toefl ibt®

Scoring and Score Interpretation

A Guide for Teachers and Partners

Understanding what TOEFL iBT scores mean, how they are produced, and how to use them

Version 1, Last updated Sept. 8, 2026



Purpose and How to Use This Guide

This guide explains what the TOEFL iBT test measures, how the test is built, how responses are scored, and, most importantly, how to interpret reported scores. It includes practical guidance on raw scores vs. scale scores and explains how score reports connect to the Common European Framework of Reference for Languages (CEFR) to support teaching, learning, and decision-making.

Use the sections below to quickly find what you need:



How the test is structured and what it measures



How responses are scored, including links to public rubrics



How to read and interpret a score report, including CEFR connections



How test quality and fairness are maintained



What scores mean for teaching and learning



Answers to common questions and misconceptions



How the TOEFL iBT Is Structured

What the Test Measures

At the heart of TOEFL iBT is a clear claim: a test taker's performance reflects their ability to use English for academic and real-world purposes. Each section of the test is designed to support this claim by targeting specific language abilities (reading, listening, speaking, and writing) that are essential for success in educational contexts.

For more detailed information, please refer to the
TOEFL iBT Test Blueprint and Specifications Document.

Structure and Task Types

The TOEFL iBT test includes a range of task types across sections.

- **Reading and Listening** tasks that measure comprehension of academic and social interpersonal texts, including
  - Complete the Words
  - Read in Daily Life
  - Read an Academic Passage
  - Listen and Choose a Response
  - Listen to a Conversation
  - Listen to an Announcement
  - Listen to an Academic Talk
- **Writing** tasks that assess sentence-level construction, multi-sentence responses, and academic discourse, including
  - Build a Sentence
  - Write an Email
  - Write for an Academic Discussion
- **Speaking** tasks that capture fluency, clarity, and organization of spoken English, including
  - Listen and Repeat
  - Take an Interview

While all forms are built to ensure comparability, the Reading and Listening sections also use multi-stage adaptive designs to measure language proficiency across a range of ability levels.

Targeted Language Use and CEFR Levels

Each task is designed with a target language-use context and a range of CEFR levels in mind, from foundational (A1) to advanced (C2), depending on the section and task type. This ensures coverage across the proficiency continuum.



How Responses Are Scored

Scoring Methods

- Selected-response tasks (e.g., multiple-choice, drag and drop, fill-in) are scored using predefined rules and answer keys.
- Constructed-response tasks (such as in Speaking and Writing) are scored against predefined rubrics using advanced scoring engines designed to evaluate construct-specific skills and include features such as organization, language control, and coherence.
- Human monitoring and review are critical to the quality of scoring for all task types, including automated scoring.

This blended approach supports both efficiency and quality and ensures the reliability and validity of the resultant scores.

Scoring Guides (Rubrics)

Scoring guides, or rubrics, describe what different performance levels look like for a single constructed-response task. They are:

- Shared with educators through technical manuals and teacher materials
- Used consistently to support score comparability across test forms

Public scoring guides are available for Writing and Speaking. Select the links that follow to review the scoring guides.

Writing — Write an Email and Write for an Academic Discussion task types
Speaking — Listen and Repeat and Take an Interview task types

Review these scoring guides with students to ensure they understand what is required to achieve their target score. Use them to teach lessons that help students reach their goals.

Additional resources to support use of scoring guides with students, available on the Preparing Students webpage:

TOEFL iBT Lesson Plans

- Use the scoring guide for Write an Email while teaching Writing, Lesson 3 to help students understand what makes a successful response to the task.
- Use the scoring guide for Write for an Academic Discussion while teaching Writing, Lesson 5 to help students understand what makes a successful response to the task.
- Use the scoring guide for Listen and Repeat while teaching Speaking, Lesson 3 to help students understand what makes a successful response to the task.
- Use the scoring guide for Take an Interview while teaching Speaking, Lesson 1 to help students understand what makes a successful response to the task.

Scored Speaking Sample Responses — Use real, scored TOEFL Practice Online (TPO) speaking responses to understand score levels and support focused classroom speaking instruction.

Rubric Raw Scores Are Not Section Scale Scores

A common misinterpretation is to read a rubric raw score on a single constructed response as if it were comparable to, or directly convertible to, a reported section-level or overall scale score. These are distinct scores and should not be conflated. The two numbers live on different scales, describe different things, and should never be compared point-for-point.

