February 2026

IELTS Research Reports

Comparing New TOEFL 2026 with Former TOEFL 2023 and IELTS

Authors: Dr. Haoshan (Sally) Ren and Dr. Andy Blackhurst
Reviewers: Dr. Leyla Karatay, Dr. Tony Clark and Dr. Jing Xu

IELTS Research

BRITISH COUNCIL idp CAMBRIDGE English

IELTS is jointly owned by the British Council; IDP IELTS; and Cambridge University Press & Assessment

Comparing New TOEFL 2026 with Former TOEFL 2023 and IELTS

Contents

List of Tables

List of Figures

Executive summary

In January 2026 ETS launched a new version of its TOEFL iBT test. Changes to the test include shorter test length, revised score reporting methods, adaptive test design, and adjustments to the numbers of questions across the four skill sections. ETS has stated that these updates are designed to make the test more efficient, accessible, and reflective of real-life communication skills. However, these changes warrant caution regarding the accurate measurement of academic language skills and existing test comparability investigations and the interpretability of scores. This report is intended to provide an in-depth analysis of the new test through a validity lens and an informed view of the new TOEFL iBT test taker's readiness for academic study in English-speaking environments.

The report focuses on construct comparison between the new TOEFL iBT, its predecessor and IELTS Academic, and on how that comparability impacts score interpretation. The authors examine test elements against a broad range of facets related to validity, such as cognitive processing, task design, scoring, alignment with CEFR, and washback to determine how the new version of TOEFL iBT compares to its predecessor. Key IELTS Academic information is included to provide a reference point for comparison.

Given the new TOEFL iBT's substantially reduced text length and complexity, its narrowing of task types, the potentially increased susceptibility to "coaching" and its new scoring system, among other considerations, this report concludes that the revised TOEFL iBT represents a substantial construct shift, undermining assumptions of score equivalence with earlier TOEFL versions. The report warns about the risk for score misinterpretation that can arise from using legacy concordance tables.

1. Introduction

ETS announced the enhanced TOEFL iBT® test, which is intended to replace the current TOEFL iBT, at the beginning of 2026. This round of change is one of the more evolutionary in its scale. The 2026 changes to the test include greater diversification of task types, shorter test length, revised score reporting methods, adaptive test design, and adjustments to the numbers of questions across the four skill sections. ETS states that these updates are designed to make the test more efficient, accessible, and reflective of real-life communication skills. This is intended to provide a more accurate measure of a test taker's readiness for academic study in English-speaking environments. The newest update reflects a convergence of recent directions seen in Pearson and the Duolingo English Test where the tasks have become shorter, more interactive, and on topics that are no longer exclusively academic. However, these changes warrant careful caution with respect to test comparability and the interpretability of scores, which raise important questions about the continued suitability of the test for the academic domain, a high-stakes context.

This report focuses on *construct* comparison and how their comparability impacts score interpretation and test equating. The concept of "test construct" has evolved through decades of inquiry and debate, reflecting shifts in how language ability has been defined and measured within the field of language assessment. A pivotal development in the broader measurement literature is Messick's (1989) unified view of construct validity, which foregrounds the interpretative basis of score meaning and the evidential rationale required to support intended uses. In language assessment, this validity tradition has been taken up and elaborated with a more explicit focus on the domain- and task-contingent nature of language use, including the view that language performance emerges from an ongoing dialectic between an individual's language ability and the contexts in which language is used (Bachman, 2007; Chapelle, 1998), with language functioning as "both the object and the instrument of our measurement" (Bachman, 1990, p. 287). These perspectives have, in turn, informed contemporary validation approaches in language testing (e.g., Bachman and Palmer, 1996; Weir, 2005; Chapelle, 2021), which interrogates how test tasks represent the construct and support meaningful score interpretations.

Guided by these theoretical approaches, this report is organized around the following core elements that collectively operationalize the construct and anchor validity arguments:

- **Content validity**: the extent to which the linguistic material, skills, and knowledge sampled in the test represent the target domain of language use. This concerns whether the tasks include appropriate and sufficiently broad content to reflect the construct being measured.
- **Context validity**: the degree to which the conditions and situational features of the test tasks reflect authentic or target-language-use contexts. This includes factors such as task purpose, input characteristics, interactional demands, and response formats. Context validity ensures that test performance meaningfully relates to real-world or domain-specific language use.

- **Cognitive processing**: the mental operations and strategic demands elicited by test tasks. This addresses whether the cognitive activities required of test takers (e.g., comprehension, inference, planning, synthesis) align with those involved in genuine language use and with the theoretical construct underlying the test. In this sense, cognitive processing evidence contributes to evaluating the degree of task authenticity, which is the extent to which test tasks approximate real-world language use conditions in terms of the cognitive demands they elicit.
- **Scoring**: the procedures, criteria, and mechanisms through which test performance is evaluated. This includes the reliability, consistency, fairness, and transparency of scoring methods, as well as the extent to which scoring practices support valid interpretations of test results.
- **CEFR coverage (B2)**: the alignment of the test's proficiency claims with the Common European Framework of Reference for Languages (CEFR) levels, including how comprehensively the test samples the linguistic competences and performance descriptors associated with each band. In this report, CEFR coverage is examined specifically at the B2 level, as this level is most commonly used as a threshold for academic admissions.
- **Washback effect:** the influence of the test on teaching and learning – what teachers teach and what learners practise. Washback reflects the test's task demands and construct emphasis, and is considered in validity work as a consequential aspect of test use (i.e., whether it promotes appropriate learning or encourages narrowing/coaching).

These domains reflect the essential mechanisms through which language ability is elicited, evaluated, and interpreted, and they provide a principled structure for comparing the tests under review.

This report examines how the revised TOEFL 2026 (hereafter New TOEFL 2026) compares with the previous version of TOEFL iBT (after the 2023 revision, hereafter, Former TOEFL 2023). Key IELTS Academic information is also included to provide a reference point for comparison. In this report, analyses of context, content, and cognitive processing for IELTS and Former TOEFL 2023 are based on the recent work of Cushing (in press), which examined the comparability of the IELTS and TOEFL tests to support the most recent concordance study (Ikeda et al., 2025). All other analyses rely on publicly available information from online sources, technical manuals, research reports published by ETS and IELTS, as well as CEFR descriptor documents published by the Council of Europe (https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions).

2. Construct Comparison by Skills

This section focuses primarily on the comparison between the two versions of TOEFL, before and after the 2026 changes. New TOEFL 2026 claims to measure “language skills and communication abilities needed in academic and daily life settings” (Manna, Li, Papageorgiou, & Gu, 2025), which correspond to the educational and public domains in CEFR terms (Manna et al., 2025, p. 2). The test is designed to cover the full proficiency range from A1 to C2 (Manna et al., 2025, p. 2). In contrast, the pre-2026 TOEFL iBT was framed more narrowly around academic English use, with ETS describing it as developed in response to institutional requests for a test measuring non-native speakers’ ability to communicate in English in an academic setting, with section definitions explicitly tied to understanding university-level academic texts and lectures, and producing spoken and written responses appropriate for college and university coursework (Chapelle, Enright, & Jamieson, 2008; Educational Testing Service, 2024a, 2024b).

Although the TOEFL Technical Manual (Manna et al., 2025, p. 1) emphasizes the use of scores for international students entering higher education, it does not specify additional uses, noting more broadly that the scores provide “an overall indication of English language proficiency” (Manna et al., 2025, p. 3). These claims, together with ETS’s newly published score-mapping information aligned to CEFR levels, also warrant a re-examination of how New TOEFL 2026 compares to IELTS Academic. For this reason, key information about IELTS Academic is also included in the accompanying tables as a reference point.

The following sections are organized by the four language skills: reading, listening, writing, and speaking. It is worth noting that, although the format of these tests follows this conventional four-skill structure, many task types within each section engage more than one skill simultaneously. Therefore, the following discussion, organized by skills, is not strictly confined to each individual skill section but also touches upon integrated task types, where multiple skills are combined to demonstrate the ability to read, listen, speak, and write.

2.1 Reading

2.1.1 Content and Context Features for Reading

Content Comparison

The Former TOEFL 2023 Reading section requires test takers to read two long academic passages and answer comprehension questions within 36 minutes. Each passage is approximately 700 words long, followed by 10 multiple-choice questions and one selected-response summary question. The New TOEFL 2026 Reading section is modified to cover non-academic reading texts as well, while adding Complete the Words (C-test) as a new question type in addition to multiple-choice items.

The three new tasks are:

- **Complete the Words (C-test)**: filling in missing letters within short passages (10 words total)
- **Read in Daily Life (RDL)**: reading everyday materials such as emails or announcements
- **Read an Academic Passage (AP)**: engaging with university-level texts

The new tasks are significantly shorter, ranging from as few as 15 words (in RDL) to around 250 words (in AP). Table 1 summarizes the content of the reading task in the three tests. For New TOEFL 2026, the C-test is listed separately.

Table 1. Content Comparison of Reading Tasks

	IELTS	**Former TOEFL 2023**	**New TOEFL 2026**	
			C-test	**RDL & AP**
Timing	60 minutes	35 minutes	27 minutes	
Reading passages	Three reading passages (700–900 words each)	Two reading passages (approximately 700 words each)	Fill in missing letters of words in a paragraph (*about 75 words)	One reading passage of about 15–150 words for RDL, and about 200 for AP, followed by comprehension questions
Number of items	Around 13 items per passage (40 total)	10 items per text (20 total)	10 truncated words in a C-test	Two–three questions per passage for RDL, five questions for AP
Test items	11 item types: see the skill measured section below	6 item types, all selected response or drag-and-drop	Single-item type	Multiple-choice questions

A notable difference between the three tests (especially for New TOEFL 2026) is the task length. This point will be elaborated in the discussion of CEFR level comparison and cognitive processing.

Skills measured:

Each test provider lists the skills their reading task intends to measure on their public-facing documents. For IELTS, skills are listed under each of their 11 item types. TOEFL lists the target skills for each task. Below is a list of reading skills that the reading tasks are intended to measure as published by each test. (See Appendix A for an itemized textual summary supporting Table 2 and Table 3 below.)

Table 2. Skills Measured by IELTS Reading Tasks

Skills measured	MC	II	WV	MI	MH	MF	ME	SC	CC	DL	SA
Understanding main ideas and overall meaning	✓						✓		✓		
Identifying specific information and detailed understanding	✓	✓						✓			✓
Recognizing opinions, views, and theories			✓			✓					
Scanning for information and relationships				✓		✓					
Identifying general topics, main ideas, and supporting ideas in paragraphs					✓						
Relating detailed descriptions to visual information										✓	

MC = Multiple choice; II = Identifying information; MI = Matching information;
MH = Matching headings; MF = Matching features; SC = Sentence completion;
ME = Matching sentence endings; DL = Diagram label completion;
WV = Identifying writer's views (Yes/No/Not Given); SA = Short-answer questions;
CC = Summary/Note/Table/Flow-chart completion

Table 3. Skills Measured by TOEFL Reading Tasks

Skills measured	Former TOEFL 2023	New TOEFL 2026	
		C-test	RDL & AP
Processing written texts for meaning and form	✓	✓	✓
Understanding factual information (linear + nonlinear texts)	✓		✓
Identifying main purpose / main idea	✓		✓
Understanding vocabulary, idioms, figurative expressions, informal language	✓	✓	✓
Making inferences and interpreting implied meaning	✓		✓
Understanding grammatical structure and complexity	✓		✓
Skimming, scanning, and locating information	✓		✓
Understanding text organization and rhetorical structure	✓		

The tables suggest that, at a broad level, the three reading components are associated with largely similar skill sets as defined in their respective specifications and mapping claims. However, this apparent comparability becomes less secure when attention shifts from overall skill coverage to the demands of individual task types. In New TOEFL 2026, for example, the C-test task contributes primarily to the assessment of fundamental lexical and syntactic processing rather than higher-level text comprehension. While this format may serve practical functions, such as screening for ability levels within an adaptive design, great caution is warranted in interpreting scores from when incorporating performance on this task:

- The C-test focuses mainly on lexical and syntactic processing, representing a relatively narrow construct that does not fully align with broader communicative testing principles. The task format does not resemble typical real-world reading, and successful performance can be achieved without establishing full passage-level meaning (Winke, Yan, & Lee, 2024).
- Earlier research highlighted the importance of pre-calibration for C-tests (Klein-Braley & Raatz, 1984), while more recent psychometric studies suggest that local item dependence can complicate stable and reliable calibration (Eckes, & Grotjahn, 2006; Baghaei, 2010).
- Some studies have reported limited effectiveness of C-tests for distinguishing proficiency levels (e.g., Roohani, 2007).
- Cognitive validation evidence also raises fairness concerns for novice and intermediate learners: eye-tracking and interview data suggest that lower-proficiency test takers may be unable to read the passage adequately, may rely on guessing strategies, and may be constrained in demonstrating reading ability (with some reports of negative affect), leading some researchers to question the ethicality of using C-tests with low-proficiency learners (Winke et al., 2024, p. 33).

