

A Cycle of Test-Teach-Test for TOEFL Preparation: A Scoping Review

Umi Dian Adhitya Wulan Ningrum¹, Margana², Yuyun Yulia³, Muhammad Kristiawan⁴

¹English Education Department, Faculty of Languages, Arts, and Culture, Universitas Negeri Yogyakarta, Indonesia

²English Education Department, Faculty of Languages, Arts, and Culture, Universitas Negeri Yogyakarta, Indonesia

³English Education Department, Faculty of Languages, Arts, and Culture, Universitas Negeri Yogyakarta, Indonesia

⁴Educational Administration Department, Universitas Negeri Bengkulu, Indonesia

HOW TO CITE:

Umi Dian Adhitya Wulan
Ningrum, Margana, Yuyun Yulia,
Muhammad Kristiawan (2026). A
Cycle of Test-Teach-Test for
TOEFL Preparation: A Scoping
Review
International Journal of Special
Education, 41(21s), 1268-1286

ABSTRACT

This systematic literature review synthesized 33 Scopus-indexed articles to construct an empirically grounded Test-Teach-Test (TTT) cycle for TOEFL preparation in higher education EFL contexts, guided by PRISMA 2020 guidelines across a decade of publications (2016–2026). Four research questions addressed the diagnostic test (RQ1), differentiated instructional interventions (RQ2), post-teaching assessments (RQ3), and digital resources as TTT enablers (RQ4), corresponding to article distributions of $n = 5$ (15.2%), $n = 10$ (30.3%), $n = 11$ (33.3%), and $n = 7$ (21.2%) respectively. The diagnostic stage findings establish that sub-skill decomposition instruments, including Cognitive Diagnostic Model (CDM) applications achieving 87, 95.8% subskill-level reliability, Automated Writing Evaluation (AWE), Lexical Feature Analysis (LFA), and usage-based validated procedures, generate the diagnostic precision that TOEFL ITP's holistic three-section scoring is structurally incapable of providing. The Teach stage evidence confirms that a tripartite instructional model, meaning-focused grammar and extensive reading, blended technology-mediated learning, and motivational self-regulatory scaffolding, is required because no single pathway achieves significant TOEFL score improvement in isolation. The second Test stage findings establish its dual function: summative benchmarking via CEFR-linked cut scores and formative feedback through AI-assisted scoring, with washback quality rather than score magnitude driving iterative cycle continuity. Digital enablers, blended learning, and CLIQ-based diagnostic construction provide scalable full-cycle delivery architecture. These findings are evidence-informed rather than causally validated, as no included study tested a complete TTT cycle against a matched control. These findings position the TTT cycle as a principled, equity-critical framework for transforming TOEFL preparation from institutional gatekeeping into diagnostically personalized language development for non-English-major students in Indonesian higher education.

Keywords: Test-Teach-Test, TOEFL preparation, EFL higher education.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.

Open access publication under the
terms and conditions
of Creative Commons Attribution (CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As the world appears to be increasingly interconnected for various reasons, the ability to communicate internationally is increasingly crucial, leading to the Test of English as a Foreign Language (TOEFL) as a gateway requirement for advanced education and secure employment. Thus, the TOEFL plays a vital role in higher education institutions (Crystal, 2003), multinational employment (Jenkins et al., 2011), cross-country scholarships (ETS, 2026; U.S. Sports Scholarships, 2026), and international professional certifications (AMINEF - American Indonesian Exchange Foundation, 2025; Indonesian International Education Foundation, 2026; International Test Center, 2026). In these contexts, the players must have an efficient interplay between students and educators to achieve the targeted TOEFL score.

This requirement disproportionately disadvantages non-English-major students whose academic and career progress can be thwarted when the preparation does not match their starting proficiency. An effective approach for students and teachers is nonetheless viewed from a critically underexplored perspective, even though it is easy to achieve a desired TOEFL score (Barkaoui, 2024). Du (2025), Sickinger et al. (2025), Yu (2026) indicate that teachers who use conventional approaches in their TOEFL classes employ non-sequential teaching practices that fail to diagnose students' prior TOEFL knowledge. In Indonesian higher education, TOEFL ITP preparation is typically delivered as a fixed, semester-long course in which every student receives the same sequence of grammar drills and practice tests. Because no diagnostic step precedes instruction, the TOEFL instructors or teachers cannot distinguish their students who need remedial grammar instruction from one who needs extensive-reading support; both receive identical treatment.

Found in many universities, TOEFL study plays a significant role in their academic programs, yet the in-charge teachers or lecturers still use inefficient, conventional teaching models. Most TOEFL preparation programs still use

conventional textbook exercises, paper-based practice exams, and goals that are focused on grammar, even though the TOEFL is a multi-sectional exam that requires cognitive ability (Illanangingtyas, 2025; Kholiti & Ghouati, 2025). These methods seldom consider that every student comes to class with their own unique strengths, weaknesses, abilities, and challenges. Therefore, students are unable to reorganize grammar rules accurately into communication products, such as writing and speaking. To tackle these issues, there must be a real change in how TOEFL preparation is taught. This change needs to be grounded in a solid theoretical framework that links teachers' adaptive duties to a systematic approach to identifying and meeting students' specific learning requirements.

Despite this recognized need, the literature on TOEFL preparation remains fragmented into isolated strands. This review's contribution is therefore not a further summary of what individual studies report, but the construction of a practical, diagnostic-driven TTT framework built from 33 Scopus-indexed empirical studies, intended to be directly adoptable by TOEFL instructors and institutions rather than serving only as a conceptual proposal. Establishing such a framework matters because it can make TOEFL preparation more personalized, effective, and data-driven. To fill this gap, a design of Test-Teach-Test can fix this structural deficiency by establishing an purposed cycle in which a diagnostic test (first Test) is aimed to call students' prior knowledge about TOEFL and finds their current TOEFL proficiency place, followed by some structured teaching practices with the reference of a diagnostic test (first Test) carried out by a post-teaching practice test (second Test) to measure the effectiveness of the structured teaching practices and students' progress. This approach ensures that every cycle of sequential teaching practices is both evidence-driven and outcome-focused.

2. Methodology

Using inclusion and exclusion criteria outlined in the PRISMA 2020 guidelines, this research conducted a Systematic Literature Review (SLR).

Focusing on research applications of a Test-Teach-Test design for TOEFL preparation, this PRISMA provides thorough guidelines for locating, filtering, and validating relevant publications. It is organized in a systematic manner. Prior to the data collection, some detailed protocols for retrieving articles from a database were formulated. Following Efendi et al. (2025), these protocols consisted of 4 main steps: 1) an arrangement of Research Questions (RQs), (2) a search strategy formulation with the database selection for the data import, (3) inclusion and exclusion criteria merged with the PRISMA 2020 guidelines with Zotero Application as the reference manager, and (4) synthesis of selected articles. This process performed a literature review with high validity and trustworthy novelty, in compliance with current scientific standards. Accordingly, the review analysis provided a broad lens on the diagnostic test, the instructional interventions based on students' needs identified by the diagnostic test, and the post-test, with formative and summative mechanisms within a Test-Teach-Test cycle.

2.1. An Arrangement of Research Questions (RQs)

As a result of the issues identified in the introduction, a critical knowledge gap emerged in this topic's knowledge base. Although previous studies reported significant results, studies focusing on the Test-Teach-Test design for TOEFL preparation are lacking in the literature due to poor analysis. In the last 5 years, studies' limited evidence on the diagnostic accuracy, instructional responsiveness, and assessment validity of TOEFL programs has persisted despite the availability of digital resources. Therefore, the existence of a Systematic Literature Review (SLR) in this research aimed to address these gaps by investigating the field-based applications of the Test (diagnostic test)-Teach (intervention)-Test (post-teaching assessments) approach for TOEFL preparation. Consequently, three Research Questions (RQs) were constructed, as follows:

- RQ1: How does the diagnostic test function as the first Test stage in a design of Test-Teach-Test in TOEFL preparation programs?
- RQ2: How is the Teach applied in a design of Test-Teach-Test in TOEFL preparation programs?
- RQ3: How is the second Test stage of a design of Test-Teach-Test positioned in evaluating teaching effectiveness in TOEFL preparation programs?
- RQ4: How do digital resources and instructional interventions persist Test-Teach-Test applications in TOEFL preparation contexts?