	Constructed-Response Raw Score	Reported Section-Level or Overall Scale Score
What it measures	Performance on a single constructed-response task	**Section-Level:** Performance across all tasks in a single section **Overall:** Performance aggregated across all four sections of the test
Scale	Whole-number levels on the task rubric (i.e., 0–5)	1–6, in increments of 0.5
Use	Diagnostic feedback on one response	Score reporting, interpretation, and decision-making
Comparability across forms	Comparable only at the individual task level	Comparable across forms and designed to carry consistent meaning across forms and administrations

What it measures:
A rubric's raw score of 4 on a single Writing task is not equivalent to a section scale score of 4, and the two should not be averaged, mapped, or compared to one another.
↳ Always interpret and communicate results using Reported Scale Scores.

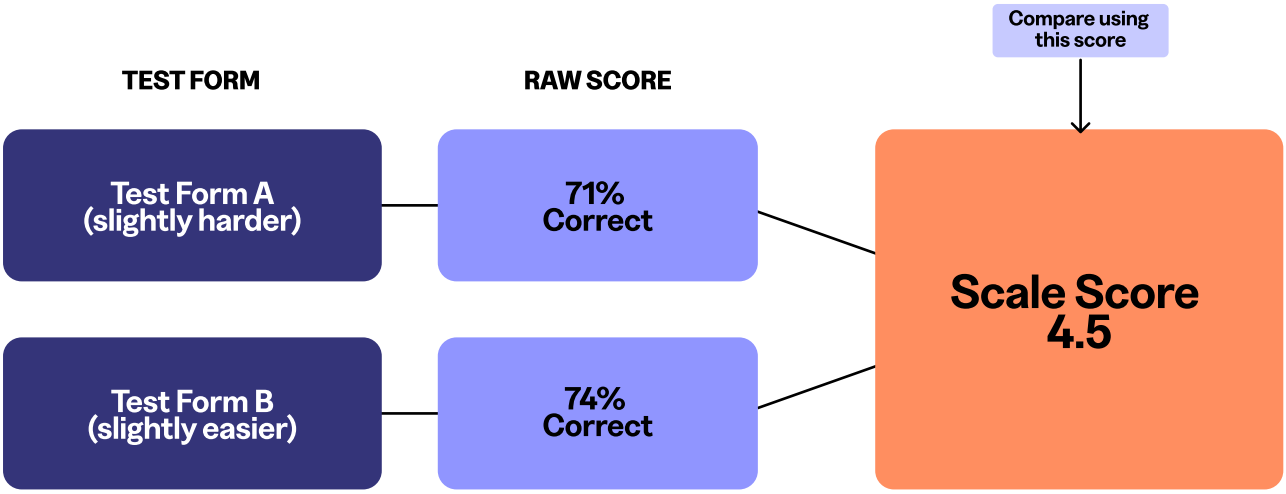
Understanding the TOEFL iBT Score Report

Raw Scores vs. Scale Scores

Raw scores are the points earned directly from performance on test tasks, such as the number of correct responses to selected-response items or rubric-based scores for constructed responses. Because different test forms (or adaptive stages) can vary slightly in difficulty, raw scores are not ideal for comparing performance across administrations. Scale scores are converted raw scores reported on a common scale so that scores carry a consistent meaning across forms and testing occasions. In general, interpretations and comparisons should be based on reported scale scores rather than raw scores.

Why Scale Score Matter

Different raw scores can represent the same level of ability when test forms dinner in difficulty.



Note: Scores shown are illustrative only and do not represent actual TOEFL iBT scale score ranges.

