

Evaluating Program Coherence and Participant Progress in a TOEFL Preparation Course Through a Logic Model

¹*Taufik Suadiyatno, ¹Ni Nyoman Padmadewi, ¹I G.A. Lokita Purnamika Utami,
¹Putu Suarcaya

¹English Education Department, Postgraduate Program, Universitas Pendidikan Ganesha, Jln. Udayana
Nomor 11, Singaraja, Buleleng District, Bali, Indonesia

*Corresponding Author e-mail: taufik.suadiyatno@student.undiksha.ac.id

Received: June 2026; Revised: July 2026; Accepted: August 2026; Published: September 2026

Abstract

Conventional evaluations of TOEFL preparation frequently emphasize end scores or participant satisfaction, which can obscure whether institutional resources, instruction, assessment, and intended longer-term benefits form a coherent program pathway. This study evaluated the coherence and participant-progress evidence of a TOEFL preparation course at Lembaga Pengembangan Pendidikan Nusa Tenggara Barat (LPPNTB) through a logic model. A qualitative case study supplemented by descriptive quantitative data integrated stakeholder interviews, classroom observation, institutional documents, try-out score records, and alumni datasets from 2021, 2022, and 2024. Qualitative evidence was coded deductively by the logic-model categories and inductively for implementation issues; descriptive numerical evidence was embedded to corroborate or complicate the qualitative interpretation. The findings indicate a generally coherent relationship among institutional support, trainer preparation, integrated reading-structure-listening instruction, guided practice, feedback, mock tests, and the intended goal of TOEFL readiness. Among 35 participants with complete Try-Out 1 and Try-Out 3 records, the mean changed descriptively from 426.00 to 474.11, and 32 participants recorded higher later scores. This pattern does not establish causation and must be considered alongside the available pretest-to-final means, which changed from 479.79 to 476.29. Alumni records suggested perceived usefulness for scholarships, English development, and career purposes, but were retrospective and self-reported. The course therefore demonstrated coherent program logic and promising, non-causal evidence of readiness. Priority improvements include diagnostic placement, differentiated digital assignments, intensive listening support, standardized assessment documentation, and longitudinal alumni tracking.

Keywords: Language program evaluation; Logic model; Participant progress; TOEFL preparation

How to Cite: Suadiyatno, T., Padmadewi, N.Y., Utami, I I G.A.L.P., & Suarcaya, P. (2026). Evaluating Program Coherence and Participant Progress in a TOEFL Preparation Course Through a Logic Model. *Journal of Language and Literature Studies*, 6(3), 1216-1228. <https://doi.org/10.36312/jolls.v6i3.5600>



<https://doi.org/10.36312/jolls.v6i3.5600>

Copyright© 2026, Suadiyatno et al.

This is an open-access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) License.



INTRODUCTION

English proficiency certification is used in many academic and professional decisions, including admissions, scholarship selection, graduation requirements, and employment. Consequently, test-preparation courses often serve two related purposes: developing the language resources needed for academic communication and familiarizing candidates with the demands of a standardized assessment (Bachman & Palmer, 2010; Brown & Abeywickrama, 2019; Yu & Green, 2021). In the TOEFL Institutional Testing Program (ITP), the assessed domains are Listening Comprehension, Structure and Written Expression, and Reading Comprehension (Educational Testing Service). A defensible preparation course should therefore extend beyond isolated grammar drills or repeated practice tests to include purposeful language development,

strategy instruction, formative feedback, learner support, and transparent assessment procedures.

Evaluating such a course solely through a final mean score or satisfaction rating is insufficient. A score may reflect prior proficiency, differences in test difficulty, repeated exposure, attendance, motivation, or score-record quality, while positive perceptions may coexist with unresolved learning needs. Language-program evaluation must therefore examine not only whether participants report progress, but also how institutional resources, instructional practices, assessment routines, and participation conditions plausibly contribute to that progress (Kiely & Rea-Dickins, 2005; Norris, 2016; Patton & Campbell-Patton, 2022). A logic model is useful for this purpose because it makes the program theory visible by linking inputs, activities, outputs, outcomes, and intended impacts (W. K. Kellogg Foundation, 2004).

The educational problem addressed in the present case is multidimensional. Participants entered with heterogeneous English proficiency and different goals, which complicated the use of a common instructional pace and uniform digital assignments. Listening remained particularly difficult because learners had to process spoken language rapidly while managing attention, vocabulary, inference, and test timing. The institution also used Google Classroom and an online TOEFL prediction system; however, digital access alone did not guarantee pedagogically appropriate task sequencing or fair testing conditions. Contemporary evidence similarly indicates that technology-supported learning depends on the quality of the learning activities it enables, while remote language testing requires attention to fairness, access, and procedural transparency (Isbell et al., 2023; Sailer et al., 2024).

