

Score comparability study between TOEIC® Listening and Reading Test Designs

Full-length linear and multistage adaptive
computer-based tests

Jaime Cid, Youhua Wei, and Feng Yu | *ETS, USA*
Chihiro Sampei | *IIBC, Japan*

August 2026

TC-26-02

Introduction

The TOEIC® Listening and Reading (TOEIC L&R) test is a standardized global English-language proficiency test for people whose native language is not English. It measures the English communication skills required in the workplace and everyday life. The TOEIC L&R test has typically been administered in a full-length linear paper-and-pencil or computer-based format (hereinafter referred to as “Linear Test”) which includes 100 items each in the Listening and Reading sections, and the total testing time is approximately two hours. The TOEIC® Listening and Reading (TOEIC L&R) Multistage Adaptive Test version (hereinafter referred to as “MSA Test”) has been operationally administered alongside with the Linear Test since 2020. The reported scores from the two versions of TOEIC L&R test are considered comparable for several reasons: a) both are built on the same framework of test design and content specifications; b) the MSA Test draws its items from an item pool in which all items have been psychometrically analyzed using an Item Response Theory (IRT) model, based on the data collected from Linear Test administrations, ensuring that all items from both test versions are indexed on a common IRT scale); and c) the reported scores from the MSA Test and the Linear Test are equated based on the same IRT scale. This study was conducted to further confirm score comparability between the MSA Test and the Linear Test.

Research design

This study examined score equivalence between the MSA Test and the Linear Test by comparing the observed scores from both test versions. Data collection followed a single group counterbalanced design. The single group aspect means that the same group of test takers completed both the MSA Test and the Linear Test, so that each test taker’s performance across the tests could be directly compared. The counterbalanced aspect was intended to mitigate potential effects of test order, such as practice or fatigue, on performance. In this study, participants were divided approximately in half: subgroup *M* took the MSA Test first, followed by the Linear Test; subgroup *L* took the Linear Test first, followed by the MSA Test. All test takers within each subgroup completed both tests in the order described above on the same day (March 2024), with a three-hour break in between. A single test form of each test version was used, meaning all participants, regardless of subgroup, took the same set of test forms.

Administering both test versions on the same day with a three-hour break involved important design tradeoffs. Conducting the two administrations in a single session reduced logistical burden and ensured that the same pool of participants contributed scores to both test versions, strengthening the within-person comparability of observed scores. In addition, the three-hour interval was intended to provide rest and reduce immediate carryover effects. However, same-day testing may still introduce some effects (e.g., fatigue, motivation) that can differentially affect performance on the second test, despite the counterbalanced order. Administering the tests on different days could have further reduced same-day fatigue and carryover, but it was not selected because it was expected to result in high participant attrition between sessions and thus compromise the single-group design.

Data collection

To ensure coverage across the full range of English proficiency on the TOEIC L&R score scale, scores from recent TOEIC L&R public test administrations in Japan (August 2023 to January 2024) were used as a proficiency proxy (hereinafter referred to as “proxy scores”) for recruiting research participants. The recruited participants were then assigned to two subgroups (*M* or *L*) based on their proxy scores, to make the proficiency distributions of the two groups as similar as possible. In this study, the proxy scores from recent administrations were used as approximate estimates of test takers’ proficiency levels to recruit and assign test takers at different ability levels.

A total of 970 individuals participated in the study, and 954 of them completed both the MSA Test and the Linear Test, as per the study design. During data cleaning, a few participant cases were identified as outliers due to potential validity issues, such as unreasonable test taking time, unusual score patterns, or low motivation to take the tests, either on Listening or Reading of the MSA or Linear Tests. These outliers were excluded from the final statistical analyses (three for the Listening score analyses, another three for the Reading score analyses, and all of these six for the total score analyses). Consequently, the final dataset includes 951 participants with Listening scores, 951 with Reading scores, and 948 with Total scores. The final sample sizes for the Listening, Reading, and Total score analyses are presented in Tables 1, 2, and 3, respectively. The tables also present subgroup summary statistics by proxy total score range. As shown in the tables, Subgroup *M* and Subgroup *L* are highly comparable in terms of average performance on the TOEIC L&R test and overall score-distribution patterns (Note: All numerical values presented in this report have been rounded to the nearest digit, one decimal place lower than the displayed precision.)

