TOEFL® Research INSIGHT

Volume 11:

TOEFL® Steps: Building the Learning Path of the TOEFL Family

ETS TOEFL®

TOEFL® Research Insight Series, Volume 11: TOEFL® Steps: Building the Learning Path of the TOEFL Family of Assessments

Preface

The TOEFL iBT® test is the world's most widely respected English-language assessment, used for admissions purposes in more than 160 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States (see test review in Alderson, 2009). Since its initial launch in 1964, the TOEFL® test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the TOEFL iBT test, was launched in 2005. It contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings, and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments.

In addition to the TOEFL iBT, the TOEFL Family of Assessments has been expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the TOEFL Primary® and TOEFL Junior® tests, which are designed to help teachers and learners of English in school settings. The TOEFL ITP® program offers colleges, universities, and others affordable tests for placement and progress monitoring within English programs that serve as a pathway to eventual degree programs.

At ETS, we understand that scores from the TOEFL Family of Assessments are used to help make important decisions about students, and we would like to keep score users and test takers up to date about the research results that help assure the quality of these scores. Through the *TOEFL® Research Insight Series* we provide institutions and English teachers with information regarding the strong research and development base that underlies the TOEFL Family of Assessments, and demonstrates our continued commitment to research.

Since the 1970s, the TOEFL test has had a rigorous, productive, and far-ranging research program. But why should test score users care about the research base for a test? In short, it is only through a rigorous program of research that a testing company can substantiate claims about what test takers know or can do based on their test scores, as well as provide support for the intended uses of assessments and minimize potential negative consequences of score use. Beyond demonstrating this critical evidence of test quality, research is also important for enabling innovations in test design and for addressing the needs of test takers and test score users. This is why ETS has established a strong research base as a fundamental feature underlying the evolution of the TOEFL Family of Assessments.

The TOEFL Family of Assessments is designed, produced, and supported by a world-class team of test developers, educational measurement specialists, statisticians, and researchers in applied linguistics and language testing. Our test developers have advanced degrees in fields such as English, language education, and applied linguistics. They also possess extensive international experience, having taught English on continents around the globe. Our research, measurement, and statistics teams include some of the world's most distinguished scientists and internationally recognized leaders in diverse areas such as test validity, language learning and assessment, and educational measurement.

To date, more than 300 peer-reviewed TOEFL Family research reports, technical reports, and monographs have been published by ETS, and many more studies on the TOEFL tests have also appeared in academic journals and book volumes. In addition, over 20 TOEFL-related research projects are conducted by ETS's Research & Development staff each year, and the TOEFL Committee of Examiners (COE), comprised of language learning and testing experts from the global academic community, funds an annual program of TOEFL Family research by independent external researchers from all over the world.

The purpose of the *TOEFL® Research Insight Series* is to provide a comprehensive yet user-friendly account of the essential concepts, procedures, and research results that help ensure the quality of scores for all members of the TOEFL Family of Assessments. Topics covered in these volumes include issues of core interest to test users, including how tests were designed; evidence for the reliability, validity, and fairness of test scores; and research-based recommendations for best practices.

The close collaboration with TOEFL score users, English language learning and teaching experts, and university scholars in the design of all TOEFL tests has been a cornerstone to their success and worldwide acceptance. Therefore, through this publication, we hope to foster an ever-stronger connection with our test users by sharing the rigorous measurement and research base and solid test development that continues to help ensure the quality of the TOEFL Family of Assessments.

Center for Language Education and Assessment Research
Research & Development Division
ETS

The following individuals contributed to this volume (in alphabetical order): Brent Bridgeman, Tim Davey, Larry Davis, Lixiong Gu, Venessa F. Manna, Spiros Papageorgiou (lead author).