A second problem concerns the tension between language development and test drilling. Test preparation can generate positive washback when it builds strategic awareness, disciplined practice, and familiarity with academic language demands. It can also generate narrow or negative washback when teaching becomes dominated by item rehearsal and shortcut strategies that do not transfer beyond the test (Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993). More recent work shows that preparation can influence how test takers understand test constructs and integrated tasks, but such changes should not automatically be equated with broad gains in proficiency (Wang & Cheng, 2025). Evaluating the balance between meaningful instruction and test familiarization is therefore central to judging the educational value of a TOEFL course.

Previous evaluations provide important but incomplete foundations. Mazidah and Masruroh (2024) used the CIPP model to evaluate a higher-education TOEFL course and identified strengths as well as continuing needs in infrastructure, resources, and engagement. Bayram and Canaran (2019) showed that homework, examinations, online programs, course materials, pacing, and stakeholder perspectives influence the quality of an English preparatory program. Primary et al. (2023) likewise demonstrated the value of stakeholder-based evaluation for identifying curriculum expectations and implementation challenges. These studies establish the importance of systematic evaluation, but CIPP-based and perception-centered designs do not necessarily make the hypothesized pathway from resources and activities to immediate outputs, participant outcomes, and longer-term uses explicit.

Logic-model research in education has examined learning support, inclusive education, curriculum implementation, and implementation processes (Abulhul, 2023; Booker, 2019; Clapham et al., 2017; Czosnek et al., 2022; Ormsby & Morrow, 2019; Smith et al., 2020). In English language teaching, Koçer and Topkaya (2023) used a logic model to examine links between activities and outcomes in a writing course. Nevertheless, comparatively little evidence has been reported on TOEFL preparation as a complete program pathway, particularly in a local Indonesian educational institution

and through the integration of interviews, observation, documents, repeated score records, and alumni data.

The present study addresses this gap by evaluating the LPPNTB TOEFL preparation course as a bounded program. LPPNTB, previously known as Rumah Bahasa, provides face-to-face instruction, Google Classroom support, institutional monitoring, TOEFL-style practice, mock tests, and an online prediction-test system. Its syllabus integrates reading and grammar development and includes later listening practice and mock tests. The online system also requires procedural preparation involving devices, listening equipment, access, supervision, and section-by-section test rules. These features make the course an appropriate case for examining both pedagogical coherence and operational readiness.

The study contributes a logic-model account that connects institutional resources, trainer practices, learning activities, immediate implementation outputs, participant-reported outcomes, descriptive score patterns, and alumni-reported uses. Unlike an evaluation that treats a higher score or positive rating as sufficient proof of effectiveness, the present analysis distinguishes internal program coherence from evidence of participant progress and from intended or perceived longer-term impacts. This distinction is necessary because the available design and records cannot establish that the course caused score change, scholarship success, or career advancement.

Accordingly, the study addressed four research questions:

1. What inputs, activities, outputs, short-term outcomes, and intended impacts characterize the TOEFL preparation course?
2. How coherently are the course components aligned within the logic model framework?
3. What non-causal evidence of participant progress and perceived longer-term usefulness can be identified across the available data sources?
4. What improvements are needed to strengthen the quality, evidentiary basis, and sustainability of the program?

LITERATURE REVIEW

TOEFL Preparation, Learning-Oriented Assessment, and Test Familiarization

Language assessment involves claims about the abilities that matter, the tasks used to elicit those abilities, and the interpretations made from scores (Bachman & Palmer, 1996, 2010). TOEFL preparation therefore requires more than knowledge of answer patterns. Learners need relevant grammar, vocabulary, reading, and listening resources as well as the ability to manage time, identify question demands, monitor comprehension, and respond under pressure. Test preparation is most educationally defensible when practice is connected to explicit learning goals, feedback, and opportunities to revise performance (Black & Wiliam, 2009; Brown & Abeywickrama, 2019; Hattie & Timperley, 2007).

Research on admissions-test preparation shows that the form and consequences of preparation vary by context, provider, learner goal, and test design (Yu & Green, 2021). Repeated exposure may increase familiarity and strategic control, but it may also create practice effects that are not equivalent to broad language development. Wang and Cheng (2025), for example, found that TOEFL iBT preparation shaped test takers' perceptions of integrated speaking and writing tasks. Such evidence supports treating familiarity and construct understanding as meaningful outcomes while maintaining a distinction between these outcomes and causal claims about proficiency growth.