Table 1
Summary Statistics of Proxy Scores Based on Listening Score Analysis Sample

Proxy Total Score Range	Frequency (%)			Mean Proxy Listening Score (<i>SD</i>)		
	Total Group	Subgroup <i>M</i>	Subgroup <i>L</i>	Total Group	Subgroup <i>M</i>	Subgroup <i>L</i>
10–340	166 (17%)	78 (17%)	88 (18%)	181 (33)	185 (29)	177 (35)
345–445	197 (21%)	95 (21%)	102 (21%)	234 (32)	234 (32)	234 (33)
450–575	197 (21%)	93 (20%)	104 (21%)	290 (33)	290 (32)	291 (34)
580–785	191 (20%)	95 (21%)	96 (20%)	369 (37)	371 (35)	367 (38)
790–990	200 (21%)	99 (22%)	101 (21%)	463 (29)	464 (28)	462 (29)
Total Group	951 (100%)	460 (100%)	491 (100%)	312 (104)	315 (104)	309 (105)

Table 2*Summary Statistics of Proxy Scores Based on Reading Score Analysis Sample*

Proxy Total Score Range	Frequency (%)			Mean Proxy Reading Score (SD)		
	Total Group	Subgroup M	Subgroup L	Total Group	Subgroup M	Subgroup L
10–340	167 (18%)	79 (17%)	88 (18%)	119 (26)	115 (25)	122 (26)
345–445	197 (21%)	95 (21%)	102 (21%)	159 (31)	160 (30)	159 (32)
450–575	197 (21%)	93 (20%)	104 (21%)	220 (36)	219 (37)	221 (36)
580–785	190 (20%)	94 (20%)	96 (20%)	314 (47)	316 (50)	311 (42)
790–990	200 (21%)	98 (21%)	102 (21%)	429 (42)	430 (44)	427 (39)
Total Group	951 (100%)	459 (100%)	492 (100%)	252 (117)	254 (120)	251 (115)

Table 3*Summary Statistics of Proxy Scores Based on Total Score Analysis Sample*

Proxy Total Score Range	Frequency (%)			Mean Proxy Total Score (SD)		
	Total Group	Subgroup M	Subgroup L	Total Group	Subgroup M	Subgroup L
10–340	166 (18%)	78 (17%)	88 (18%)	299 (36)	300 (35)	299 (36)
345–445	197 (21%)	95 (21%)	102 (21%)	393 (31)	393 (31)	394 (31)
450–575	196 (21%)	92 (20%)	104 (21%)	511 (38)	510 (38)	512 (37)
580–785	190 (20%)	94 (21%)	96 (20%)	683 (60)	688 (60)	678 (60)
790–990	199 (21%)	98 (21%)	101 (21%)	892 (62)	895 (64)	889 (60)
Total Group	948 (100%)	457 (100%)	491 (100%)	564 (216)	569 (218)	559 (214)

Analyses and results

Score comparability between the MSA Test and the Linear Test was evaluated from the following perspectives: a) overall mean score, b) score distribution, c) correlation of scores between the MSA Test and the Linear Test, and d) mean score differences between the two test versions by proficiency level group.

Overall mean score

Tables 4, 5, and 6 present summary statistics by test version and subgroup for Listening, Reading, and Total scores, respectively. “MSA – Linear” denotes the score difference between the MSA Test and the Linear Test, calculated by subtracting the Linear Test score from the MSA Test score.