TOEFL® Steps: Building the learning path of the TOEFL Family

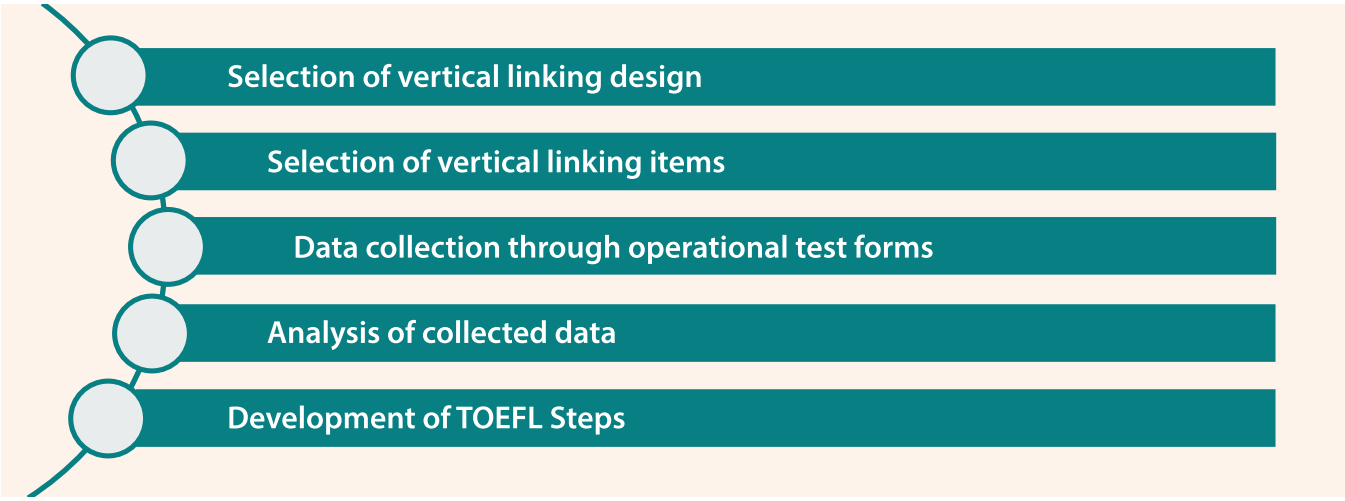
The primary motivation for administering a language test is to use its scores to facilitate decisions of various kinds about language proficiency (Bachman & Palmer, 2010). For example, score-based decisions in educational and workplace contexts relate to admission of international students to higher education institutions, placement into different levels of language courses, monitoring of achievement of learning goals, and professional certification and promotion for which proof of language proficiency is required. Decisions made on the basis of test scores can be extremely consequential, both for test takers but also for score users, such as universities and employers, and society overall.

The different TOEFL tests are used sequentially along the educational continuum from primary school through college and in a variety of contexts, with the implicit assumption that scores or "achievements" on one test are related to scores on the next. Although the various members of the TOEFL Family of Assessments were developed at different time points and target different groups of test takers, it is nonetheless important to show the relationship between scores of different tests, even when they are intended for different proficiency levels and test-taker ages.

Tests in the TOEFL Family of Assessments use different scales for reporting scores. For example, the score scale range for the reading and listening sections of the TOEFL Primary tests is 100 to 115 with one-point increments, whereas the score scale ranges for the same sections of the TOEFL Junior test is 200 to 300 with five-point increments. The score scale range of the reading section of the TOEFL ITP Level 1 test is 31 to 67 with one-point increments, whereas the score scale range for the reading section of the TOEFL iBT test is 0 to 30 with one-point increments. The use of distinct score scales was grounded in the desire to avoid confusion of the scores of one test with the scores of another test. In addition to having unique reporting score scales, each test also has a unique set of language proficiency descriptors to facilitate score interpretation (see Papageorgiou, Morgan, et al., 2015, Powers et al., 2017, Wang & Papageorgiou, in press). Score interpretation for each test is also facilitated through studies (e.g., Papageorgiou, Tannenbaum, et al., 2015) that map the scores of each test to levels in well-known language proficiency frameworks such as the Common European Framework of Reference (CEFR; Council of Europe, 2001). However, score interpretation through performance descriptors and score mapping is based on separate studies for each individual test in the TOEFL Family, and such studies provide only indirect evidence for comparisons between tests. Directly establishing the relationship of the score scales of the different TOEFL tests empirically can help test users decide when a student who took one test might be better served by taking a test that provides more information at a different proficiency level.

To establish the relationship of the score scales of the different TOEFL tests empirically, vertical linking (or scaling) procedures can be employed, whereby the scores of all tests can be expressed in a common metric (Kolen, 2006). This volume in the *TOEFL® Research Insight Series* reports on a multiyear vertical linking research project for the TOEFL Family of Assessments. The purpose of the vertical linking project was to support decisions about readiness to take each TOEFL test by empirically linking the different score scales. Based on the results of this project, a visual tool called "TOEFL Steps" was created to help score users decide when language learners at different ages and proficiency levels are ready to take a test in the TOEFL Family. The major stages of the project are shown in Figure 1 and are discussed in the remainder of this volume. A detailed account of the project is provided in three chapters (Gu et al., in press; Monfils & Manna, in press; Papageorgiou et al., in press) in the book *Meaningful language test scores: Research to enhance score interpretation* (Papageorgiou & Manna, in press).

Figure 1: Major stages of the vertical linking project



Selection of vertical linking design

The TOEFL Family of Assessments consists of the TOEFL iBT test, the TOEFL ITP Assessment Series, the TOEFL Junior tests, and TOEFL Primary tests (for details see www.ets.org/toefl). As shown in Figure 2, the first three are intended for adolescents and adults, whereas the TOEFL Junior and TOEFL Primary tests are intended for young learners (11 years or older and 8 years or older, respectively).