Washback and the Language-Development/Test-Drilling Tension

Washback refers to the influence of testing on teaching and learning (Alderson & Wall, 1993). In test-preparation settings, positive washback can include disciplined study,

clearer understanding of target skills, and purposeful feedback. Negative washback can arise when teaching narrows to predicted item types, memorized tricks, or short-term score maximization (Alderson & Hamp-Lyons, 1996). A logic-model evaluation can examine this tension by asking whether activities plausibly build durable language resources or primarily support familiarity with a specific test procedure.

Listening deserves particular attention because difficulties may arise from limited aural vocabulary, rapid processing, attention, inference, and insufficient metacognitive control. Process-oriented listening instruction can combine repeated exposure with prediction, verification, reflection, and strategy discussion. A recent EFL case study reported that multimedia input, collaboration, and metacognitive instruction supported learners' listening comprehension and strategic awareness (Shamsi & Bozorgian, 2024). This evidence supports intensive listening cycles that diagnose specific breakdowns rather than simply increasing the number of practice items.

Logic Models and Improvement-Oriented Language Program Evaluation

A logic model represents the expected pathway through which a program produces change. Inputs are the resources and initial conditions available to the program; activities are what the program does; outputs are the immediate products of implementation; outcomes are short- to medium-term changes; and impacts are broader or longer-term changes that the program intends to contribute to (W. K. Kellogg Foundation, 2004). The framework clarifies assumptions and helps evaluators locate both supporting evidence and breaks in the expected pathway.

The value of a logic model lies in analysis rather than diagram production. Booker (2019) cautions that logic models should be validated against stakeholder experience and contextual evidence rather than treated as static representations. Smith et al. (2020) and Czosnek et al. (2022) similarly show that logic-based evaluation can connect determinants, implementation strategies, mechanisms, and outcomes. In the present case, heterogeneous proficiency, digital access, listening difficulty, and incomplete records operate as determinants; instruction, feedback, practice, and try-outs operate as strategies; familiarity and strategic awareness are plausible mechanisms; and readiness is a short-term outcome.

Language-program evaluation is most useful when it is contextual, participatory, and oriented toward decisions for improvement (Kiely & Rea-Dickins, 2005; Norris, 2009, 2016; Patton & Campbell-Patton, 2022). A logic model complements this orientation because it can integrate stakeholder accounts with implementation records and outcome indicators. Koçer and Topkaya (2023) demonstrated that a logic model can reveal alignments and discrepancies between course activities and perceived outcomes. For the LPPNTB course, the same approach permits positive evidence to be interpreted alongside continuing listening problems, uneven task difficulty, incomplete assessment records, and the limitations of retrospective alumni reports.

Digital Support and Differentiation

Digital platforms can expand access to materials, practice, feedback, and testing, but their pedagogical value depends on the activities they enable. A second-order meta-analysis of technology-enhanced learning found that learning benefits are associated with how technology supports active, constructive, or interactive activity rather than with technology use alone (Sailer et al., 2024). In the present context, Google Classroom should therefore be evaluated as a structured learning pathway, not merely as a repository.

Heterogeneous readiness also creates a case for differentiated assignments. Differentiated instruction adjusts task complexity, support, pacing, or product expectations in response to learner readiness and needs. Recent EFL evidence indicates

that differentiated instruction can support performance and positive learner perceptions when tasks are intentionally matched to readiness levels (Mohamed et al., 2025). For a TOEFL course, differentiation may include tiered listening passages, targeted grammar pathways, optional extension tasks, and feedback linked to diagnostic profiles.

RESEARCH METHOD

Research Design and Integration Strategy

This study employed a qualitative case study supplemented by descriptive quantitative data. The TOEFL preparation course at LPPNTB was treated as a bounded case defined by its institutional setting, course cycle, participants, trainer, coordinator, curriculum, digital support, assessment procedures, and intended outcomes. The qualitative strand was primary because the central evaluative task was to explain program coherence, implementation, stakeholder experience, and improvement needs. Descriptive quantitative evidence from score records and alumni ratings was embedded within the case to corroborate, qualify, or complicate the qualitative interpretation rather than to test causal effects.

The design is consistent with case-study mixed-methods research, in which qualitative and quantitative evidence is integrated to obtain a more complete account of a bounded case (Guetterman & Fetters, 2018; Yin, 2018). Integration occurred at three points. At the design level, all sources were mapped to the same logic-model components. At the analysis level, interview, observation, document, score, and alumni evidence were compared in a joint evidence matrix. At the interpretation level, convergence, complementarity, and dissonance were examined. For example, participant reports of readiness were considered together with the Try-Out 1-to-Try-Out 3 pattern and the inconsistent available pretest-to-final pattern. The logic model therefore functioned as both the program theory and the integration framework.

