A banded score scale for the TOEFL iBT test: Rationale, development, and score linking

April 2026

Spiros Papageorgiou, Shuhong Li, Lixiong Gu, & Karen Barton

This report provides information about the development of the TOEFL iBT score scale introduced in 2026, and score mappings intended to help support score-based decisions. We first describe the rationale behind the introduction of a banded score scale (1–6, in 0.5 increments), intended to support proficiency-focused score reporting aligned with the Common European Framework of Reference (CEFR) levels (Council of Europe, 2001, 2020). We then describe the procedures followed to develop mapping tables between the banded score scale and the original TOEFL iBT score scale, mapping tables between the banded score scale and the CEFR levels, and score comparison tables between the TOEFL iBT banded score scale and IELTS Academic band scores. Limitations and plans for ongoing monitoring using operational data are discussed.

Key words: TOEFL iBT, language proficiency levels, score scale, score mapping

Updates to the TOEFL iBT test and rationale behind the development of a new score scale

The primary motivation for administering a test is to use its scores to make decisions. These score-based decisions can have important consequences for individuals, institutions, and society overall. In the case of the TOEFL iBT test, scores are primarily used to inform high-stakes decisions about the language proficiency level of international students who seek to pursue higher education in institutions where English is the language of instruction.

Since its launch in 1964, the TOEFL test has undergone several revisions motivated by advances in theories of language ability and changes in English teaching practices (Taylor & Angelis, 2008). The most recent iteration of the TOEFL test, the TOEFL iBT test, was launched in 2005 with the goal of evaluating language proficiency through test tasks that reflect the reading, listening, speaking, and writing demands of real-world academic environments. The TOEFL iBT test was subsequently updated in 2019 and in 2023. These updates aimed to improve the test-taking experience by keeping the same overall design and reducing overall testing time without compromising measurement quality (for details on the updates, see ETS, 2025; Gu et al., 2023). Through those updates, the reported score scale for the test (0-30 for each of the four test sections for Reading, Listening, Writing, and Speaking, and 0-120 for the summed total score) remained unchanged.

As the TOEFL iBT evolved, additional updates to the test content were introduced in January 2026. These updates aimed to maintain the purpose, targeted population and score use of the TOEFL iBT test, while introducing new task types to increase the ways and opportunities for test takers to show what they know. The updates also aimed to operationalize the same underlying construct of academic language ability. For example, the Write for Academic Discussion task (Davis & Norris, 2023; Hsieh & Ohta, in press) remained unchanged. In addition, similar reading and listening skills are tested, such as

identifying the main idea and understanding the important details in academic passages, and listening to classroom announcements and talks (Manna et al., 2025). A new banded and CEFR-aligned score scale was also introduced, ranging from bands 1 to 6, in increments of 0.5, for the four test sections and the total score.

The 2026 updates were intended to increase the variety and volume of evidence available to support interpretations about test taker's English language abilities, through the introduction of new task types and multi-stage testing (MST). With additional numbers and types of items and tasks, as Manna et al. (2025) explain, the 2026 updates aimed to provide efficient measurement of both foundational aspects of language proficiency (lexical and grammatical competence) as well as the ability of language learners to communicate in English through a range of language knowledge activities and communicative language tasks. These activities and tasks aim to provide test takers with brief but informative opportunities to demonstrate their language skills, and with a banded score scale, as explained next, for score users to more easily interpret test taker performance in terms of the CEFR levels.

The CEFR is the de facto standard for setting language policy and proficiency goals around the world, and virtually all major international, as well as many local test providers present score information in relation to its language proficiency levels (Deygers et al., 2018; Fulcher, 2016; Papageorgiou, 2022; Read, 2019). The updated TOEFL iBT test intends to measure aspects of English language proficiency described in the CEFR levels, from A1 to C2. The same range of scores (1-6) is reported for the four test sections and the overall score (which is the rounded average of the section scores). In addition, the same numeric score value across the test sections and the total score is used for a given CEFR level. For example, a score of 4 aligns with CEFR level B2 for Reading, Listening, Writing, Speaking, as well as the overall score (see also discussion of CEFR thresholds later in this section). This consistent score mapping is intended to offer more intuitive score interpretation in relation

to the CEFR levels, compared to the original score scale. In the original TOEFL iBT scale of 0-30 for section scores and 0-120 for total scores, the minimum score for CEFR level B2 varied across test sections (18 for Reading, 17 for Listening, 17 for Writing, and 20 for Speaking, see Papageorgiou et al., 2015). By contrast, the banded score scale uses consistent thresholds, such that a score of 4 represents CEFR level B2 for all test sections. In addition, because the overall score is the average, rather than the sum of the four section scores as in the original scale, it is possible to use the same CEFR thresholds for the overall score. To sum up, the CEFR levels were incorporated into the development of the banded score scale, with the intention to use the same numeric score for each of the six CEFR levels (A1-C2) across the four test section scores, as well as the overall score.