2.2. The Search Strategy Formulation

This search method used core search strings and keyword combinations with Boolean operators (such as "AND," "OR," and "NOT") to get specific articles while remaining focused on a Test-Teach-Test approach for TOEFL preparation. The goal was to minimize data collection while remaining effective. Therefore, the combinations were "Pretest" OR "Teach" OR "Posttest" AND "TOEFL" OR "TOEFL ITP" OR "TOEFL preparation" OR "TOEFL training" AND "EFL" OR "ESL" OR "higher education" OR "English language teaching" OR "language instruction". Scopus was selected as the exclusive database for three reasons. Initially, the rigorous quality screening described in Table 1 was facilitated by Scopus's uniform indexing and SJR quartile-ranking algorithm. Scopus covers papers in applied linguistics and language assessment on a far wider scale than the single-publisher index, thereby minimizing the risk of publisher-induced bias. Finally, all of the results from each screening step, including the search string, Boolean logic, and record counts, could be independently recreated thanks to the use of a single auditable database (Figure 2). This decision is still seen as a constraint; as mentioned again in the Conclusion, papers indexed only in ERIC, Web of Science, or regional databases may not have been included in Scopus-only searches.

2.3. Inclusion and Exclusion Criteria

To ensure the retrieved articles' specificity, the 843 documents were imported from the Scopus database, imported into Zotero as the reference

manager, and analyzed using inclusion and exclusion criteria in line with the PRISMA 2020 guidelines. Table 1 summarizes the inclusion and exclusion criteria.

Table 1. Inclusion and Exclusion Criteria

Inclusion	Exclusion
Scopus-indexed database publications with SJR recognition in Q1-Q4 quartiles	Non-Scopus-indexed database publications with no SJR recognition
Focusing on pretest, teach, posttest applications for TOEFL preparation	Not focusing on pretest, teach, posttest applications for TOEFL preparation
Addressing higher education (undergraduates, masters, doctoral degrees) students	Not addressing higher education (undergraduates, masters, doctoral degrees) students
English as a Foreign Language (EFL) and as a Second Language (ESL) contexts	Others than English as a Foreign Language (EFL) and as a Second Language (ESL) contexts
Field-based research	Non-field-based research
No duplicate documents	Duplicate documents
Publications between 2016 and 2026 (a decade)	Publications outside the range of 2016 to 2026
Open access	Close access
Available full-text format	Unavailable full-text format
English written	Not English written

According to Table 1, the data collection protocol used predetermined inclusion and exclusion criteria to double-check the specificity of the retrieved articles. The retrieved documents were published in the Scopus-indexed database with SJR recognition in Q1-Q4 quartiles between 2016 and 2026 (a decade), excluding those not published in the Scopus-indexed database with SJR recognition in Q1-Q4 quartiles. In addition, the data collection focused on pretest, teach, posttest applications for TOEFL preparation,

addressing higher education students (undergraduates, master's, doctoral) in English as a Foreign Language (EFL) and as a Second Language (ESL) contexts. The documents included English-language open-access field-based research articles in full text, with no duplication. These empirical research articles were imported into Zotero to be merged with the PRISMA 2020 analysis and citation tracking. Figure 1 presents the PRISMA 2020 analysis.

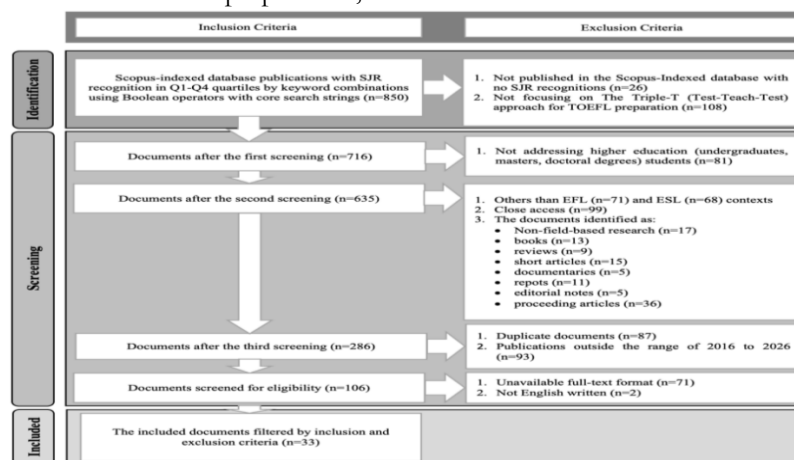


Figure 1. PRISMA 2020 analysis

The PRISMA 2020 standards in Figure 1 consist of three linked phases of identification, screening, and inclusion of publications with specific inclusion and exclusion criteria. The retrieval of documents was performed using the Scopus database, with keyword combinations, Boolean operators, and core search strings, with predetermined Boolean strings and other

keywords, then extracted to the Zotero application for management. Scopus-indexed publications with SJR recognition in Q1-Q4 quartiles. 26 papers were omitted as not published in a Scopus-indexed database with SJR recognition, and 108 were deleted from Zotero as not focusing on pretest, teach and posttest applications for TOEFL preparation.

Table 2. A list of included articles

Stage	Article	Journal	Publisher
Test (diagnostic test)	Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings (Bridgeman & Ramineni, 2017)	Assessing Writing	Elsevier
	Placement of multilingual writers: Is there a role for student voices? (Ferris et al., 2017)	Assessing Writing	Elsevier
	Assessment from the ground up: Developing and validating a usage-based diagnostic assessment procedure in a graduate EAP context (Dolgova & Siczek, 2019)	Journal of English for Academic Purposes	Elsevier
	Use of lexical features in non-native academic writing (Vo, 2019)	Journal of Second Language Writing	Elsevier
	Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context (Mirzaei et al., 2020)	Studies in Educational Evaluation	Elsevier
Teach (differentiated Interventions based on the needs from diagnostic test)	Effects of extensive reading and translation activities on grammar knowledge and attitudes for EFL adolescents (Lee et al., 2016)	System	Elsevier
	Students' perceptions and attitudes towards asynchronous technological tools in blended-learning training to improve grammatical competence in English as a second language (Pinto-Llorente et al., 2017)	Computers in Human Behavior	Elsevier
	The effect of blended learning on EFL students' grammar performance and attitudes: An investigation of Moodle (Bader Al Bataineh et al., 2019)	Arab World English Journal	AWEJ Publishing
	Task-supported language teaching to enhance young EFL adolescent learners' comprehension and production of English phrasal verbs in Taiwan (Chou, 2020)	Education 1268-1286	Routledge
	Hong Kong secondary students' self-regulated learning strategy use and English writing: Influences of motivational beliefs (Bai & Wang, 2021)	System	Elsevier
	The impact of game-based learning (GBL) on students' motivation, engagement and academic performance on an Arabic language grammar course in higher education (Eltahir et al., 2021)	Education and Information Technologies	Springer

