0.6)	4.17	< .001

Note. Ratings used a five-point scale (1 = strongly disagree, 5 = strongly agree).

DISCUSSION

Overall effects of the two instructional conditions

The first research question concerned total criterion-referenced performance. Both conditions produced large pretest-posttest gains, indicating that a short, needs-based TBLT module can improve learners' completion of an occupationally relevant direction-giving task in either environment. This result is consistent with research showing that technology-mediated tasks can sustain meaningful language use when task goals, information gaps, and interaction are deliberately designed (González-Lloret, 2020; Lai & Li, 2011). It also extends prior Thai research on online TBLT engagement and computer-assisted task work by using a workplace-derived performance measure (Durongbhandhu & Suwanasilp, 2023; Soongpankhao et al., 2023).

The on-site condition nevertheless produced a higher posttest mean after pretest performance was taken into account. This difference should be interpreted at the level of the implemented package. Physical co-presence, printed maps, immediate access to gesture, and fewer connectivity demands occurred together. The study cannot determine which of these components produced the difference. Accordingly, the evidence supports the more limited conclusion that the on-site package was more effective for this implementation of the task, not that face-to-face TBLT is categorically superior to online TBLT.

This bounded interpretation also avoids a false opposition between task design and delivery mode. The module design was common across conditions, but design is realized through tools, interactional arrangements, and materials. A well-designed task can be implemented more or less successfully online, and an on-site task can lose its communicative value if learners simply point to a route. The results therefore show that the pedagogic core and the environment must be considered together. Task design does not make modality irrelevant; it determines how modality-specific resources are incorporated or controlled.

Criterion-level performance as the main contribution

The second research question reveals why a total-score comparison is insufficient. Four criteria favored the on-site condition, one favored the online condition, and three showed no reliable difference. The profile suggests that task performance is multidimensional: an instructional

environment may support some components while constraining others. Criterion-referenced assessment therefore contributes more than a pass-fail judgment at the level of the whole task. It identifies the particular points at which task resources, interactional demands, and the instructional environment align.

This profile can be understood as a decomposition of the total effect. If only the overall 0.88-point difference had been reported, a teacher might conclude that the online module should simply be replaced. The item pattern supports a different response: retain the online features that appear to assist precision, but redesign the interactional stages that depend on acknowledgement, partner monitoring, and comprehension checks. The criterion analysis therefore turns a comparative finding into a set of design hypotheses.

The on-site advantages clustered around interaction management: acknowledging the customer's problem, clarifying transportation mode, explaining a route through landmarks, and confirming understanding. These steps require close monitoring of the partner, rapid response to signals of comprehension, and coordination between talk and spatial information. The questionnaire data offer a convergent explanation: on-site learners rated social interaction and immediate feedback more highly and described gesture and direct partner access as useful. The largest criterion-level difference, landmark-based explanation, is especially compatible with the availability of embodied spatial cues. This interpretation remains cautious because the study did not video-code gesture or interactional repair directly.

The online advantage for distance clarification points in a different direction. Digital maps can make numerical distances, stops, and route segments visually salient through zooming and screen-based inspection. Online learners also identified screen-shared maps as useful. Yet the study design does not allow a clean causal attribution to communication mode because map format changed with condition. The logical implication is methodological as well as pedagogical: future modality research should cross communication mode with material format so that digital-map affordances can be separated from the effects of video-mediated interaction.

No significant differences appeared for greetings, additional-support offers, or closings. These steps were relatively bounded and could be prepared as short conventional sequences. Their stability across conditions suggests that the most consequential condition effects occurred in criteria that required adaptation to the partner and the spatial problem rather than in routine opening or closing moves. This distinction gives the criterion-level analysis a clear theoretical role: it differentiates task components by their interactional and resource demands instead of treating the task as a unitary outcome.

Learner experience and mixed-methods integration

The third research question concerned learner experience. On-site learners valued social interaction, immediate feedback, and technical ease, while online learners valued flexibility. The open-ended responses clarify the trade-off. Online participation reduced travel and supported access to digital maps, but connectivity and device-management problems interrupted some task performances. These experiences matter because a learner who is

coordinating audio, video, file windows, and an unstable connection has fewer attentional resources available for monitoring the partner and managing the route explanation. Conversely, the flexibility of online access may make participation possible where physical attendance is difficult.