The banded score scale is less granular than the original scale it replaced. A more granular scale might be helpful when the goal is to rank students by language ability. However, the primary use of TOEFL iBT test scores is to inform "yes/no" types of decisions about admission of international students to a degree program or placement into a language class, whereby a single language proficiency cut score is determined and considered by each program. For example, university admissions criteria for international students typically include minimum score requirements, rather than credit for higher-scoring applicants, suggesting that the primary use of the scores is to determine whether a student meets a desired language proficiency threshold (see discussion of university requirements in Papageorgiou et al., 2015). A band level score scale can inform such yes/no decisions, while at the same time it can offer a more practical approach to score interpretation than the original score scale, through improved consistency across test sections and alignment to a widely used language proficiency framework.

Given the relative ease of interpretability and use, the band scale approach to score reporting has long been used in language assessments, including the FSI scale developed by the U.S. Foreign Service Institute for oral assessment, the ACTFL Oral Proficiency

Interview (OPI) administered by the American Council on the Teaching of Foreign Languages, the Eiken test administered by the Eiken Foundation of Japan, and the IELTS test administered jointly by the IELTS partners (for a review of such scales, see North & Schneider, 1998). The popularity of score bands for reporting test results reflects their ease of use when the assessment goal is to certify that a candidate meets a particular standard. Additionally, when accompanied by descriptors of the expected skills and abilities for each band or level, such score levels simplify interpretation of test taker performance. For the TOEFL iBT test, performance descriptors (Basic, Low-Intermediate, High-Intermediate, and Advanced), corresponding to CEFR levels A2, B1, B2 and C1 respectively, were developed previously for the four test sections to facilitate score interpretation of the original TOEFL iBT score scale (Wang & Papageorgiou, 2023). The availability of the performance descriptors in four levels effectively provided users with score bands aligned to the CEFR levels, despite the more granular 30-point scale used for the four test sections.

To support score interpretation for the banded score scale and the decisions made from the TOEFL iBT test, performance descriptors were also created and made available on the TOEFL website (see Appendix A). These descriptors have been selected from the CEFR (Council of Europe, 2001, 2020) with minor modifications so that they are more relevant to test content. The procedures for mapping the banded scores on the CEFR proficiency levels and the scores of the original scale are described next.

Alignment of the TOEFL iBT banded score scale to the original score scale and the CEFR levels

CEFR score mapping was integrated into the development of the banded score scale, as opposed to treating it as a subsequent step, to support the interpretation of the test scores in relation to the CEFR levels for students, score users, and other stakeholders. This is in contrast to the original TOEFL iBT scale, which was established before the CEFR was widely adopted around the world. Additionally, score users, accustomed to making decisions

using the original score scale, needed to understand the alignment of the banded scores to the original scale. In this section we describe the parallel efforts to map the banded scale scores to the original TOEFL iBT score scale (0–30 per section, 0–120 total) and the CEFR levels, achieved through a combination of linking and score mapping methodologies. Because the task types and number of items available in each test section varies, we describe the methodology separately for test sections evaluating receptive language skills (Reading and Listening), and test sections evaluating productive language skills (Writing and Speaking).

Reading and Listening tasks were administered in a field test to over 5,000 test takers, as part of the development of the TOEFL Essentials test (Papageorgiou et al., 2021). The participants were recruited from over 80 countries in Asia, Europe, Africa, and the Americas, and represented a diverse range of first languages as shown in Table 1. Teachers of students who participated in the field trials were asked to provide an overall estimate of each of their students' CEFR level (A1-C2) as an approximate indicator of the participants' proficiency level. Having such estimates of the students' language ability were used during data collection for monitoring the proficiency distribution of test takers, to help ensure the sample included individuals across a broad range of language proficiency. Table 2 presents the teachers' estimates in the three broad categories of Basic (A1-A2), Independent (B1-B2), and Proficient (C1-C2) language user (see Council of Europe, 2001, p. 23). It should be noted that test takers at lower proficiency levels, although not the primary target population, were included to help collect data for operationalizing the MST design of the updated test.

Table 1. Top ten first language of the field test participants

First Language	Number of participants	Percent
Spanish	1062	18.49
South Asian Languages*	1074	18.70
Chinese	914	15.91
Portuguese	633	11.02
Korean	604	10.52
Japanese	517	9.00
Turkish	153	2.66
Arabic	145	2.52
German	70	1.22
French	39	0.68
Other**	533	9.28
Total	5744	100.00

*South Asian languages include major native languages in South Asian countries, including Hindi, Telugu, Gujarati, etc.

**Other first languages include Vietnamese, Russian, etc.

Table 2. Teachers' estimated CEFR levels of the field test participants

CEFR Level	Participants by CEFR Level	
	Number	Percent
C1-C2	1830	31.86
B1-B2	3114	54.21
A1-A2	799	13.91
Unknown	1	0.02
Total	5744	100

Each test taker responded to reading and listening tasks typical of the prior versions of the TOEFL iBT test, that is, the version using the original score scale, along with the task types included in the test version using the banded score scale. By collecting responses from participants who completed tasks across both test versions, the study employed a non-equivalent groups with anchor test (NEAT) design (previously employed in a TOEFL vertical linking project, see Papageorgiou et al., 2023) to establish a robust, empirically grounded link between the underlying psychometric scales for the reading and listening items in the two versions of the TOEFL iBT test.