	Promoting EFL learners' pragmatic awareness through Test-Teach-Test (TTT) method via Telegram and Edmodo (Niaghi, 2023)	Mextesol Journal	Asociacion Mexicana de Maestros de Ingles MEXTESOL A.C
	Students' performance in English: effects of teachers' leadership behavior and students' motivation (Igwe & Ligaya, 2025)	Journal of Education and Learning	Intelektual Pustaka Media Utama
	Interplay of mindfulness, motivation, and self-regulation in predicting L2 achievement: A mixed-methods study (Fathi & Aghamirzaei, 2026)	Acta Psychologica	Elsevier
	Formative assessment as a self-regulation engine in blended language learning: A case study with top-tier engineering talents (Ren at al., 2026)	Acta Psychologica	Elsevier
Test (post-teaching assessments)	Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features (Riazi, 2016)	Assessing Writing	Elsevier
	The TOEFL iBT writing: Korean students' perceptions of the TOEFL iBT writing test (Kim, 2017)	Assessing Writing	Elsevier
	Does English proficiency affect academic performance? (Geide-Stevenson, 2018)	International Review of Economics Education	Elsevier
	Examining the applications and opinions of the TOEFL ITP® assessment series test scores in three countries (Golubovich at al., 2018)	ETS Research Report Series	John Wiley & Sons
	The relationship between test preparation and state test performance: Evidence from the Measure of Effective Teaching (MET) project (Li & Xiong, 2018)	Education Policy Analysis Archives	Arizona State University
	Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study (Fleckenstein at al., 2020)	Assessing Writing	Elsevier
	Effects of notetaking instruction on intermediate and advanced L2 English learners: A quasi-experimental study (Siegel, 2020)	Journal of English for Academic Purposes	Elsevier
	The washback effect of TOEFL iBT and a local English Proficiency Exam on students' motivation, autonomy and language learning strategies (Sadeghi at al., 2021)	Heliyon	Elsevier
	Integrating task and game-based learning into an online TOEFL preparatory course during the COVID-19 outbreak at two Indonesian higher education institutions (Pratiwi & Waluyo, 2022)	Malaysian Journal of Learning and Instruction	Universiti Utara Malaysia Press
	Exploring the potential of using an AI language model for automated essay scoring (Mizumoto & Eguchi, 2023)	Research Methods in Applied Linguistics	Elsevier
	Beyond word count: The role of note quality on EFL students' lecture listening comprehension (Sakurai & Spring, 2026)	System	Elsevier

Teaching Media and Learning Technologies	Washback: Exploring what constitutes “good” teaching practices (Barnes, 2017)	Journal of English for Academic Purposes	Elsevier
	Are TOEFL iBT® writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers (Ling, 2017)	Assessing Writing	Elsevier
	Voice in timed L2 argumentative essay writing (Zhao, 2017)	Assessing Writing	Elsevier
	A corpus-based investigation of noun phrase complexity in the L2 writings of a first-year composition course (Lan & Sun, 2019)	Journal of English for Academic Purposes	Elsevier
	CLIQ! Intelligent Question Classifier for the elaboration of exams (Jardim et al., 2022)	Software Impacts	Elsevier
	Enhancing students’ learning outcomes through MALL in English intensive program for railway mechanical technology (Pratiwi et al., 2023)	International Journal of Language Education	Universitas Negeri Makassar
	Private English tutors’ agency amid China’s ‘double reduction’ policy: a Bourdieusian perspective (Lyu & Lam, 2025)	Journal of Education Policy	Routledge

According to Table 2, the 4 stages include Test (diagnostic test) with 5 articles, Teach (differentiated Interventions based on the needs from diagnostic test) with 10 articles, Test (post-teaching assessments) with 11 articles, and Teaching Media and Learning Technologies with 7 articles. Those articles were published between 2016 and 2026 with a vacuum in 2024. The Assessing Writing held the most published articles, with Elsevier as the top 1 publisher with 23 publications. Furthermore, these 33 included articles were extracted in RIS format, then imported into VOSviewer to generate a keyword occurrence network revealing their interconnections, highlighting the dominant research trends and uncovering potential gaps. Afterward, the gaps guided the analysis of the 4 predetermined stages in the results section. Figure 2 illustrates a network visualization of keyword occurrences across 33 included articles.

2.4. Synthesis of selected articles

The study addressed four research questions (RQs) and synthesized thirty-three papers on TOEFL test preparation, classifying them according to research aims, methodologies, outcomes, and implications. Experimental and quasi-experimental research were the most

common types of field-based designs. The diagnostic test found students' gaps in previous knowledge and performance (RQ1), according to the basic results of 5 articles included in the review, and to fill these gaps, 10 publications provided differentiated instructional strategies (RQ2). Eleven articles formed the basis of the post-teaching evaluation, which used a Test-Teach-Test approach (RQ3) and, in seven of those articles, digital resources (RQ4) to determine how successful these interventions were. In all 33 documents, one drawback that kept coming up was the lack of longitudinal research that followed students' progress through full Test-Teach-Test cycles. There is a dearth of longitudinal research that follows students' progress throughout Test-Teach-Test treatments, and 33 gaps were found in the 33 documents that were considered for inclusion in the review. Consequently, the purpose of this research was to plot information from real-world EFL and ESL classrooms in order to solve this problem.

3. Results

33 included articles from Scopus publications are synthesized into 4 interconnected stages for a Test-Teach-Test design. The first Test stage is

defined as the diagnostic test to answer RQ1; the Teach stage is designed as differentiated interventions based on the needs identified in the diagnostic test to answer RQ2; and the second Test stage is established as the post-teaching assessments to answer RQ3. Additionally, the teaching media and learning technologies are included as part of the test and intervention tools to answer RQ4. All these designs are based on the findings of the included articles, as they have high accuracy in field-based data. Additionally, four Research Questions (RQs) addressing the Test-Teach-Test cycle for TOEFL preparation yielded

33 included articles. Figure 3 illustrates the distribution.

3.1. Functions of the First Test or Diagnostic Test in TOEFL Preparation (RQ1)

Based on the analysis of 33 included articles, the first Test or the diagnostic test picks up 5 ones which construct three classifications, namely: 1) diagnostic approach; 2) addressed skill Areas; 3) directions for the Teach stage to be merged in the lesson plans. In summary, Table 3 presents the classifications of the first Test (diagnostic test) approach in TOEFL preparation.

Table 3. Classifications of the first Test (diagnostic test) stage for TOEFL preparation

Diagnosics Approach	TOEFL Skill Areas	Directions for Teach Stage
Automated Writing Evaluation (AWE) (Bridgeman & Ramineni, 2017)	Writing Skill in TOEFL iBT	Lesson plans with scalable automated diagnostic test or pre-test for TOEFL preparation program
Placement test by the directions of students' prior knowledge of TOEFL ITP (Ferris et al., 2017)	All skills in TOEFL ITP	Lesson plans with diagnostic tests for students' multilingual backgrounds
Usage-based diagnostic assessment (Dolgova & Siczek, 2019)	Writing and Reading Skills in TOEFL iBT	Lesson plans with validated procedures for diagnostic tests based on students' needs analysis not on only standardized procedures.
Lexical Feature Analysis (LFA) (Vo, 2019)	Writing in TOEFL iBT	Lesson plans with FLA diagnostic test for EFL students.
Cognitive Diagnostic Model (CDM) (Mirzaei et al., 2020)	Reading in TOEFL ITP or iBT	Lesson plans with CDM as the framework for diagnostic test.

To replace human raters in a diagnostic test, Automated Writing Evaluation (AWE) can reliably serve as a judgment mechanism in realistic writing settings, making it an adaptable pre-test tool for TOEFL preparation classes (Bridgeman & Ramineni, 2017). In a multilingual setting with heterogeneous levels of English proficiency among students, student voice is incorporated into the diagnostic test for pre-TOEFL classes to determine placement (Ferris et al., 2017). As a reference for placement decisions, Dolgova and Siczek (2019) have established a diagnostic assessment using a bottom-up mechanism with responsive instruments. Furthermore, Vo (2019) adds this mechanism to

lexical feature analysis in EFL academic writing to serve as a reliable marker of proficiency gaps in TOEFL preparation classes. Meanwhile, Mirzaei et al. (2020) used a Cognitive Diagnostic Model (CDM) to diagnose the TOEFL reading section. As a transferable methodological test, the CDM model can identify students' sub-skill scarcities.

