

**Score Reliability, Classification Decision Accuracy, and Interrater Reliability of TOEFL Primary®
Tests**

Feifei Li

ETS Research Institute, Princeton, New Jersey, United States

Corresponding author: Feifei Li, Email: fli@ets.org

Suggested citation: Li, F. (In Press). *Score Reliability, Classification Decision Accuracy, and Interrater Reliability of TOEFL Primary® Tests*. ETS Research Report Series. <https://doi.org/10.64634/ys398g44>

Abstract

This study evaluates the score reliability, classification decision accuracy, and interrater reliability of TOEFL Primary® test scores. Using large operational datasets from Reading, Listening (Step 1 and Step 2), and Speaking tests, reliability was estimated via Cronbach's alpha and standard error of measurement (SEM). Classification decision accuracy for CEFR levels was estimated using the Livingston and Lewis (1995) method. Interrater reliability for Speaking was evaluated through percentage agreement, correlation, and quadratic weighted kappa based on a double-scoring design. Results indicate strong internal consistency across all test sections ($\alpha = .83\text{--}.90$) and relatively small SEMs (1.19–1.61), supporting score precision. Classification decision accuracy was high for Reading and Listening (approximately .78–.82) and somewhat lower for Speaking (.76), reflecting greater complexity in performance-based assessment. Interrater reliability for Speaking demonstrated moderate to strong agreement, with correlations and weighted kappa values ranging from .68 to .88 and adjacent agreement exceeding 96%. Overall, findings provide strong evidence supporting the reliability and interpretability of TOEFL Primary scores for reporting CEFR-aligned proficiency levels and making educational decisions about young English learners.

Keywords: TOEFL Primary®; score reliability; classification decision accuracy; interrater reliability; CEFR levels; Reading assessment; Listening assessment; Speaking assessment; English learners; educational decision-making

Introduction

The TOEFL Primary® tests are developed by Educational Testing Service (ETS) as an English-language proficiency test for young learners age 8 and above who are learning English as a foreign language. The tests assess students' knowledge, skills and abilities to fulfill core communication goals in English.

A team of language and assessment specialists develop the materials for the TOEFL Primary tests. Test specifications, test questions, and test forms are rigorously reviewed, following the ETS Standards for Quality and Fairness (ETS, 2014). Test questions are carefully checked to ensure that the content is appropriate, according to established ETS procedures.

A fundamental part of test score quality is ensuring that scores reported on different versions (e.g., test forms) of the same test can consistently distinguish more and less proficient test takers. In other words, to be useful to score users, test scores must be reliable. As an ETS assessment program, the TOEFL Primary test program strives to ensure score reliability through strict adherence to the guidelines and practices established for the development and operational implementation of the TOEFL Primary tests. Evidence of score reliability is important because it supports the interpretation that test scores carry consistent meaning across different test forms.

Research Purpose

Teachers and parents use the TOEFL Primary tests to learn about young students' developing English skills, for a variety of purposes. This memorandum aims to help users of the TOEFL Primary tests to determine whether the reported test results are reliable enough to support their intended purposes. It covers the following areas:

- Reliability coefficients and the standard error of measurement (SEM) of the TOEFL Primary Reading and Listening test scores, including Step 1 and Step 2 tests, and scores for the TOEFL Primary Speaking test
- Decision accuracy of classifications into Common European Framework of Reference (CEFR¹) levels for the TOEFL Primary Reading and Listening tests, including Step 1 and Step 2 tests, and the TOEFL Primary Speaking test
- Interrater reliability of the TOEFL Primary Speaking test

Research Questions

This study is guided by five key research questions concerning the psychometric quality and interpretability of the TOEFL Primary Reading, Listening, and Speaking assessments. First, we examine score reliability – specifically, the extent to which different forms of the TOEFL Primary Reading, Listening, and Speaking tests maintain high internal consistency (as indexed by Cronbach's alpha). Second, we investigate measurement precision by estimating the SEM for each test section, thereby determining the typical magnitude of measurement error associated with scores in Reading, Listening, and Speaking. Third, we evaluate classification decision

accuracy, asking how accurately the observed test scores classify learners into the appropriate proficiency levels of the CEFR compared to those learners' true proficiency levels (their true scores in classical test theory terms). Fourth, for the Speaking test we assess interrater reliability, analyzing the degree to which different trained human raters consistently assign the same scores to the same speaking responses. Finally, we compare the performance across test sections, determining whether there are notable differences in reliability estimates or in classification accuracy among the Reading, Listening, and Speaking assessments.