Twelve parallel field test forms were developed using a balanced incomplete block (BIB) design so that enough test questions (items) could be field tested, and so that the parallel forms could be linked psychometrically. Each block consisted of items of the same type and of similar difficulty levels (e.g., low, medium, high). Each form included multiple blocks from all item types, with individual blocks repeating across forms in differing combinations. In this manner each test taker completed all item types, while keeping the test length manageable. This design also provided sufficient numbers of common anchor items across the twelve test forms to enable precise estimation and linking of item parameters, estimated via item response theory (IRT), with the added benefit of creating a solid foundation for mapping the original scale to the banded score scale.

The IRT-based analyses allow for estimation of item parameters (such as difficulty) and test taker proficiency levels on a common underlying measurement scale. The original psychometric scale underlying the TOEFL iBT scale was established through field trials in 2003-2004 (Wang et al., 2008) and updated with operational data. The anchor items in the reading and listening field test (see Papageorgiou et al., 2021) were used to place all items back onto the original measurement scale via the IRT test characteristic curve (TCC) linking method (Stocking & Lord, 1983). This is a mathematical procedure where anchor items are used to compute transformation constants (slope A; intercept B), which are then used to transform item parameters of new items so that TCCs of new items align to the TCC of the items on the existing scale. The resulting transformation constants for placing the field-tested item parameters onto the original scale are found in Table 3. Visually, the TCCs demonstrate strong alignment of anchor items between the transformed item parameters and those on the original (reference) scale. As the TCCs in Figures 1 and 2 indicate, the anchor items were estimated to be slightly more difficult in the field test data compared to those on the reference scale for both Reading and Listening.

Table 3. Transformation constants to place field test item parameters onto TOEFL iBT scale

	A	B
Reading	1.025981	-0.442808
Listening	0.954680	-0.490414

Figure 1. TCC plots for Reading anchor items (transformed vs reference)

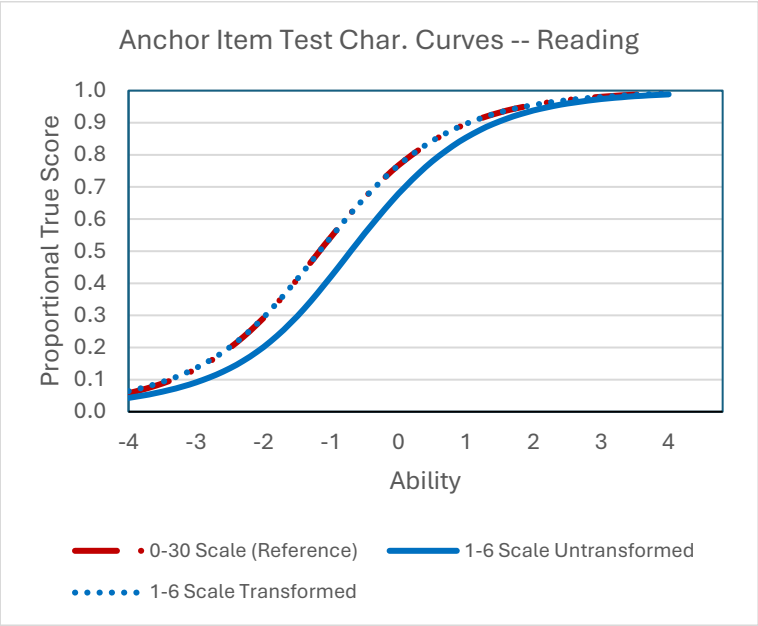
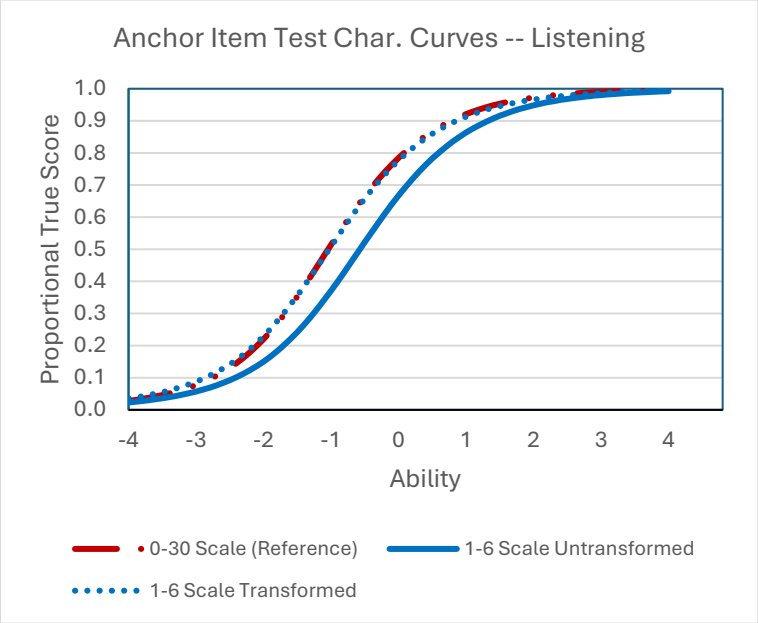