How to Interpret Scale Scores

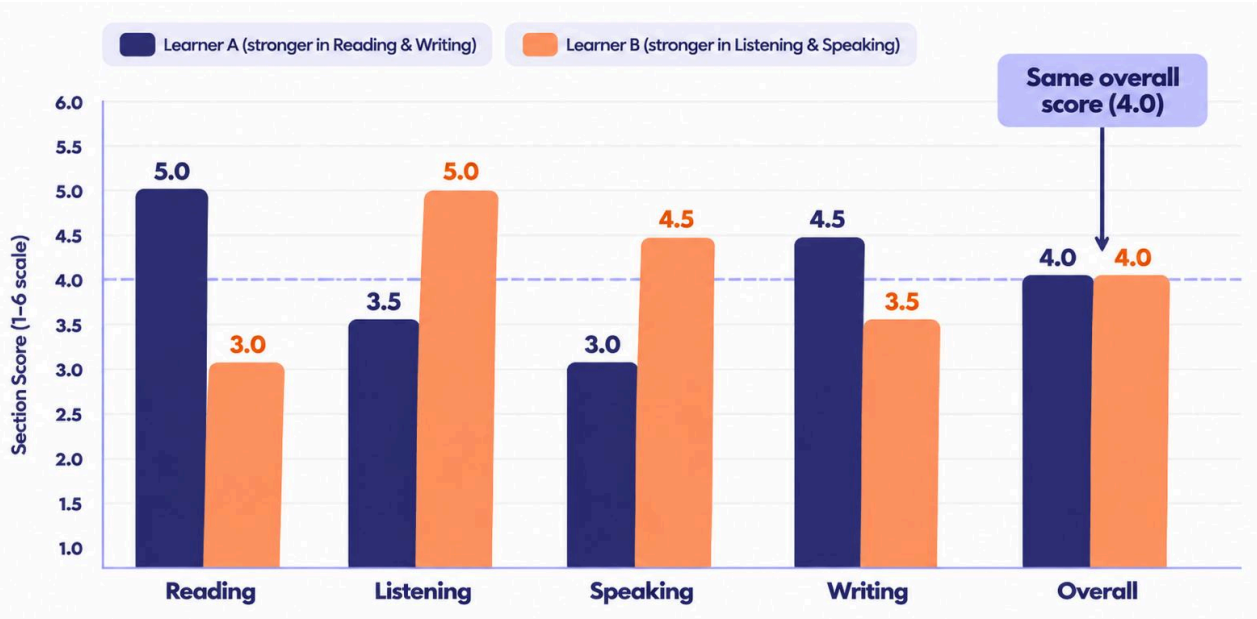
- Higher scores indicate stronger evidence of the targeted language abilities for that skill area (reading, listening, speaking, writing) in the contexts represented by the test.
- Use score ranges or performance bands when possible. Small score differences may not represent meaningful differences in real-world ability.
- Consider measurement uncertainty. Any test score is an estimate; interpret results as "about this level" rather than as a precise point.

From Scores to Instruction

- Connect scores to descriptors and examples. Use rubric language and CEFR-oriented "can do" statements (where provided) to translate numbers into actionable feedback.
- Interpret section scores first, then the overall picture. Different learners can have the same overall score with different skill profiles that imply different instructional needs.

Same Overall Score, Different Skill Profiles

Two learners with the same overall score of 4.0 may have very different strengths and instructional needs.



Note: Scores shown are illustrative only. TOEFL iBT section scores are reported on a scale of 1-6 in half-point increments

How CEFR Alignment Is Established

The TOEFL iBT scores are aligned to the Common European Framework of Reference for Languages (CEFR), an international standard used by educators and institutions worldwide.

Alignment studies examine:

- Task content and language demands
- Scoring rubrics in relation to CEFR descriptors
- Statistical relationships between test scores and CEFR levels

Score reports indicate how a test taker's performance maps to CEFR levels, helping teachers and learners understand what a score means in practical terms.

Connecting Scores to the CEFR

TOEFL iBT score reports may include a CEFR-related interpretation to help users understand what performance typically looks like at different points on the score scale. The table below provides a summary of the "can do" statements in each skill by CEFR level to help interpret scores relative to typical CEFR performance. Using these kinds of descriptors is helpful in ensuring items are written to target CEFR levels. In addition, scale-score-to-CEFR alignment is established through a combination of expert judgment and empirical evidence (for example, examining how test tasks and scoring rubrics relate to CEFR descriptors and analyzing statistical relationships between test scores and external indicators).