Overview

Following this introduction, the remainder of the memorandum is organized as follows. Section 2 provides an overview of the TOEFL Primary tests, describing the Reading and Listening assessments (including the Step 1 and Step 2 levels) and the Speaking assessment, and outlining their purposes, score scales, and score reporting procedures. Section 3 details the methodological approach, defining key psychometric concepts and explaining how we evaluated internal consistency reliability, the standard error of measurement, the accuracy of CEFR classification decisions, and the consistency of ratings for the Speaking test. Section 4 presents the results of these analyses, summarizing the chief findings for score reliability, classification accuracy, and inter-rater consistency for each test section. Finally, Section 5 offers a discussion and conclusion, integrating the findings and considering their implications for the use and interpretation of TOEFL Primary scores in educational settings.

Overview of Tests

Description of Tests

The TOEFL Primary tests are intended to provide teachers and parents with accurate and reliable insights into young learners' developing English proficiency. Information from TOEFL Primary score reports helps teachers and parents track progress over time, identify strengths and areas for improvement in English, plan future instruction tailored to individual needs, and place learners in classes that match their proficiency levels (ETS, 2023).

The TOEFL Primary tests are intended for young learners, typically starting at age 8, who have been studying English for about one year or longer. However, the suitability of the tests

should ultimately be determined by each learner's level of English proficiency. These tests are not tied to any specific curriculum. Instead, they assess learners' knowledge, skills, and abilities in achieving fundamental English communication goals, which are common across educational contexts worldwide (ETS, 2023).

Reading and Listening Tests

TOEFL Primary Reading and Listening tests are designed to assess young learners' knowledge, skills, and abilities for fulfilling core communication goals in English. Young learners acquire English proficiency at different rates and through different pathways, influenced by factors such as instructional context, curriculum and exposure to language. To accommodate the developmental variability, the TOEFL Primary Reading and Listening tests are offered at two levels, Step 1 and Step 2, allowing teachers and parents to select the test that more closely aligns with students' current learning experience and levels of English proficiency.

The two levels are defined as follows

- Step 1 Tests measure reading and listening proficiencies and are intended for young learners at the early stages of English language development.
- Step 2 Tests measure more advanced reading and listening proficiencies and are designed for young learners who have acquired some communicative skills in English.

Speaking Test

The TOEFL Primary Speaking test measures young learners' ability to communicate orally in routine, age-appropriate social situations related to their daily lives. The test is suitable for learners across a broad range of English proficiency levels and is designed to capture students' emerging and developing speaking skills through structured, performance-based tasks.

Test Scores and Score Reporting

Reading and Listening Tests

As described in the *TOEFL Primary® Test Handbook* (ETS, 2023), each TOEFL Primary Reading and Listening score report (for both Step 1 and Step 2) provides the following information:

- Scale scores that help show progress within a level when students take the test again
- Raw scores (the number of correct responses in each section) are converted to scale scores ranging from 100 to 115, reported in one-point increments.
- Step 1 scale scores range from 101–109.
- Step 2 scale scores range from 104–115.
- CEFR levels, which describe students' English-language proficiency using an internationally recognized framework for language ability.

The relationship between TOEFL Primary Reading and Listening scale scores and CEFR levels is summarized in Table 1.

Table 1 TOEFL Primary Reading and Listening Scale Scores and CEFR Levels

CEFR	Reading scale score	Listening scale score
Below A1	100–101	100–101
A1	102–106	102–104
A2	107–110	105–111
B1	111–115	112–115

Speaking Test

According to the *TOEFL Primary Test Handbook* (ETS, 2023), each TOEFL Primary Speaking score report provides:

- Scale scores that indicate progress within a proficiency level when students take the test multiple times

- CEFR levels that show student ability in comparison to a widely used international framework for describing language proficiency

Speaking score ranges from 0 to 27, where a score of 0 is reported as *No Score*. The relationship among Speaking scores and CEFR levels is summarized in Table 2. Additional details regarding the interpretation of speaking scores can be found in ETS (2025).