Context Features

Table 4 below compares the relevant context factors that contribute to reading text processing. IELTS and Former TOEFL 2023 are broadly aligned in targeting an academic university-level domain, using longer continuous passages (e.g., articles/textbook-style prose). These longer passages predominantly instantiate expository and argumentative discourse and draw on general academic content with minimal specialist knowledge, covering both concrete and abstract information. In contrast, New TOEFL 2026 includes a mixed domain with two main text types: functional daily-life materials (e.g., notices, emails) and academic passages. This is accompanied by a shift in genre and discourse characteristics, and a greater emphasis on concrete information in the daily-life texts, even though the academic passages remain largely expository/argumentative. Overall, these differences indicate a change in the contextual conditions under which reading ability is elicited, with implications for construct coverage and the comparability of score interpretations across test versions.

Table 4. Context Comparison of Reading Tasks

Context validity factors	**IELTS (Academic Reading)**	**Former TOEFL 2023**	**New TOEFL 2026**
Domain	Social and academic	Primarily academic university-level textbook style	Social and academic
Genre	Longer continuous academic passages (essays, articles) with occasional diagrams	Academic articles, textbook style, continuous prose	Two main genres: daily-life texts (notices, emails) + academic passages.
Discourse mode	Mostly expository, argumentative and descriptive academic prose	Expository, descriptive, academic argumentation	Daily-life texts often functional / explanatory; academic passages remain expository / argumentative
Content knowledge	General academic topics, minimal specialist prior knowledge required	Academic topics common in university reading, minimal prior subject-specific knowledge required	Daily-life texts largely real-world, non-specialist; academic passages similar to current TOEFL
Nature of information	Concrete and abstract	Concrete and abstract	Mostly concrete

Lexical Complexity

To describe the linguistic complexity across reading passages of the three tests, several quantitative indices were examined, following Cushing (in press). *Words per sentence* provides an estimate of syntactic complexity, with higher values generally indicating more embedded clauses and greater processing loads for readers. The *Flesch–Kincaid Reading Grade* (Flesch, 1948; Kincaid, Fishburne, Rogers, & Chissom, 1975) estimates the U.S. school grade level required for comprehension, based on sentence length and syllable density. The *Academic Word List (AWL) percentage* (Coxhead, 2000) reflects the proportion of mid-frequency academic vocabulary in a text, which is often associated with advanced reading proficiency. The *B2 and C1 lexical type percentages* show the proportion of vocabulary items belonging to higher CEFR lexical bands, offering an additional indicator of lexical sophistication.

Table 5. Comparison of IELTS Academic and Current TOEFL iBT Reading Passages Metrics (from Cushing, 2025)

Metric	**IELTS 1 (Rockets)**	**IELTS 2 (Traffic)**	**IELTS 3 (Dung beetles)**	**TOEFL 1 (By-catch)**	**TOEFL 2 (Rome)**
CEFR level	C2	C2+	C1+	D1**	C1+
Total words	550*	842	332*	714	688
Words per sentence	23.08	30.39	19.94	26.52	18.73
Flesch–Kincaid Reading Grade	12.45	16.15	11.19	15.53	11.98
Academic Word List %	12.45	16.15	11.19	15.53	11.98
B2 type %	11.07	14.83	10.38	14.79	15.32
C1 type %	2.77	4.44	2.19	2.67	4.94

* The complete text is not publicly available; only an extract was included in the sample materials

** D1 level in Text inspector is a proprietary grade, not part of the original CEFR, used to describe a very high level of academic language proficiency that goes beyond the standard C2 level

Table 6. Metrics for Texts in New TOEFL 2026

Metric	Read in Daily Life (RDL)			Academic Passage (AP)_	
	Farmer's market	Email invitation	Inventory audit	Paradox of choice	Mirror test
CEFR level	C2+	C2	C2	D1	C2
Total words	142	129	131	207	196
Words per sentence	12.91	16.12	18.71	18.82	19.30
Flesch–Kincaid Reading Grade	8.31	11.37	11.44	15.06	10.54
Academic Word List %	6.42	12.87	13.98	23.19	7.27
B2 type %	10.28	6.93	12.09	18.12	4.50
C1 type %	1.87	1.98	3.30	5.80	4.50

As shown in the table above, the IELTS Academic passages show consistently C1 to C2 profiles, with relatively long sentences and high proportions of academic vocabulary (AWL roughly 11–16%). This tendency is closely mirrored in the former TOEFL (2025) texts, which display similar syntactic length and lexical density, and both tests target comparable reading demands.

The New TOEFL 2026 texts, by contrast, span a wider range of genres and show greater variability in the lexical complexity measures. Specifically, texts in RDL have noticeably shorter sentences and lower readability grades, reflecting lighter processing demands despite differences in their CEFR classifications. The APs align more closely with the IELTS and current TOEFL, showing higher readability grades and longer sentences, but variability exists in difficulty levels between the two passages in the sample test. Taken together, New TOEFL 2026 reading scores may reflect a somewhat different configuration of reading demands, combining academic reading with brief, functional real-world texts. This configuration may shift construct representation relative to earlier versions and therefore warrants caution when interpreting New TOEFL 2026 reading scores as directly comparable to IELTS or Former TOEFL 2023.

2.1.2 Cognitive Processes for Reading

Two frameworks are used for this comparison of the cognitive processes engaged by the three tests for reading. Khalifa and Weir (2009) illustrates the reading types, mental processes, and knowledge sources engaged for authentic reading activities (for a detailed recap see Cushing and Ren, 2023). Liu and Read (2023) developed their framework based specifically on reading skills required for university study. The two tables below show how the three tests compare within both frameworks.

Table 7. Comparison of Cognitive Processes for Reading (Khalifa & Weir, 2009)

Cognitive processes **Khalifa and Weir (2009)**	**IELTS**	**Current TOEFL**	**New TOEFL 2026**		
			C-test	RDL	AP
Word recognition	x	X	x	x	x
Lexical access	x	X	x	x	x
Syntactic parsing	x	X	x	x	x
Establishing propositional meaning at clause and sentence levels	x	X		x	x
Inferencing	x	x		x	x
Building a mental model	x	x			x
Creating a text-level representation	x	x			(x)
Creating an intertextual representation		(x)			

Processes that appear to be only partially represented—or comparatively less emphasized—are indicated in parentheses.

Table 8. Comparison of Cognitive Processes for Reading (Liu & Read, 2023)

Liu and Read (2023) for Academic Reading	**IELTS**	**Former TOEFL 2023**	**New TOEFL 2026**		
			C-test	RDL	AP
Core academic language knowledge					
Understand general academic vocabulary	X	x	x	x	x
Careful reading for intra-textual model building					
Integrating textual information across sentences	X	x	x	x	x
Inferring the situation implied in a text	X	x		x	x
Understanding author's point of view	X	x		x	x
Inferring the contextual meaning of figurative language	X	x		x	x
Careful reading for intertextual model building					
Understanding the relationships between multiple texts		(x)			
Drawing implications/conclusions based on multiple texts		(x)			
Expeditious reading					
Searching for specific meaning	X	x		x	x
Skimming for general idea	X	x		x	x

Processes that appear to be only partially represented—or comparatively less emphasized—are indicated in parentheses.

Results from the comparisons in Table 8 indicate that the overall coverage of cognitive processing is broadly comparable across the three tests. However, it is important to note that text length is not explicitly included as a key parameter in these frameworks. The recent shift toward shorter reading tasks across several test providers warrants renewed discussion about the impact of text length on construct coverage, particularly in relation to the reading skills required for postgraduate study or migration contexts. In most university settings, there is currently little evidence to support that the literacy demands placed on students have materially shifted toward consistently shorter or simpler texts; postgraduate study in particular continues to require sustained engagement with extended academic readings (e.g., journal articles, book chapters, research reports) and the associated endurance- and integration-related processes. If test content is increasingly shaped around brief passages, this may be less an indication of a genuine domain shift in academic literacy requirements than a design choice driven by testing constraints, with potential consequences for the representativeness of the construct sampled and the defensibility of academic-readiness interpretations.

2.1.3 Text Length and CEFR Coverage

Concerns about text length relate to both practicality and authenticity. From a practical perspective, it is difficult for item writers to generate a full range of question types that tap different cognitive processes when working with very short passages. From an authenticity standpoint, it is uncommon for real-world readers to engage in all levels of processing (e.g., identifying main ideas, locating details, making inferences, and analysing arguments) within a single brief text. Expecting such comprehensive processing from minimal input likely reduces the authenticity of the reading demands and risks narrowing the construct being assessed.

Lastly, the length of texts used in reading tasks reflects how closely a test can elicit the targeted reading skills described in the CEFR descriptors, which in turn supports the validity of standard setting results where test scores are mapped onto the CEFR levels. Several descriptors at B2 and C1 explicitly reference text length as part of reading ability. For example:

- Can scan quickly through long and complex texts, locating relevant details (B2).
- Can understand in detail lengthy, complex texts, whether or not these relate to their own area of speciality, provided they can reread difficult sections (C1).
- Can understand in detail a wide range of lengthy, complex texts likely to be encountered in social, professional or academic life, identifying finer points of detail including attitudes and implied as well as stated opinions (C1).

In this context, shorter texts, even when they contain advanced vocabulary or complex syntax, tend to engage a different set of reading processes than longer passages. Brief texts generally require less integration of information across multiple sections, involve fewer shifts in argument or perspective, and offer fewer opportunities to track cohesion or follow how ideas develop over time. Shorter texts also place lower demands on sustained attention and on maintaining coherence across larger stretches of discourse.

Longer texts naturally prompt readers to manage these additional layers of processing, such as resolving references across paragraphs, synthesizing information as it accumulates, and interpreting how earlier sections shape meaning later on. These discourse-level processes are central to the CEFR descriptors that highlight the ability to handle “long,” “lengthy,” or “complex” texts. Thus, CEFR mappings call for cautious interpretation in contexts where variations in skill coverage may shape the kinds of evidence that test performance can provide.