Setting, Participants, and Data Sources

The case was the LPPNTB TOEFL preparation course delivered through face-to-face instruction with Google Classroom support and online test procedures. The course included integrated instruction in Reading Comprehension, Structure and Written Expression, and Listening Comprehension, followed by practice and mock-test activities. The institutional records contained 41 participants distributed across morning and afternoon classes; 35 had complete Try-Out 1 and Try-Out 3 records. Two participants were interviewed; both are 20-year-old postgraduate students enrolled in the morning class of the 2026 course. Stakeholder-selection rationale: Two key stakeholders—one trainer and one program coordinator—were selected, both of whom had been involved in the program for more than four years.

The participant interview sample was purposively selected to represent variation in class membership, learning needs, and course experience. The trainer was included because he was directly responsible for lesson preparation, instructional adaptation, feedback, and assessment practice. The coordinator was included because she held information on institutional goals, planning, trainer selection, resources, monitoring, records, and intended long-term use. No standardized entry-proficiency categories were available in the retrieved dataset; heterogeneity was therefore treated as a documented contextual condition rather than a formally classified variable.

Table 1. Participants and Data Sources

Source/group	Number	Selection or coverage	Evidence contributed
TOEFL trainer	1	Purposive; direct responsibility for instruction and	Interview; classroom implementation evidence

Source/group	Number	Selection or coverage	Evidence contributed
Program Administrator	1	assessment practice Institutional goals, planning, trainer selection, resources, monitoring, records, and intended long-term use	Interview; institutional documents
Current course participants	2	Purposive variation by class and learning experience	Interviews; observation participation
Course score records	41 total; 35 complete T1-T3 pairs	Available institutional records from morning and afternoon classes	Pretest/try-out records; missingness noted
Alumni dataset, 2021-2022	255 records	Available institutional alumni records	Retrospective self-reported uses and benefits
Alumni dataset, 2024	5 records	Available follow-up records	Ratings, reported official scores, certificate use

Instruments

Six instruments or structured extraction tools were used to align evidence with the logic model. The interview protocols addressed the same five program components while using stakeholder-specific prompts. The observation guide documented enacted instruction and participation; the document matrix compared planned and implemented features; the score form recorded available testing points and missingness; the alumni form summarized perceived uses and ratings; and the coding framework defined the logic-model categories and allowed emergent subthemes.

Table 2. Instruments and Analytical Purpose

Instrument	Main content	Logic-model use
Trainer interview protocol	Preparation, materials, adaptation, instruction, feedback, assessment, participant progress, constraints	Inputs; activities; outputs; outcomes
Participant interview protocol	Goals, experience, perceived learning, confidence, difficulties, digital tasks, recommendations	Activities; outcomes; improvement needs
Coordinator interview protocol	Institutional goals, planning, staffing, resources, monitoring, records, intended impacts	Inputs; outputs; intended impacts
Classroom-observation guide	Instructional sequence, engagement, interaction, feedback, practice, test procedures, digital support	Activities; outputs
Document-analysis matrix	Schedule, syllabus, online-test tutorial, materials, assessment and alumni records	Planned logic; implementation evidence
Score/alumni extraction forms and coding framework	Pairing, means, change, missingness, ratings, reported uses; deductive codes and emergent themes	Outputs; outcomes; perceived impacts; integration

Data Collection Procedures

Semi-structured interviews were used to obtain stakeholder accounts of program goals, implementation, perceived progress, continuing difficulties, and improvement priorities. The participants and coordinator were interviewed on 20 May 2026. The interview lasted 30 minutes for each participant. With consent, the interviews were not recorded or transcribed. The field notes were prepared immediately after the interview. Classroom observation was conducted on the fourth of May 2026 and focused on instructional activities, participant engagement, feedback, interaction, strategy explanation, practice, and test preparation. The duration of observations was 60 minutes. Institutional documents comprised the course schedule, TOEFL syllabus, online-test tutorial, assessment records, and alumni datasets. Document analysis was used to compare the written program design with observed or reported implementation (Bowen, 2009).

The score file included 41 participants across the two classes. Records were retained in their original form, and no missing score was imputed. A valid Try-Out 1-to-Try-Out 3 pair required a numeric score at both points for the same participant. Six records failed this pairing criterion because one or both scores were absent. The available pretest-to-Try-Out 3 means were retained as a cautionary comparison, although the number of valid pairs and the distributional statistics must be confirmed from the raw file. Alumni evidence comprised 255 records from 2021-2022 and five records from 2024. The institutional files did not provide a complete sampling frame or invitation denominator; alumni response rates could therefore not be calculated and the evidence was treated as supplementary, retrospective, and self-reported.