Figure 2: Intended test takers for the TOEFL Family of Assessments

TOEFL iBT TOEFL ITP	TOEFL Junior	TOEFL Primary
Adolescents and adults who plan to use English in a higher education context	Young learners (11 years old or older) using English in school settings	Young learners (8 years old or older) using English in school settings

TOEFL ITP Assessment Series refers to two separate tests — one that is intended for students with beginning to intermediate English-language skills (Level 2) and the other designed for those having intermediate to advanced skills (Level 1). The TOEFL ITP Level 1 test was the focus of the research project reported in this volume, because its design, as explained later, allowed for vertical linking between the TOEFL Junior tests and the TOEFL iBT test. Data were collected for the TOEFL ITP Level 2 test at a later point for subsequent analysis in a separate study; this effort is not discussed in this volume. Therefore, in this *TOEFL® Research Insight Series* volume we report on the vertical linking study for the reading and listening sections of the TOEFL iBT test, the TOEFL ITP Level 1 test (henceforth TOEFL ITP), the TOEFL Junior tests, and the TOEFL Primary tests. At the time of writing, not all tests in the TOEFL Family included speaking and writing sections. However, as these sections are being added to the different tests in the TOEFL Family, further work is planned to vertically link their scores.

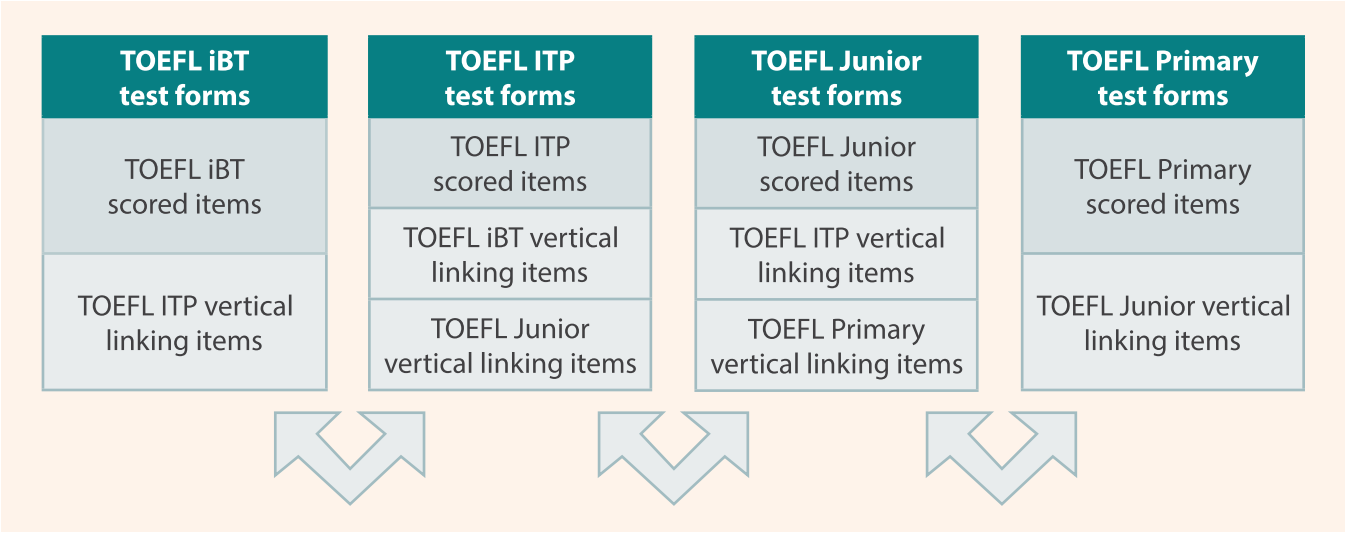
A common item linking design (also called nonequivalent groups anchor test linking, NEAT) was chosen for the TOEFL vertical linking project. By employing such a design, some test items function as vertical linking items. The vertical linking items appear not only in the test form for which they were originally designed, but also in the test form of another test, designed for an adjacent higher- or lower-proficiency level. In other words, the vertical linking items used in the project were administered in two adjacent tests. The use of adjacent tests was considered suitable for the purposes of this project so that test takers did not have to encounter vertical linking items that were too difficult, too easy, or inappropriate for their age. For example, it would be more appropriate for TOEFL Primary test takers to take only vertical linking items from TOEFL Junior than taking items from tests intended for adult language learners.

The data collection was conducted through operational (live or official) test administrations, which meant that the test takers did not know that some test items belonged to a different test in the TOEFL Family of Assessments. The decision to collect data through operational test administrations was made because test takers were expected to be motivated to perform according to their best ability, which may not always be the case when data are collected as part of a research study. The vertical linking items of the test forms used in the project occupied slots typically reserved for test items that are not scored, but instead are administered for quality control or other purposes. Using these slots for the vertical linking items had the advantage of avoiding disruption to the administration of the operational tests or their scoring.

Figure 3 illustrates how vertical linking items for reading and listening were embedded in each of the TOEFL tests. Each test form contained both the operational items (items used for scoring) as well as the vertical linking items (from adjacent tests, which were not used for scoring) as follows:

- Operational TOEFL iBT test forms contained the TOEFL iBT test items used for scoring and TOEFL ITP vertical linking items.
- Operational TOEFL ITP test forms contained the TOEFL ITP items used for scoring and also vertical linking items from either adjacent test, TOEFL iBT or TOEFL Junior.
- Operational TOEFL Junior test forms contained TOEFL Junior test items used for scoring and also vertical linking items from the two adjacent tests, the TOEFL ITP test, and the TOEFL Primary test.
- Operational TOEFL Primary tests contained TOEFL Primary items used for scoring and TOEFL Junior vertical linking items.