2.2 Listening

2.2.1 Content and Context Features for Listening Tasks

Content Comparison

In IELTS, the Listening test is identical for both Academic and General Training formats, containing items geared towards both domains. It consists of four recordings: two set in everyday social contexts, and two related to education or training, each including one monologue and one conversation. Test takers answer 40 questions using formats such as multiple choice, matching, labelling tasks, and various completion or short-answer items. These tasks are designed to assess a wide range of listening abilities, from grasping main ideas and key details to following directions, interpreting descriptions, and extracting specific factual information.

The Former TOEFL 2023 Listening section comprises three lectures and two conversations, each followed by a set of comprehension questions. The former TOEFL iBT Listening section focused primarily on academic input: test takers listened to extended university-style lectures as well as campus-based conversations between students and university staff. Questions assessed skills such as understanding main ideas and details, pragmatic meaning, speaker purpose, and the organization of information.

The New TOEFL 2026 Listening section introduces several major changes. First, it adopts an adaptive design: test takers will complete two modules, and the difficulty level of the second module will depend on their performance in the first. Second, the range of input types has been broadened beyond traditional academic lectures and campus conversations. The new format incorporates four task types designed to reflect a wider variety of real-world listening situations:

- **Listen and Choose a Response (LCR)**: selecting the most appropriate reply to short utterances.
- **Listen to a Conversation (CV)**: interpreting idiomatic expressions, tone, and speaker intentions.
- **Listen to an Announcement (AN)**: inferring meaning from short campus-style messages.

- **Listen to an Academic Talk (AC)**: analysing structure, main points, and key details in short content-focused talks.

At the level of task format, this set-up is reminiscent of TOEFL Essentials Listening tasks and has some surface similarities to the Duolingo English Test's approach to listening, which includes short, interaction-oriented response selection within its Interactive Listening component (alongside dictation-style listening tasks). It is worth noting that for New TOEFL 2026, Listening is also tested in the "Listen and Repeat" task in the Speaking section. In the listen and repeat task, test takers listen to a series of sentences within a visual scenario to repeat each sentence. Listening is required to process the sentence meaning and form, track increasing complexity, and follow the visual progression of the scenario.

The table below compares the test content of the listening sections between the three tests.

Table 9. Content Comparison of Listening Tasks

	IELTS	**Former TOEFL 2023**	**New TOEFL 2026**
Task description	Listen to a text and answer written comprehension questions	Listen to a text and answer written comprehension questions	Selecting short replies to short utterances. Listen to a text and answer written comprehension questions
Number of tasks	Four (two dialogues, two monologues)	Five (two conversations, three lectures)	Four short tasks *Two modules (where the second four-task module is adaptive)
Number of items per task	10	Five to six	Unspecified
Item type	Selected and constructed response	Selected-response comprehension questions (multiple choice)	Selected response
Length of listening passage	Listening passages are approximately 270 words in length on average. Passages are around two minutes each	Conversations: 500–600 words; around three minutes Lectures:700–800 words; around five minutes	LCR: 10 words CV: 100 words 50 words AN & AC (175–250 words)
Total time for section	30 minutes	36 minutes	27 minutes
Total listening time	12 minutes	16 minutes	Unknown

Across the three tests, the listening content differs notably in task design, passage length, and amount of auditory input. IELTS and Former TOEFL 2023 both use extended spoken texts followed by written comprehension questions. New TOEFL 2026 combines very short utterances requiring brief responses and shorter comprehension tasks, resulting in substantially less sustained listening. This redesign narrows the amount of linguistic input test takers engage with and places greater emphasis on brief, targeted comprehension rather than the integration of information across longer stretches of speech.

Context Features

To provide a fuller picture of the constructs targeted by IELTS, the current TOEFL, and New TOEFL 2026, it is important to consider the contextual features of their listening tasks. These design features shape the kinds of comprehension processes that test takers are expected to engage in, and therefore contribute directly to how each test operationalizes its listening construct. Table 10 summarizes several of these contextual characteristics across the three tests, offering a basis for understanding how differences in task design may influence what aspects of listening proficiency are elicited and evaluated. A key difference, as shown in the table, is that the task length and reading input is much shorter for New TOEFL 2026 test.

Table 10. Context Comparison of Listening Tasks

Context factor	IELTS	Former TOEFL 2023	New TOEFL 2026
Domain	Academic + everyday	Mainly academic	Academic + everyday
Genre	Dialogues, monologues, announcements	Conversations, lectures	Short replies, dialogues, announcements, academic talks
Discourse mode	Dialogue + lecture	Dialogue + lecture	Dialogue, short informational audio, academic talk (short lecture)
Content knowledge	General + academic topics; low prior knowledge	Academic topics; low prior knowledge	Everyday + academic; low prior knowledge
Transmission	Audio. Context and topic provided in instructions; some items have diagrams or charts to complete	Audio. Some content and context visuals	Audio
Times heard	Once	Once	Once (expected)
Length	Longer passages	Conversations: 500–600 words; Lectures: 700–800 words	Very short (10–250 words)
Accent	Mixed accents	Mainly North American	Likely mixed
Gender	Mixed	Mixed	Mixed

Lexical Complexity

As with reading texts, Cushing (in press) compared the lexical complexity of IELTS Listening and the listening task in the former TOEFL, and the same procedure was applied here to examine the lexical profiles of the listening input for New TOEFL 2026. Each task was processed using TextInspector to generate its lexical profile. Because individual LCR and CV items were too short to yield stable results, they were analyzed in aggregated form: for LCR items, all eight sentences were combined together with the response options presented to test takers, and for CV items, the two sample conversations were merged. For all remaining tasks, the longest available sample texts were selected. The resulting lexical

profiles are summarized in the table below, alongside Cushing's (in press) results for IELTS Listening texts and the current TOEFL Listening texts.

Table 11. Comparison of IELTS Academic and TOEFL iBT Listening Metrics

	IELTS		Former TOEFL 2023		New TOEFL 2026			
Metric	**Library Map**	**Arts Center**	**1**	**2**	**LCR**	**CV**	**AN**	**AC**
CEFR level	B2+	C1+	C2	C1+	B1+	B2	C1+	C2
Words per sentence	25.4	22.67	27.67	18.95	5.09	6.75	13.67	17.90
Flesch–Kincaid Reading Grade	10.64	9.82	11.95	10.06	2.64	2.99	9.5	15.45
Academic Word List %	2.42	5.12	6.67	8.77	3.7	1.83	9.89	9.09
B2 type %	3.73	5.09	9.66	10.05	3.15	NA	7.44	9.22
C1 type %	0.62	0.54	3.86	3.31	NA	NA	NA	4.35

LCR = Listen and Choose a Response CV = Listening to a Conversation
AN = Listen to an Announcement AC = Listen to an Academic Talk

Across metrics, the New TOEFL 2026 Listening texts are more varied than those in the former TOEFL and IELTS, with some tasks showing higher lexical sophistication and others being considerably simpler due to their short length or item-based format (e.g., LCR, CV). The IELTS and current TOEFL passages generally appear more consistently aligned with upper-B2 to C1 profiles, whereas New TOEFL 2026 spans a slightly wider range from B1+ to C2.

The IELTS Listening passages show moderately long sentence structures (22–25 words per sentence) but relatively low AWL percentages (2.4–5.1%). The former TOEFL samples exhibit even longer syntactic units in one passage (27.67 words per sentence) and higher AWL densities (up to 8.77%). This suggests a somewhat more academic lexical profile. In contrast, the proposed New TOEFL 2026 tasks vary by task type. Short response–based items (LCR and CV) show very short sentences (five to seven words), resulting in low Flesch–Kincaid scores, though this may reflect task format rather than textual simplicity. More extended tasks (AN and AC) contain longer sentences (13.67–17.90 words per sentence) and higher AWL proportions (9.09–9.89%), comparable to or exceeding the current TOEFL and noticeably higher than IELTS. The C1-type lexical percentages follow a similar pattern, with the AC passage showing the highest proportion (4.35%).

2.2.2 Cognitive Processes for Listening

For cognitive processing, this report adopts the same framework used in Cushing (in press) in comparing communication goals and underlying skills relevant to listening in academic contexts outlined in Papageorgiou et al. (2021). Table 12 below summarizes the subskills that each test aims to assess, based on the information provided in their respective online materials for test takers. To maintain clarity, the table follows each test provider's own

labelling conventions: IELTS organizes its tasks by item type, the current TOEFL groups its multiple-choice items by their interpretive focus, and New TOEFL 2026 (same as the abbreviation used in the table above) defines its categories in terms of task types.

Table 12. Comparison of Skills Measured by Listening Tasks According to Papageorgiou et al.'s (2021) Framework

	IELTS	**Former TOEFL 2023**	**New TOEFL 2026**
Communication goals			
Main ideas & supporting details	Multiple choice Matching Summary completion	Gist Detail	CV AN AC
Relationships among ideas	Matching Sentence completion Plan/map/diagram labelling Summary completion	Organization Connecting content	AC
Inferences & opinions	---	Inference	CV
Speaker purpose & attitude	---	Function, Attitude	LCR CV
Foundational skills			
Process extended speech in real time	All item types	All item types	All item types
Make use of phonological information	All item types	All item types	All item types
Make use of lexical and grammatical meaning	All item types	All item types	All item types
Make use of pragmatic information	--	Inference, Function, Attitude	LCR
Process organizational devices	All item types	Organization, Connecting content	AC

LCR = Listen and Choose a Response CV = Listening to a Conversation
AN = Listen to an Announcement AC = Listen to an Academic Talk

As discussed in Cushing (in press), IELTS item types place greater emphasis on identifying ideas and understanding the relationships between them, while giving comparatively less attention to inferences, opinions, and speaker intention. In contrast, TOEFL iBT includes item categories that explicitly address these higher-level interpretive skills. Meanwhile for New TOEFL 2026, as shown in the table, the revised tasks remain similarly focused on skills assessed as the current TOEFL with its academic focus.

2.2.3 Text Length and CEFR Coverage

With respect to CEFR coverage, the New TOEFL 2026 tasks broaden coverage of campus-related genres, such as announcements, that align well with B2 descriptor domains. However, the shortened academic lecture component no longer fully reflects the B2 descriptors requiring the ability to process extended, complex oral input, for example:

- Following extended discourse and complex lines of argument when the topic is familiar and signposting is clear.
- Following complex argumentation in a clearly articulated lecture.
- Understanding the main ideas of propositionally and linguistically complex discourse on concrete and abstract topics, including technical discussions in one's field.

As with reading, score interpretation should therefore consider the extent to which the revised tasks represent the full range of CEFR constructs at each level, rather than assuming uniform coverage across levels.

2.3 Writing

2.3.1 Content and Tasks for Writing

The IELTS Writing section differs across the Academic and General Training (GT) versions. The Academic test focuses more on higher-education related skills and the GT test focuses more on everyday or workplace communication, rather than academic description.

The Writing section in the New TOEFL 2026 test now contains three short tasks to be completed in about 12 minutes:

- **Build a Sentence (BAS)**: ordering scrambled words into a grammatically correct sentence.
- **Write an Email (Email)**: composing a response to an everyday situation.
- **Write for an Academic Discussion (WAD)**: contributing opinions or ideas to a classroom-style online discussion.

The table below summarizes the content of the Writing task in the three tests. It is worth noting that the first two new and shorter Writing tasks are identical to those in the TOEFL Essentials test, which was designed as a lower-stakes, shorter assessment; this overlap raises questions about the extent to which these task formats are optimally suited to supporting high-stakes interpretations of academic writing proficiency in New TOEFL 2026.