The quantitative and qualitative results converge most clearly around interaction-dependent criteria. The lower online scores for landmark explanations and comprehension checks occurred alongside lower ratings for social interaction and immediate feedback. This convergence strengthens the interpretation that the online package required additional interactional scaffolding. The distance criterion shows the complementary pattern: an online advantage occurred alongside positive comments about screen-shared maps. The mixed-methods contribution is therefore explanatory rather than decorative; the experience data help identify plausible mechanisms while remaining distinct from causal evidence.

Theoretical implications

The study advances a criterion-sensitive account of modality in TBLT. Technology-mediated and face-to-face tasks should not be compared only by a single achievement score because task criteria differ in their dependence on social cues, spatial resources, temporal coordination, and interface management. The results support a modality-criterion alignment principle: the effect of an instructional condition is partly a function of how its available resources match the demands of each task component. This principle is consistent with technology-mediated TBLT research emphasizing the inseparability of task design, tools, and interaction (González-Lloret & Ortega, 2014; Lai & Li, 2011), but the present study operationalizes that relationship through needs-based, criterion-referenced performance.

The principle is deliberately framed as an interpretive proposition rather than a newly named hypothesis. A single short intervention cannot establish a general law of modality. It can, however, demonstrate a replicable analytic strategy: derive criteria from target-task evidence, compare conditions at both total and criterion levels, collect experience data relevant to the same task demands, and use the convergent pattern to formulate the next experimental test. This strategy strengthens theory by making claims more specific and more readily falsifiable.

The connection to TBNA is also important. The criteria were not invented after the intervention to explain differences; they were derived from workplace informants and target discourse before the module was evaluated (Soongpankhao & Lambert, 2026). This sequence strengthens the validity of the analysis because the observed differences concern components that matter to successful workplace performance. Criterion-level modality research is most useful when the criteria themselves are externally grounded in target-task needs.

Practical implications

For online TBLT, interaction-dependent criteria should be made visible in the task instructions and deliberately scaffolded. Learners can be required to pause after each route segment, use explicit comprehension checks, ask the customer to restate a key landmark, and use shared annotation or chat as a supplementary channel. A brief technical orientation, downloadable

low-bandwidth maps, audio-only contingency procedures, and advance device checks can reduce avoidable disruption. These supports should be treated as part of task design, not as peripheral technical arrangements.

For on-site instruction, teachers should preserve the communicative challenge rather than allow physical pointing to replace oral explanation. The instruction used in this study—both partners have maps, but the employee does not trace the route for the customer—is one way to maintain the information-transfer demand. Digital maps can also be used on site so that material format is aligned across instructional environments and learners practice the tools they are likely to use in contemporary service work.

A blended sequence may combine the complementary strengths observed here. Learners could use online materials for input, route research, and distance estimation, then complete on-site interactional rehearsals focused on acknowledging needs, landmark-based explanation, repair, and comprehension checking. Assessment should retain the same workplace-derived criteria across phases so that learners and teachers can see which components have developed and which require further practice.

The same criterion set can support formative feedback. Rather than returning only a total of six or seven, a teacher can show that a learner communicated distance accurately but omitted transport-mode clarification or failed to check understanding. Learners can then repeat a shortened task focused on the missing function. Because the criteria come from workplace evidence, this feedback remains connected to the target domain and can be used consistently across online, on-site, and blended phases.

CONCLUSION, LIMITATIONS, AND FUTURE RESEARCH

This study aimed to compare online and on-site implementations of a needs-based TBLT module for an occupationally relevant giving-directions task among Thai EFL learners. Specifically, it examined whether instructional condition affected learners' overall criterion-referenced posttest performance after accounting for pretest scores, which individual task-performance criteria differed between the two conditions, and how learners evaluated their learning experiences.

The findings showed that both instructional conditions produced substantial improvements in learners' ability to complete the target task. Although the on-site group achieved a higher overall posttest score, the criterion-level analysis revealed a more differentiated pattern. On-site learners performed better in acknowledging the customer's problem, clarifying the mode of transportation, explaining routes through visible landmarks, and confirming understanding. In contrast, online learners performed better in clarifying distance, while no significant differences were found for greetings, offers of additional support, or closings. These results indicate that the effects of instructional condition were not uniform across all components of the task.