Table 4

Summary Statistics of MSA and Linear Test Listening Scores

Group	Frequency	Mean Score (SD)		
		MSA	Linear	MSA – Linear
Subgroup <i>M</i>	460	325 (102)	332 (97)	–7 (45)
Subgroup <i>L</i>	491	327 (102)	325 (95)	2 (42)
Total Group	951	326 (102)	328 (96)	–2 (44)

Table 5

Summary Statistics of MSA and Linear Test Reading Scores

Group	Frequency	Mean Score (SD)		
		MSA	Linear	MSA – Linear
Subgroup <i>M</i>	459	259 (108)	264 (118)	–5 (46)
Subgroup <i>L</i>	492	260 (111)	257 (115)	2 (44)
Total Group	951	259 (110)	260 (117)	–1 (45)

Table 6*Summary Statistics of MSA and Linear Test Total Scores*

Group	Frequency	Mean Score (SD)		
		MSA	Linear	MSA – Linear
Subgroup <i>M</i>	457	584 (201)	596 (208)	–12 (67)
Subgroup <i>L</i>	491	587 (203)	582 (205)	5 (63)
Total Group	948	586 (202)	589 (206)	–3 (65)

As shown in the three tables, the overall mean scores for the total sample were very similar between the MSA Test and the Linear Test, with mean score differences of 2 points, 1 point, and 3 points for the Listening, Reading, and Total scores, respectively. These score differences were negligible, indicating that the two test versions are highly comparable from an average score perspective.

Further analysis for each subgroup revealed that the second administered test consistently produced higher average scores than the first, regardless of which test version was taken first. Subgroup *M*, which took the MSA Test first, had higher mean scores on the Linear Test for both the Listening and Reading sections (by 7 points and 5 points, respectively). Likewise, Subgroup *L*, which took the Linear Test first, had higher mean scores on the MSA Test for both the Listening and Reading sections (by 2 points each). This consistent pattern suggested a potential test-taking order effect that contributed to enhanced performance on the second administered test, i.e., the effect of practice and familiarity on the test. Nevertheless, on average, the order effect was small and negligible. Therefore, the subsequent comparability analyses do not account for test-taking sequence.

Score distribution

Table 7 presents summary statistics for score distributions by test version (the MSA Test and the Linear Test) and section (Listening, Reading, and Total): minimum score (Min), 25th, 50th, and 75th percentile scores, and maximum score (Max). As shown in the results, the two test versions had closely aligned distributions of Listening, Reading, and Total scores. This finding further supports the score comparability between the two test versions, as suggested by the mean score comparison (see the previous section).

Table 7
Percentiles of MSA and Linear Test Scores

Percentile	Listening Score		Reading Score		Total Score	
	MSA	Linear	MSA	Linear	MSA	Linear
Min	75	75	55	60	165	205
25 th (25%)	245	255	175	155	430	415
50 th (50%)	315	320	250	245	565	565
75 th (75%)	410	405	350	355	755	760
Max	495	495	495	495	990	990

Score correlation and consistency

Correlation coefficients and scatterplots were used to further evaluate score comparability between the MSA Test and the Linear Test. Based on the full study sample, the observed correlations between the two test versions were 0.90 for Listening scores, 0.92 for Reading scores, and 0.95 for Total scores, each indicating a strong positive correlation. Figures 1, 2, and 3 present scatterplots for the Listening section, Reading section, and the entire test, respectively. Across all three scatterplots, most test takers’ scores cluster closely along the identity line (i.e., the diagonal line representing equal scores on both tests), suggesting a high level of agreement between scores from the two test versions.