Data Analysis

Qualitative analysis combined deductive and inductive coding. In the first cycle, interview transcripts, observation notes, and documents were coded according to inputs, activities, outputs, short-term outcomes, and intended or perceived impacts. In the second cycle, inductive subthemes were developed within and across these categories, including heterogeneous readiness, listening difficulty, feedback, digital-task fit, procedural test familiarity, documentation gaps, and differentiation needs. Analysis followed iterative data condensation, display, comparison, and conclusion verification (Miles et al., 2014; Saldaña, 2021).

Cross-source comparison examined whether evidence converged, complemented, or contradicted other evidence. A claim of program coherence required support from more than one source or a direct match between a documented program element and enacted or reported practice. Negative and cautionary evidence was retained, including missing score records, the inconsistent pretest-to-final pattern, persistent listening problems, and the limitations of alumni self-report. This procedure reduced the risk of constructing a success narrative from positive comments alone.

Quantitative evidence was analyzed descriptively. For each valid pair, change was calculated as Try-Out 3 minus Try-Out 1. Group means were calculated as the sum of valid scores divided by the number of valid cases. The number and proportion of participants with higher, equal, or lower later scores were summarized. Alumni ratings were reported as means only where the institutional dataset provided numerical responses. No inferential test was conducted; therefore, the manuscript does not use “statistically significant” or interpret mean change as a causal effect. Missing cases were excluded only from analyses requiring the unavailable score and were not listwise deleted from qualitative or other descriptive analyses.

Integration was completed through a logic-model joint display that placed qualitative findings, documents, score patterns, and alumni evidence in the same

analytical matrix. Quantitative findings were used in three ways: to corroborate qualitative reports, to add magnitude to implementation patterns, and to identify dissonance. The difference between the Try-Out 1-to-Try-Out 3 pattern and the available pretest-to-final pattern was treated as dissonant evidence requiring explanation rather than as an error to be ignored.

Discussion

The findings indicate that the LPPNTB TOEFL preparation course has a coherent pathway in which institutional resources, trainer preparation, integrated instruction, guided practice, feedback, digital support, and repeated testing plausibly contribute to participant readiness. This supports the logic-model principle that outcomes should be interpreted in relation to the resources and activities that precede them rather than inferred from a final score alone (W. K. Kellogg Foundation, 2004; Booker, 2019). This aligns with Koçer and Topkaya (2023), who linked activities and outcomes through logic-model evaluation. The present study extends that work to TOEFL preparation by integrating interviews, observation, documents, score records, and alumni evidence across multiple sources.

The alignment among the syllabus, trainer practices, classroom participation, feedback, assignments, and mock testing confirms earlier research showing that program quality depends on the interaction of curriculum, assessment, materials, pacing, technology, and stakeholder experience. Bayram and Canaran (2019) identified these dimensions as important to program quality, while Mazidah and Masruroh (2024) highlighted infrastructure, resources, and engagement in TOEFL-course evaluation. Primary et al. (2023) likewise showed the value of stakeholder evidence for identifying implementation problems. The current findings extend these studies by organizing those elements into a sequence of inputs, activities, outputs, and outcomes. Institutional resources matter because trainers convert them into learning opportunities through explanation, structured practice, feedback, and assessment.

The score evidence offers promising but cautious support for participant progress. Among 35 participants with complete Try-Out 1 and Try-Out 3 records, the mean increased from 426.00 to 474.11, and 32 participants obtained higher later scores. However, the available pretest-to-final means decreased slightly from 479.79 to 476.29. This inconsistency prevents a simple claim of sustained improvement and shows that interpretation depends on the selected baseline. Research indicates that test preparation may improve familiarity, strategic control, and test-taking efficiency without demonstrating broad language-proficiency gains (Yu & Green, 2021). Wang and Cheng (2025) similarly found that TOEFL preparation can reshape test takers' understanding of task constructs. Thus, the findings support distinguishing test readiness and familiarity from causal improvement in general English proficiency.

This distinction is also important for washback. The course combined grammar explanation, reading subskills, listening practice, discussion, feedback, and TOEFL-style exercises, suggesting that instruction was not limited to mechanical drilling. Such practices are consistent with positive washback when test demands encourage systematic learning, strategy development, and attention to academic language (Alderson & Wall, 1993; Alderson & Hamp-Lyons, 1996). Feedback may further support learning when it helps learners identify current performance and needed improvement (Black & Wiliam, 2009; Hattie & Timperley, 2007). Nevertheless, repeated TOEFL-item exposure may create narrow washback if short-term answering strategies replace deeper language development. The findings therefore suggest that positive and potentially negative washback can coexist, making language development, strategic awareness, test familiarity, and procedural readiness useful as separate indicators.