Table 13. Content Comparison of Writing Tasks

Characteristic	IELTS	Former TOEFL 2023	New TOEFL 2026
Number of tasks	Two	Two	Three
Task description	Task 1: Describe or explain information presented in a chart, graph, or table Task 2: Write an essay in response to a point of view, argument, or problem.	Task 1: Integrated writing. Summarize information from reading and listening to short texts Task 2: Academic discussion. State and support an opinion in an online classroom discussion.	Task 1: Build a sentence in context (BAS) Task 2: Write an email Task 3: Academic discussion (WAD)
Timing	60 minutes total (tasks not timed separately) Recommended: 20 minutes on Task 1, 40 minutes on Task 2	30 minutes total: 20 minutes for integrated writing, 10 minutes for academic discussion	23 minutes in total
Functions elicited	Describe from visual information Support an opinion	Synthesize information from reading & listening Support an opinion	Online communication Support an opinion
Text length of expected response	Task 1: at least 150 words Task 2: at least 250 words	Integrated writing: 150–225 words Academic discussion: At least 100 words	BAS: 10 sentences, <10 words each Email: “write as much as possible” in 7 minutes WAD: at least 100 words in 10 minutes
Scoring	0–9 (analytic)	0–5 (holistic)	BAS: dichotomous Email: 0–5 (holistic) AD: 0–5 (holistic)

* Passage length was not mentioned in the technical manual. The number of words for each task is estimated based on sample tests.

The Write for an Academic Discussion (WAD) task, introduced in 2023 to replace the original independent writing task, measures integrated reading and writing skills, while being significantly shorter in its input and output requirement than the original independent task

and the integrated task. This WAD task retains the function of assessing test takers' ability to "express their ideas in a clear and coherent way" (Davis & Norris, 2023, p. 9) while avoiding the mechanistic "five-paragraph essay" format. Davis and Norris (2023) note that this task simulates a common academic genre and provides test takers with a clear audience and communicative purpose. The task format and scoring criteria remain unchanged in the 2026 TOEFL revision. Although the WAD task captures several important aspects of academic writing, its construct coverage is different from the previous integrated writing task which demands in-depth engagement with various modes of source materials. Removing the integrated writing task also removes the section's clearest operationalisation of source-based academic writing, arguably the most consequential writing ability for tertiary study, and eliminates what has historically been viewed as one of TOEFL iBT writing assessment's signature strengths: assessing integrated, meaning-focused academic composing rather than decontextualised production.

The newly added BAS task introduces short, conversational exchanges in which test takers produce a sentence-level response to a contextualized prompt. Although it is a writing task, its content is more reminiscent of spoken discourse or text-based communication than of traditional writing tasks (see Figure 1 below from the technical manual). This represents a mismatch in task mode, and it suggests that the BAS draws on simpler cognitive and linguistic processes than those typically associated with extended writing required in higher education, such as planning, organizing ideas, and producing cohesive written text. In addition, the dichotomous scoring approach (where a response is considered correct only if all words are filled in exactly as intended) places strong emphasis on precise form. This design likely only captures a narrow slice of performance compared to tasks that allow for partial success or alternate formulations. Taken together, these features raise questions about how closely the task reflects real academic writing situations, and whether it is a good fit for the kinds of writing students are expected to produce in university settings.

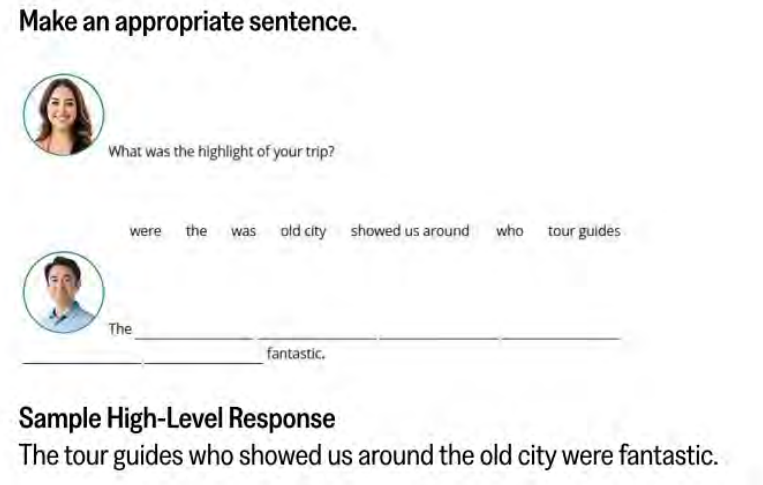

Figure 1. New TOEFL 2026 "Build a Sentence" Sample Item

The added Email task broadens the domain coverage of the writing component by introducing a non-academic, genre-specific communicative context. However, the way the task is framed (particularly the instruction to "*write as much as you can*" for an email) differs

from common expectations in authentic email communication, where clarity, conciseness, and relevance are typically valued more than length. It is also not fully clear how such factors are weighted within the automated scoring system. This task therefore is likely to capture only rushed sentence-level construction rather than the planning and audience-oriented discourse management that characterise effective email writing. As a result, performance may reflect "testwise" strategies (e.g., maximising length through loosely connected sentences) more than the ability to produce purpose-driven, pragmatically appropriate written communication in realistic settings, and it may also influence preparation strategies (see Section 5 on Washback Effects).

These differences in task design suggest that the Email task may engage aspects of writing ability that are not typically emphasized in real-world email practices, which may also influence the kinds of writing behaviours the task encourages (i.e., potential washback effects). As a result, score users may need to interpret performance on this task with additional caution, particularly when drawing inferences about a test taker's proficiency in authentic email communication or their broader writing competence.

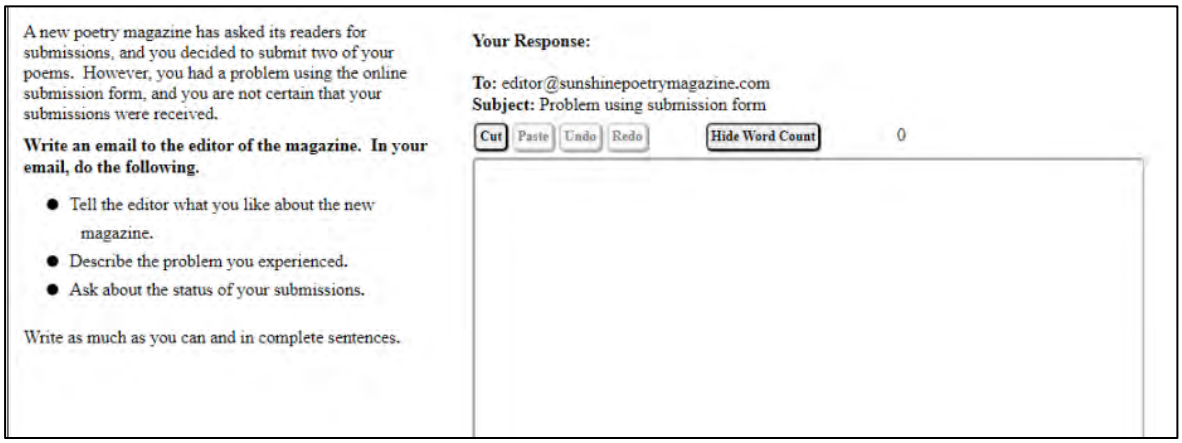

Figure 2. New TOEFL 2026 "Write an Email" Sample Item

2.3.2 Cognitive Processes for Writing

The analysis of cognitive processing follows the framework outlined in Cushing (in press), drawing on the cognitive validity model proposed by Shaw and Weir (2007). Each test is compared against the following six criteria:

- **Macro-planning** involves setting the overall direction of the text by generating ideas and clarifying key task parameters, including genre, purpose, and audience.
- **Organization** focuses on shaping the internal structure of the text by establishing relationships between ideas, arranging them in logical sequence, and prioritizing relevant ideas according to their contribution the thesis.
- **Micro-planning** is to make specific linguistic choices, selecting words to form sentences and paragraphs.
- **Translation** refers to converting abstract ideas into concrete linguistic forms that accurately express the intended meaning.
- **Monitoring** entails continually reviewing the accuracy, flow, and consistency of the text with the writer's intentions, ensuing alignment with the communicative goal.

- **Revising** involves making changes as a result of monitoring, enhancing the wording, content, organization, clarity, and coherence throughout the text.

Table 14. Comparison of Cognitive Processes for Writing

Cognitive processes	IELTS	Former TOEFL 2023		New TOEFL 2026		
		Integrated writing	WAD	BAS	Email	WAD
Macro-planning	x	x	X		x	x
Organization	x	x	(x)		(x)	(x)
Micro-planning	x	x	X	X	x	x
Translation	x	x	X			x
Monitoring	x	x	X	X		x
Revising	x	x	(x)			(x)

Processes that appear to be only partially represented—or comparatively less emphasized—are indicated in parentheses.

A broad comparison of the six cognitive processes, as summarized in Table 14, indicates that the overall coverage across writing tasks is largely comparable when the three New TOEFL 2026 tasks are considered collectively. It can be observed that, between IELTS and Former TOEFL 2023 with the integrated writing task, the cognitive processes are generally consistent. The introduction of the new tasks and the removal of the integrated writing task in New TOEFL 2026 have shifted the coverage slightly.

For New TOEFL 2026, the required output length in all the writing tasks is much shorter than before, which affects the extent to which key aspects of the academic writing construct can be represented and assessed. As shown in the table above, the writing tasks together in New TOEFL 2026 still engage most stages of cognitive processing, particularly micro-planning, translation, and monitoring, supported by shorter responses. However, processes such as macro-planning, organization, and revising are only partially represented in some tasks and tend to be more fully demonstrated in extended writing. For example, the short time given to the Email and WAD tasks leaves little to no space for macro-planning or revision. This reduced textual space also constrains the stability of linguistic indicators (such as lexical diversity or syntactic complexity) that are typically more reliable in longer samples (Koizumi, 2012), especially when marked by automarkers. As a result, while the tasks continue to capture certain aspects of the writing construct, score interpretation may require greater caution due to the narrower performance samples on which inferences about broader writing proficiency are based.

2.3.3 Scoring

To review the writing construct as operationalized in the scoring criteria, we consider the following rubric comparison. As the sentence-completion task is scored dichotomously, it is not included in the subsequent comparison.

Rubric comparison:

Table 15. Comparison of Writing Scoring Criteria (Highest Possible Score)

<table>
<tr><th>Crite-rion</th><th>IELTS Band 9</th><th>Former TOEFL 2023 Score 5</th><th>New TOEFL 2026 Score 5</th></tr>
<tr><td rowspan="2">Task Response</td><td rowspan="2">T1: All the requirements of the task are fully and appropriately satisfied.

There may be extremely rare lapses in content.

T2: The prompt is appropriately addressed and explored in depth.

A clear and fully developed position is presented which directly answers the question/s.

Ideas are relevant, fully extended and well supported.

Any lapses in content or support are extremely rare.</td><td>IW: A response at this level successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading. The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.</td><td>Email: The response is effective, is clearly expressed, and shows consistent facility in the use of language. Consistent use of appropriate social conventions (e.g., politeness, register, organization of information and formulation of actions such as requests, refusals, criticisms, etc.)</td></tr>
<tr><td colspan="2">WAD: The response is a relevant and very clearly expressed contribution to the online discussion, and it demonstrates consistent facility in the use of language.</td></tr>
<tr><td rowspan="2">Coherence & Cohesion / Topic development</td><td rowspan="2">The message can be followed effortlessly.

Cohesion is used in such a way that it very rarely attracts attention.

Any lapses in coherence or cohesion are minimal.

Paragraphing is skillfully managed.</td><td>IW: The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.</td><td>Email: Elaboration that effectively supports the communicative purpose.</td></tr>
<tr><td colspan="2">WAD: Relevant and well-elaborated explanations, exemplifications, and/or details.</td></tr>
</table>

<table>
<tr><td rowspan="2">Lexical Resource</td><td rowspan="2">T1: Full flexibility and precise use are evident within the scope of the task.