Figure 3: Data collection design for the TOEFL vertical linking project



Selection of listening and reading vertical linking items

The project team relied on general criteria for selecting vertical linking items based on good practice in vertical linking design (Kolen & Brennan, 2014). Based on these criteria, vertical linking items needed to:

1. Represent content overlap or skill progression along the language proficiency continuum (i.e., the construct of measurement interest) for adjacent tests in the TOEFL Family of Assessments
2. Have a similar "look and feel" to items in the operational test
3. Require minimal or no changes to test directions in the operational test
4. Be of appropriate difficulty for the target student population of the operational test
5. Be representative of the content of the operational test
6. Be placed in a similar position in adjacent tests

As discussed later, the principle psychometric assumption of common-item linking is that the vertical linking items will function in similar ways when taken by language learners at the same level of language ability, irrespective of the test in which the items appear. However, in the real world we recognize that several factors related to test design might result in differences in the way the vertical linking items function when they appear in different tests. Such design factors include the relevance of the topics of the reading passages or listening input, which might vary across the different tests because of the target test-taker population, or the length of input, or number of response options, which might vary from the surrounding items. Even the position of test tasks in a test form might impact test-taker perceptions of difficulty or fatigue. In the case of the TOEFL vertical scaling project, the overall goal was to minimize the impact of such test design factors and make the vertical linking items look "at home" when they appeared in tests other than the test for which they were originally designed.

As discussed in detail in Papageorgiou et al. (in press), the project team and experienced assessment developers at ETS reviewed all listening and reading test tasks to identify those that were suitable for providing vertical linking items shared by adjacent tests (see Figure 3), based on the above criteria. Table 1 presents the number of test forms used in the project and the number of listening and reading vertical linking items for each pair of adjacent tests. The operational test forms were administered between January 2019 and February 2020.

Table 1: Details of test forms containing vertical linking items

Test	Number of test forms	Test section	Vertical linking items	Number of tasks	Number of items
TOEFL Primary	1	Listening	TOEFL Junior	2	8
		Reading	TOEFL Junior	2	8
TOEFL Junior	4	Listening	TOEFL Primary	8	12
		Reading	TOEFL Primary	5	12
		Listening	TOEFL ITP	6	21[a]
		Reading	TOEFL ITP	3	24
TOEFL ITP	10	Listening	TOEFL Junior	5	20
		Reading	TOELF Junior	3	23
		Listening	TOEFL iBT	5	25
		Reading	TOEFL iBT	4	44
TOEFL iBT	2	Listening	TOEFL ITP	7	30
		Reading	TOEFL ITP	4	37[b]

[a] 24 items selected originally, but 21 were used during form assembly.
[b] 44 items selected originally, but 37 were used during form assembly.

Data collection through operational test forms

Responses from 163,209 test takers on the 17 test forms were collected as shown in Table 2. As all these test forms were administered to large samples of test takers, the samples for this study are considered representative of the test-taking populations of each test in the TOEFL Family.

Table 2: Test forms used in data collection

Test Form	Listening vertical linking items	Reading vertical linking items	Test takers
TOEFL Primary Form 1	TOEFL Junior	TOEFL Junior	2,777
TOEFL Junior Form 1	TOEFL ITP	TOEFL ITP/Primary	16,219
TOEFL Junior Form 2	TOEFL ITP	TOEFL ITP/Primary	6,659
TOEFL Junior Form 3	TOEFL Primary	TOEFL ITP/Primary	2,638
TOEFL Junior Form 4	N/A	TOEFL Primary	1,701
TOEFL ITP Form 1	TOEFL Junior	N/A	12,142
TOEFL ITP Form 2	TOEFL iBT	TOEFL iBT	11,612
TOEFL ITP Form 3	TOEFL Junior	TOEFL Junior	10,682
TOEFL ITP Form 4	TOEFL iBT	TOEFL iBT	10,285
TOEFL ITP Form 5	TOEFL Junior	TOEFL Junior	9,835
TOEFL ITP Form 6	TOEFL Junior	TOEFL Junior	9,709
TOEFL ITP Form 7	TOEFL Junior	N/A	9,660
TOEFL ITP Form 8	TOEFL iBT	TOEFL iBT	9,429
TOEFL ITP Form 9	TOEFL iBT	TOEFL iBT	8,765
TOEFL ITP Form 10	TOEFL iBT	N/A	9,493
TOEFL iBT Form 1	TOEFL ITP	TOEFL ITP	16,419
TOEFL iBT Form 2	TOEFL ITP	TOEFL ITP	15,184