Figure 2. TCC plots for Listening anchor items (transformed vs reference)



Once the items were equated to the underlying TOEFL iBT scale, a point-to-point correspondence between the 0-30 section scores and the 1-6 section scores was conducted, which also allowed for mapping the banded scale scores to the CEFR. Specifically, the minimum ability level (i.e. the latent ability represented by the IRT theta value) for a given CEFR level was estimated for the Reading and Listening scores from the original scale and used as the lower boundary of score band corresponding to that CEFR level. Tables 4 and 5 present the theta value for each CEFR level, along with the corresponding 0-30 TOEFL iBT section score and the 1-6 TOEFL iBT band score for Reading and Listening respectively.

Table 4. CEFR mapping for the TOEFL iBT 0-30 and 1-6 Reading score scales

		TOEFL iBT Scaled Score	
CEFR Level	Theta	Original score scale (0-30)	Banded score scale (1-6)
C2	2.22	29	6
C1	1.25	27	5.5
	0.52	24	5
B2	0.11	22	4.5
	-0.41	18	4
B1	-1.10	12	3.5
	-1.85	6	3
A2	-1.90	4	2.5
	-2.16	3	2
A1	-2.10	2	1.5
	-4.00	0	1

Table 5. CEFR mapping for the TOEFL iBT 0-30 and 1-6 Listening score scales

		TOEFL iBT Scaled Score	
CEFR Level	Theta	Original score scale (0-30)	Banded score scale (1-6)
C2	1.53	28	6
C1	0.90	26	5.5
	0.08	22	5
B2	-0.17	20	4.5
	-0.65	17	4
B1	-1.16	13	3.5
	-1.52	9	3
A2	-1.79	6	2.5
	-2.07	4	2
A1	-2.40	2	1.5
	-4.20	0	1

The mapping of the test scores in the TOEFL Writing and Speaking sections was established by combining information from several separate steps that included an evaluation of the alignment between test tasks and CEFR descriptors, linking of the original TOEFL iBT score scale to the CEFR levels (Papageorgiou et al., 2015), an additional standard setting study for the updated test tasks (described in detail in Davis et al., 2023), and an equipercentile linking study. First, as described in Davis et al. (2023), to ensure the

content of the updated Writing and Speaking sections was relevant to language ability as described in the CEFR, the writing and speaking task requirements and scoring rubrics were compared to various CEFR scales and level descriptors. This systematic comparison was conducted by one senior research scientist and one senior assessment developer, who were heavily involved in the development of the original test tasks. This step helped ensure that the alignment of test scores to CEFR levels was justified from a content perspective.

After the alignment evaluation, an ETS-internal standard setting study was conducted on the Writing and Speaking test forms assembled from the field test tasks (Davis et al., 2023). The study employed a performance profile method to identify minimum raw scores (sum of task scores) for each CEFR level. In this study, test takers with response profiles across different levels of performance were selected (based on the total score across the Writing or Speaking test tasks), after which a portfolio was constructed for each individual containing the test taker's written or spoken responses. Language experts then compared the portfolios to the CEFR level descriptors to establish the minimum speaking or writing score for each CEFR level.

The field test forms used in the standard setting study (Davis et al., 2023) were similar to the TOEFL iBT Writing and Speaking base forms created for the 2026 version of the test except for small differences in the number of items, a result of the standard setting data being collected in the development of the TOEFL Essentials test. The base forms were the reference test forms on which the newly developed banded scale was established. Equipercentile linking analyses were conducted to convert the raw scores of the 2026 TOEFL iBT Writing and Speaking base forms to those of the forms used in the standard setting study (Davis et al., 2023).

The equipercentile linking established the mapping between the minimum raw scores on the 2026 base forms and the corresponding CEFR levels, because the forms used in the standard setting by Davis et al. (2023) had already been mapped to the CEFR levels. The linking was performed using data from approximately 500 Writing test takers and 700 Speaking test takers, a subset of the field test participants as described in Davis et al. (2023), who completed all tasks included in the respective TOEFL iBT Writing and Speaking field test forms.

Once the Writing and Speaking banded scores were aligned to the CEFR levels based on the steps described above, it was possible to align band scores to the original TOEFL iBT score scale using the minimum scores for each CEFR level on the original scale. (For the mapping of the original score scale to the CEFR levels, see Papageorgiou et al., 2015). Consequently, a mapping between the original 0-30 scores and the 1-6 banded scores for the Writing and Speaking sections was created based on the separate CEFR mapping for the each of the two score scales.

Resultant banded TOEFL iBT scales

The score mapping from the original score scale to the banded score scale is provided in Table 6, and the mapping of the banded score scale to the CEFR levels is provided in Table 7. As mentioned in a previous section, the 1-6 overall score is calculated as the average of the four section scores—Reading, Listening, Writing, and Speaking, as opposed to a summed total score as for the original score scale. This average is then rounded to the nearest half-band increment. Note that to ensure comparability between the 0–120 total score and the 1-6 overall band score, the minimum section scores (0–30) corresponding to a given 1-6 band score were summed across all four sections to produce the total 0-120 score in the last column of Table 6. Descriptive statistics (mean and standard deviation (SD) based on 2024 TOEFL iBT test takers are provided for the original and banded scales in Table 8. For both scales, the Reading and Listening sections show relatively larger means

and standard deviations, whereas the Writing and Speaking sections show smaller values. The section scores have similar order of difficulty and spread across the two score scales.