Table 2. TOEFL Primary Speaking Scores and CEFR Levels

CEFR	Score
Below A1	1–9
A1	10–15
A2	16–21
B1	22–25
B2	26–27

Method

Score Reliability

The TOEFL Primary tests are administered in multiple equivalent forms—parallel editions of the test built to the same specifications with different questions. Because examinees’ scores may vary across forms, score reliability is an important indicator of test quality. In classical test theory, reliability reflects the degree to which test takers would receive consistent scores and maintain the same score rank ordering if they were tested repeatedly with different forms of the assessment. The two statistics commonly used to quantify score reliability are the reliability coefficients and the standard error of measurement.

Reliability Coefficient

Internal consistency reliability for each test section was evaluated using Cronbach’s alpha (Cronbach, 1951). Cronbach’s alpha is a reliability coefficient that measures how closely related the test items are as a group, indicating the extent to which they all assess the same underlying construct. In classical test theory terms, Cronbach’s alpha can be interpreted as the expected correlation between scores on two parallel forms of the test. Its values range from

0.00 (indicating no consistency beyond random chance) to 1.00 (perfect consistency). In practice, the following guidelines are commonly used for interpreting Cronbach's alpha:

- $\geq .90$: Excellent reliability
- .80–.89: Good reliability
- .70–.79: Acceptable reliability
- .60–.69: Questionable reliability
- $< .60$: Poor reliability

These threshold values are not absolute; they should be considered in the context of the testing scenario. For example, appropriate reliability expectations may depend on factors such as the stakes of the decisions being made and the diversity or homogeneity of the content and test-taking population. Higher-stakes tests or those intended to cover a broad domain typically warrant higher reliability coefficients, whereas slightly lower alphas may be tolerated for shorter subtests or preliminary assessments.

Standard Error of Measurement

The standard error of measurement (SEM) provides a complementary index of score precision by quantifying the amount of random error in test scores. Conceptually, a test taker's *true score* is defined as the average score that the person would earn over an infinite number of parallel test administrations (Lord & Novick, 1968). The difference between an examinee's observed score on a single test form and their true score represents measurement error. The SEM is an estimate of the standard deviation of these hypothetical observed score errors. It is expressed on the same scale as the test scores and is calculated as:

$$SEM = SD \times \sqrt{1 - r} \quad (1)$$

where SD is the standard deviation of observed test scores for a group of test takers, and r is the reliability coefficient for the test.

A smaller SEM indicates greater score precision (less score fluctuation due to measurement error), whereas a larger SEM indicates more uncertainty in an individual's

observed score. Importantly, the magnitude of the SEM should be interpreted relative to the score scale and the critical score thresholds (such as cut scores for proficiency levels like CEFR levels; AERA/APA/NCME, 2014). For example, an SEM of 2 points may be considered small if the score scale spans 100 points, but less acceptable if the score scale spans only 10 points or if decisions hinge on small score differences near a cut score.

Under the assumption of normally distributed measurement error, approximately 68% of observed scores fall within ± 1 SEM of the true score. Approximately 95% fall within ± 2 SEM. These rules of thumb illustrate how SEM can be used to create confidence intervals around observed scores. In practice, reasonably small SEM values (relative to the score scale and the importance of decisions being made) are desired to ensure that test scores are precise and reliable for their intended purposes (Lord & Novick, 1968; Brennan, 2001).

Classification Decision Accuracy

Because the TOEFL Primary assessment reports performance levels (e.g., CEFR levels) in addition to scale scores, a key quality indicator is classification decision accuracy (Livingston & Lewis, 1995). Classification accuracy refers to the extent to which test takers are correctly classified into proficiency categories based on their observed scores, as compared to the categories they would be placed in if their true proficiencies were known. In other words, it quantifies how often an examinee's observed level (for example, "Meets CEFR Level A1") matches their true level of English proficiency.

To estimate classification decision accuracy, we followed the methodology of Livingston and Lewis (1995). This approach uses information about the observed score distribution, the score scale's cutoff scores for each proficiency level, and the test's reliability coefficient to estimate the likelihood of accurate classifications. In this study, estimates were obtained using the ETS program RELCLASS-COMP (Version 4.14). For each proficiency level threshold (e.g., the cut score between CEFR levels A1 and A2), a two-by-two contingency table was constructed, comparing each examinee's observed classification (whether the examinee's score *reaches* the threshold for that level) with the examinee's true classification (whether their true score would place them at or above that level). Table 3 illustrates these two-by-two contingency tables. The cells on the main diagonal of the table (labeled "consistent classification" in Table 3) represent

cases where the observed classification agrees with the true classification (i.e., consistent classification). The off-diagonal cells (labeled “inconsistent classification”) indicate misclassification relative to the examinee’s true level. The proportion of correctly classified examinees at a given threshold is computed by summing the proportions in the two consistent (diagonal) cells. This proportion can be interpreted as the probability that an examinee will be placed into the correct category on the basis of a single test administration.