Learners' reported experiences provided further insight into these patterns. On-site participants reported stronger social interaction, more immediate feedback, and greater technical ease,

whereas online participants valued the flexibility of the online environment and the usefulness of screen-shared digital maps. Taken together, the performance and experience data suggest that interaction-dependent task components may benefit from the immediacy and interpersonal cues available in face-to-face settings, while some information-focused components, such as distance estimation, may be supported by digital resources. However, these relationships should be interpreted as plausible explanations rather than direct causal effects.

The findings therefore support a criterion-sensitive approach to evaluating TBLT across instructional environments. Rather than asking whether online or on-site TBLT is generally more effective, it may be more useful to examine how the resources and constraints of each environment correspond to the demands of particular task components. From a pedagogical perspective, online TBLT may require additional scaffolding for interaction-dependent functions, such as explicit comprehension checks, opportunities for clarification, and procedures for monitoring the interlocutor. On-site instruction, meanwhile, should preserve the communicative demands of the task and avoid allowing physical pointing or other nonverbal shortcuts to replace meaningful oral interaction. A blended approach may also be useful by combining digital resources for route planning and distance estimation with face-to-face practice for interaction management and comprehension checking.

Several limitations should be considered when interpreting these findings. First, the intervention was conducted within a single day, so the study cannot determine whether the observed gains would be retained over time or transferred to other occupational tasks. Future research should therefore incorporate longer interventions, delayed posttests, and multiple target tasks to examine retention and transfer. Second, the participants were drawn from one English for International Communication program and represented a relatively narrow CEFR A2–B1 proficiency range. Replication across institutions, proficiency levels, and occupational contexts would strengthen the generalizability of the findings.

Thirdly, communication mode and material format were confounded because the online group used digital maps whereas the on-site group used printed maps. Consequently, the observed differences cannot be attributed solely to online or face-to-face communication. Future studies should use factorial designs that systematically vary communication mode and material format, for example by comparing online and on-site interaction with both digital and printed maps. Such designs would allow researchers to distinguish the effects of interactional mode from those of the learning materials.

Fourthly, individual speaking time and interactional behavior were not recorded in sufficient detail to determine whether the two conditions differed in task duration, amount of talk, clarification, repair, gesture, or comprehension checking. Future studies should record task performances and conduct more detailed interactional analyses to identify the processes associated with criterion-level differences. Fifth, the pass-fail assessment checklist captured functional task completion but did not measure differences in fluency, accuracy, complexity, pragmatic appropriateness, or interactional quality. Future research could retain the workplace-derived criterion checklist while incorporating additional validated analytic measures. Binary criterion-level outcomes could also be examined using logistic or mixed-effects models when repeated tasks, multiple raters, or longitudinal observations are available.

Finally, the learner-experience data were based primarily on immediate self-report. Although these responses helped identify possible explanations for the observed performance patterns, they do not provide direct evidence of the mechanisms underlying learning. Future research could strengthen this aspect of the design by incorporating interviews, stimulated recall, platform logs, classroom observations, and recordings of learner interaction.

Despite these limitations, the study demonstrates the value of linking task-based needs analysis, instructional design, and criterion-referenced assessment. When performance criteria are derived from workplace needs, criterion-level analysis can reveal not only whether learners improve, but also which components of a target task are better supported or constrained by particular instructional environments. Such evidence can provide a more useful basis for designing online, on-site, and blended TBLT than broad comparisons of modality alone.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Rajamangala University of Technology Phra Nakhon, Thailand, for institutional support in the completion of this research. During the preparation of this manuscript, the authors used OpenAI ChatGPT (GPT-5.5) and Grammarly solely for language editing and proofreading purposes, including a thorough review of the grammar, spelling, punctuation, and clarity of expression throughout the entire text. The authors have reviewed and verified all suggested revisions, and hereby take full responsibility for the content of this manuscript.