Table 6. Score mapping of the TOEFL iBT banded score scale to the original TOEFL iBT scale

Overall and Section Band Scores (1-6)	Reading (0-30)	Listening (0-30)	Writing (0-30)	Speaking (0-30)	Total (0-120)
6	29-30	28-30	29-30	28-30	114+
5.5	27-28	26-27	27-28	27	107+
5	24-26	22-25	24-26	25-26	95+
4.5	22-23	20-21	21-23	23-24	86+
4	18-21	17-19	17-20	20-22	72+
3.5	12-17	13-16	15-16	18-19	58+
3	6-11	9-12	13-14	16-17	44+
2.5	4-5	6-8	11-12	13-15	34+
2	3	4-5	7-10	10-12	24+
1.5	2	2-3	3-6	5-9	12+
1	0-1	0-1	0-2	0-4	0+

Table 7. Mapping of the TOEFL iBT banded score scale to CEFR Llevels

CEFR Level	Reading	Listening	Writing	Speaking	Overall
C2	6	6	6	6	6
C1	5 - 5.5	5 - 5.5	5 - 5.5	5 - 5.5	5 - 5.5
B2	4 - 4.5	4 - 4.5	4 - 4.5	4 - 4.5	4 - 4.5
B1	3 - 3.5	3 - 3.5	3 - 3.5	3 - 3.5	3 - 3.5
A2	2 – 2.5	2 – 2.5	2 – 2.5	2 – 2.5	2 – 2.5
A1	1 – 1.5	1 – 1.5	1 – 1.5	1 – 1.5	1 – 1.5

Table 8. Average and standard deviation of TOEFL iBT scores on the banded and original scales

Section	TOEFL iBT banded Scale (1-6)		TOEFL iBT original Scale (0-30/120)	
	Mean	SD	Mean	SD
Reading	4.59	1.01	21.93	6.50
Listening	4.77	1.02	22.05	6.16
Writing	4.37	0.82	21.09	4.63
Speaking	3.96	0.99	20.66	4.61
Total	4.40	0.81	85.74	18.62

In addition, score comparison tables between the TOEFL iBT and IELTS tests were developed, to help score users set comparable score requirements across the two tests (see Appendix B for a summary).

Conclusion

In this report, we presented the rationale behind the introduction of a reported score scale in the form of band levels (1–6, in 0.5 increments), as part of updates to the TOEFL iBT test in January 2026. We then described the procedures followed to develop mapping tables between the banded score scale and the original TOEFL iBT score scale (0-30 for each of the four test sections, and 0-120 for the summed total score), mapping tables between the banded score scale and the CEFR levels, and score comparison tables between the TOEFL iBT banded score scale and IELTS Academic band scores.

Given that data collection took place prior to the launch of the updated TOEFL iBT test, we acknowledge some anticipated limitations. First, the field test taking population might differ to some extent from the operational test population. In addition, mapping of the banded Reading and Listening scores to the CEFR levels was indirect through linking to the original TOEFL iBT score scale. Similarly, mapping of the banded Speaking and Writing scores to the original (0-30) score scale for Speaking and Writing was indirect through separate CEFR score mapping studies. As operational data become available, ongoing analyses will examine the stability of the banded score scale, and evaluate the expected performance of items and tasks. It will also be useful to collect additional validity evidence, for example by comparing the banded scores with other, CEFR-aligned measures.

References

Council of Europe. (2001). *The Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages/

Davis, L., Garcia Gomez, P., Li, S., Manna, V. (2023). Mapping TOEFL Essentials speaking and writing scores to the CEFR levels. In S. Papageorgiou & V. Manna (Eds.), *Meaningful language test scores: Research to enhance score interpretation* (pp. 120-140). John Benjamins. https://doi.org/10.1075/illa.1.07dav

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

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly, 15*(1), 3–15. https://doi.org/10.1080/15434303.2016.1261350

ETS. (n.d.). *Compare TOEFL iBT scores*. ETS. https://www.ets.org/toefl/institutions/ibt/compare-scores.html

ETS. (2025). *TOEFL iBT® test framework and test development. TOEFL Research Insight Volume 1*. ETS. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v1.pdf

Fulcher, G. (2016). Standards and frameworks. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment* (pp. 29-44). De Gruyter Mouton.