CEFR	Reading	Listening	Speaking	Writing
A1	Can understand very short, simple texts (signs, notices, basic messages) with familiar words and repeated phrases.	Can understand very basic words and phrases about self, family, and immediate needs when speech is slow and clear.	Can use simple phrases to introduce self, ask/answer basic questions, and interact in a very limited way with help.	Can write very short messages (forms, simple notes) using familiar words and basic sentence patterns.
A2	Can read short, simple texts and find predictable information in social interpersonal materials (emails, schedules, ads).	Can understand the main point in short, clear messages and announcements on familiar topics.	Can communicate in simple routines (asking for information, describing daily activities) with pauses and limited range.	Can write short, connected notes and simple emails about immediate needs, using basic connectors (and, but, because).
B1	Can understand texts that consist mainly of social interpersonal or school-related language; can identify main ideas and key details.	Can follow the main points of clear standard speech on familiar matters; can understand simple explanations and instructions.	Can describe experiences and plans, give simple reasons, and handle most predictable situations, though accuracy may vary.	Can write straightforward connected text (emails, short essays) on familiar topics, with generally clear organization.
B2	Can read articles and reports concerned with contemporary issues; can understand viewpoints and infer meaning from context.	Can understand extended speech and lectures and follow complex lines of argument when the topic is familiar.	Can interact with fluency and spontaneity; can present and defend views, using appropriate detail and organization.	Can write clear, detailed texts (reports, argument essays) and synthesize information with good control of tone and structure.
C1	Can understand long, complex texts, including implicit meaning and nuanced argumentation across academic genres.	Can understand extended speech even when not clearly structured; can follow specialized discussion and implied meaning.	Can express ideas fluently and precisely, adapt register, and manage complex discussion with effective discourse control.	Can write well-structured, detailed texts on complex subjects, integrating sources and maintaining a consistent style.
C2	Can read virtually all forms of written language with ease, including highly abstract or stylistically complex texts.	Can understand virtually everything heard, including fast, idiomatic speech; can interpret subtle meaning and stance.	Can participate effortlessly in any conversation; can convey fine shades of meaning precisely and adjust style strategically.	Can write complex, well-crafted texts (e.g., critical analyses) with sophisticated control of organization, nuance, and style.

Note: *These descriptors are provided as a quick, teacher-friendly reference of typical performance expectations. When TOEFL iBT score reports provide CEFR-related statements for specific sections or score ranges, those official report statements should be used for decisions and communication. For more descriptors, please refer to the TOEFL Score Scale Update.*

How to Use CEFR Levels

- Treat CEFR levels as interpretive categories, not as exact equivalents. A test score supports an inference of performance that is consistent with a CEFR level, but it does not "prove" a level in all contexts.

- Use skill-specific links. When CEFR information is provided by section, interpret reading/listening/speaking/writing separately; do not assume the same CEFR level across all skills.

- Communicate what the learner can do next. Pair CEFR labels with descriptors (e.g., academic discussion, email writing, lecture comprehension) to guide instruction and goal setting.

- Document local decisions. Institutions that use CEFR-related interpretations for placement or progression should define how they will use the information (and review outcomes over time).

Interpreting a Score Report

To interpret a learner's results:

1. Start with the section scale scores to identify strengths and needs (e.g., stronger receptive skills than productive skills).
2. Use any provided performance descriptors (rubric language, "can do" statements, or CEFR-related descriptors) to translate the numbers into what the learner can do.
3. Choose instructional next steps that match the observed profile (e.g., note-taking from lectures, synthesizing sources in writing, improving spoken organization).

How Test Quality Is Maintained

A Rigorous Development Process

Every TOEFL iBT test item goes through a multi-step quality assurance process, including:

- **Design and authoring** by trained assessment specialists
- **Alignment checks** to test specifications, claims, evidence, and CEFR targets
- **Multiple content and fairness reviews** to ensure accessibility, fairness, and appropriateness
- **Pretesting and data review** to evaluate how items function with real test takers
- **Ongoing monitoring** after administration to confirm scoring and performance remain stable

Human expertise plays a central role throughout development, even as technology is used to support efficiency and consistency.

Evidence and Validity

Test tasks are selected and designed to elicit observable evidence of language use, such as understanding written texts, following spoken information, organizing ideas in writing, or communicating meaning orally. Scores are based on patterns of performance across multiple tasks rather than a single question, which strengthens the quality of the evidence.

Validity means that the test results are appropriate for their intended interpretations and uses. For TOEFL iBT, validity evidence comes from:

- Careful alignment between tasks and real-world language use
- Consistent scoring processes
- Ongoing research and review to ensure scores reflect meaningful differences in proficiency

Validity is treated as a continuous process, not a one-time check, and we monitor validity throughout the development, scoring, and administration each time and over time.

For more detailed information, please refer to the TOEFL iBT Technical Manual.

Why Validity Matters

This careful development and ongoing validity research is why test takers can trust that their scores are fair, and why institutions can rely on those scores to make sound enrollment and placement decisions.