Note: TOEFL ITP Level 1 only

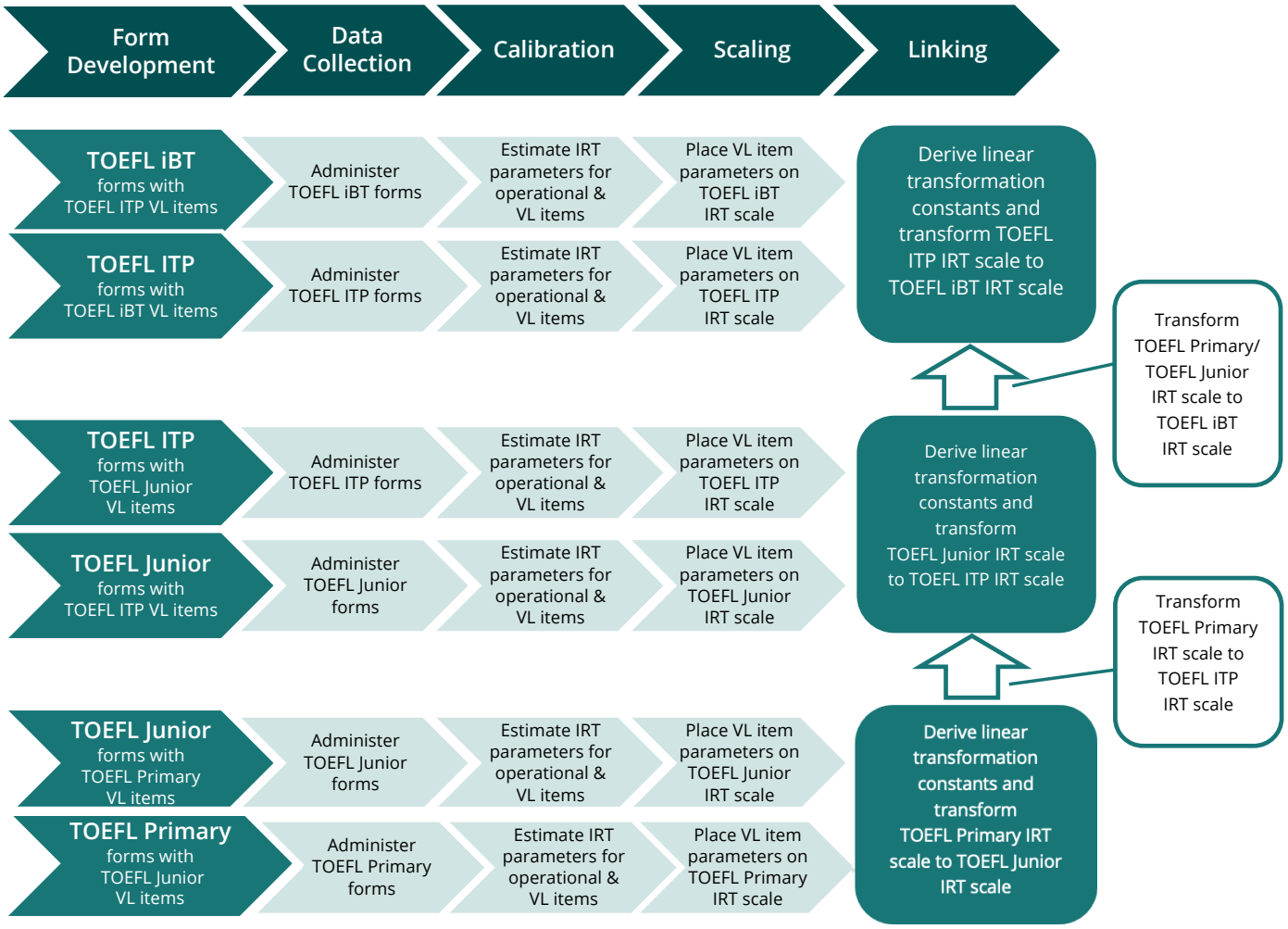
Analysis of the collected data

This section describes the various analyses performed by the psychometric and assessment development teams (described in detail in Gu et al., in press).

Selecting the base test

The scaling approach for the collected reading and listening data was based on the same item response theory (IRT) statistical model used for the reading and listening scores of all tests in the TOEFL Family of Assessments. In the IRT scaling approach, a base score scale should be selected to link the statistical properties of all test questions (for example, difficulty) onto a common underlying scale. The IRT score scale for the TOEFL iBT test was selected as the base scale. This decision was made because the TOEFL iBT test, which was launched in 2005 (ETS, 2020), is a well-established test in the TOEFL Family in terms of consistency of design, characteristics of the test-taking population, and stability of the score scale. Therefore, linking between adjacent tests started with linking the TOEFL ITP test to the TOEFL iBT test. The TOEFL Junior test was then linked to the TOEFL ITP test. Last, the TOEFL Primary test was linked to the TOEFL Junior test. Thus, through this process, eventually all the tests were linked to the underlying IRT scale of the designated base test of the vertical scale, the TOEFL iBT test. This process is shown in Figure 4.

Figure 4: Linking procedure to develop the TOEFL vertical scales for reading and listening



Note: VL=vertical linking

Calibration and scaling of vertical linking items

The collected data were calibrated, which is a process to estimate statistical properties for the test questions. In the IRT approach these properties are called item parameters, and the model used for the study estimated two parameters, the *a*-parameter for discrimination (how well a test question distinguishes between low- and high-ability test takers) and *b*-parameter (how difficult the test question is).