Gu, L., Li, S., Li, T., & Norris, J. M. (2023). *Maintaining score quality on the enhanced TOEFL iBT® test* (Research Memorandum No. RM-23-05). ETS. https://www.ets.org/Media/Research/pdf/RM-23-05.pdf

Hsieh, C.-N., & Ohta, R. (in press). *Examining L2 writing processes in the TOEFL iBT Write for an Academic Discussion task using eye-tracking, keystroke logging, and stimulated recall*. ETS Research Report Series. ETS

IELTS. (n.d.). *IELTS and the CEFR*. IELTS. https://ielts.org/organisations/ielts-for-organisations/compare-ielts/ielts-and-the-cefr

Ikeda, N., Clark, T., Papageorgiou, S., Gu, L., Ohta, R., Blackhurst, A., & Bruce, E. (2025a). *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, and Cambridge University Press & Assessment. https://ielts.org/researchers/our-research/research-reports

Ikeda, N., Clark, T., Papageorgiou, S., Gu., L., Ohta, R., Blackhurst, A., & Bruce, E. (2025b). *Aligning scores of language proficiency tests: A score concordance study between IELTS Academic and TOEFL iBT* (TOEFL Research Report No. RR-105). ETS. https://www.ets.org/Media/Research/pdf/RR-25-02

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement, 25*(2), 97–110. https://doi.org/10.2307/1434746
North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-262. https://doi.org/10.1177/026553229801500204

Manna, V. F., Li, S., Papageorgiou, S., & Gu, L. (2025). *TOEFL iBT® technical manual* (TOEFL Research Report No. RR-106). ETS. https://doi.org/10.64634/eje8f497

Papageorgiou, S. (2022). Still deluded by artifices? The role of the Common European Framework of Reference in facilitating test score interpretation. *Language Teaching Research Quarterly, 29*(1), 57-64. https://doi.org/10.32038/ltrq.2022.29.04

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the* TOEFL® Essentials™ *test 2021* (Research Memorandum No. RM-21-03). ETS. https://www.ets.org/content/dam/ets-org/pdfs/toefl/RM-21-03.pdf

Papageorgiou, S., Ginsburgh, M., & Garcia Gomez, P. (2023). Assessment design issues in developing vertical scales for language tests. In S. Papageorgiou & V. F. Manna (Eds.), *Meaningful language test scores: Research to enhance score interpretation* (pp. 35-60). John Benjamins.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (ETS Research Memorandum RM-15-06). ETS. https://www.ets.org/Media/Research/pdf/RM-15-06.pdf

Read, J. (2019). The influence of the Common European Framework of Reference (CEFR) in the Asia-Pacific region. *LEARN Journal: Language Education and Acquisition Research Network, 12*(1), 12–18. https://files.eric.ed.gov/fulltext/EJ1225686.pdf

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in Item Response Theory. *Applied Psychological Measurement*, 7(2), 201–210.

Taylor, C., & Angelis, P. (2008). The evolution of TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 27–54). Routledge.

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis: TOEFL iBT and CBT. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 259–318). Routledge.

Wang, L., & Papageorgiou, S. (2023). Scale anchoring methodology for developing revised performance level descriptors for the TOEFL iBT test. In S. Papageorgiou & V. F. Manna (Eds.) *Meaningful language test scores: Research to enhance score interpretation* (pp. 80-98). John Benjamins.

Appendix A

TOEFL iBT performance descriptors

CEFR	Section Score	Listening	Reading	Writing	Speaking
C2	6	Can make appropriate inferences when links or implications are not made explicit in a listening sample.	Can understand a broad range of long and complex texts, appreciating subtle distinctions of style and identifying both implicit and explicit meaning.	Can use a comprehensive & unrestricted mastery of a wide range of language to formulate thoughts precisely, give emphasis, and eliminate ambiguity. Can convey finer shades of meaning by using, with reasonable accuracy, a wide range of qualifying devices (e.g., adverbs showing degree or clauses showing limitations). Has strong command of a broad lexical repertoire, including idiomatic expressions. Shows awareness of connotative levels of meaning.	Can express themselves at length with a natural, effortless, smooth flow. Pauses only to reflect on precisely the right words or find an appropriate example. Can use a full and reliable mastery of a broad range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No signs of having to restrict what they want to say. Can provide clear, smoothly flowing, elaborate and often memorable descriptions.
C1	5 – 5.5	Skilled at using contextual, grammatical and lexical cues to infer attitude, mood & intentions and anticipate what's next. Can recognize a wide range of idiomatic expressions, but may need to confirm occasional details, especially if the accent is unfamiliar. Can follow most lectures and discussions with ease.	Skilled at using contextual, grammatical and lexical cues to infer attitude, mood & intentions and anticipate what's next. Can understand lengthy, complex texts from social, professional or academic life, identifying finer points of detail, including attitudes & implied and stated opinions.	Can use a broad range of complex grammatical structures appropriately and with flexibility. Layout, paragraphing and punctuation are consistent and helpful. Can precisely qualify opinions in relation to degrees of certainty / uncertainty. Can express strong disagreement diplomatically.	Can express themselves fluently and spontaneously, almost effortlessly. Only a difficult subject can hinder a smooth flow of language. Can use a broad range of complex grammatical structures and less common vocabulary appropriately. Can use the full range of phonological features in the target language with sufficient control to ensure intelligibility.
B2	4 – 4.5	Can follow extended discourse and complex lines of argument, provided	Can use several strategies, including identifying main points	Can produce text that is well-organized and coherent, using a range	Can produce stretches of language with an even tempo, but can be hesitant while searching for patterns