T2: Full flexibility and precise use are widely evident.

A wide range of vocabulary is used accurately and appropriately with very natural and sophisticated control of lexical features.

Minor errors in spelling and word formation are extremely rare and have minimal impact on communication.</td><td>WAD: Precise, idiomatic word choice.

Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like there/their).</td><td>Email & WAD: Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like there/their).</td></tr>
<tr><td rowspan="2">Grammatical Range & Accuracy</td><td rowspan="2">A wide range of structures is used with full flexibility and control.

Punctuation and grammar are used appropriately throughout.

Minor errors are extremely rare and have minimal impact on communication.</td><td>IW: Occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.</td><td>Email: Effective syntactic variety and precise, idiomatic word choice.</td></tr>
<tr><td colspan="2">WAD: [The response] demonstrates consistent facility in the use of language.

Effective use of a variety of syntactic structures.

Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like there/their).</td></tr>
</table>

As shown in Table 15 above, replacing the integrated writing task with email writing and sentence completion introduces a modest shift in the skills assessed. In Former TOEFL 2023, the integrated task evaluates the ability to select, synthesize, and summarize information from source texts. These are skills that are no longer directly measured in the revised section. In their place, the new Email task brings in elements of sociopragmatic competence, particularly familiarity with conventions of email communication.

These changes also correspond to adjustments in the cohesion, lexical, and grammatical dimensions of the scoring criteria. As both the email writing task and the academic discussion prompt shorter responses, the revised rubrics deemphasize discourse-level cohesion and instead focus on the clarity and completeness of idea presentation. Similarly, the lexical and grammatical criteria now lean more toward accuracy and consistency in error-free production, whereas IELTS maintains broader attention to the range and variety of vocabulary and structures.

Importantly, these proposed criteria should be considered together with the functionalities and scoring mechanisms of the automated scoring engine used by TOEFL. With limited human involvement, the linguistic features and thus the constructs represented in scores depend heavily on the design and capabilities of the automarking algorithm.

Automated Scoring:

Available documentation for the New TOEFL 2026 format indicates that human raters continue to play a role in scoring speaking responses, although the extent of their involvement is not fully specified. The Technical Manual notes that automated scoring engines are used and that human ratings are employed for quality assurance, model validation, and sampling. However, it does not clearly state whether every response is double-scored or whether some are scored solely by automation. Secondary sources also suggest that human involvement may not be uniform across all responses. At the same time, ETS maintains transparent scoring practices by publishing the constructs measured by the scoring engine in its technical manual, allowing users to understand the basis on which scores are generated.

According to the technical manual, when focusing on the features targeted by the scoring engine, content relevance in the Email task appears to be operationalized largely through surface-level indicators such as the number of sentences, basic discourse coherence, and lexical similarity to the prompt. As these features can be increased through strategies like producing a longer but only loosely related response or inserting formulaic cohesive expressions, they do not necessarily demonstrate whether the message meaningfully addresses the communicative purpose or topic focus specified in the scoring rubric.

Score	General Description
5	**A fully successful response** The response is effective, is clearly expressed, and shows consistent facility in the use of language. A typical response displays the following: • Elaboration that effectively supports the communicative purpose • Effective syntactic variety and precise, idiomatic word choice • Consistent use of appropriate social conventions (e.g., politeness, register, organization of information and formulation of actions such as requests, refusals, criticisms, etc.) • Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like *there/their*)

Figure 3. Rubric Descriptors for Score 5 – Write an Email

Table 3. Writing Section–*Write an Email*

Scoring dimensions	Feature examples
Content "elaboration ... supports the communication purpose"	• Number of sentences • Discourse coherence • Similarity to question prompt
Syntactic/Lexical variety "syntactic variety ... idiomatic word choice"	• Sentence variety • Word frequency • Correctness of collocations
Social Conventions "politeness, register, organization ... formulation of actions"	• Use of politeness indicators (e.g., modals, hedge words)
Accuracy/Errors "Almost no lexical or grammatical errors"	• Grammaticality • Grammatical errors • Word or usage errors • Mechanical errors (e.g., spelling or interpunctuation errors)

Figure 4. Automarker Scoring Dimensions and Features – Write an Email

Even if the reliance on surface-level lexical indicators in automated scoring is viewed as an inherent constraint of current technology, it is still important to consider how this applies to the Academic Discussion task. In this task, a central construct-relevant requirement is the ability to engage with the prompt and contribute appropriately to an academic-style exchange. The scoring features reported for the automated system, however, do not explicitly reflect all elements highlighted in the rubric, for instance, the expectation for "relevant and well-elaborated explanations, exemplifications, and details." This indicates that some aspects of the intended construct may be represented more indirectly within the automated features, suggesting an area where closer examination could support clearer alignment and more transparent score interpretation.

Score	Description
5	**A fully successful response** The response is a relevant and very clearly expressed contribution to the online discussion, and it demonstrates consistent facility in the use of language. A typical response displays the following: • Relevant and well-elaborated explanations, exemplifications and/or details • Effective use of a variety of syntactic structures and precise, idiomatic word choice • Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like *there/their*)

Figure 5. Rubric Descriptors for Score 5 – Academic Discussion

Table 4. Writing Section–*Write for an Academic Discussion*

Scoring dimensions	Feature examples
Content "Relevant and well-elaborated explanations … details"	• Number of sentences • Discourse coherence • Similarity to question prompt
Syntactic/Lexical variety "variety of syntactic structures and precise, idiomatic word choice"	• Sentence variety • Word frequency • Correctness of collocations

Figure 6. Automarker Scoring Dimensions and Features – Academic Discussion

2.4 Speaking

2.4.1 Content and Context Validity for Speaking

The IELTS Speaking test is a face-to-face interview (online or in-person) conducted by a trained examiner. The interview includes a range of tasks, such as personal questions, a short individual long turn, and a two-way discussion. These types of tasks elicit both everyday and more abstract language use.

Former TOEFL 2023 includes a set of structured tasks, including both independent questions—where test takers give their opinions or describe personal experiences—and integrated questions, which require listening to a brief lecture or reading a short text before responding.

In New TOEFL 2026, the Speaking section has been revised from multiple integrated and independent tasks to two practical ones: *Listen and Repeat (LRP)*, where students repeat seven short sentences based on visual cues to assess pronunciation and rhythm, and *Take an Interview (INT)*, where they answer a few conversational questions on familiar topics within 45 seconds each.

Table 16. Content Comparison of Speaking Tasks

	IELTS	Former TOEFL 2023	New TOEFL 2026
Number of tasks	Three	Four	Two
Task description	Task 1: General questions on familiar topics Task 2: Long turn (candidate is asked to speak for one to two minutes on a topic) Task 3: Discussion (elaborate on issues related to Task 2)	One is an independent / personal opinion task Three are integrated, i.e., based on listening and/or reading	LRP: Repeat a series of sentences within a scenario in an academic or daily life setting. INT: Take part in a simulated interview, moving from personal/factual questions to opinions.
Preparation time	Yes	Yes	No
Delivery mode	All three tasks are embedded within an oral interview with a human examiner.	All tasks are computer-delivered.	On computer, with visual assistance.
Functions elicited	Providing personal information, expressing and justifying opinions, explaining, suggesting, speculating, expressing preferences, comparing, summarizing, narrating.*	Expressing and justifying opinions, summarizing. synthesizing information from multiple sources.	LRP: ability to accurately and intelligibly reproduce heard sentences INT: ability to answer general and academic questions with coherent elaboration, intelligible delivery, and effective use of vocabulary, grammar, and prosody.
Timing	11–14 minutes total, including one minute of	Approximately 15 minutes	Approximately eight minutes

	preparation time (Task 2)		

2.4.2 Cognitive Processes for Speaking

A common framework for analyzing cognitive processing in speaking tasks is the model used in Taylor and Chan's (2015) test comparison work, drawing on Weir's (2005) cognitive validity framework, and Field's (2011) insights on speech production. This same framework was applied in Cushing and Ren (2022) in their comparison of IELTS and the Duolingo English Test. However, as Cushing and Ren (2022) note, a key limitation of this framework is that it focuses primarily on the internal processes involved in producing an utterance and does not account for interactional competence, which is an essential component of communicative performance. This limitation becomes evident in the present comparison: *all three tests meet the six cognitive-processing stages identified in the framework*, making it insufficient for distinguishing how the tests differ in eliciting interactional skills. For this reason, a detailed stage-by-stage breakdown is not provided here.

What is worth foregrounding, however, is that the "Listen and Repeat" task type appears to tap only minimally into higher-level cognitive processing and is weakly aligned with meaning-oriented, authentic oral communication. The task primarily requires accurate phonological encoding and immediate reproduction under time pressure, thereby placing substantial demands on attention and working memory rather than on message construction, pragmatic choice-making, or interactive management. In this respect, "Listen and Repeat" resembles earlier automated speaking formats (e.g., PhonePass / Versant) that have been criticised on authenticity grounds, particularly because successful performance can be achieved through faithful repetition without demonstrating the ability to formulate, negotiate, or adapt meaning for a communicative purpose (Chun, 2006, 2008). Related critiques have also been raised in evaluations of highly constrained, technology-mediated speaking tasks, where limited opportunity for interaction and contingent response can restrict construct coverage of communicative speaking ability (Wagner, 2020). Taken together, these concerns suggest that, from a communication-oriented assessment perspective, "Listen and Repeat" may contribute only narrow evidence about oral ability and should be interpreted cautiously if used to support broad claims about communicative speaking proficiency.

2.4.3 Scoring

IELTS Speaking is scored by a trained examiner who rates a test taker's performance across four analytic criteria: (1) Fluency and Coherence, (2) Lexical Resource, (3) Grammatical Range & Accuracy, and (4) Pronunciation. Each criterion receives a band score from 0–9, and the four scores are averaged to produce the overall Speaking band.

The Former TOEFL 2023 Speaking responses are scored using a combination of trained human raters and automated scoring. Responses are evaluated on three analytic dimensions: Delivery (clarity and intelligibility), Language Use (lexical and grammatical performance),

and Topic Development (coherence and completeness of the response). Each task receives a score on a 0–4 scale, which is then converted into the final scaled Speaking score (0–30).

In New TOEFL 2026, responses to all speaking tasks are assigned scores from 0 to 5 score points based on criteria defined in the respective scoring rubric according to the TOEFL 2025 Technical Manual.

As noted in the earlier analysis of writing, an important consideration in scoring productive tasks is the role of the automated scoring engine. A review comparing the full human-scoring rubrics for the interview tasks with the scoring features used by the automated system indicates that the rubric's criterion for topical relevance and elaboration (i.e., whether the response directly addresses the prompt and develops ideas appropriately) is not explicitly represented among the features used by the scoring engine.

Score	Description
5	**A fully successful response** The response fully addresses the question, and it is clear and fluent. A typical response exhibits the following: • The response is on topic and well elaborated. • Good conversational speaking pace is maintained with appropriate and natural use of pauses. • Pronunciation is easily intelligible; rhythm and intonation effectively convey meaning. • A range of accurate grammar and vocabulary allows clear expression of precise meanings.