A major weakness was the mismatch between heterogeneous learner readiness and relatively uniform pacing and task demands. This supports differentiated-instruction research showing that task complexity, scaffolding, and pacing should respond to learner readiness (Mohamed et al., 2025). Google Classroom expanded access to materials, but access alone did not ensure that tasks were appropriate for learners at different proficiency levels. This agrees with Sailer et al. (2024), whose meta-analytic evidence shows that technology-enhanced learning depends on the quality of the activities it supports rather than technology itself. In a TOEFL context, digital platforms should therefore function as differentiated learning pathways. Diagnostic profiles could guide tiered listening, structure, reading, and vocabulary activities.

Persistent listening difficulty exposed another break in the pathway from instruction to readiness. Learners continued to experience problems involving rapid processing, vocabulary, inference, attention, and test timing. This is consistent with Shamsi and Bozorgian (2024), who found that multimedia input, collaboration, metacognitive instruction, and strategy awareness can support EFL listening comprehension. The comparison suggests that increasing the number of listening items may strengthen familiarity without addressing why comprehension fails. A stronger approach would use diagnostic listening, transcript-supported noticing, prediction and verification, vocabulary preparation, repeated listening for different purposes, and learner reflection. This extends TOEFL evaluation by treating listening difficulty as a mechanism-level problem.

The digital testing component also broadens the concept of TOEFL readiness. Participants needed suitable devices, audio equipment, internet access, procedural knowledge, and familiarity with online testing. These demands show that computer-mediated assessment requires technical as well as linguistic preparedness. However, they also introduce concerns about access, privacy, supervision, and fairness. Isbell et al. (2023) argue that technology-mediated language testing raises justice issues that cannot be treated as purely technical. The present findings therefore extend conventional program evaluation by incorporating technical preparedness and equitable access into the logic model. Pre-test technology checks, alternatives for access problems, clear privacy information, and documented procedures for technical incidents should be considered part of program quality.

The alumni evidence suggested perceived usefulness for scholarships, certification, English development, and career purposes, but these longer-term claims require caution. The records were retrospective and self-reported, and response rates could not be established. Utilization-focused and language-program evaluation emphasize credible evidence for program decisions rather than favorable testimony alone (Norris, 2016; Patton & Campbell-Patton, 2022). Alumni responses should therefore be interpreted as indications of perceived utility, not evidence that the course caused later academic or career outcomes. A longitudinal system recording response rates, verified official scores where consent is provided, certificate use, and subsequent outcomes would strengthen future impact evidence.

The findings confirm previous research that language-program quality depends on curriculum, instruction, assessment, technology, learner characteristics, and institutional conditions, while extending the literature by connecting these elements through a logic-model pathway. The strongest evidence supports internal program coherence and short-term readiness; the weakest concerns causal score improvement and long-term impact. Contradictory score patterns, incomplete records, unavailable test-form equivalence evidence, and the absence of a comparison group make causal attribution inappropriate. Standardized participant identifiers, fixed assessment points, section-level scores, attendance records, diagnostic placement, differentiated digital assignments, intensive

listening support, and systematic alumni tracking would strengthen future evaluation. The study therefore shows that TOEFL preparation should be judged not only by score change, but by whether resources, activities, outputs, outcomes, and intended impacts form a plausible and evidence-supported chain.

CONCLUSION

The study concludes that the LPPNTB TOEFL preparation course is characterized by a generally coherent relationship among institutional support, trainer preparation, integrated instruction in reading, structure, and listening, guided practice, feedback, digital learning support, and repeated try-out activities. Through the logic-model perspective, these inputs and activities were plausibly connected to immediate outputs and short-term outcomes such as test familiarity, strategic awareness, confidence, and procedural readiness. The convergence of syllabus evidence, stakeholder accounts, classroom observation, and assessment records indicates that the intended program was substantially enacted rather than existing only as a formal plan. Nevertheless, program coherence was not uniform across all components. Heterogeneous entry proficiency, relatively undifferentiated pacing and digital assignments, persistent listening difficulties, and incomplete assessment documentation weakened the pathway from common instructional inputs to equitable learner outcomes. These findings indicate that stronger diagnostic placement, differentiated learning pathways, process-oriented listening instruction, and improved documentation are necessary to increase both instructional responsiveness and evaluative credibility.