		the topic is reasonably familiar and the argument is guided by explicit markers. Can distinguish main themes from asides, as long as the lecture is delivered in standard or familiar language. Can recognize the point of view expressed and distinguishes this from facts being reported. Can identify the main reasons for and against an argument or idea in a discussion conducted in clear standard or familiar language.	and using context clues. Can adapt style and speed of reading to different texts and using appropriate reference sources selectively. Has a broad active vocabulary, but may have some difficulty with uncommon idioms. Can understand articles concerned with contemporary problems in which specific viewpoints are adopted. Can recognize when a text provides factual information and when it makes an argument.	of linking expressions and devices. Has a good command of simple language structures and some complex grammatical forms, although tends to use complex structures rigidly with some inaccuracy. Can develop a clear description or narrative, expanding and supporting main points with relevant supporting detail and examples. Lexical accuracy is generally high; occasional incorrect word choice doesn't hinder communication.	and expressions. Few noticeably long pauses. Shows a fairly high degree of grammatical control. Does not make mistakes that lead to misunderstanding; intelligible throughout, despite a few systematic mispronunciations. Can explain a viewpoint on a topical issue, giving the pros and cons of various options. Can develop a clear description or narrative, expanding and supporting main points with relevant detail and examples.
B1	3 – 3.5	Can understand the main points made in clear standard language on familiar matters. Can extrapolate the meaning of unknown words from the context and deduces sentence meaning, if the topic is familiar. Can understand clear factual information about common topics, identifying general messages and specific details, if speaker articulates clearly in a familiar manner. Can follow a lecture within their own field, if the subject matter is familiar and clearly structured. Can follow much of everyday conversation, if	Can follow a line of argumentation or the sequence of events in a story by focusing on common logical and temporal connectors. Can deduce the probable meaning of unknown words by identifying their parts (e.g., roots, lexical elements, suffixes and prefixes). Can understand descriptions of places, events, explicitly expressed feelings and perspectives in narratives, guides and magazine articles that use everyday language. Can find and understand relevant information in everyday material (letters, brochures and short official documents).	Can write basic emails/ letters of a factual nature (e.g., to request information or ask for and give confirmation). Can use a wide range of simple language to flexibly express most intended meanings. Can express the main point comprehensibly. Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.	Can speak comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production. Can express a main point comprehensibly. Is generally intelligible, despite regular mispronunciation of individual sounds and less familiar words. Can give brief explanations for opinions, plans and actions. Has sufficient vocabulary to express themselves with some circumlocutions on most topics relevant to everyday life, like family, hobbies and interests, work, travel and current events.

		clearly articulated in standard language.			
A2	2 – 2.5	Can understand phrases related to areas of immediate priority (e.g., basic personal and family information, shopping, local geography, jobs), if speaker articulates clearly and slowly. Can understand the essential information from short, recorded passages dealing with predictable everyday matters. Can use their recognition of known words to deduce the meaning of unfamiliar words in short expressions used in routine contexts.	Can understand texts describing people, places, everyday life etc., if given in simple language. Can find specific, predictable information in everyday material, like advertisements, menus, reference lists and timetables. Can use recognition of known words to deduce the meaning of unfamiliar words in short expressions used in routine contexts. Can understand short, simple texts on familiar matters of a concrete type, which consist of common or job-related language.	Can write brief, everyday expressions to satisfy simple needs of a concrete type (e.g., personal details, daily routines, wants and needs, requests for information). Can use some simple structures correctly, but still makes basic mistakes; nevertheless, it is usually clear what they are trying to communicate. Has sufficient vocabulary for the expression of basic communicative needs.	Can construct phrases on familiar topics with enough ease to handle short exchanges, despite noticeable hesitation and false starts. Pronunciation is generally intelligible when communicating in simple everyday situations. Can give short, basic descriptions of events and activities. Can explain likes or dislikes about something and why they prefer one thing over another, making simple, direct comparisons.
A1	1 – 1.5	Can recognize concrete information (e.g., places and times) on familiar topics encountered in everyday life, provided the information is delivered slowly and clearly. Has a basic vocabulary repertoire of words and phrases related to particular concrete situations.	Can get an idea of the content of simpler informational material and short, simple descriptions, especially if there is visual support. Recognizes familiar names, words and very basic phrases on simple notices in the most common everyday situations. Finds and understands simple, important information in short texts.	Can compose a short, very simple message (e.g., a text message) to friends to give them a piece of information or ask them a question. Has a very basic range of simple expressions about personal details and needs of a concrete type. Can produce simple isolated phrases and sentences.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words and repair communication. Has a very basic range of simple expressions about personal details and needs of a concrete type. Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by those used to dealing with speakers of the language group.