Table 3. Decision Accuracy at Each Classification Level

Classification Level Status	Does Not Reach a Classification Level Based on True Score	Reaches a Classification Level Based on True Score
Does not reach a classification level	Consistent classification	Inconsistent classification
Reaches a classification level	Inconsistent classification	Consistent classification

There are no rigid cutoffs in the field for what constitutes “acceptable” classification accuracy; the required precision of classifications depends on the testing context and purposes. However, in practice, the following general benchmarks are often cited:

- $\geq .85$: High classification accuracy (strong support for decision-making)
- $.80 - .84$: Acceptable accuracy for most educational uses
- $.70 - .79$: Moderate accuracy; caution warranted, especially near cut scores
- $< .70$: Low accuracy; substantial misclassification risk

These criteria offer a rough heuristic — in any specific use case, the tolerance for misclassification error should be weighed against the stakes and purposes of the test (Livingston & Lewis, 1995).

Interrater Reliability for Speaking Test

The TOEFL Primary® Speaking test is scored by human raters. Although each student’s Speaking responses are typically scored by a single rater in operational testing, a system of ongoing *double-scoring* is used for quality control. In the double-scoring sample (about 5% of Speaking responses), two trained raters independently score the same responses. We used these double-scored responses to evaluate *interrater reliability*, which reflects the consistency

of scoring between different raters. This interrater reliability is described in three ways: percentage agreement, Quadratic-weighted Kappa, and Correlation (Pearson's r). Each of these indices provides insight into a different aspect of rater consistency, as described below.

Percentage Agreement

Percentage agreement between two raters is frequently defined as the percentage of exact score agreement and adjacent score agreement. The percentage of exact score agreement tends to decrease with increasing numbers of item score points. The fewer the item score points, the fewer degrees of freedom on which two raters can vary, and the higher the percentage of agreement. Adjacent agreement (± 1 point) refers to the percentage of ratings in which two raters differ by no more than one score point.

Quadratic Weighted Kappa

Quadratic-weighted Kappa takes into account the degree of agreement between raters where large disagreements are weighted more heavily than small disagreements. The range of the quadratic weighted kappa is from 0.0 to 1.0, with perfect agreement being equal to 1.0 (Cohen, 1968). For a human-scored item with m categories ($m=1,2,\dots,m$), one can construct an $m \times m$ rating table with scores provided by two raters noted as A and B. Suppose m is the maximum obtainable score for each item, n_{ij} is the number of responses for which rater A's score = i and rater B's score = j , n_{i+} is the number of responses for which rater A = i , n_{+j} is the number of responses for which rater B = j , and n_{++} is the number of all responses from either rater A or rater B. The quadratic weighted kappa coefficient is defined as (Cohen, 1968):

$$\kappa_{ij} = \frac{\left(\sum_{i=0}^m \sum_{j=0}^m w_{ij} \frac{n_{ij}}{n_{++}} \right) - \left(\sum_{i=0}^m \sum_{j=0}^m w_{ij} \frac{n_{i+} n_{+j}}{n_{++}^2} \right)}{1 - \left(\sum_{i=0}^m \sum_{j=0}^m w_{ij} \frac{n_{i+} n_{+j}}{n_{++}^2} \right)} \quad (2)$$

For quadratic weighted kappa, the weights are

$$w_{ij} = 1 - \frac{(i-j)^2}{m^2} \quad (3)$$

In practice, quadratic weighted kappa often yields values similar to the intraclass correlation for two ratings. It is widely used for evaluating rater agreement on ordinal scales.

Correlation

Interrater reliability or consistency is an indicator of whether test takers are ranked in the same way and is also frequently measured using correlation between two sets of ratings provided by different raters. As a general guideline in practice, whole-test correlations of .70 or higher are considered acceptable, while values of .80 or above indicate strong consistency.

Data and Sample Characteristics

The data were drawn from a representative operational form of the TOEFL Primary® tests. Table 4 summarizes the key descriptive statistics for each test section, including the sample size (N), mean observed score, standard deviation (SD), minimum score, and maximum score. As shown in the table, the sample sizes for the Reading and Listening sections were substantial (over 26,000 for Step 1 and around 17,000 for Step 2), and the Speaking section included data from 8,744 test takers. The mean scores reflect mid-range performance on the Step 1 and Step 2 scales (which range up to 109 and 115, respectively) and on the Speaking scale (which ranges from 0 to 27).