Figure 1
Scatterplot of Listening Scores: MSA Test vs. Linear Test

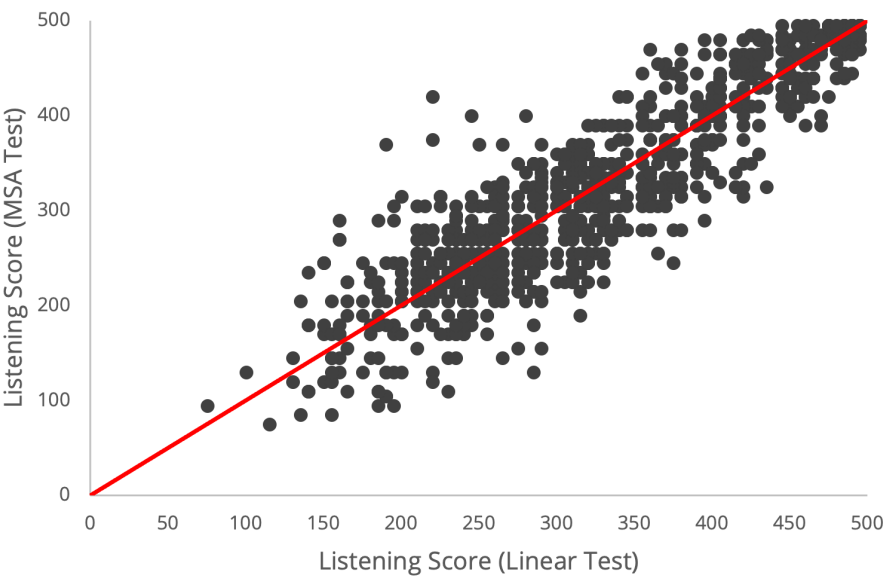
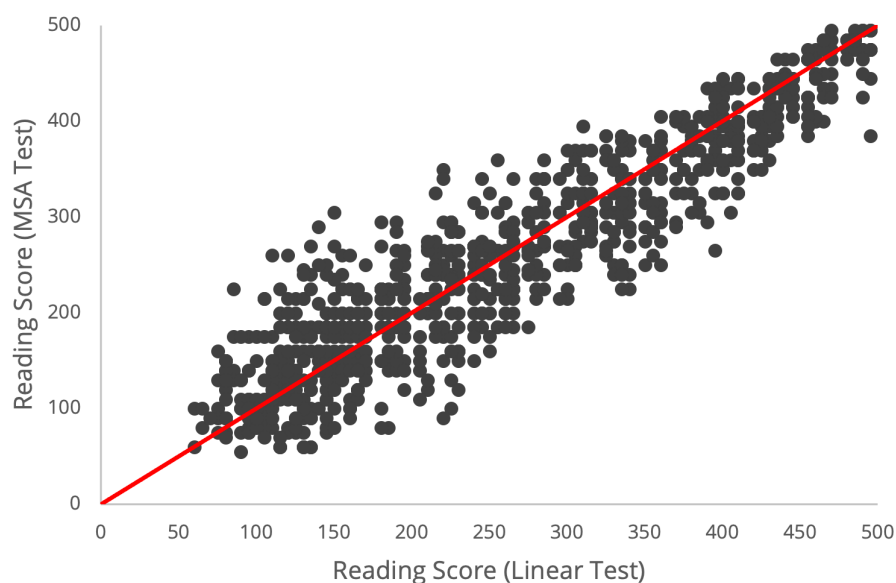
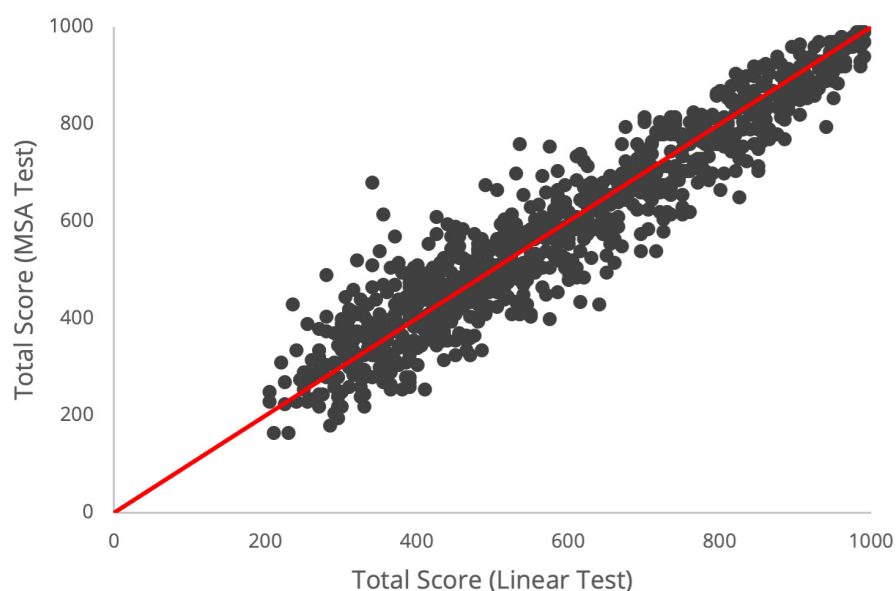