A1	1 – 1.5	Can recognize concrete information (e.g., places and times) on familiar topics encountered in everyday life, provided the information is delivered slowly and clearly. Has a basic vocabulary repertoire of words and phrases related to particular concrete situations.	Can get an idea of the content of simpler informational material and short, simple descriptions, especially if there is visual support. Recognizes familiar names, words and very basic phrases on simple notices in the most common everyday situations. Finds and understands simple, important information in short texts.	Can compose a short, very simple message (e.g., a text message) to friends to give them a piece of information or ask them a question. Has a very basic range of simple expressions about personal details and needs of a concrete type. Can produce simple isolated phrases and sentences.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words and repair communication. Has a very basic range of simple expressions about personal details and needs of a concrete type. Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by those used to dealing with speakers of the language group.

Appendix B

Score comparison with the IELTS test

Score users rely on different language tests to make important decisions about language proficiency. To be fair to test takers and to help support decisions about language proficiency irrespective of which test was taken, score requirements across different language tests should be comparable. This section describes the process used to produce a score comparison table between the TOEFL iBT 1-6 banded score scale and the IELTS Academic 0-9 band scores, to help score users decide on the most appropriate equivalent scores across the two tests.

For the receptive skills sections (Reading and Listening), two data sources were used. First, the results of a 2025 score concordance study between IELTS Academic and TOEFL iBT were consulted. This study was conducted collaboratively between ETS, and the three owners of IELTS (British Council, IDP: IELTS Australia, and Cambridge University Press & Assessment), with analysis of the collected data performed by a third party. The technical report is available on both the ETS website, as well as the IELTS website (Ikeda et al., 2025a; 2025b), offering concordance tables between the original TOEFL iBT score scale and the IELTS nine band levels. Second, the mapping of the original TOEFL score scale for the Reading and Listening test sections to the banded score scale were examined. Given the known relationship between the original TOEFL iBT scores (0-30 score scale for Reading and Listening sections) and IELTS nine band levels, as well as the known relationship between the TOEFL iBT 0-30 score scale and 1-6 score scale for Reading and Listening sections, the score comparison between the TOEFL iBT 1-6 score scale and the IELTS bands was established.

For the productive skills sections (Writing and Speaking scores), the results of the 2025 score concordance study between IELTS Academic and TOEFL iBT were consulted first,

followed by review of the mapping of the 0–30 score scale to the 1-6 score scale for the Writing and Speaking test sections. The score comparison between the TOEFL iBT banded score scale and the IELTS bands was created using these two data sources.

For the total score, the average of the 4 section scores was first estimated and then rounded up to the next half band on the 1-6 score scale. The decision to round up was made to minimize chances for false positive classifications, that is, classifying a TOEFL iBT test taker at a higher IELTS band score than their true language ability level. The results for the section and total scores are presented in Table B1.

Table B1. Score comparison table between IELTS band scores and TOEFL iBT banded score scale

IELTS (0-9)	TOEFL Reading (1-6)	TOEFL Listening (1-6)	TOEFL Writing (1-6)	TOEFL Speaking (1-6)	TOEFL Overall (1-6)
9	6	6	6	6	6
8.5	5.5	6	6	6	6
8	5.5	5.5	6	6	6
7.5	5	5	5.5	5	5.5
7	4.5	5	5	4.5	5
6.5	4	4.5	4.5	4	4.5
6	3.5	3.5	4	3.5	4
5.5	3.5	3	3	3	3.5
5	3	2.5	2	2.5	2.5
4.5	2.5	1.5	1.5	2	2
4	1	1	1	1.5	1.5

Note: Total scores in the table are based on the average of the sections scores, rounded up to the next highest half band, to minimize false positive classifications. Total scores awarded to test takers can be obtained with different section score combinations.

To evaluate the score comparisons, the CEFR level of the equivalent test scores across the two tests were examined (see Table B2), based on the information found on the website of

each test (ETS, n.d.; IELTS, n.d.). Overall, the CEFR level of the equivalent test scores converged. Across the 11 IELTS half-band scores compared (Band 4 to Band 9), the corresponding TOEFL 1-6 score was mapped to the same CEFR level for eight of them, whereas for the remaining three bands, the corresponding score was mapped to an adjacent band.

Table B2. Comparison of CEFR classifications based on scores from IELTS and TOEFL iBT

IELTS claimed CEFR level	IELTS (0-9)	TOEFL							
		Reading (1-6)	Listening (1-6)	Writing (1-6)	Speaking (1-6)	Total unrounded (1-6)	Total rounded (1-6)	Total rounded up (1-6)	Total rounded up CEFR level
C2	9	6	6	6	6	6.000	6	6	C2
C2	8.5	5.5	6	6	6	5.875	6	6	C2
C2	8	5.5	5.5	6	6	5.750	6	6	C2
C1	7.5	5	5	5.5	5	5.125	5	5.5	C1
C1	7	4.5	5	5	4.5	4.750	5	5	C1
B2	6.5	4	4.5	4.5	4	4.250	4.5	4.5	B2
B2	6	3.5	3.5	4	3.5	3.625	3.5	4	B2
B2	5.5	3.5	3	3	3	3.125	3	3.5	B1
B1	5	3	2.5	2	2.5	2.500	2.5	2.5	A2
A2	4.5	2.5	1.5	1.5	2	1.875	2	2	A2
A2	4	1	1	1	1.5	1.125	1	1.5	A1