Table 4. Summary Statistics for Section Scale Scores

	Test Section	N	Mean	SD	Minimum	Maximum
Step 1	Reading	26,488	105.67	2.44	100	109
	Listening	26,453	105.21	2.36	100	109
Step 2	Reading	16,914	108.03	3.88	100	115
	Listening	16,917	107.48	3.86	100	115
Speaking	Speaking	8,744	15.48	5.55	0	27

In addition to test scores, demographic information was examined to characterize the test-taking populations. For the Step 1 Reading and Listening sample, 53.3% of test takers were female and 46.7% were male. Approximately one-third (31.2%) of these test takers were 10 years old, and the remaining 68.8% were between 8 and 14 years old. The largest national groups in the Step 1 sample were from Japan (26.8% of test takers), Turkey (24.3%), Brazil (23.5%), and Korea (14.2%). The Step 2 Reading and Listening sample had a similar gender

distribution (52.3% female, 47.7% male) and age distribution (33.6% were 10 years old; 66.4% were 8–14 years old). The Step 2 test takers primarily came from Vietnam (57.9%), with substantial representation from Korea (11.5%) and Jordan (10.6%). The Speaking test sample was 48.8% female and 51.3% male; about 27.1% of examinees were 10 years old and 72.9% were between 8 and 14. The majority of the Speaking test takers were from Turkey (73.1%), followed by a smaller group from Mexico (10.5%), with the rest distributed across various other countries. This demographic information indicates that the samples for each test level were large and diverse, providing a robust basis for the reliability and validity analyses presented in this memorandum.

Results

Score Reliability

The Cronbach's Alpha estimates of the 5 TOEFL Primary scale scores based on data samples as described above are estimated and shown in Table 5. Cronbach's Alpha estimates were consistently high, with values of .83 and .87 for Reading (Step 1 and Step 2, respectively), .85 and .89 for Listening (Step 1 and Step 2), and .90 for Speaking.

Table 5. Reliability Estimates of the TOEFL Primary® Test Scores

Test	Scale score range	Reliability
Reading – Step 1	101–109	.83
Reading – Step 2	104–115	.87
Listening – Step 1	101–109	.85
Listening – Step 2	104–115	.89
Speaking	1–27	.90

The standard error of measurement for each of the TOEFL Primary® test scores is presented in Table 6. SEM values ranged from 1.19 to 1.50 for Reading, 1.22 to 1.36 for Listening, and 1.61 for Speaking.

Table 6. Standard Error of Measurement for the TOEFL Primary® Test Scores

Test	Scale score range	Standard Error of Measure
Reading – Step 1	101–109	1.50
Reading – Step 2	104–115	1.19
Listening – Step 1	101–109	1.36
Listening – Step 2	104–115	1.22
Speaking	1–27	1.61

Classification Decision Accuracy

Reading and Listening Step 1 Tests

Table 7 and Table 8 summarize the classification decision accuracy for CEFR levels of Step 1 tests for Reading and Listening sections respectively. The classification accuracy for Reading CEFR levels is .82, and the classification accuracy for Listening CEFR levels is also 0.82.

Table 7. Classification Decision Accuracy of Reading CEFR Levels—Step 1

		True Score				
		Band Level	A2 (107-109)	A1 (102-106)	Below A1 (100-101)	Total
Observed Score	A2 (107-109)		0.35	0.06	0.00	0.41
	A1 (102-106)		0.07	0.45	0.02	0.54
	Below A1 (100-101)		0.00	0.02	0.02	0.05
	Total		0.42	0.54	0.04	

Note. Estimated Proportion Correctly Classified: Total = 0.82

Table 8. Classification Decision Accuracy of Listening CEFR Levels—Step 1

		True Score				
		Band Level	A2 (105-109)	A1 (102-104)	Below A1 (100-101)	Total
Observed Score	A2 (105-109)	0.54	0.04	0.00	0.58	
	A1 (102-104)	0.09	0.25	0.02	0.35	
	Below A1 (100-101)	0.00	0.04	0.03	0.07	
	Total	0.63	0.32	0.05		

Note. Estimated Proportion Correctly Classified: Total = 0.82

Reading and Listening Step 2 Tests

Table 9 and Table 10 summarize the classification decision accuracy of CEFR levels of Step 2 tests for Reading and Listening sections respectively. The classification decision accuracy for Reading CEFR levels is 0.82, and the classification decision accuracy for Listening CEFR levels is 0.78.