Evidence of participant progress should be interpreted as promising but non-causal. Among the 35 participants with complete Try-Out 1 and Try-Out 3 records, the mean increased from 426.00 to 474.11, with 32 participants obtaining higher later scores; however, the available pretest-to-final means decreased slightly from 479.79 to 476.29. Accordingly, the study cannot conclude that the course caused stable score improvement or broader English-proficiency gains. Rather, the available evidence supports a narrower conclusion that participants demonstrated encouraging within-course readiness patterns alongside reported gains in familiarity and confidence. Alumni records also suggested perceived usefulness for scholarships, certification, English development, and career purposes, but their retrospective and self-reported nature prevents firm conclusions about long-term impact. Future program cycles should therefore combine diagnostic profiling, differentiated digital tasks, stronger listening support, standardized assessment and attendance records, fair online-testing procedures, and longitudinal alumni tracking. With these improvements, the logic model can function not only as an evaluation framework but also as a practical monitoring system for continuous program improvement.

RECOMMENDATION

LPPNTB should use the revised logic model as an operational monitoring tool rather than as a one-time evaluation product. Each course cycle should begin with a diagnostic profile, connect differentiated activities to identified needs, and maintain a complete implementation and assessment record. A short dashboard can report participation, task completion, section-level performance, listening needs, technical incidents, and follow-up outcomes without converting descriptive trends into causal claims.

Future research should employ prospectively designed mixed-methods evaluation with clearly documented instruments, complete paired records, equivalent or linked test forms, delayed follow-up, and, where feasible, a comparison group or interrupted time-series structure. Multi-site studies could examine which program mechanisms are transferable and which depend on local institutional conditions. Comparative use of

logic-model, CIPP, CIPPO, or Kirkpatrick frameworks may also clarify which approach produces the most actionable evidence for language-program improvement.

ACKNOWLEDGMENT

The authors gratefully acknowledge the TOEFL trainer, program coordinator, course participants, alumni respondents, and administrators of Lembaga Pengembangan Pendidikan Nusa Tenggara Barat (LPPNTB) for their valuable cooperation throughout this study. Appreciation is also extended to the personnel who facilitated classroom observation, access to institutional documents, score records, alumni datasets, and other supporting information required for the evaluation. Their assistance made it possible to examine the program from multiple perspectives and to develop a comprehensive logic-model interpretation. All contributions are acknowledged in a non-identifying manner to preserve participant confidentiality and to respect fully the ethical principles applied throughout the research process.

DECLARATION OF USING AI TOOLS

OpenAI ChatGPT was used as an AI-assisted writing tool during manuscript development. Its use was limited to supporting structural revision, language refinement, synthesis of the discussion, drafting assistance, and checking the consistency of citations and academic expression. The tool was not used to generate, fabricate, modify, or replace primary data; conduct interviews or observations; determine participant scores; or independently make final research decisions. All AI-assisted text was critically reviewed, edited, and verified by the authors against the original data and cited literature. The authors accept full responsibility for the accuracy, interpretation, originality, ethical compliance, and integrity of the final manuscript.

REFERENCES

- Abulhul, Z. (2023). A logic model conceptualizes the initiative of applying for an inclusive education program. *Journal of Practical Studies in Education*, 4(6), 15-19. <https://doi.org/10.46809/jpse.v4i6.74>
- Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, 13(3), 280-297. <https://doi.org/10.1177/026553229601300304>
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115-129. <https://doi.org/10.1093/applin/14.2.115>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Bayram, I., & Canaran, O. (2019). Evaluation of an English preparatory program at a Turkish foundation university. *Journal of Language and Linguistic Studies*, 15(1), 48-69.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5-31. <https://doi.org/10.1007/s11092-008-9068-5>
- Booker, Z.-Z. (2019). A revised logic model for educational program evaluation. *Journal of Modern Applied Statistical Methods*, 18(2), Article eP3047. <https://doi.org/10.22237/jmasm/1604190300>

- Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27-40. <https://doi.org/10.3316/QRJ0902027>
- Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson.
- Clapham, K., Manning, C., Williams, K., O'Brien, G., & Sutherland, M. (2017). Using a logic model to evaluate the Kids Together early education inclusion program for children with disabilities and additional needs. *Evaluation and Program Planning*, 61, 96-105. <https://doi.org/10.1016/j.evalprogplan.2016.12.004>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
- Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE.
- Czosnek, L., Zopf, E. M., Cormie, P., Rosenbaum, S., Richards, J., & Rankin, N. M. (2022). Developing an implementation research logic model: Using a multiple case study design to establish a worked exemplar. *Implementation Science Communications*, 3, Article 90. <https://doi.org/10.1186/s43058-022-00337-8>
- Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods* (2nd ed.). McGraw-Hill.
- Educational Testing Service. (n.d.). The TOEFL ITP Assessment Series. <https://www.ets.org/toefl/itp.html>
- Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs: Principles and practices. *Health Services Research*, 48(6 Pt 2), 2134-2156. <https://doi.org/10.1111/1475-6773.12117>
- Guetterman, T. C., & Fetters, M. D. (2018). Two methodological approaches to the integration of mixed methods and case study designs: A systematic review. *American Behavioral Scientist*, 62(7), 900-918. <https://doi.org/10.1177/0002764218772641>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>
- Isbell, D. R., Kremmel, B., & Kim, J. (2023). Remote proctoring in language testing: Implications for fairness and justice. *Language Assessment Quarterly*, 20(4-5), 469-487. <https://doi.org/10.1080/15434303.2023.2288251>
- Kiely, R., & Rea-Dickins, P. (2005). *Program evaluation in language education*. Palgrave Macmillan.
- Koçer, P., & Topkaya, E. Z. (2023). Evaluating an English language writing course through logic model: Links between outcomes and activities. *Innovational Research in ELT*, 4(1), 14-31. <https://doi.org/10.29329/irelt.2023.558.2>
- Lembaga Pengembangan Pendidikan Nusa Tenggara Barat. (2024). Data alumni peserta Rumah Bahasa LPPNTB tahun 2021, 2022, dan 2024 [Unpublished institutional dataset].
- Lembaga Pengembangan Pendidikan Nusa Tenggara Barat. (n.d.-a). Silabus TOEFL [Unpublished institutional document].
- Lembaga Pengembangan Pendidikan Nusa Tenggara Barat. (n.d.-b). Tutorial mengakses dan mengerjakan soal TOEFL di Web Rumah Bahasa [Unpublished institutional document].
- Mayer, R. E. (2020). *Multimedia learning* (3rd ed.). Cambridge University Press.
- Mazidah, E. N., & Masruroh, L. (2024). CIPP model evaluation on TOEFL course at higher education. *Edulitics Journal*, 9(2), 112-128.

- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE.
- Mohamed, F. T., Olamo, T. G., & Yemiru, M. A. (2025). Effect of differentiated instruction on primary EFL students' writing performance and perception: The case of Grade 7 students in Hawassa City, Ethiopia. *Social Sciences & Humanities Open*, 11, Article 101230. <https://doi.org/10.1016/j.ssaho.2024.101230>
- Norris, J. M. (2009). Understanding and improving language education through program evaluation: Introduction to the special issue. *Language Teaching Research*, 13(1), 7-13. <https://doi.org/10.1177/1362168808095520>
- Norris, J. M. (2016). Language program evaluation. *The Modern Language Journal*, 100(S1), 169-189. <https://doi.org/10.1111/modl.12307>
- Ormsby, S. M., & Morrow, J. A. (2019). Development of a logic model for use in evaluation of learning support programs. *Journal of Student Success and Retention*, 6(1), 1-26.
- Patton, M. Q., & Campbell-Patton, C. E. (2022). *Utilization-focused evaluation* (5th ed.). SAGE.
- Primary, M. A., Emilia, E., & Sukyadi, D. (2023). Program evaluation of English language learning for EYL curriculum development in Indonesia: Teachers' perception, challenges, and expectation. *Journal on English as a Foreign Language*, 13(1), 314-338. <https://doi.org/10.23971/jeft.v13i1.6068>
- Sailer, M., Maier, R., Berger, S., Kastorff, T., & Stegmann, K. (2024). Learning activities in technology-enhanced learning: A systematic review of meta-analyses and second-order meta-analysis in higher education. *Learning and Individual Differences*, 112, Article 102446. <https://doi.org/10.1016/j.lindif.2024.102446>
- Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). SAGE.
- Shamsi, E., & Bozorgian, H. (2024). Collaborative listening using multimedia through metacognitive instruction: A case study with less-skilled and more-skilled EFL learners. *Asian-Pacific Journal of Second and Foreign Language Education*, 9, Article 25. <https://doi.org/10.1186/s40862-023-00248-8>
- Smith, J. D., Li, D. H., & Rafferty, M. R. (2020). The Implementation Research Logic Model: A method for planning, executing, reporting, and synthesizing implementation projects. *Implementation Science*, 15, Article 84. <https://doi.org/10.1186/s13012-020-01041-8>
- W. K. Kellogg Foundation. (2004). *Logic model development guide*. W. K. Kellogg Foundation.
- Wang, P., & Cheng, L. (2025). The impact of TOEFL iBT preparation on Chinese test-takers' perceptions of integrated speaking and writing design. *Critical Inquiry in Language Studies*, 22(1), 64-87. <https://doi.org/10.1080/15427587.2024.2448815>
- Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE.
- Yu, G., & Green, A. (2021). Preparing for admissions tests in English. *Assessment in Education: Principles, Policy & Practice*, 28(1), 1-12. <https://doi.org/10.1080/0969594X.2021.1880120>