Figure 2*Scatterplot of Reading Scores: MSA Test vs. Linear Test***Figure 3***Scatterplot of Total Scores: MSA Test vs. Linear Test*

Tables 8 and 9 present the frequency counts and percentages of test takers by MSA–Linear score differences for the Listening and Reading sections, respectively. The test takers were classified based on the standard error of measurement for repeater test takers' score differences (SE_{diff}). Technically speaking, test takers are expected to obtain slightly different scores when taking the TOEIC L&R test repeatedly. If other factors such as change in ability, motivation, fatigue, and practice are not considered, repeaters'

score differences should follow a normal distribution with SE_{diff} . That is, approximately 68% should fall within one SE_{diff} and 95% within two SE_{diff} . The SE_{diff} between the MSA Test and the Linear Test is 40 for both the Listening and Reading sections. In the present study, for the Listening section, 70.6% and 93.7% of participants fell within one and two SE_{diff} , respectively (Table 8). A similar pattern was observed for the Reading section: 66.2% fell within one SE_{diff} and 92.6% within two SE_{diff} (Table 9). These findings indicate that, for the participant group, the score differences between the MSA Test and the Linear Test aligned closely with the theoretically expected normal distribution.

Table 8

Distribution of MSA–Linear Listening Score Differences Based on SE_{diff}

MSA–Linear Listening Score Differences	Frequency	%
Within 1 SE_{diff} (Score Difference ≤ 40)	671	70.6%
Within 2 SE_{diff} (Score Difference ≤ 80)	891	93.7%
Total Group	951	100.0%

Table 9

Distribution of MSA–Linear Reading Score Differences Based on SE_{diff}

MSA–Linear Reading Score Differences	Frequency	%
Within 1 SE_{diff} (Score Difference ≤ 40)	630	66.2%
Within 2 SE_{diff} (Score Difference ≤ 80)	881	92.6%
Total Group	951	100.0%

Score comparison by proficiency level

Score comparability across proficiency levels was also analyzed to evaluate whether the two test versions yield comparable performance throughout the full range of TOEIC L&R scores. Table 10 presents the mean Listening scores for the MSA Test and the Linear Test by proxy Listening score range, along with the mean score differences (MSA – Linear: MSA minus Linear). For example, 335 study participants had proxy Listening scores between 200 and 295; their average Listening scores were 265 on the MSA Test and 272 on the Linear Test. The mean difference was 7, which was considered negligible. Similarly, small score differences were observed across all proxy score range groups except the lowest range group (5–95), which showed a mean difference of 63. Given that only two participants were in this group, further evaluation with a larger sample is warranted. Collectively, the results indicate that the MSA Test and the Linear Test produce comparable Listening scores across a wide range of proficiency levels.