3.2. The Teach Stage: Differentiated Instructional Interventions (RQ2)

The study of the 33 included articles shows that 10 articles discuss the Teach stage as the second and most distinct stage of the TTT cycle. These 10 articles form three classifications: (1)

grammar-focused and extensive reading interventions, (2) technology-mediated and blended learning approaches, and (3) motivational and self-regulatory instructional strategies. These classifications are separate but complementary teaching paths selected by

instructors in response to proficiency gaps identified in the initial Test (diagnostic) stage. In summary, Table 4 shows the distribution of the 10 articles across the three Teach stage classifications.

Table 4. Classifications of the Teach stage for TOEFL preparation

Teach Approach	TOEFL Skill Areas	Directions for Teach Stage
Extensive reading + translation (Lee et al., 2016)	Reading Skill in TOEFL ITP	Meaning-focused input reinforces implicit grammar acquisition.
Task-supported language teaching (TSLT) (Chou, 2020)	Reading Skill in TOEFL iBT	Task-based production improves comprehension of target grammatical structures.
Asynchronous blended learning (Bader Al Bataineh et al., 2019; Pinto-Llorente et al., 2017)	Writing Skill in TOEFL iBT	Self-paced LMS instruction accommodates diverse proficiency levels in TOEFL classes.
Game-based learning (GBL) (Eltahir et al., 2021)	All TOEFL Skills in TOEFL ITP	GBL significantly enhances student motivation, engagement, and TOEFL performance.
TTT via Telegram and Edmodo (Niaghi, 2023)	All TOEFL Skills	Technology-enhanced TTT cycles improve pragmatic awareness and sectional skill gains
Formative assessment in blended learning (Ren et al., 2026)	All TOEFL Skills	Embedded formative assessment builds learner self-regulation capacity
Self-regulated learning strategies (Bai & Wang, 2021)	Writing Skill in TOEFL ITP	Motivational beliefs drive writing performance in TOEFL preparation contexts
Teacher leadership and student motivation (Igwe & Ligaya, 2025)	All TOEFL Skills	Teacher facilitative leadership and student motivation jointly predict English performance
Mindfulness and self-regulation (Fathi & Aghamirzaei, 2026)	All TOEFL Skills	Mindfulness and self-regulation jointly predict L2 achievement outcomes

Students taking the TOEFL ITP structure and reading sections often have diagnostic exams that reveal gaps in their understanding of grammar and reading comprehension; in such circumstances, teachers are directed to evidence-based grammar and extensive reading therapies with translation. It suggests that the TOEFL's structure and reading comprehension sections, which include implicit grammar learning, may be strengthened through meaning-focused instruction. The extended reading technique is more in line with the integrated reading and

structure requirements of the TOEFL ITP and iBT exams than the usual grammar exercises that condition separate rule memory. Instead, it exposes students to grammatical structures in communicatively meaningful contexts. Moreover, TSLT can be used to teach EFL students to improve their understanding and ability to produce the target grammar structures (Chou, 2020). Structured yet meaning-embedded grammar instruction, complemented by extensive reading as an input-rich mechanism for consolidating implicit grammatical knowledge,

should be the focus of the Teach stage when diagnostic tests reveal students' specific sub-skill weaknesses, such as sentence completion, error identification, or reading comprehension questions in the TOEFL ITP's structure section.

Meanwhile, blended learning using asynchronous technology tools significantly enhanced ESL students' grammatical ability (Pinto-Llorente et al., 2017). Autonomous learning that accommodates varying skill levels is highly regarded by students, especially in TOEFL prep courses with a diverse student body, since diagnostic scores can vary widely. LMS are effective platforms for providing diverse Teach stage material to address the diagnostic gaps discovered in the first Test stage. Similarly, EFL students' grammatical performance and attitudes were improved through Moodle-based blended learning (Bader Al Bataineh et al., 2019). Given that students' disengagement and lack of motivation are well-established outcomes of traditional instruction, GBL should be incorporated into TOEFL preparation programs to demonstrate that students' motivation, engagement, and academic performance are greatly enhanced by interactive game technologies (Eltahir et al., 2021). The TTT design, using the messaging app Telegram and the educational portal Edmodo, greatly enhanced EFL students' pragmatic awareness through the Test-Teach-Test method (Niaghi, 2023). With limited time for face-to-face teaching, technology-enhanced TTT cycles work well for English as a foreign language (EFL) students in higher education (Ren et al., 2026).

Given that there are often motivational, affective, and metacognitive gaps in TOEFL preparation in addition to linguistic ones, the third sub-classification of the Teach stage focuses on the self-regulatory and motivating components of differentiated education. Among EFL students, students' ability to self-regulate their learning and

their ideas about what motivates them were strong predictors of their success in English writing (Bai & Wang, 2021). Thus, it is imperative that students' self-regulation abilities, along with their understanding of the topic and test-taking strategies, be central to TOEFL preparation classes. The significance of the Teach stage and the delivery of instructional content, along with the motivating and facilitative roles that teachers play as orchestrators of student engagement within TTT cycles. Students' motivation and teachers' leadership behaviours interacted to predict English achievement across all skill areas. The most thorough evidence for this classification used a mixed-methods research design to show how motivation, self-regulation, and mindfulness all contribute to predicting L2 performance (Fathi & Aghamirzaei, 2026). Students who have struggled with anxiety, disengagement, or attributional helplessness may benefit greatly from these motivational-regulatory interventions as they prepare for the TOEFL.

3.3. The Second Test Stage: Post-Teaching Assessments (RQ3)

11 of the 33 articles reviewed focus on the second Test stage, which evaluates students after teaching to assess growth and the success of the interventions implemented in the Teach stage. There are three main groups into which these publications fall: (1) methods for evaluating TOEFL writers' abilities and performance; (2) assessments of post-test proficiency and washback effects; and (3) strategies for both formative and summative assessment of TOEFL training programs. The second Test stage is not only a summative endpoint but also indicates progress and determines whether the cycle should continue or end. In summary, Table 5 summarizes the distribution of the 11 articles across the three classifications of the second Test stage.

Table 5. Classifications of the second Test (post-teaching assessment) stage for TOEFL preparation

Assessment Approach	TOEFL Skill Areas	Assessment Function
---------------------	-------------------	---------------------

Writing performance measurement by textual features (Kim, 2017; Riazi, 2016)	Writing Skill in TOEFL iBT	As a post-Teach writing skill measurement to TOEFL task demands
CEFR-linked standard setting for writing post-tests (Fleckenstein et al., 2020)	Writing Skill in TOEFL iBT	As a validated proficiency benchmark for interpreting post-test results
Voice and rhetorical analysis (Geide-Stevenson, 2018)	Writing and Reading Skills in TOEFL iBT	As a capture of rhetorical developments as a dimension of post-teaching progress
TOEFL ITP/iBT score outcomes for proficiency post-test results (Golubovich dkk., 2018; Li & Xiong, 2018)	All Skills in TOEFL ITP or iBT	As TOEFL score validations for credible post-Teach outcome measures
Washback assessment of TOEFL on autonomy (Sadeghi et al., 2021)	All Skills in TOEFL ITP or iBT	As a document of motivational and strategic behavioural changes from post-intervention
Formative post-tests by notetaking and listening comprehension (Sakurai & Spring, 2026; Siegel, 2020)	Listening Skill in TOEFL ITP or iBT	As parts of a fine-grained skill-specific assessment from Teach stage interventions
Game-based post-assessment (Pratiwi & Waluyo, 2022)	All Skills in TOEFL ITP or iBT	As a document in blended/emergency remote teaching contexts
AI-automated essay scoring for post-teaching feedback (Mizumoto & Eguchi, 2023)	Writing Skill in TOEFL ITP	As a scalable automated scoring within the iterative cycles

According to Table 5 under the second Test stage, the first classification is the TOEFL writing performance assessment. The integrated and autonomous writing tasks included in TOEFL iBT preparation programs will find this particularly helpful. Writing performance measurement by textual features should be adjusted to mirror the writing requirements of the TOEFL exam rather than the general academic writing tasks on the TOEFL iBT on academic assignments (Riazi, 2016), and influenced by their prior knowledge of writing requirements (Kim, 2017), which affected their test performance and how they interpreted their scores. This highlights the importance of feedback systems in the second stage of the test. To create cut scores for the TOEFL iBT writing rubrics that are connected to the CEFR (Fleckenstein et al., 2020). In the context of internationally recognized competence standards, their work provides a validated benchmark framework for teachers to interpret post-training test scores. In contrast to the standard TOEFL post-tests, additional research on voice in timed L2 argumentative essays in

TOEFL iBT writing and rhetorical analysis provides diagnostic insights into students' development of rhetorical abilities (Geide-Stevenson, 2018).