Figure 7. Rubric Descriptors for Score 5 – Take an Interview

Table 6. Speaking Section–*Take an Interview*

Scoring dimensions	Feature examples
Fluency	• Speaking rate • Length of uninterrupted *runs* (word sequences without pauses) • Number of pauses • Number of hesitations
Intelligibility	• Correctness of pronunciation • Naturalness of speech rhythm • Naturalness of prosody (e.g., syllable stress)
Language Use: Vocabulary and Grammar	• Vocabulary diversity (using a wide range of words that are distinct from one another) • Vocabulary richness (use of words which are less common) • Grammaticality • Grammatical accuracy (few grammar errors)
Organization	• Discourse coherence • Use of discourse connectives

Figure 8. Automarker Scoring Dimensions and Features – Take an Interview

The absence of a clear match between rubric requirements and the features captured by the automated scoring system has several implications for score interpretation. When construct-relevant qualities such as topical relevance or elaboration are not directly assessed by the scoring features, there is a risk that the resulting scores may place greater weight on linguistic form than on substantive communicative performance. This does not necessarily invalidate the scores but narrows the aspect of performance that are being consistently measured. As the extent of human rater involvement in the 2026 scoring model is not publicly specified, it remains difficult to determine the extent to which human judgment compensates for gaps in automated scoring. This uncertainty complicates any analysis of how the scoring system operationalizes the construct and underscores the need for careful and transparent score interpretation.

3. Adaptive Testing

Adaptive testing has been widely adopted in language proficiency assessment because it improves measurement efficiency and enhances score precision by aligning test difficulty with the test taker's ability level. While the most common form of adaptation occurs at the item level, this approach requires extremely large, finely calibrated item banks, which can be challenging to maintain in tests that rely on complex task formats. To address these constraints, the TOEFL 2025 Technical Manual outlines the advantages of using a multi-stage adaptive (MST) format as an alternative.

In the MST design, modules are built as coherent task sets rather than as individual items, allowing the test to preserve balanced content coverage, uphold a task-based test structure, and ensure that all materials undergo expert review before administration. At the same time, adaptive routing within the multi-stage framework still personalizes difficulty based on test taker performance, improving efficiency and keeping testing time manageable without compromising reliability. MST also eases demands on item pools and provides greater control in assembling test forms than item-level adaptive systems, making it a practical and psychometrically robust approach for measuring language proficiency.

In New TOEFL 2026, all test takers begin with a routing module made up of moderately challenging tasks (generally aligned with CEFR B1–B2 levels). A test taker's performance on this initial module determines which follow-up module they receive: either a lower-difficulty or higher-difficulty version. The graphs below are taken from the TOEFL 2025 Technical Manual (Manna et al., 2025) to illustrate the test taking paths for listening and reading modules. Final scores reflect not only the number of correct answers but also the difficulty level of the modules completed. This design enables the test to adjust to a test taker's proficiency while maintaining coherent, task-based modules.

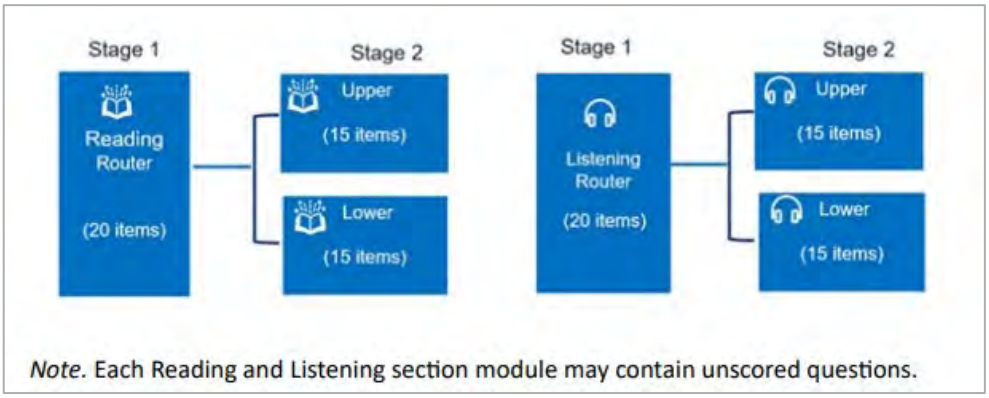

Figure 9. TOEFL Reading and Listening Multi-stage Adaptive Test Methodology (reprinted from the TOEFL 2025 Technical Manual)

Section	Task type	Number of scored questions in stages			Number of scored questions in paths	
		Stage 1	Stage 2 lower	Stage 2 upper	Easy path	Hard path
Reading	*Complete the Words*	10	10	10	20	20
	Read in Daily Life	5	5	0	10	5
	Read an Academic Passage	5	0	5	5	10
	Total	20	15	15	35	35
Listening	*Listen and Choose a Response*	8	7	3	15	11
	Listen to a Conversation	4	4	4	8	8
	Listen to an Announcement	4	4	0	8	4
	Listen to Academic Talk	4	0	8	4	12
	Total	20	15	15	35	35

Note. Each Reading and Listening section module may contain extra unscored questions.

Figure 10. TOEFL MST Content Design for Reading and Listening Sections (reprinted from the TOEFL 2025 Technical Manual)

In the multi-stage adaptive design, modules are built as coherent task sets rather than as individual items, allowing the test to preserve balanced content coverage, uphold a task-based test structure, and ensure that all materials undergo expert review before administration. At the same time, adaptive routing within the multi-stage framework still personalizes difficulty based on test taker performance, improving efficiency and keeping testing time manageable without compromising reliability. MST also eases demands on item pools and provides greater control in assembling test forms than item-level adaptive systems, making it a practical and psychometrically robust approach for measuring language proficiency.

While the multi-stage adaptive design strengthens the efficiency and measurement precision of New TOEFL 2026, caution should be taken when interpreting scores as these psychometric gains do not necessarily compensate for shifts in the underlying construct being sampled. Adaptive routing increases reliability for the sampled behaviours, but score interpretations must acknowledge that the meaning of the score is tied to the sampled construct. Thus, while adaptive testing enhances precision, it also requires careful attention to construct coverage to ensure that efficiency gains do not come at the expense of representativeness of the academic proficiency the test is intended to measure.

While the multi-stage adaptive design may strengthen the *efficiency* and *measurement precision* of New TOEFL 2026, these psychometric gains should not be treated as a remedy for the more fundamental issue identified in this comparison—namely, that the revised assessment samples a different configuration of construct-relevant behaviours than the prior test. Adaptive routing can improve reliability and reduce measurement error *for what is being sampled*, but it does not, in itself, guarantee that the sampled domain is sufficiently aligned with the intended construct, nor that it is comparable to the construct operationalised by the legacy assessment. In other words, greater precision is not equivalent to greater validity: a

test can estimate a score very precisely while still estimating a *different* underlying attribute than stakeholders assume.

This distinction is particularly consequential in comparability arguments. The interpretability of New TOEFL 2026 scores ultimately depends on the construct representation embedded in its task set and routing logic—what language use behaviours are elicited, which skills and processes are foregrounded, and which are attenuated or omitted. Multi-stage adaptivity may enhance the consistency of measurement within the revised framework, yet score meaning remains inseparable from construct coverage. If adaptive design achieves shorter testing times by narrowing the range of content, reducing task diversity, or prioritising behaviours that are more readily scalable, then efficiency gains risk coming at the expense of representativeness of the academic proficiency domain the test purports to measure. Accordingly, any claims that adaptivity makes the revised test "effectively" equivalent should be tempered: improved efficiency and precision do not offset construct discrepancies, and comparability of score interpretations cannot be assumed without explicit evidence that the revised test maintains appropriate construct coverage and alignment with the intended target domain.

4. CEFR Comparison

This section presents CEFR B2 descriptors that are most directly aligned with areas affected by the transition from the Former TOEFL 2023 format to New TOEFL 2026. By focusing on descriptors linked to extended discourse, processing of complex information, and production of structured written and spoken output, this section provides a basis for interpreting how the test's construct coverage may shift under the new design.

In terms of constructs that are no longer targeted in New TOEFL 2026 (marked as removed, in the first half of Table 17), one of the most consequential changes is the removal of the integrated writing task. The revised test therefore no longer assesses the skills captured by eight B2 descriptors that focus on processing or producing longer, extended texts. For the other skills, the shortened tasks affect skills associated with handling extended or complex input, with the impact most evident in demands related to sustained comprehension, discourse-level integration, and the management of cumulative information across longer stretches of text. For Listening (italicised), the relevant skills remain within the intended assessment scope; however, given the very short lecture stimuli, they are unlikely to be elicited to the same depth as in the previous listening tasks. In terms of the added tasks, the new Listening "announcement" task introduces CEFR coverage associated with understanding announcements. The Email writing task similarly aligns, in principle, with the CEFR correspondence category; however, as discussed earlier, the evaluation criteria may not operationalise these skills appropriately. As a result, any claims of added CEFR coverage should be treated with caution, and the extent to which these tasks provide meaningful evidence for the intended correspondence-related construct warrants closer scrutiny.

Table 17. CEFR B2 Coverage Analysis

B2 levels					
Tasks that elicit the following: **Removed**					
CEFR descriptor scheme		**Mode of communication**	**Activity, strategy or competence**	**Scale**	**Descriptor**
Listening	Communicative language activities	Reception	Oral comprehension	Overall oral comprehension	Can follow **extended discourse** and complex lines of argument, provided the topic is reasonably familiar, and the direction of the argument is signposted by explicit markers.
				Understanding as a member of a live audience	Can follow **complex lines of argument** in a clearly articulated lecture, provided the topic is reasonably familiar.
				Overall oral comprehension	Can understand the main ideas of **propositionally and linguistically complex** discourse on both concrete and abstract topics delivered in standard language or a familiar variety, including technical discussions in their field of specialisation.
Reading	Communicative language activities	Reception	Reading comprehension	Reading for orientation	Can scan quickly through **long and complex texts**, locating relevant details.

Speaking	Communicative language activities	Production	Oral production	Sustained monologue: putting a case (e.g., in a debate)	Can construct **a chain of reasoned argument**.
	Communicative language strategies	Interaction		Co-operating	Can **summarise the point reached at a particular stage in a discussion** and propose the next steps.
Writing	Communicative language activities	Production	Written production	Overall written production	Can produce clear, detailed texts on a variety of subjects related to their field of interest, **synthesising and evaluating information and arguments from a number of sources.**
			Written production	Reports and essays	Can produce an **essay or report** which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options.
			Written production	Reports and essays	Can **synthesise** information and arguments from a number of sources.
		Mediation	*Mediating concepts*	*Collaborating to construct meaning*	*Can consider **two different sides of an issue**, giving arguments for and against, and propose a solution or compromise.*

			Mediating communication	Facilitating communication in delicate situations and disagreements	Can outline the main points in a disagreement with reasonable precision and **explain the positions** of the parties involved.
			Mediating communication	Facilitating communication in delicate situations and disagreements	Can **summarise** the statements made **by the two sides**, highlighting areas of agreement and obstacles to agreement.
	Communicative language competences		Pragmatic competence	Thematic development	Can develop a clear argument, expanding and supporting their points of view **at some length** with subsidiary points and relevant examples.
			Pragmatic competence	Coherence and cohesion	Can structure **longer texts in clear, logical paragraphs**.
Listening/ wri	*Communicative language activities*	*Mediation*	*Mediating a text*	*Note-taking (lectures, seminars, meetings, etc.)*	*Can understand a clearly structured lecture on a familiar subject, and can take notes on points which strike them as important, even though they tend to concentrate on the actual formulation and therefore to miss some information.*

ting	*Communicative language activities*	*Mediation*	*Mediating a text*	*Note-taking (lectures, seminars, meetings, etc.)*	*Can take accurate notes in meetings and seminars on most matters likely to arise within their field of interest.*
Added tasks that elicit the following (some are not scored)					
Listening	Communicative language activities	Reception	Oral comprehension	Understanding announcements and instructions	Can understand **announcements** and messages on concrete and abstract topics delivered in standard language or a familiar variety at normal speed.
W (partial)	Communicative language activities	Interaction	Written interaction	Correspondence	Can use formality and conventions appropriate to the context when writing personal and professional letters and **e-mails**.
					Can obtain, by letter or e-mail, information required for a particular purpose, collate it and forward it by **e-mail** to other people.
					Can compose non-routine professional letters, using **appropriate structure and conventions**, provided these are restricted to matters of fact.