Table 10*MSA–Linear Listening Score Comparison by Proxy Listening Score Range*

Proxy Listening Score Range	Frequency (%)	Mean Listening Score (<i>SD</i>)		
		MSA	Linear	MSA – Linear
5–95	2 (0.2%)	165 (60)	228 (38)	–63 (23)
100–195	139 (14.6%)	218 (59)	221 (47)	–2 (50)
200–295	335 (35.2%)	265 (56)	272 (51)	–7 (47)
300–395	234 (24.6%)	350 (60)	351 (59)	–1 (42)
400–495	241 (25.3%)	451 (51)	446 (52)	5 (35)
Total Group	951 (100.0%)	326 (102)	328 (96)	–2 (44)

Likewise, Table 11 presents the mean Reading scores for the MSA Test and the Linear Test, along with the mean differences, by proxy Reading score range. Table 12 summarizes the corresponding results for the Total scores. Although the average score differences were relatively larger in certain score ranges, they remained well within the standard error of measurement for repeaters' score differences (SE_{diff}) between the MSA Test and the Linear Test. Nevertheless, further analysis with a larger sample would be preferable to investigate score differences across proficiency levels.

Table 11*MSA–Linear Reading Score Comparison by Proxy Reading Score Range*

Proxy Reading Score Range	Frequency (%)	Mean Reading Score (<i>SD</i>)		
		MSA	Linear	MSA – Linear
5–95	36 (3.8%)	131 (59)	134 (51)	–3 (49)
100–195	353 (37.1%)	171 (58)	162 (53)	10 (47)
200–295	237 (24.9%)	254 (62)	252 (65)	2 (47)
300–395	168 (17.7%)	329 (57)	347 (61)	–18 (44)
400–495	157 (16.5%)	420 (55)	432 (50)	–12 (29)
Total Group	951 (100.0%)	259 (110)	260 (117)	–1 (45)

Table 12*MSA-Linear Total Score Comparison by Proxy Total Score Range*

Proxy Total Score Range	Frequency (%)	Mean Total Score (SD)		
		MSA	Linear	MSA – Linear
5–95	0 (0%)	-	-	-
100–195	3 (0.3%)	352 (113)	355 (32)	-3 (84)
200–295	65 (6.9%)	344 (86)	340 (74)	4 (63)
300–395	208 (21.9%)	403 (91)	399 (82)	4 (69)
400–495	164 (17.3%)	479 (95)	474 (93)	4 (75)
500–595	137 (14.5%)	562 (82)	569 (85)	-7 (68)
600–695	93 (9.8%)	670 (91)	676 (101)	-6 (66)
700–795	91 (9.6%)	740 (94)	760 (90)	-20 (59)
800–895	90 (9.5%)	820 (70)	829 (72)	-8 (60)
900–990	97 (10.2%)	921 (76)	927 (73)	-6 (39)
Total Group	948 (100.0%)	586 (202)	589 (206)	-3 (65)

Overall, for the Listening, Reading, and Total scores, the average differences between the MSA Test and the Linear Test were consistently small across different proficiency ranges. Moreover, there was no consistent pattern indicating that either test version produced higher or lower scores at any given proficiency level. These findings indicate consistent score comparability between the MSA Test and the Linear Test throughout the full range of TOEIC L&R scores.

Conclusion

The TOEIC L&R MSA Test was developed to enhance test-taking efficiency by reducing testing time while maintaining measurement quality comparable to the full-length Linear Test. Empirical evidence from the comparability study confirms that scores from the MSA Test and Linear Test are highly comparable and strongly correlated across all proficiency levels. Specifically, using a single-group counterbalanced design, this study demonstrated that scores from the MSA Test and the Linear Test were highly consistent and strongly correlated. Across all analyses, including means, score distributions, correlations, and measures of consistency, the two versions exhibited a high degree of score comparability. Both tests also showed consistent performance across various proficiency levels.

In conclusion, the evidence affirms that the MSA Test and the Linear Test offer comparable reliability in measuring English language proficiency within the TOEIC L&R framework. At the same time, the study provides strong support for the design of the MSA Test, confirming that it yields scores highly comparable to those of the Linear Test while substantially improving measurement efficiency through reduced testing time.