The calibration process can be conducted either concurrently across all proficiency levels or separately by test form and within level (see Monfils and Manna (in press) for detailed comparisons of the two methods). Given that each test in the TOEFL Family of Assessments has its own established score scale and test forms are administered with different schedules, it was more practical to use the separate calibration approach. Therefore, during data analysis, the vertical linking items, together with the scored items of the test forms in which they were embedded, were calibrated by test to estimate their IRT item parameters. The item parameters of the vertical linking items were also placed onto the underlying IRT scale of the test forms in

which they were embedded. This procedure is typically called "scaling." It is a routine procedure for tests whose scores are calculated using IRT methods, where item parameters for items that do not count toward scoring are placed onto the base scale of the test following calibration.

Because the vertical linking items were calibrated and scaled separately by test form (i.e., both the test form in which they originally appeared, and the test form in which they were embedded as vertical linking items), they had two sets of item parameters, one on the scale of the original test form and another on the scale of the embedded test form. For example, for the items used to vertically link the TOEFL Junior test and TOEFL ITP test, one set of item parameters was based on the responses of the TOEFL Junior test takers and another set of item parameters was based on the responses of the TOEFL ITP test takers.

Content analysis of flagged vertical linking items

As explained earlier, the vertical linking items were selected to reflect the overlap between adjacent tests along the language proficiency continuum. Therefore, the assumption is that the vertical linking items will function similarly when taken by individual language learners at the same level of language ability, irrespective of the TOEFL test they took. For example, although TOEFL Junior and TOEFL ITP test-taker groups might differ overall in their language ability, it is reasonable to assume that those individual test takers of either test with the same level of language ability would have the same probability of answering an item correctly. We note that test takers of different ages are likely to differ cognitively even if their overall language proficiency is similar. The use of adjacent tests is intended to some extent to address cognitive differences of the target test-taking populations, especially for young language learners taking the TOEFL Primary tests. These test takers took vertical linking items from the TOEFL Junior test only, intended for learners at the age of 11 or older, and not the TOEFL ITP or TOEFL iBT tests intended for adolescents and adult learners.

After the items had been placed on a common scale, psychometric analysis of the vertical linking items examined their difficulty and discrimination as estimated from both their original test form and the test form in which they were embedded. Based on this analysis, vertical linking items that varied in their difficulty and/or discrimination when appearing in different tests were flagged as "outlier items." All flagged outlier items were passed to the Assessment Development team with the following question: Are factors irrelevant to language proficiency the reason why specific vertical linking items were flagged for content analysis? Depending on the outcome of the content analysis, some vertical linking items were dropped from further analysis.

Table 3 shows the number of vertical linking items selected after dropping flagged items.

Table 3: Details of vertical linking items used in the study

Test	Test section	Adjacent test	Number of vertical linking items in test forms	Number of vertical items used for the vertical scales
TOEFL Primary	Listening	TOEFL Junior	8	8
	Reading	TOEFL Junior	8	8
TOEFL Junior	Listening	TOEFL Primary	12	10
	Reading	TOEFL Primary	12	12
	Listening	TOEFL ITP	21[a]	20
	Reading	TOEFL ITP	24	22
TOEFL ITP	Listening	TOEFL Junior	20	18
	Reading	TOEFL Junior	23	22
	Listening	TOEFL iBT	25	21
	Reading	TOEFL iBT	44	41
TOEFL iBT	Listening	TOEFL ITP	30	27
	Reading	TOEFL ITP	37[b]	33

[a] 24 items selected originally, but 21 were used during form assembly.
[b] 37 items selected originally, but 33 were used during form assembly.

Finalization of the underlying vertical scales for reading and listening

Following removal of flagged items, another run of the linking analyses with the remaining vertical linking items was conducted so that the linking across tests was free of differential performance caused by construct-irrelevant factors. Once the underlying IRT scales of all the tests in the test series were transformed to a common base scale, a decision was made not to transform the underlying scale to a numerical scale as is typically the case with vertical scales used for score reporting purposes. Instead, the project team decided to use the CEFR levels in lieu of an integer scale, given that all tests in the TOEFL Family of Assessments were previously mapped to the CEFR levels. The results from the score mapping studies and the vertical linking project converged, with only minor adjustments made to the CEFR score mapping for the TOEFL Primary, TOEFL Junior, and TOEFL ITP tests. Using the CEFR levels instead of transforming the vertical scale into an integer score scale helps avoid the complication of adding yet another numeric scale alongside those for each test in the TOEFL Family.

Examination of measurement precision

Precision of test scores was also examined by estimating the conditional standard error of measurement (CSEM). The CSEM is a statistical index of the precision of a particular test score expressed in the common, underlying IRT scale. Therefore, the CSEM allows us to compare the relative precision of the reported scale scores across the full range of test-taker ability, for all tests in the TOEFL Family. Through this comparison, decisions can be made about which test in the TOEFL Family will provide more precise measurement for an individual at a given level of reading and listening proficiency. This notion of precision for measuring language proficiency levels across the TOEFL Family of Assessments is conveyed through the focus levels shown in Figure 5, where focus level corresponds to the range of ability most precisely measured by each test.