The second Test stage consists of two sections: evaluations of students' level of competence after the test and the impact the TOEFL has on their motivation and study habits. The relevance of TOEFL results as post-teaching outcome measures in the cycles was confirmed by Geide-Stevenson (Geide-Stevenson, 2018), who found a strong correlation between English ability, as measured by post-TOEFL exams, and academic performance in higher education. The widespread opinion that post-test TOEFL ITP scores are reliable indicators of language ability for decisions at the institutional and programmatic levels, as reflected in their use in three countries (Golubovich et al., 2018). A well-designed intervention during the teaching stage, based on diagnostic data from the first exam, resulted in measurable improvements in post-test scores (Li & Xiong, 2018). (Sadeghi et al., 2021)

examined the impact of a local English Proficiency Exam and the TOEFL iBT on students' motivation, autonomy, and language-learning strategies. Those studies are connected to the discussion of proficiency-based post-test evaluations and washback effects. Students' performance on subsequent exams improved significantly after a well-planned intervention in the classroom that was grounded on diagnostic data from the first assessment (Li & Xiong, 2018). Research by Sadeghi et al. (2021) looked at how students' motivation, autonomy, and language-learning practices were affected by a local English Proficiency Exam and the TOEFL iBT. Discussions of washback effects and proficiency-based post-test ratings are related to those studies.

The further division concerns the formative and summative functions of the second Test stage, which is a component of many preparation programs. The academic performance of advanced and intermediate L2 learners improved significantly after receiving explicit instruction in note-taking (Siegel, 2020). While the COVID-19 pandemic was underway, two colleges in Indonesia offered an online TOEFL prep course that used a mix of task and game-based learning (Pratiwi & Waluyo, 2022). Improving students' TOEFL scores is still possible with well-designed post-teaching evaluations within a TTT framework, even when teaching remotely is an

emergency. Mizumoto and Eguchi (2023) noted that automated second-test assessments may be employed to enable standard artificial intelligence language models to be used for automated essay grading. These evaluations might provide quick, scalable feedback. AI-generated scores are highly correlated with the scores provided by human raters. The existing ways of post-lesson assessment in TOEFL listening preparation are useless, and the quantity of notes was not a predictor of the post-scores for students' lecture listening comprehension, but the quality of the notes was (Sakurai & Spring, 2026).

3.4. Digital Resources and Teaching Media as TTT Enablers in TOEFL Preparation (RQ4)

From the 33 articles that were considered, 7 addressed the use of instructional media and digital resources in TOEFL preparation settings as aids for the Test-Teach-Test cycle. Digital assessment and testing media for both the Test and Teach phases, and technology-enhanced instructional media for the Teach stage, are the two main categories into which these articles fall. Digital technologies are becoming more integral to the design, delivery, and assessment of the TTT cycle. This result means that they are no longer an add-on to TOEFL preparation programs. Table 6 lists the classifications of digital resources and teaching media.

Table 6. Classifications of Digital Resources and Teaching Media

Digital Resource / Technology	Applied Stage in a TTT Cycle	TOEFL Skill Areas	Information
Technology-mediated teaching & feedback practices (Barnes, 2017)	Teach and second Test	All Skills	Quality washback significantly linked to learner-centered, technology-mediated teaching practices; positive feedback loops reinforce TTT cycle continuity.
Keyboard type at TOEFL testing centres (Ling, 2017)	Second Test	Writing Skill	Digital test delivery is format-neutral for writing performance.
Timed digital essay writing platforms (Zhao, 2017)	Teach	Writing Skill	Voice development in timed L2 writing is sensitive to instructional scaffolding.
Corpus analysis software for academic writing (Lan & Sun, 2019)	Second Test	Writing Skill	As TOEFL score validations for credible post-Teach outcome measures.

CLIQ: Intelligent Question Classifier, AI-powered Software Impacts (Jardim dkk., 2022)	All Stages	All Skills	AI classifier automates cognitive-level categorization of TOEFL-aligned items; enables efficient, scalable diagnostic and post-teaching test construction.
MALL (Mobile-Assisted Language Learning) (Pratiwi dkk., 2023)	All Stages	All Skills	MALL significantly improves L2 learning outcomes in intensive English program.
Private tutoring digital platforms (Lyu & Lam, 2025)	Teach	All Skills	Private tutor agency using digital tools mediates the cycle's effectiveness under regulatory constraints.

According to Table 6, digital technologies contribute to all three stages of the TOEFL preparation cycle. CLIQ presented an intelligent question classifier (Jardim et al., 2022) to automatically classify test questions by cognitive level and subject region. The classification is based on Bloom's taxonomy, which is consistent with the multi-sectional cognitive standards of TOEFL. The Mobile-Assisted Language Learning (MALL) was shown to be very beneficial for students in an intensive English program (Pratiwi et al., 2023). This finding lends credence to the idea that mobile and digital platforms can play a significant role in facilitating the implementation of all three stages of the TTT, even in universities with limited resources, such as those in Indonesia. A Bourdieusian approach was designed to look at private English tutor agencies in China under the 'double reduction' policy (Lyu & Lam, 2025).

The results demonstrated that outstanding teaching methods, i.e., those that are learner-responsive and feedback-oriented, were substantially associated with positive learning outcomes in the case of TOEFL iBT (Lyu & Lam, 2025). This finding adds to the evidence that, for the cycle to be effective, teachers must change from a content-oriented to a more facilitative, technology-augmented role. The elements of digital administration do not substantially impact the total scores, but the keyboard type in the TOEFL iBT writing results indicates the growing integration of digital interfaces into TOEFL assessment scenarios (Ling, 2017). (Lan & Sun, 2019; Zhao, 2017) examined how writing works by demonstrating

that instructional scaffolding affected the voice aspect of timed second-language argumentative writing, an important component of the TOEFL iBT writing section.

The most sophisticated digital TTT implementations in TOEFL preparation include AI-based assessment systems, mobile learning platforms, and corpus-based diagnostic instruments. Automated essay scoring systems based on LLMs may reach the dependability of trained human raters for TOEFL writing assignments, offering scalable, instantaneous response in second Test cycles for digital TTT implementations (Mizumoto & Eguchi, 2023). The strongest washback impact occurs when instructors utilize post-test data for continual modification of educational procedures. This result implies that technology-mediated formative assessment dashboards might be a means to dramatically increase the pedagogical responsiveness of the cycle (Barnes, 2017). Also, the continuous self-regulation engine, enabled by the systematic inclusion of digital formative assessment in the Teach stage, is very helpful for the outcomes of TOEFL listening and integrated skills, particularly for engineering and EAP students who need TOEFL scores for academic promotion (Ren dkk., 2026; Sakurai & Spring, 2026). The convergence of the findings points to the future of Test-Teach-Test for TOEFL preparation: the development of fully digitalized, adaptive TTT systems with AI-driven diagnostics, need-responsive teaching, materials, and automated scoring of post-teaching evaluations.