What This Means for Teaching and Learning

For Teachers

TOEFL iBT provides:

- Transparent links between classroom instruction and assessment
- Scores that can be interpreted using CEFR-based descriptions of ability
- Rubrics and examples that support teaching and feedback

For Learners and the Public

TOEFL iBT offers:

- Clear explanations of what scores represent
- Confidence that results are based on rigorous, research-based practices
- A fair and meaningful measure of English proficiency



Common Questions and Misconceptions

Teachers and partners often have questions about how TOEFL iBT scoring works. Below are answers to the most common ones.

"Can I compare raw scores from different test dates?"

No. Because different test forms (or adaptive stages) can vary slightly in difficulty, raw scores are not ideal for comparing performance across administrations. Scale scores represent the converted raw scores onto a common scale so that scores have a consistent meaning across forms and testing occasions. Always interpret and compare results using the reported scale scores, not raw scores or points.

"Does a small score difference mean one student is stronger than another?"

Not necessarily. Any test score is an estimate that has some measurement error around it. It is best to use score ranges or performance bands when possible, because small score differences may not represent meaningful differences in real-world ability. Interpret results as "about this level" rather than as a precise point.

"Is the overall score just an average of the four section scores?"

The overall score is an average of the four section scores, but that doesn't mean every overall score represents the exact same mix of skill-level performance. When interpreting overall performance, keep in mind that different learners can have the same overall score with very different skill profiles. That is why it is important to interpret section scores first, then consider the overall picture. Two learners with the same overall score may have quite different instructional needs depending on their relative strengths and weaknesses across reading, listening, speaking, and writing.

"Can TOEFL iBT reported scores have any decimal value, such as 4.2?"

No. Overall and section-level TOEFL iBT scores are reported only in the score increments used in the score scale (1–6 in increments of 0.5). They are not reported as any possible decimal value. Teachers should use the scores exactly as they appear on the score report. Reported scale scores are designed to support consistent interpretation across test forms and administrations; they should not be recalculated, averaged into unsupported decimal values, or treated as more precise than the score report indicates.

"Can constructed-response tasks receive half-point rubric scores, such as 1.5, 2.5, 3.5, or 4.5?"

No. The scoring rubrics for individual constructed-response tasks use whole-number score levels. A response is evaluated holistically, meaning the scorer considers the response as a whole and assigns the score level that best fits the overall performance.

Half-point scores may appear in reported section or overall scale scores, but those are different from rubric scores. Reported scale scores are converted scores used for score reporting and interpretation; they should not be read as half-point rubric scores.

"Where can I find the scoring rubrics?"

The scoring guides for TOEFL iBT are publicly available. The scoring guides for the Write an Email and Write for an Academic Discussion task types are available in the Writing Rubrics document. The scoring guides for the Listen and Repeat and Take an Interview task types are available in the Speaking Rubrics document. These rubrics describe what different score levels look like and are directly linked to CEFR descriptors.

"Are Speaking and Writing responses scored by humans or by computers?"

Both—automated scoring is paired with human monitoring and review. Constructed-response tasks (such as in Speaking and Writing) are scored using advanced scoring engines designed to evaluate construct-specific skills, including features such as organization, language control, and coherence. Human monitoring and review are critical to the quality of scoring for all task types, including responses that leverage automated scoring. This blended approach supports both efficiency and quality.

"What do the CEFR levels on the score report mean?"

CEFR levels are interpretive categories that describe typical performance and provide a helpful way to bring meaning to the numeric scale score. Note that when TOEFL iBT score reports provide CEFR-related statements for specific sections or score ranges, those official report statements should be used for decisions and communication.

“If a learner scores at B2 in Reading, does that mean they are B2 in all skills?”

No. When CEFR information is provided by section, interpret reading, listening, speaking, and writing separately. Do not assume the same CEFR level across all skills. Different learners may show different profiles, and using skill-specific interpretations provides more useful guidance for instruction.

“How do I know the test is fair and accurate?”

Validity means that the test results are appropriate for their intended interpretations and uses. For TOEFL iBT, validity evidence comes from (1) careful alignment between the claims of the assessment, the tasks developed, and real-world language use; (2) consistent scoring processes; (3) psychometric analyses; and (4) ongoing research and review to ensure scores reflect meaningful differences in proficiency. Every test item goes through a multi-step quality assurance process, including multiple content and fairness reviews, and ongoing monitoring. Validity is treated as a continuous process, not a one-time check, and we monitor validity throughout the development, scoring, and administration each time and over time.

This guide is grounded in official TOEFL iBT test specifications, technical manuals, and internal validity and scoring documentation.