5. Consequential / Washback Effects

Several features of the revised task types may also have unintended washback effects on learning and preparation. The substantially shorter reading passages have already prompted social-media commentary suggesting that the new TOEFL reading tasks are “easier,” (https://www.youtube.com/watch?v=2uSYEekG97s) which may shift preparation toward superficial strategies rather than developing the extended reading skills required for academic study. Such perceptions risk creating a mismatch between what learners practice for the test and the demands they will encounter in university settings, where managing long and complex texts remains essential.

A similar concern arises in several features of the revised task types may also generate unintended washback effects on learning and preparation. In high-stakes contexts, washback is often mediated less by technical design intentions than by how key stakeholders (learners, teachers, and the preparation industry) perceive test demands and respond strategically (Alderson & Wall, 1993; Cheng, Watanabe, & Curtis, 2004). In the case of New TOEFL 2026, public-facing discussions already foregrounded the reduced reading load and the inclusion of non-academic, everyday materials in the reading section (Educational Testing Service [ETS], 2025a, 2025b), with commentaries explicitly characterising the revised reading component as “easier” because candidates “have to read less” (e.g., TOEFL with Juva, n.d.). Such perceptions—irrespective of ETS’s intended construct claims—may redirect preparation toward short-text routines (e.g., rapid scanning, keyword matching, option elimination, and format-specific heuristics) rather than the sustained academic reading capacities needed for university study (e.g., maintaining comprehension across extended discourse, integrating ideas across paragraphs, and tracking complex argument structure). This matters because when test demands are perceived as narrowly defined or readily coachable, preparation may shift toward score-maximising routines rather than broader literacy development, reducing alignment between test-focused practice and the academic reading demands encountered in tertiary study (Green, 2007).

Concerns about washback effects also come up in some of the specific task types discussed in the previous sections. In writing, for example, the new Email task explicitly asks test takers to “write as much as possible,” which does not conform to real-world expectations for emails to be concise and purpose-driven. This kind of instruction may encourage inappropriate email-writing behaviours that develop through preparation, such as writing for length rather than for clarity, appropriateness, or audience. A related issue also arises in the Read Aloud speaking task where test takers are asked to read sentences aloud, which places more emphasis on fluent oral reading than on the kinds of speaking that matter in academic or professional communication.

The main washback risk is that the revised task types may be perceived as increasingly coachable while being less directly relevant to the critical communicative and academic literacy demands of target university study. If so, preparation may shift toward optimising

short-horizon, format-specific routines rather than developing durable language capacities that support academic participation. In high-stakes use, this would risk positioning the test as a nominal hurdle to clear—useful for selection, but less effective as a credible indicator of readiness or as a mechanism that meaningfully supports positive outcomes for both universities and prospective students.

6. Summary and Discussion

The revisions introduced in New TOEFL 2026 modify construct coverage across skills by reshaping task formats, input characteristics, and scoring procedures. Overall, the test continues to assess key aspects of English proficiency, but it does so with a different emphasis compared to Former TOEFL 2023. A summary table is presented in Appendix C.

The new tasks broaden the range of social and interpersonal contexts represented in the test, offering wider coverage of everyday communication. At the same time, the reduced presence of integrated academic tasks—such as source-based writing and speaking—means that fewer opportunities are provided to elicit higher-order academic skills, including synthesis and source integration. This results in a construct that is more general-purpose in orientation, with comparatively less focus on academic depth.

Across skills, input and output lengths are shorter than in the previous test design. Shorter texts and briefer responses allow for efficient administration and adaptive delivery, but they also narrow the sampling of linguistic and cognitive processes, particularly those associated with extended comprehension, sustained production, or engagement with complex academic materials. As a result, the test captures a more concise sample of language performance, which shapes the types of inferences that can be made about test takers' abilities.

Within this broader shift, the writing section illustrates how changes in task design influence content, processing, and scoring. The inclusion of new genres such as email writing extends the range of communicative contexts represented, while the removal of the integrated writing task reduces direct assessment of expository writing and source integration—skills that were central to the previous construct definition but are not referenced in the new rubrics. From a cognitive-processing perspective, shorter response lengths and tighter time constraints limit observable planning and revision processes and may make certain linguistic indicators, such as lexical diversity, less stable. These changes are reflected in the updated scoring framework: criteria that previously required selecting and integrating information from source texts no longer appear in the 2026 rubrics, aligning the scoring with the shift toward non–source-based writing tasks. In addition, while automated scoring plays a primary role, it remains unclear how human raters are involved in operational scoring and how their judgments interact with the automated system.

Some of the newer item types assess more specific linguistic skills—for example, lexical-syntactic processing in C-tests or sentence-level formulation tasks in writing. These tasks broaden task variety but emphasize focused, lower-level linguistic operations rather than

discourse-level communicative abilities. In speaking, repetition-based tasks primarily tap perceptual and short-term memory demands, representing a different construct emphasis from tasks requiring idea development or extended discourse.

The shift to primarily automated scoring for productive skills brings consistency and efficiency to rating processes. However, automated systems typically rely on quantifiable linguistic features, which influences the aspects of performance that can be evaluated directly and shapes the operational definition of writing and speaking proficiency within the test.

Finally, while multi-stage adaptive testing improves measurement precision within the available item pool, it does not alter the underlying construct represented by the tasks themselves. Because the revisions collectively redefine the test's construct coverage, scores from New TOEFL 2026 are not directly comparable to results from earlier TOEFL versions, and previous concordance tables are not transferable.

Given the substantive differences between New TOEFL 2026 and its earlier versions—including changes to task types, response lengths, scoring procedures, and the balance between academic and general-purpose language use—the basis for interpreting score equivalence has fundamentally shifted. As a result, existing concordance studies are unlikely to remain meaningful for the New TOEFL 2026 test, and new concordance efforts would need to account carefully for the altered construct coverage. Institutions and stakeholders should therefore exercise caution when applying legacy concordance tables or assuming comparability with scores from Former TOEFL 2023.

These changes carry significant implications for receiving organizations, particularly higher education institutions. Stakeholders may need to consider whether reduced assessment of certain academic skills will manifest as challenges for students once enrolled, or whether institutions will instead take on the responsibility of adjusting curricula, support structures, or preparatory programs to address skills that are no longer directly measured. This may include revisiting academic support courses, refining expectations for incoming students' skill profiles, or adapting instructional approaches in areas such as international teaching assistant (ITA) training. Institutions' decisions in this regard will shape how the revised test is integrated into admissions and how students are supported after matriculation.

References

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. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and context in defining constructs in language assessment. In J. Fox , M. Wesche & D. Bayless (Eds.), *What are we measuring? Language testing reconsidered* (pp. 41–72). University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Baghaei, P. (2010). *An investigation of the invariance of Rasch item and person measures in a C-test.* In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung / The C-Test: Contributions from Current Research* (pp. 101–112). Peter Lang.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge University Press.

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Sage.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Lawrence Erlbaum Associates.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly*, *3*(3), 295–306. https://doi.org/10.1207/s15434311laq0303_4

Chun, C. W. (2008). Comments on 'Evaluation of the usefulness of the Versant for English Test: A response': The author responds. *Language Assessment Quarterly, 5*(2), 168–172. https://doi.org/10.1080/15434300801934751

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Cushing, S. T. (in press). *Content Comparison of IELTS and TOEFL iBT.*

Cushing, S. T. (2025). *Testing academic language proficiency: Comparing the TOEFL iBT® test with the Duolingo English Test (TOEFL Research Report No. RR-104)*. ETS.

Cushing, S. T., & Ren, H. (2022). Comparison of IELTS Academic and Duolingo English Test. *IELTS Partnership Research Papers: Studies in Test Comparability Series, No. 1*. IELTS Partners: British Council/Cambridge Assessment English/IDP: IELTS Australia. Available at https://www.ielts.org/for-researchers/research-reports

Davis, L., & Norris, J. M. (2023). *A comparison of two TOEFL writing tasks* (Research Memorandum No. RM-23-06). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-23-06.pdf

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290–325. https://doi.org/10.1191/0265532206lt330oa (Original work published 2006)

Educational Testing Service. (n.d. [a]). *TOEFL IBT listening section*. TOEFL. https://www.ets.org/toefl/test-takers/ibt/about/content/listening.html

Educational Testing Service. (n.d. [b]). *TOEFL IBT reading section*. TOEFL. https://www.ets.org/toefl/test-takers/ibt/about/content/reading.html

Educational Testing Service (n.d. [c]). *TOEFL IBT speaking section*. https://www.ets.org/toefl/test-takers/ibt/about/content/speaking.html

Educational Testing Service (n.d. [d]). *TOEFL IBT writing section*. https://www.ets.org/toefl/test-takers/ibt/about/content/writing.html

Educational Testing Service. (2024a). *TOEFL iBT® test framework and test development* (TOEFL® Research Insights Series, Volume 1; updated November 2024). ETS.

Educational Testing Service. (2024b). *TOEFL iBT® test and score data summary 2024*. ETS.

Educational Testing Service. (2025a). *TOEFL transformation announcement* [Press release].

Educational Testing Service. (2025b). *TOEFL iBT® practice test 1* [Practice test].

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (Studies in Language Testing, Vol. 30, pp. 65–111). Cambridge University Press.

Flesch, R. (1948). *A new readability yardstick. Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy & Practice, 14*(1), 75–97. https://doi.org/10.1080/09695940701272880

IELTS, n.d [a]. *The IELTS Guide for Test Takers.* https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/ielts-guides/ielts-guide-for-test-takers.pdf

IELTS, n.d. [b]. *IELTS Academic format: Reading.* https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-reading

IELTS, n.d. [c]. *IELTS Academic test—sample test questions*. https://ielts.org/take-a-test/preparation-resources/sample-test-questions/academic-test

IELTS, n.d. [d]. *IELTS Academic format: Listening*. https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-listening

Ikeda, N., Clark, T., Papageorgiou, S., Gu, L., Ohta, R., Blackhurst, A., & Bruce, E. (2025). *Aligning scores of language proficiency tests: A score concordance study between IELTS Academic and TOEFL iBT*. IELTS Partnership Research Papers: Studies in Test Comparability Series, No. 1/25. British Council/IDP: IELTS Australia/Cambridge University Press & Assessment.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger.

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and practice in second language reading.* UCLES/Cambridge University Press.