4. Discussion

A review of 33 articles from the Scopus database has shown that the Test-Teach-Test (TTT) strategy is not just a list of processes, but a systematic and evidence-based approach to teaching. Therefore, the first Test is diagnostic testing, the Teach is instruction, and the second Test is post-teaching evaluation. They are not independent. They are interrelated and sequential. A significant trend is also evident in the distribution of the research across the phases. Most of the articles are about the second Test stage with $n = 11$, 33.3%, followed by the Teach stage with $n = 10$, 30.3%, teaching media and technology with $n = 7$, 21.2%, and the first Test stage with $n = 5$, 15.2%. This discrepancy is indicative of a broader trend in language assessment research: post-instruction assessment is more common than diagnostic assessment. But this trend also points to a glaring gap. The diagnostic stage should form the basis of the entire TTT cycle and be the least investigated. This has substantial consequences for teaching practice. Instruction may still be differentiated, but without a significant diagnostic step, there are no obvious, distinct beginning points.

4.1. Functions of the First Test or Diagnostic Test in TOEFL Preparation (RQ1)

An important but controversial argument is made across the five sections on the diagnostic stage: that the first Test stage is more successful when language ability is broken down into distinct sub-skills. The diagnostic function becomes almost useless without this degree of specificity. Nevertheless, it is important to investigate this allegation further, particularly given that the TOEFL ITP score report remains the most popular method for university placement in Indonesia. The TOEFL ITP is not designed to be a reliable tool for diagnosing certain subskills. When put into reality, this causes serious issues. As an example, students may have vastly different underlying ability profiles if they achieve the same CEFR B2 benchmark score of 543, which is essentially the sum of these three section scores (Global Exam, 2020). While one may do very well with inferential reading but poorly with grammar,

the other may exhibit the reverse trend (Barnes, 2017). Despite these distinctions, this study found that students are considered on equal footing and usually take the same standardized tests (Kim, 2017). As the answer, the Cognitive Diagnostic Model (CDM) method zeroes in on whether each student has mastered the specific cognitive sub-skills required by TOEFL tests.

There is good reason to think that this is the best way to go. An independent multi-CDM research with 886 Chinese university EFL learners obtained a test-level classification accuracy of 72.1% (Wang et al., 2025). This may not sound like a big deal, but it is important when contrasted to a baseline classifier, which would generally only get 1268-1286% accuracy in EFL groups like this. But more importantly, the sub-skill reliability of each of the six inferential reading sub-skills ranged from 87% to 95.8%, well above the 0.80 level commonly applied in applied language testing to ensure the reliability of diagnostic categorization conclusions (Dolgova & Siczek, 2019; Fan et al., 2021; Helleve et al., 2023; Mirzaei et al., 2020; Treadaway & Read, 2024). And there's more to this data than just the numbers. Their findings give significant evidence that CDM-based diagnostic instruments may be more effective in characterizing the cognitive strengths and limits of pupils than holistic placement scores.

4.2. The Teach Stage: Differentiated Instructional Interventions (RQ2)

10 articles covering the Teach stage all agree on one thing: improving students' TOEFL scores requires more than a single type of education. Lee et al. (2016) discovered that on the grammar and reading route, there were statistically significant improvements in reading comprehension and grammar when extended reading (ER) was combined with translation. Scalable differentiation is enabled by a technology-mediated approach that helps overcome these restrictions. The meta-analytic analysis showed an effect size of $ES = 2.338$ across 28 studies for blended learning grammar training. However, the methodological heterogeneity of the control circumstances warrants cautious interpretation of

this finding. A big meta-analysis of 49 studies with 5,919 people found that ER doesn't make any difference in reading competence overall.

In contrast to more complicated structures, an improvement of just 39.7 %, increases for simpler structures reached 44.1% (Ananda & Asiza, 2025). Found in this study on grammaticality judgment assessments, conditions that included explicit instruction with within-task feedback also routinely performed better than task-only conditions (Lee dkk., 2016). Taken as a whole, these results indicate that TSLT works better when targeted at a specific, well-defined sub-skill rather than as a blanket approach. However, the overarching trend is evident: as compared to traditional classroom teaching, blended learning regularly and significantly outperforms it. This is particularly true for the mixed-proficiency, non-English-major TOEFL groups that are prevalent in Indonesian universities.

4.3. The Second Test Stage: Post-Teaching Assessments (RQ3)

Measuring score improvement alone is insufficient to determine the efficacy of a TTT cycle, according to the 11 studies focusing on the second Test stage, the largest cluster in this evaluation, which together account for 33.3% of all included research. Found in this study regarding the validity of writing, post-lesson evaluations using general academic writing assignments, rather than assignments that mimic the TOEFL style, may skew assessments (Kim, 2017; Riazi, 2016). This is not a minor technological issue at Indonesian colleges, as post-test results are used to determine whether students meet the graduation criteria for CEFR B2 or TOEFL ITP 543 (Global Exam, 2020). Using non-aligned activities poses a significant validity threat with substantial institutional effects since students' use of TOEFL-specific rhetorical tactics relies on their knowledge of the rubric (Fleckenstein et al., 2020).

The validity of the post-test as an indicator of competence is bolstered by two further studies. Nevertheless, Li and Xiong's most practically relevant discovery is that diagnostic anchoring, not the amount of test prep, is the main factor in

students' post-exam score improvement (Li & Xiong, 2018). An essential caveat is introduced by the washback results of Sadeghi et al. (2021). There was no discernible washback impact on student motivation from the TOEFL iBT, but a locally developed English proficiency test showed varying effects, depending on students' levels of competence. Any TTT cycle design that depends only on score improvement to maintain student interest across iterations is thus called into question. In other words, for the cycle to keep going, the second stage of the test must record more than just students' scores; it must show how their autonomy, strategy use, and self-efficacy have changed. Pratiwi and Waluyo's findings are among the most applicable to the given setting (Pratiwi & Waluyo, 2022). These results, when considered as a whole, demonstrate that the second Test stage serves as more than a simple final assessment of the lesson. Institutional accountability demands are met by summative score reporting, and students' continued growth is supported through formative feedback on method, strategy, and effect. Neither can be missing from the TTT cycle if it is to serve as an evidence-driven and truly iterative method of TOEFL preparation.

4.4. Digital Resources and Teaching Media as TTT Enablers in TOEFL Preparation (RQ4)

7 articles on educational media and online resources show that they offer the proliferation of standalone platforms and the advent of technological arrangements that can cater to all three TTT phases concurrently, which gives this change its analytical significance. The most prominent instances of this are the protocols MALL for content distribution, CLIQ for diagnostic item generation, and the full-cycle integration between Telegram and Edmodo (Jardim et al., 2022). Out of all these cross-stage enablers, MALL is the most flexible. Depending on students' initial proficiency levels, these increases might correspond to 1268-1286 score points on the TOEFL ITP. Teachers sometimes make the mistake of oversampling lower-order recall questions when creating first-test instruments without a taxonomy-aligned

categorization system. In doing so, they overlook the higher-order inferential and evaluative items that have a greater impact on TOEFL reading and writing results. Digital technologies may actually transform post-test data into improved training, according to Barnes' study on technology-mediated washback (Barnes, 2017; Sadeghi et al., 2021). What matters most is not the platform a teacher uses, but rather how often they incorporate post-test findings into their teaching and use evidence. The most powerful positive washback effects were produced by learner-responsive, feedback-oriented practices.

In a TTT cycle, the true benefit of technology is realized only when instructors actively respond to what students show. The LLM-human agreement ranges from $r = 0.30$ to 0.80 . As high student-to-teacher ratios and sluggish turnaround times for writing feedback are two well-known challenges to conducting effective iteration cycles in Indonesian TOEFL preparation programs, this scalability is especially advantageous (Jardim et al., 2022). The integration of MALL-based delivery with $g = 0.88$, CLIQ-based diagnostic construction, and AI-assisted formative scoring indicates the direction of a fully digitalized adaptive TTT system. This system would undergo iterative cycles that are progressively faster, more accurately calibrated to individual sub-skill profiles, and more scalable than they were before. In terms of providing comprehensive and fair TOEFL preparation at the institutional level, this is the best and most practical option for Indonesian institutions.