Figure 5: Focus proficiency levels in the TOEFL Family of Assessments

TOEFL iBT	TOEFL ITP	TOEFL Junior	TOEFL Primary
Test takers: Adolescents and adults who plan to use English in a higher education context	Test takers: Adolescents and adults who plan to use English in a higher education context	Test takers: Young learners (11 years old or older) using English in middle or high school	Test takers: Young learners (8 years old or older) using English in middle or high school
Focus proficiency levels CEFR B2–C1	Focus proficiency levels CEFR B2–C1	Focus proficiency levels CEFR B1–B2	Focus proficiency levels CEFR A1–B1
	A score at C1 level might indicate readiness to take the TOEFL iBT test	A score at B2 level might indicate readiness to take the TOEFL iBT or TOEFL ITP test	A score at B1 level might indicate readiness to take the TOEFL Junior test

Development of TOEFL Steps

Once data analysis was concluded, the project team focused on how to best convey the key points about the language learning path established through the TOEFL vertical linking project. Following several rounds of consultation with ETS staff from the R&D division and the TOEFL program, as well as the members of the TOEFL Committee of Examiners in 2019 (www.ets.org/toefl/score-users/ibt/about/board), the project team developed TOEFL Steps, a set of two visual tools, one for reading (Figure 6), and one for listening (Figure 7).

Figure 6: TOEFL Steps for Reading

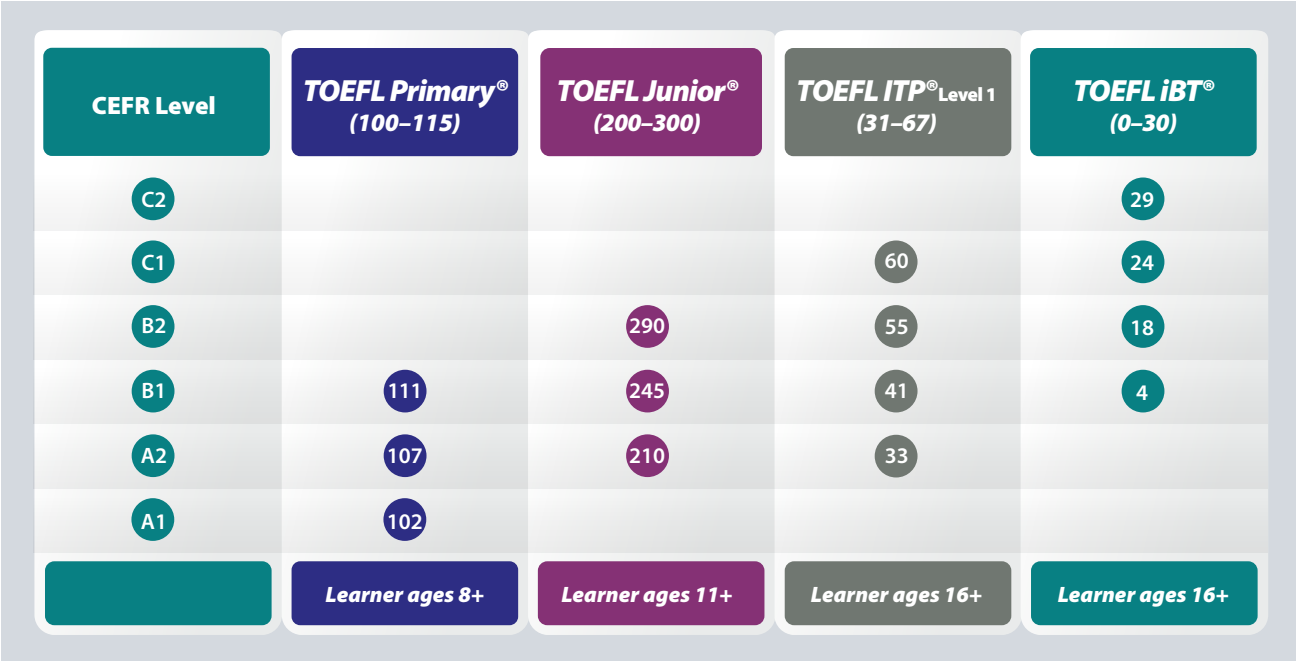
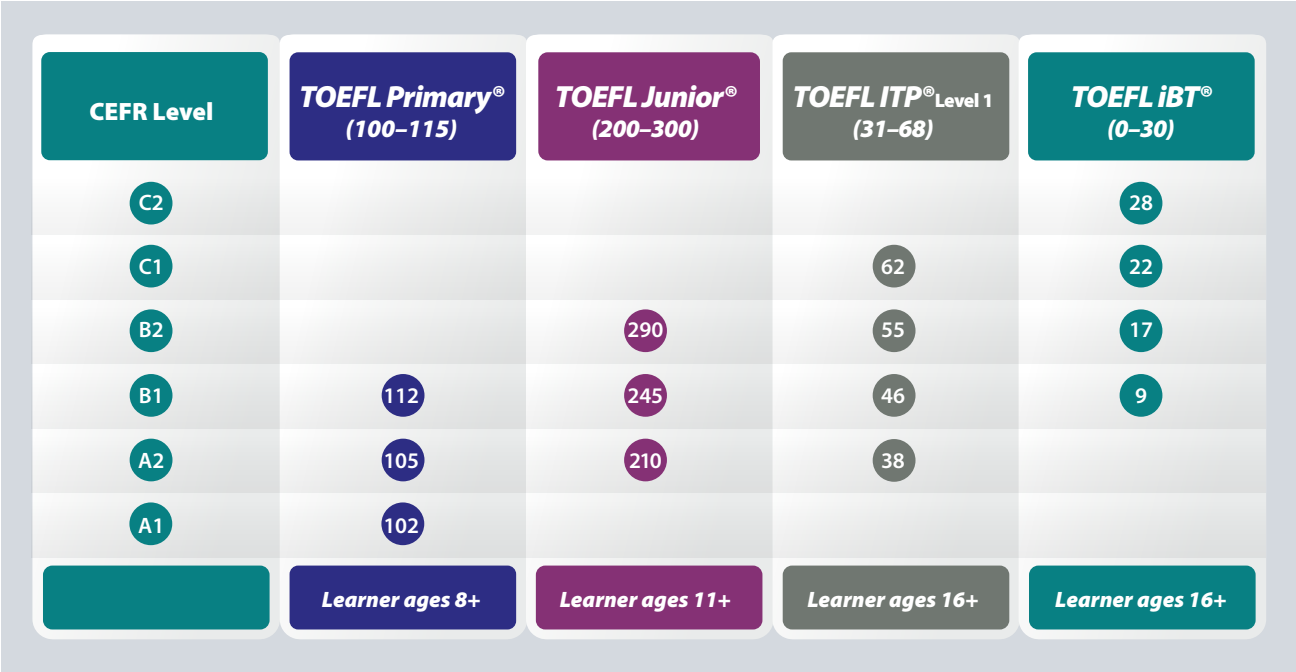