THE AUTHORS

Watcharaphong Soongpankhao is an Assistant Professor in the English for International Communication Program at the Faculty of Liberal Arts, Rajamangala University of Technology Phra Nakhon, Bangkok, Thailand. He received his Ph.D. in Education from Curtin University, Western Australia. His research interests include task-based language teaching, task engagement, ESP, and Thai performing arts.

watcharaphong.s@rmutp.ac.th

Suwanna Kemdang is a lecturer in the English for International Communication Program at the Faculty of Liberal Arts, Rajamangala University of Technology Phra Nakhon, Bangkok, Thailand. She received her Ph.D. in Curriculum and Instruction from the University of Arkansas, Fayetteville, Arkansas, USA. Her research interests lie in ESL, EFL, and adult language teaching and learning.

suwanna.k@rmutp.ac.th

Gary Edward Gray is a lecturer in the English for International Communications Program at the Faculty of Liberal Arts, Rajamangala University of Technology Phra Nakhon, Bangkok, Thailand. He holds a B.S. from the University of Massachusetts in Lowell, Massachusetts, USA. His teaching interests include English speaking and listening, especially with regards to communicative skills.

gary.g@rmutp.ac.th

REFERENCES

- Aubrey, S., & Philpott, A. (2023). Second language task engagement in face-to-face and synchronous video-based computer-mediated communication modes: Performances and perceptions. *System*, 115, Article 103069. <https://doi.org/10.1016/j.system.2023.103069>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE.
- Durongbhandhu, N., & Suwanasilp, D. (2023). Computer-assisted English lessons for logistics with task-based language teaching for enhancing receptive and productive vocabulary knowledge of EFL Thai learners. *LEARN Journal: Language Education and Acquisition Research Network*, 16(1), 47–74. <https://so04.tci-thaijo.org/index.php/LEARN/article/view/263432>
- Ellis, R. (2019). Towards a modular language curriculum for using tasks. *Language Teaching Research*, 23(4), 454–475. <https://doi.org/10.1177/1362168818765315>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Giraldo, F. (2020). Task-based language assessment: Implications for the language classroom. *GIST – Education and Learning Research Journal*, 21, 209–224. <https://doi.org/10.26817/16925777.828>
- González-Lloret, M. (2020). Collaborative tasks for online language teaching. *Foreign Language Annals*, 53(2), 260–269. <https://doi.org/10.1111/flan.12466>
- González-Lloret, M., & Ortega, L. (Eds.). (2014). *Technology-mediated TBLT: Researching technology and tasks*. John Benjamins.
- Lai, C., & Li, G. (2011). Technology and task-based language teaching: A critical review. *CALICO Journal*, 28(2), 498–521. <https://doi.org/10.11139/cj.28.2.498-521>
- Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Wiley-Blackwell.
- McDonough, K., & Chaikitmongkol, W. (2007). Teachers' and learners' reactions to a task-based EFL course in Thailand. *TESOL Quarterly*, 41(1), 107–132. <https://doi.org/10.1002/j.1545-7249.2007.tb00042.x>
- McNamara, T. (1996). *Measuring second language performance*. Longman.
- Norris, J. M. (2009). Task-based teaching and testing. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 578–594). Wiley-Blackwell. <https://doi.org/10.1002/9781444315783.ch30>
- Soongpankhao, W. (2023). *Developing fully-online TBLT for customer relation skills in a university in Thailand: From needs analysis to evaluation* [Doctoral dissertation, Curtin University]. Curtin Research Publications. <https://hdl.handle.net/20.500.11937/95489>
- Soongpankhao, W., Aubrey, S., & Lambert, C. (2023). Impact of goal-tracking on engagement in language use in an online TBLT module for Thai university students. *System*, 119, Article 103184. <https://doi.org/10.1016/j.system.2023.103184>
- Soongpankhao, W., & Lambert, C. (2026). Task-based needs analysis as a foundation for TBLT course design: Evidence from Thai university EFL learners. *International Journal of TESOL Studies*, 8(1), 157–179. <https://doi.org/10.58304/ijts.260508>
- Stickler, U. (2022). *Technology and language teaching*. Cambridge University Press. <https://doi.org/10.1017/9781108874403>
- Winke, P., & Goertler, S. (2008). Did we forget someone? Students' computer access and literacy for CALL. *CALICO Journal*, 25(3), 482–509.