5. Conclusion

This systematic review of 33 Scopus-indexed studies synthesizes an empirically grounded Test-Teach-Test cycle for TOEFL preparation. Its four stages, diagnostic testing, differentiated teaching, post-teaching assessment, and digital enablement, are interdependent rather than sequential add-ons: diagnostic data give the Teach stage its targets, and the second Test stage determines whether the cycle should end or continue. The central finding is a structural imbalance in the field; diagnostic-stage research is markedly under-represented relative to post-

teaching assessment research, which matters because the diagnostic stage is the foundation on which the rest of the cycle depends. Digital tools (mobile-assisted learning, AI-assisted scoring, automated item classification) emerge as the infrastructure that makes full-cycle, personalized delivery feasible at an institutional scale. Practically, this framework offers Indonesian TOEFL preparation programs a concrete alternative to uniform, grammar-centred instruction: formal diagnostic screening as the entry point for teaching decisions, and post-teaching assessment that tracks learner autonomy and self-efficacy alongside scores. This has particular urgency for non-English-major students, for whom a TOEFL ITP score of 543 (or CEFR B2 equivalent) is often a graduation gate rather than a chosen goal as a diagnostically personalized approach can make that requirement attainable through targeted instruction rather than repeated, undifferentiated test drilling.

It is important to state plainly what this review's evidence does, and does not, support. This collection of 33 papers varies considerably in student demographics, methodology (descriptive, experimental, mixed-methods), and TOEFL version (ITP and iBT), which precludes a single aggregated cycle-level effect size and limits direct cross-study comparison. No included study compared a complete TTT cycle against a matched non-TTT control using longitudinal experimental methods; the cycle-level claim that TTT outperforms conventional instruction is therefore theoretically well-supported but not yet experimentally confirmed, and future randomized or quasi-experimental longitudinal designs tracking full cycle iterations across semesters are needed to close this gap. Finally, the reviewed literature gives limited attention to how teacher identity, institutional policy, and resource constraints affect the fidelity of TTT implementation, factors that matter directly for Southeast Asian and Indonesian higher education, where regulatory frameworks, staffing ratios, and digital-resource inequality shape what is achievable in practice, and which future mixed-methods research should address directly.

Despite these limitations, the pattern across all four research questions is consistent, and it points toward the same practical priority: Indonesian and Southeast Asian higher education institutions preparing non-English-major students for

TOEFL need a diagnostic-first, cycle-based instructional model, not merely faster or more digital versions of the conventional, undifferentiated approach.

6. References

- AMINEF - American Indonesian Exchange Foundation. (2025). Fulbright Master's Degree Scholarship. *Fulbright Indonesia*. <https://www.aminef.or.id/grants-for-indonesians/fulbright-programs/scholarship/fulbright-masters-degree-scholarship/>
- Ananda, J., & Asiza, N. (2025). From theory to practice: The impact of task-based learning on grammar proficiency in EFL teaching. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, 13(1), 1622–1637. <https://doi.org/10.24256/ideas.v13i1.6424>
- Bader Al Bataineh, K., Abdullah Ahmed Banikalef, A., & H. Albashtawi, A. (2019). The effect of blended learning on EFL students' grammar performance and attitudes: An investigation of Moodle. *Arab World English Journal*, 10(1), 324–334. <https://doi.org/10.24093/awej/vol10no1.27>
- Bai, B., & Wang, J. (2021). Hong Kong secondary students' self-regulated learning strategy use and English writing: Influences of motivational beliefs. *System*, 96, 102404. <https://doi.org/10.1016/j.system.2020.102404>
- Barkaoui, K. (2024). The academic achievement of undergraduate students with different TOEFL iBT score profiles: A replication study. *Language Assessment Quarterly*, 21(4–5), 350–378. <https://doi.org/10.1080/15434303.2024.2438142>
- Barnes, M. (2017). Washback: Exploring what constitutes “good” teaching practices. *Journal of English for Academic Purposes*, 30, 1–12. <https://doi.org/10.1016/j.jeap.2017.10.003>
- Bridgeman, B., & Ramineni, C. (2017). Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings. *Assessing Writing*, 34, 62–71. <https://doi.org/10.1016/j.asw.2017.10.001>
- Chou, M.-H. (2020). Task-supported language teaching to enhance young EFL adolescent learners' comprehension and production of English phrasal verbs in Taiwan. *Education* 1268-1286, 48(4), 455–470. <https://doi.org/10.1080/03004279.2019.1617328>
- Crystal, D. (2003). *English as a global language* (2nd ed). Cambridge University Press.
- Dolgova, N., & Siczek, M. (2019). Assessment from the ground up: Developing and validating a usage-based diagnostic assessment procedure in a graduate EAP context. *Journal of English for Academic Purposes*, 41, 100771. <https://doi.org/10.1016/j.jeap.2019.100771>
- Du, Q. (2025). How artificially intelligent conversational agents influence EFL learners' self-regulated learning and retention. *Education and Information Technologies*, 30(15), 21635–21701. <https://doi.org/10.1007/s11268-1286-11268-1286>
- Efendi, A., Shao, Y.-H., & Huang, C.-Y. (2025). Technological development and optimization of pushing and grasping functions in robot arms: A review. *Measurement*, 242, 115729. <https://doi.org/10.1016/j.measurement.2024.115729>
- Eltahir, Mohd. E., Alsalthi, N. R., Al-Qatawneh, S., AlQudah, H. A., & Jaradat, M. (2021). The impact of game-based learning (GBL) on students' motivation, engagement and academic performance on an Arabic language grammar course in higher education. *Education and Information Technologies*, 26(3), 3251–3278. <https://doi.org/10.1007/s11268-1286-10396-w>
- ETS. (2026). *The TOEFL ITP Assessment Series*. Test Content. <https://www.ets.org/toefl/itp/test-content.html>
- Fan, T., Song, J., & Guan, Z. (2021). Integrating diagnostic assessment into curriculum: A theoretical framework and teaching practices. *Language Testing in Asia*, 11(1), 2. <https://doi.org/10.1186/s41268-1286-00117-y>
- Fathi, J., & Aghamirzaei, A. (2026). Interplay of mindfulness, motivation, and self-regulation in predicting L2 achievement: A mixed-methods study. *Acta Psychologica*, 262, 106059. <https://doi.org/10.1016/j.actpsy.2025.106059>
- Ferris, D. R., Evans, K., & Kurzer, K. (2017). Placement of multilingual writers: Is there a role for student voices? *Assessing Writing*, 32, 1–11. <https://doi.org/10.1016/j.asw.2016.10.001>
- Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, 43, 100420. <https://doi.org/10.1016/j.asw.2019.100420>
- Geide-Stevenson, D. (2018). Does English proficiency affect academic performance? *International Review of Economics Education*, 28, 41–48. <https://doi.org/10.1016/j.iree.2018.04.002>
- Global Exam. (2020, Februari 6). TOEFL ITP: Score conversion into CEFR levels. *GlobalExam Blog*. <https://global-exam.com/blog/en/toefl-ity-score-conversion-into-cefr-levels/>