Figure 7: TOEFL Steps for Listening



We conclude this volume with a few key points about the use of TOEFL Steps for Reading and Listening.

- **Target age group:** The target age group for each test is clearly shown at the bottom, because it is very important to remind score users that how we interpret scores and language proficiency levels should be based on the age of the test takers.
- **Use of the CEFR levels:** The CEFR levels are used as a common reference point to help interpret the proficiency level of a test taker. Scores are displayed for each test that indicate the minimum score needed to be placed at each CEFR level.
- **Test selection and readiness to take a TOEFL test:** Teachers and students can use TOEFL Steps to decide when it is a good time to take a test in the TOEFL Family based on performance on another test or the test taker's proficiency level. For example, if a teacher thinks that a TOEFL Primary test taker is at B1 level, and then the test taker gets a score of 112 or higher, then this test taker might be ready to take TOEFL Junior. High school students who have taken TOEFL Junior and wish to take one of the tests intended for higher education might need to wait until they achieve a relatively high score first (for example 290).

References

Alderson, J. C. (2009). Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing*, *26*(4), 621–631. https://doi.org/10.1177/0265532209346371

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

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

ETS. (2020). *TOEFL® program history.* Retrieved from www.ets.org/toefl/research/insight-series

Gu, L., Li, Y., Monfils, L., & Papageorgiou, S. (in press). Statistical methodology for developing vertical scales for language tests. In S. Papageorgiou & V. F. Manna (Eds.), *Meaningful language test scores: Research to enhance score interpretation*. John Benjamins.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 156–186). Praeger Publishers.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Monfils, L., & Manna, V. F. (in press). Considerations in developing vertical scales for language tests (in press). In S. Papageorgiou & V. F. Manna (Eds.), *Meaningful language test scores: Research to enhance score interpretation*. John Benjamins.

Papageorgiou, S., & Manna, V. F. (Eds.) (in press). *Meaningful language test scores: Research to enhance score interpretation*. John Benjamins.

Papageorgiou, S., Ginsburgh, Mitch, & Garcia Gomez, Pablo (in press). 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*. John Benjamins.

Papageorgiou, S., Morgan, R., & Becker, V. (2015). Enhancing the interpretability of the overall results of an international test of English language proficiency. *International Journal of Testing*, *15*(4), 310–336. https://doi.org/10.1080/15305058.2015.1078335

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* (Research Memorandum No. RM-15-06). ETS. www.ets.org/Media/Research/pdf/RM-15-06.pdf

Powers, D., Schedl, M., & Papageorgiou, S. (2017). Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies. *Language Testing*, *34*(2), 175–195. https://doi.org/10.1177/0265532215623582

Wang, L., & Papageorgiou, S. (in press). 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*. John Benjamins.

www.ets.org

Copyright © 2023 by ETS. All rights reserved. ETS, the ETS logo, TOEFL, TOEFL IBT, TOEFL ITP, TOEFL JUNIOR and TOEFL PRIMARY are registered trademarks of ETS in the United States and other countries. The T logo is a trademark of ETS. 1116533517