Table 9 Classification Decision Accuracy of Reading CEFR Levels—Step 2

		True Score				Total
Band Level		B1 (111-115)	A2 (107-110)	A1 (104-106)	Below A1 (100)	
Observed Score	B1 (111-115)	0.47	0.05	0.00	0.00	0.52
	A2 (107-110)	0.05	0.24	0.03	0.00	0.32
	A1 (104-106)	0.00	0.04	0.11	0.00	0.16
	Below A1 (100)	0.00	0.00	0.00	0.00	0.00
	Total	0.52	0.34	0.14	0.00	

Note. Estimated Proportion Correctly Classified: Total = 0.82

Table 10 Classification Decision Accuracy of Listening CEFR Levels—Step 2

		True Score				Total
Band Level		B1 (112-115)	A2 (105-111)	A1 (104)	Below A1 (100)	
Observed Score	B1 (112-115)	0.27	0.05	0.00	0.00	0.32
	A2 (105-111)	0.04	0.53	0.02	0.00	0.59
	A1 (104)	0.00	0.03	0.06	0.00	0.09
	Below A1 (100)	0.00	0.00	0.00	0.00	0.00
	Total	0.31	0.61	0.08	0.00	

Note. Estimated Proportion Correctly Classified: Total = 0.78

Speaking Test

Table 11 present the classification decision accuracy of CEFR levels respectively for TOEFL Primary® Speaking Test. The classification decision accuracy for the classification decision accuracy for the CEFR levels is .76.

Table 11 Classification Decision Accuracy of Speaking CEFR Levels

		<i>True Score</i>					Total
CEFR		B2 (26-27)	B1 (22-25)	A2 (16-21)	A1 (10-15)	Below A1 (1-9)	
Observed Score	B2 (26-27)	0.00	0.00	0.00	0.00	0.00	0.01
	B1 (22-25)	0.00	0.09	0.03	0.00	0.00	0.12
	A2 (16-21)	0.01	0.04	0.31	0.07	0.02	0.44
	A1 (10-15)	0.00	0.00	0.03	0.23	0.00	0.26
	Below A1 (1-9)	0.00	0.00	0.00	0.04	0.13	0.18
	Total	0.01	0.14	0.38	0.34	0.15	

Note. Estimated Proportion Correctly Classified: Total = 0.76

Classification decision accuracy was generally high across the TOEFL Primary tests, with rates around .82 for Step 1 Reading and Listening, .82 for Step 2 Reading, slightly lower at .78 for Step 2 Listening, and .76 for Speaking.

Interrater Reliability for the Speaking Test

Table 12 summarizes the means and standard deviations for Rater 1 and Rater 2, along with indices of inter-rater agreement, including percentage of exact agreement, percentage of adjacent agreement, Pearson correlations, and quadratic weighted kappa. Across all speaking tasks, the mean score differences between raters were minimal (≤ 0.04). Inter-rater correlations ranged from .68 to .88, adjacent agreement exceeded 96% for all items, and quadratic weighted kappa values ranged from .68 to .88, indicating moderate to strong agreement.

Overall, the results suggest that the rating process demonstrates a high level of reliability, particularly when considering adjacent agreement and weighted kappa. Items with broader score scales (0–5) showed slightly lower exact agreement than those with narrower scales (0–3), which is expected given the increased number of response categories. At the item level, Item 5 exhibited the lowest exact agreement, likely reflecting greater score variability associated with the wider scale. Item 6 showed the lowest quadratic weighted kappa value, suggesting relatively lower consistency in rating patterns; however, given the high adjacent agreement, most disagreements between raters were minor and confined to adjacent score categories.

Table 12. Rater 1 and Rater 2 Item Ratings and Inter-Rater Reliability for the Double-Scoring Sample

Item	Rating Scale	Double Scoring N	Rater 1		Rater 2		% Exact Agreement	% Adjacent Agreement	Correlations (R1,R2)	Weighted Kappa
			Mean	SD	Mean	SD				
1	0-5	434	2.76	1.27	2.72	1.25	57%	97%	0.84	0.84
2	0-3	388	1.95	1.01	1.95	1.02	76%	100%	0.88	0.88
3	0-5	399	2.58	1.04	2.58	1.06	58%	96%	0.76	0.76
4	0-3	437	1.93	0.77	1.89	0.79	70%	100%	0.75	0.75
5	0-5	430	2.65	1.18	2.64	1.12	51%	97%	0.78	0.78
6	0-3	407	1.68	0.76	1.66	0.73	65%	100%	0.68	0.68
7	0-3	429	2.11	0.80	2.08	0.81	67%	100%	0.74	0.73

Note. Double-scored responses were randomly selected at task level.

Discussion

The results indicate that TOEFL Primary test scores demonstrate strong reliability across all sections. Cronbach's Alpha estimates suggest a high degree of internal consistency, indicating that the test items within each section reliably measure the intended constructs.

Measurement precision, as reflected by the standard error of measurement (SEM), was also found to be acceptable across all test sections. SEM values are small relative to the score range between the CEFR level cuts. These relatively small SEM values indicate that observed scores are close approximations of test takers' true scores, supporting claims regarding the precision and stability of the reported scores.

Classification decision accuracy analyses showed that the proportion of correctly classified test takers was generally high. For Step 1 Reading and Listening, classification

accuracy was approximately .82. For Step 2, accuracy remained at .82 for Reading and was slightly lower at .78 for Listening. The Speaking test demonstrated a classification accuracy of .76. Overall, these results suggest that classification into CEFR levels is reasonably accurate, particularly for Reading and Listening. The somewhat lower accuracy observed for Speaking likely reflects the inherent variability associated with human scoring and the complexity of performance-based tasks.

Interrater reliability for the Speaking test further supports the consistency of scoring. Correlations between raters ranged from .68 to .88, and quadratic weighted kappa values fell within the same range, indicating moderate to strong agreement. In addition, adjacent agreement exceeded 96% across all items, suggesting that most discrepancies between raters were minor and limited to adjacent score categories rather than substantial differences.

Finally, patterns across test sections reveal meaningful differences between selected-response and constructed-response formats. Reading and Listening, which are selected-response assessments, exhibited slightly higher classification accuracy. In contrast, the Speaking test, a constructed-response assessment, demonstrated the highest reliability coefficient ($\alpha = .90$) but somewhat lower classification accuracy. This latter result is expected given the higher number of score points in Speaking, which leads to a higher Cronbach's Alpha estimate, and the additional variability introduced by human scoring along with the more complex nature of spoken language performance, which likely lowers the classification accuracy.

Overall, the evidence from reliability, measurement precision, classification accuracy, and interrater agreement provides a coherent and complementary picture of test quality. These results support the interpretation that TOEFL Primary scores are sufficiently consistent and precise for reporting CEFR-aligned proficiency levels and for informing educational decisions about young learners. While some variability is observed in performance-based scoring, it is well controlled and does not substantially undermine the overall dependability of score-based classifications.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.
- Brennan, R. L. (2001). *Generalizability theory*. Springer.
<https://psycnet.apa.org/doi/10.1007/978-1-4757-3456-0>
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.
<https://psycnet.apa.org/doi/10.1037/h0026256>
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16802fc1bf>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Educational Testing Service. (2014). *ETS standards for quality and fairness*.
<https://www.ets.org/pdfs/about/standards-quality-fairness.pdf>
- Educational Testing Service. (2023). *TOEFL Primary® test taker handbook*.
<https://www.ets.org/pdfs/toefl/toefl-primary-handbook.pdf>
- Educational Testing Service. (2025). *Understanding your Speaking score report*.
<https://www.ets.org/pdfs/toefl/toefl-primary-understand-speaking-score-reports.pdf>
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, 32(2), 179–197.
<https://doi.org/10.1111/j.1745-3984.1995.tb00462.x>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Notes

¹ The CEFR is a widely used international standard for describing language proficiency across six levels (Council of Europe, 2001): A1–A2 (Beginner): Basic ability to understand and use simple expressions and everyday phrases; B1–B2 (Intermediate): Ability to communicate independently in familiar situations and handle more complex language; and C1–C2 (Advanced/Proficient): Ability to use language fluently, accurately, and effectively in a wide range of contexts.

Copyright © 2026 by Educational Testing Service. All rights reserved.

ETS, TOEFL, TOEFL Primary, and the ETS logo are registered trademarks of Educational Testing Service (ETS). All other trademarks are the property of their respective owners.