Kim, T., & Lee, B. (2025). The Passage is Too Brief for Comprehension: On the Construct Validity of the English Reading Section in the Korean College Scholastic Ability Test. *Language Assessment Quarterly*, *22*(2), 138–163. https://doi.org/10.1080/15434303.2025.2497815

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel* (Research Branch Report 8–75). Naval Air Station Memphis, Research Branch.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-Test1. *Language Testing*, *1*(2), 134–146. https://doi.org/10.1177/026553228400100202

Knoch, U., & Fan, J. (2024). Test score comparison tables: How well are they serving test users?. *Language Testing*, *41*(3), 681–693. https://doi.org/10.1177/02655322241239348

Koizumi, R., & In'nami, Y. (2012). WITHDRAWN: Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 522–532. https://doi.org/10.1016/j.system.2012.10.017

Liu, X., & Read, J. (2023). Designing a new diagnostic reading assessment for a local post-admission assessment program: A needs-driven approach. In *Diagnostic assessment of second language reading* (pp. xx–xx). Springer. https://doi.org/10.1007/978-3-031-33541-9_5

Manna, V. F., Li, S., Papageorgiou, S., & Gu, L. (2025). *TOEFL iBT® technical manual* (TOEFL Research Report No. TOEFL-RR-109; ETS Research Report No. RR-25-12). Educational Testing Service. https://doi.org/10.64634/eje8f497

McCall, W. A. (1922). *How to measure in education*. Macmillan.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

North, B., & Piccardo, E. (2016). Developing illustrative descriptors of aspects of mediation for the CEFR. *Language Teaching, 49*(3), 455–459. https://doi.org/10.1017/S0261444816000100

Papageorgiou, S., Schmidgall, J., Harding, L., Nissan, S., & French, R. (2021). Assessing academic listening. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 61–106). Routledge. https://doi.org/10.4324/9781351142403-3

Roohani, A. (2007). Validity and Discriminatory Power of the C-Test as a Measure of General Language Proficiency. *Teaching English Language, 1*(Special Issue 2), 1–28. doi: 10.22132/tel.2007.113455

Rouhani, A. (2007). Validity and Discriminatory Power of the C-Test as a Measure of General Language Proficiency. *Teaching English Language, 1*(Special Issue 2), 1–28. 10.22132/tel.2007.113455

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing: Research and practice in second language writing.* UCLES/Cambridge University Press.

Taylor, L., & Chan, S. (2015). *IELTS equivalence research project (GMC 133): Final report*. Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire.

TOEFL with Juva. (n.d.). *New TOEFL iBT test 2026 – Everything you MUST know* [Video]. YouTube. Retrieved January 18, 2026, from https://www.youtube.com/watch?v=2uSYEekG97s

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Wagner, E. (2020). Duolingo English Test, Revised Version July 2019. *Language Assessment Quarterly, 17*(3), 300v315. https://doi.org/10.1080/15434303.2020.1771343

Winke, P., Yan, X., & Lee, S. (2024). What does the cloze test really test? A cognitive validation of a French cloze test with eye-tracking and interview data. In G. Yu & J.

Xu (Eds.), *Language test validation in a digital age* (pp. 17–42). Cambridge University Press & Assessment.

Zhang, Y. (2025). *Validity Verification of the New TOEFL Writing Task Based on Classical Test Theory*. 10.48550/arXiv.2509.05347

Appendix A - Reading Skill Coverage:

IELTS Academic Reading:

https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-reading

- Detailed understanding of specific points or general understanding of the main points of the text (Multiple choice)
- Ability to recognise specific information given in the text (Identifying information)
- Ability to recognise opinions or ideas (yes/no/not given for identifying writer's views)
- Ability to scan a text in order to find specific information (Matching information)
- Ability to identify the general topic of a paragraph (or section) and to recognise the difference between the main idea and a supporting idea (Matching headings)
- Ability to recognise relationships and connections between facts in the text (Matching features)
- Ability to recognise opinions and theories (Matching features)
- Ability to understand the main ideas in the text (Matching sentence endings)
- Ability to find detail/specific information in a text (Sentence completion)
- Ability to understand details and/or the main ideas of a part of the text (Summary/note/table/flow-chart completion)
- Ability to understand a detailed description in the text, and then relate that description to information given in a diagram (Diagram label completion)
- Ability to find and understand specific information in the text (Short-answer questions)

Old TOEFL reading: https://www.ets.org/toefl/test-takers/ibt/about/content/reading.html

- Recognize factual information
- Recognize implied information
- Infer writer's purpose
- Identify the meaning of words as they are used in the text
- Understand the logical order of ideas
- Recognize the major ideas and relative importance of information in the text

New TOEFL 2026 reading: (Manna et al., 2025)

- The ability to process written texts for meaning and form (C-test)
- Understand information in common, nonlinear text formats (RDL)
- Identify the main purpose of a written communication (RDL)
- Understand informal language, including common idiomatic expressions (RDL)
- Make inferences based on text (RDL)

- Understand telegraphic language (RDL)
- Skim and scan for information (RDL)
- Identify the main ideas and basic context of a short, linear text (AP)
- Understand the important details in a short text (AP)
- Understand the range of grammatical structures used by academic writers (AP)
- Infer meaning from information that is not explicitly stated (AP)
- Understand a broad range of academic vocabulary (AP)
- Understand a range of figurative and idiomatic expressions (AP)
- Understand ideas expressed with grammatical complexity (AP)
- Understand the relationship between ideas across sentences and paragraphs (AP)
- Recognize the rhetorical structure of all or part of a written text (AP)

Appendix B - Listening Skill Coverage

IELTS: https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-listening

Merged Skill / Ability	**MC**	**MI**	**LB**	**FN**	**SC**	**SA**
General understanding of the main point (main idea, author's intention, etc.)	✓			✓		
Understanding factual details (types, places, prices, times)	✓	✓				✓
Recognising relationships between pieces of information (connections, cause–effect)		✓		✓	✓	
Understanding descriptions and spatial/directional information (visuals, directions)			✓			
Identifying key points / important information (note-worthy ideas, essential info)	✓			✓	✓	✓

MC = Multiple choice MI = Matching information

LB = Plan/map/diagram labelling FN = Form/note/table/flow chart/summary completion

SC = Sentence completion SA = Short-answer questions

The table is a summary of the bullet points below:

- A detailed understanding of specific points, or general understanding of the main points of the recording (MCQ)
- Listen for detailed information. For example, whether you can understand information about the type of hotel or guest house accommodation in an everyday conversation (Matching)
- Recognise how facts in the recording are connected to each other (Matching)
- ability to understand descriptions and how the descriptions relate to a visual (Plan/map/diagram labelling)
- ability to understand explanations of where things are and follow directions (e.g., straight on/through the far door) (Plan/map/diagram labelling)
- main points the person listening would naturally write down (note taking) (Form/note/table/flow chart/summary completion)
- ability to identify the important information (Sentence completion)
- Understand relationships between ideas/facts/events, such as cause and effect (Sentence completion)
- ability to listen for facts, such as places, prices or times, heard in the recording (Short-answer questions)

Current TOEFL: https://www.ets.org/toefl/test-takers/ibt/about/content/listening.html

- Understanding main ideas and supporting details in a spoken text (lectures, conversations)
- Recognising speaker attitude, intent or degree of certainty
- Connecting information across speakers/turns and across the discourse (e.g., lecture with discussion)
- Interpreting academic/campus-context talk (lectures, student-staff conversations)
- Using lexical and grammatical knowledge in a listening context (part of foundational skills)
- Inferring meaning that is not explicitly stated, and distinguishing speaker roles or context

New TOEFL 2026: (Manna et al., 2025)

- Understand common vocabulary and formulaic phrases (Listen and Choose a Response)
- Understand simple grammatical structures, including question-formation patterns (Listen and Choose a Response)
- Recognize socially appropriate responses in short spoken exchanges (Listen and Choose a Response)
- Recognize and distinguish English phonemes and the use of common intonation and stress patterns to convey meaning in carefully articulated speech (Listen and Choose a Response)
- Infer implied meaning, speaker role, or context in short spoken exchanges (Listen and Choose a Response)
- Identify the main ideas and basic context of a conversation (Listen to a Conversation)
- Understand the important details in a conversation (Listen to a Conversation)
- Understand the range of grammatical structures used by proficient speakers (Listen to a Conversation)
- Understand a wide range of vocabulary including idiomatic and colloquial expressions (Listen to a Conversation)
- Infer meaning from information that is not explicitly stated (Listen to a Conversation)
- Recognize the purpose of a speaker’s utterance (Listen to a Conversation)
- Make simple predictions about the further actions of the speakers (Listen to a Conversation)
- Follow the connection between ideas across speaker turns (Listen to a Conversation)

Appendix C – Leading Conclusions on Changes to New TOEFL 2026

Leading conclusions	**Key cause: skill/task**	**Explanation**
1. Broader social focus, narrower academic depth	Reading Writing Speaking	New tasks increased coverage of social and interpersonal topics, but at the cost of diminishing the assessment of higher-level academic skills. For example, the removal of the integrated writing and speaking task reduced opportunities to assess important academic skills such as source integration. This change alters the interpretation of these scores, especially in academic admissions contexts. For example, many B2 descriptors that were covered by the current TOEFL test no longer accurately reflect abilities measured by New TOEFL 2026 (see Section 4).
2. Input/output length restrains language sample and risk construct-underrepresentation	Reading input Listening input Writing output Speaking output	Both the new task types and those retained from the previous TOEFL now feature much shorter texts across all four skills. This change raises several concerns: - A narrower construct coverage of cognitive processing (higher-level processes are not or barely covered) - Reduced reliability and generalizability of ability estimate due to smaller language samples - Limited target domain representation. Academic contexts require engaging with much longer and more complex materials. Also required in B2+ CEFR descriptors - Shortened output for evaluation makes estimates of linguistic features such as lexical diversity highly unreliable.
3. Problematic item types **For example:**	Writing: *Build a sentence*	- Assesses writing through spoken discourse - Covers only lexical and syntactic skills; narrow construct, not aligned with communicative testing principles - Rigid dichotomous *scoring* does not capture test-taker proficiency accurately/reliably

		- Would be a nice item type for item-level adaptive testing, but TOEFL is not using it that way for writing
	Reading: *C-test*	- Covers only lexical and syntactic skills; narrow construct, not aligned with communicative testing principles - Early literature on C-tests emphasizes the need for pre-calibration (Klein-Braley & Raatz,1984). But modern psychometric theories reveal that local item dependence makes reliable calibration challenging (Eckes & Grotjahn, 2006; Baghaei, 2010) - Unreliable for distinguishing proficiency levels (e.g., Roohani, 2007)
	Speaking: *Listen and Repeat*	- Accesses short-term memory and auditory perception, does not test oral communicative ability (no idea development, no communication goals, no discourse planning) - A common task type for low-proficiency learners or children in non-high-stakes tests - Low discriminatory power at higher proficiency levels
4. Scoring ***According to ETS's technical manual, all writing and speaking responses are rated by automakers. Human raters are used for setting standards and interfere when the automarker reports issues.***	Writing Speaking	**Discrepancies** between the features rated by the scoring engines and the full rubrics for writing and speaking tasks. For example, in the interview task, the rubric includes dimensions such as task response and topic relevance, but the automated marking criteria does not mention any measures for this specifically for the automarker. In addition, all the common downsides of automarking need to be justified and validated for such a scoring method, including (but not limited to): - Higher-level ability constructs measured by surface-level and quantifiable features - Bias issues inherited from training data - Low understanding of nuanced meaning, context, pragmatic features (tend to downgrade creative or rhetorical uses of language) - Easy to game (potential manipulation by test takers)

		- Potential negative washback effects – test takers will learn how to game the system rather than how to speak/write more authentically
5. Multi-stage adaptive test muddies reliability	Listening Reading	In this context, adaptive testing offers a more reliable measure only of what remains measurable, which is not the same construct that was originally intended to be assessed (as mentioned above). In other words, while multi-stage adaptive testing (MST) enhances efficiency and measurement precision, it cannot offset the reduction in construct coverage caused by constraints in item types and scoring methods.
6. No longer warrant a concordance study because of construct differences	All skills	The substantial revisions to task type and task length in the enhanced New TOEFL 2026 results in a shift in construct coverage. As a result, direct score concordance with earlier versions of the test is unlikely to be valid, and any claims of score equivalence should be interpreted with great caution.