- Golubovich, J., Tolentino, F., & Papageorgiou, S. (2018). Examining the applications and opinions of the TOEFL ITP® assessment series test scores in three countries. *ETS Research Report Series*, 2018(1), 1–30. <https://doi.org/10.1002/ets2.12231>
- Helleve, I., Eide, L., & Ulvik, M. (2023). Case-based teacher education preparing for diagnostic judgement. *European Journal of Teacher Education*, 46(1), 50–66. <https://doi.org/10.1080/02619768.2021.1900112>
- Igwe, M. C. P., & Ligaya, M. A. D. (2025). Students' performance in English: Effects of teachers' leadership behavior and students' motivation. *Journal of Education and Learning (EduLearn)*, 19(1), 91–102. <https://doi.org/10.11591/edulearn.v19i1.21680>
- Illananingtyas, T. (2025). The implementation of the direct method in teaching speaking at university level. *International Conference on English Language Teaching (ICON-ELT)*, 6(1), 310–324. <https://new-conference.unisma.ac.id/index.php/ICON/article/view/1999>
- Indonesian International Education Foundation. (2026). *Testing & Certification*. <https://www.iief.or.id/testing-certification>
- International Test Center. (2026). *TOEFL®*. <https://itc-indonesia.com/toefl/>
- Jardim, R. R. J., Delgado, C., & Silva, M. F. (2022). CLIQ! Intelligent Question Classifier for the elaboration of exams. *Software Impacts*, 13, 100345. <https://doi.org/10.1016/j.simpa.2022.100345>
- Jenkins, J., Cogo, A., & Dewey, M. (2011). Review of developments in research into English as a lingua franca. *Language Teaching*, 44(3), 281–315. <https://doi.org/10.1017/S0261444811000115>
- Kholiti, Y. E., & Ghouati, A. E. (2025). From chalk and talk to interactive learning: A study on the evolution of EFL teaching methods. *Journal of English Language Teaching and Applied Linguistics*, 7(4), 41–46. <https://doi.org/10.32996/jeltal.2025.7.4.5>
- Kim, E.-Y. J. (2017). The TOEFL iBT writing: Korean students' perceptions of the TOEFL iBT writing test. *Assessing Writing*, 33, 1–11. <https://doi.org/10.1016/j.asw.2017.02.001>
- Lan, G., & Sun, Y. (2019). A corpus-based investigation of noun phrase complexity in the L2 writings of a first-year composition course. *Journal of English for Academic Purposes*, 38, 14–24. <https://doi.org/10.1016/j.jeap.2018.12.001>
- Lee, J., Schallert, D. L., & Kim, E. (2016). Effects of extensive reading and translation activities on grammar knowledge and attitudes for EFL adolescents. *System*, 52, 38–50. <https://doi.org/10.1016/j.system.2015.04.016>
- Li, H., & Xiong, Y. (2018). The relationship between test preparation and state test performance: Evidence from the Measure of Effective Teaching (MET) project. *Education Policy Analysis Archives*, 26, 64. <https://doi.org/10.14507/epaa.26.3530>
- Ling, G. (2017). Are TOEFL iBT® writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing*, 31, 1–12. <https://doi.org/10.1016/j.asw.2016.04.001>
- Lyu, M., & Lam, R. (2025). Private English tutors' agency amid China's 'double reduction' policy: A Bourdieusian perspective. *Journal of Education Policy*, 40(4), 667–688. <https://doi.org/10.1080/02680939.2025.2479151>
- Mirzaei, A., Heidari Vincheh, M., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation*, 64, 100817. <https://doi.org/10.1016/j.stueduc.2019.100817>
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Niaghi, M. R. (2023). Promoting EFL learners' pragmatic awareness through Test-Teach-Test (TTT) method via Telegram and Edmodo. *Mextesol Journal*, 46(4), 1–11. <https://doi.org/10.61871/mj.v46n1268-1286>
- Pinto-Llorente, A. M., Sánchez-Gómez, M. C., García-Peñalvo, F. J., & Casillas-Martín, S. (2017). Students' perceptions and attitudes towards asynchronous technological tools in blended-learning training to improve grammatical competence in English as a second language. *Computers in Human Behavior*, 72, 632–643. <https://doi.org/10.1016/j.chb.2016.05.071>
- Pratiwi, D. I., Amumpuni, R. S., Fikria, A., & Budiastuti, R. E. (2023). Enhancing students' learning outcomes through MALL in English intensive program for railway mechanical technology. *International Journal of Language Education*, 7(2). <https://doi.org/10.26858/ijole.v7i2.22839>
- Pratiwi, D. I., & Waluyo, B. (2022). Integrating task and game-based learning into an online TOEFL preparatory course during the COVID-19 outbreak at two Indonesian higher education institutions. *Malaysian Journal of Learning and Instruction*, 19. <https://doi.org/10.32890/mjli2022.19.2.2>
- Ren, H., Zhang, W., & Cao, H. (2026). Formative assessment as a self-regulation engine in blended language learning: A case study with top-tier engineering talents. *Acta Psychologica*, 264, 106505. <https://doi.org/10.1016/j.actpsy.2026.106505>
- Riazi, A. M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing*, 28, 15–27. <https://doi.org/10.1016/j.asw.2016.02.001>

- Sadeghi, K., Ballıdağ, A., & Mede, E. (2021). The washback effect of TOEFL iBT and a local English Proficiency Exam on students' motivation, autonomy and language learning strategies. *Heliyon*, 7(10), e08135. <https://doi.org/10.1016/j.heliyon.2021.e08135>
- Sakurai, S., & Spring, R. (2026). Beyond word count: The role of note quality on EFL students' lecture listening comprehension. *System*, 137, 103927. <https://doi.org/10.1016/j.system.2025.103927>
- Sickinger, R., Pill, J., & Brunfaut, T. (2025). Exploring the scoring validity of holistic and dimension-based Comparative Judgements of young learners' EFL writing. *Assessing Writing*, 66, 100986. <https://doi.org/10.1016/j.asw.2025.100986>
- Siegel, J. (2020). Effects of notetaking instruction on intermediate and advanced L2 English learners: A quasi-experimental study. *Journal of English for Academic Purposes*, 46, 100868. <https://doi.org/10.1016/j.jeap.2020.100868>
- Treadaway, M., & Read, J. (2024). Setting standards for a diagnostic test of aviation English for student pilots. *Language Testing*, 41(3), 557–583. <https://doi.org/10.1177/02655322231224051>
- U.S. Sports Scholarships. (2026). *TOEFL Requirements for International Athletes*. TOEFL Requirements for International Athletes. <https://www.ssusa.co.uk/toefl-exam>
- Vo, S. (2019). Use of lexical features in non-native academic writing. *Journal of Second Language Writing*, 44, 1–12. <https://doi.org/10.1016/j.jslw.2018.11.002>
- Wang, X., Zhang, G., & Chen, L. (2025). How teachers scaffold student discourse moves and reading comprehension through Knowledge Building pedagogy: Insights from Chinese EFL reading classrooms. *Reading and Writing*. <https://doi.org/10.1007/s11268-1286-10733-w>
- Yu, M. (2026). Optimizing EFL vocabulary acquisition: A randomized controlled mixed-methods investigation of artificial intelligence-driven incidental, contextual, and multimodal strategies. *Education and Information Technologies*, 31(1), 53–97. <https://doi.org/10.1007/s11268-1286-11268-1286>
- Zhao, C. G. (2017). Voice in timed L2 argumentative essay writing. *Assessing Writing*, 31, 73–83. <https://doi.org/10.1016/j.asw.2016.08.004>

7. Acknowledgements

The authors declare that they use generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the authors used Grammarly Pro and Quillbot Pro to improve the language, grammar, and readability of the manuscript, Zotero to manage the references, and VOSviewer to explore knowledge networks. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the final contents of the publication. The authors affirm that no generative AI tools were used to create, interpret, or analyse scientific data, figures, or core ideas.

8. Author Declaration

The authors declare that they have no conflict of interest. To be eligible for authorship, each author of this research has contributed to at least one of the following: conceptualization, methodology, formal analysis, or investigation, as well as at least one aspect of writing (either original draft preparation or writing reviews and editing).

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Umi Dian Adhitya Wulan Ningrum	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
Margana		✓				✓		✓	✓	✓	✓	✓		
Yuyun Yulia	✓		✓	✓			✓			✓	✓		✓	✓
Muhammad Kristiawan					✓		✓			✓		✓		✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition