Examining Score Comparability of the Online and Paper-Based TOEIC L&R

Masaya Kanzaki
Kanda University of International Studies

Reference Data:

Kanzaki, M. (2025). Examining score comparability of the online and paper-based TOEIC L&R. In B. Lacy, M. Swanson, & P. Lege (Eds.), *Moving JALT into the Future: Opportunity, Diversity, and Excellence*. JALT. https://doi.org/10.37546/JALTPCP2024-09

With the introduction of the online version of the Test of English for International Communication Listening and Reading (TOEIC L&R) in April 2020, concerns emerged regarding the comparability of scores between the online and traditional paper-based formats. This study investigated the extent of score comparability between the two versions. Two online tests and two paper-based tests were administered to 127 university students over a five-week period, and their scores were analyzed. Score differences between the two formats were most pronounced in the listening section and least pronounced in the reading section, with total scores falling in between. Although statistically significant differences were found in total scores across formats, their practical significance is debatable, as the effect sizes were small. This paper supports the use of the online version, which offers several advantages over the paper-based format.

2020年4月にTest of English for International Communication Listening and Reading（TOEIC L&R）のオンライン版が導入されたことにより、オンライン版と従来の紙版との得点の同等性に対する懸念が生じた。本研究では、両形式間の得点の同等性の程度を検証した。127名の大学生に対し、5週間の期間内にオンライン版テスト2回と紙版テスト2回を実施し、その得点を分析した。その結果、リスニングでは2形式間の得点差が最も大きく、リーディングでは最も小さく、合計点はその中間であった。合計点には形式間で統計的な有意差があったが、効果量は小さく、実用的な意義については議論の余地がある。本稿では、紙版と比べていくつかの利点を持つオンライン版の活用を支持する。

The Test of English for International Communication Listening and Reading (TOEIC L&R) is a major English proficiency assessment in Japan that is widely used in both business and academic settings. In 2023, the total annual number of test administrations exceeded 1.76 million (IIBC, n.d.-a). Many Japanese companies use TOEIC L&R scores for recruitment, promotion, and internal benchmarking, often setting specific score requirements (IIBC, n.d.-b). Universities and other academic institutions also use the TOEIC L&R for admissions, course placement, and graduation requirements (IIBC, n.d.-c). Its popularity stems from its focus on practical, workplace-related English, aligning well with Japan's emphasis on English for professional communication. As a result, achieving a high TOEIC score is widely regarded as an asset that can significantly enhance an individual's employment prospects and career advancement opportunities.

In April 2020, the Institute for International Business Communication (IIBC), the Japanese operator of the TOEIC program, introduced an online version of the TOEIC L&R as part of its Institutional Program (IP). This launch coincided with the COVID-19 pandemic, which disrupted in-person language testing nationwide, including the traditional paper-based TOEIC L&R. The online version, which can be taken anywhere with an internet connection, was well suited to meet testing demands during that period. Consequently, many schools and companies that had previously used the paper-based TOEIC L&R IP test transitioned to the online version. However, administrators monitoring TOEIC scores at these institutions unexpectedly observed higher scores on the online version compared to established norms for the paper-based test. These observations raised concerns about score comparability, despite IIBC's assurance that scores from both versions are equivalent (IIBC, n.d.-d, 2020). Given the importance of the TOEIC L&R in Japan, empirical research is needed to investigate this issue. The present study aims to address this gap.

Literature Review

Research on score comparability between the online and paper-based versions of the TOEIC L&R is limited. The only published studies found online at the time of writing are by Richard (2021, 2023). He examined the scores of 56 students in 2021 and 54 students in 2022 and found that the average scores in both cohorts were higher on the online version. He also analyzed university-wide TOEIC L&R scores from 2018 to

2022, alongside results from the Computerized Assessment for English Communication (CASEC). While CASEC scores remained relatively stable over the years, TOEIC scores spiked in 2020 when the online version was introduced, then declined over the following two years, returning to pre-2020 levels in 2022. Another key finding was that although the effect sizes for score differences between the two versions were small to medium, individual score differences varied considerably.

Beyond the TOEIC L&R, researchers have examined mode effects on test performance in various fields, including language proficiency (Mohammadi & Barzgaran, 2012; Yu & Iwashita, 2021), medicine (Hochlehnert et al., 2011; Karay et al., 2015), college admissions (Li et al., 2016; Steedle et al., 2020), and high school subjects (Boo & Vispoel, 2012; Pengelley et al., 2024). Overall, these studies suggest that mode effects between paper-based and computer-based tests are generally small or negligible. However, Pengelley et al. (2024) cautioned that computer-based tests may increase cognitive load, potentially making them more mentally demanding for some learners.

Since the test length of the online TOEIC L&R is less than half that of the paper-based version, test length may also influence score differences. Studies on this topic (e.g., Ackerman & Kanfer, 2009; Jensen et al., 2013; Laitusis et al., 2007; Liu et al., 2004) indicate that longer tests do not necessarily lead to performance declines in high-stakes situations but may do so in low-stakes settings.

The primary aim of the present study is to examine score comparability between the online and paper-based versions of the TOEIC L&R. To this end, it compares scores from two forms of each test version so that score comparability across formats can be evaluated against score comparability within the same format.

Methodology

Materials

The online and paper-based versions of the TOEIC L&R were used in this study. The online version was introduced in Japan in April 2020, while the paper-based version has been administered since its initial launch in 1979, with major revisions implemented in 2006 and 2016 (Powers & Schmidgall, 2018).

Both versions include the same question types: four in the listening section (photographs, question–response, conversations, and talks) and three in the reading section (incomplete sentences, text completion, and reading comprehension).

Table 1 summarizes the main differences between the two formats. Notably, the online version incorporates a computer-adaptive component in which the difficulty of the final 20 questions in both the listening and reading sections adjusts based on the test taker's performance on the first 25 questions. According to IIBC (n.d.-d, 2020), this adaptive design enables the online test to assess English proficiency using fewer items and less time than the paper-based version. For additional details on the online version, see IIBC (n.d.-d, 2020); for further information on the paper-based version, refer to Educational Testing Service (2022a, 2022b).

Table 1
Differences Between the Online and Paper-Based TOEIC L&R

	Online	Paper-based
Number of questions	Listening: 45 questions Reading: 45 questions	Listening: 100 questions Reading: 100 questions
Test time	Listening: About 25 minutes Reading: 37 minutes	Listening: About 45 minutes Reading: 75 minutes
Delivery	On computer	On paper
Answering	Choose answers on the computer screen	Mark answers on a marksheet with a pencil
Adaptive	Yes	No
Audio	Audio device connected to a PC (test taker adjusts the volume)	Speaker connected to a CD player (proctor adjusts the volume)
Place	Anywhere with internet access	A designated room with other test takers
Proctoring	AI (optional)	Human proctoring

Educational Testing Service (2022a, 2022b) reports that the KR-20 reliability coefficients for both the listening and reading sections of the paper-based version are approximately 0.90. However, reliability data for the online version have not yet been publicly disclosed.

All four tests in this study were administered through the Institutional Program, which allows institutions to independently organize test sessions. Tests from the Public

Program—fully managed by IIBC in Japan—were not included. For further details on these two testing programs, see Educational Testing Service (2024).

Test Administrations

The participants took two online and two paper-based TOEIC L&R tests between July and August 2023. The first online test was taken between July 18 and July 28, followed by the first paper-based test on August 1. The second online test was administered between August 2 and August 22, and the second paper-based test was held on August 23. AI remote proctoring was used for the online tests, and participants could choose the time and location for taking the test within the 11-day window. In contrast, all participants took the paper-based tests together in the same room under the supervision of a human proctor.

Participants

The participants in this study were students at a private university located in Japan's Kanto region, known for its strong emphasis on foreign language education. Ethical approval was obtained from the university's institutional review board prior to participant recruitment, which was conducted via the university's web portal in May and June 2023.

To participate, students were required to register for the online version of the TOEIC L&R administered by the university in July 2023. A registration fee of 3,500 yen was required, except for fourth-year students in certain departments. This requirement was intended to ensure that participants were genuinely motivated to take the TOEIC L&R. The costs for the remaining three tests were covered by a research grant, ensuring no additional financial burden on participants. As an additional incentive, each participant received an Amazon gift coupon worth 3,000 yen upon completing all four tests.

Students who expressed interest were asked to submit an informed consent form, and 175 provided consent. Of these, 129 completed all four tests. However, data from two participants were excluded due to exceptionally large discrepancies between their two online listening scores. One participant scored 355 on the first test and 85 on the second (a difference of 270 points), while another scored 190 on the first and 455 on the second (a difference of 265 points). Since listening scores are measured on a 5-to-495 scale, these anomalies were deemed abnormal, and their data were excluded from further analysis.

As a result, the final analysis included data from 127 participants. The academic year breakdown was as follows: nine first-year students, 24 second-year students, 44 third-year students, and 50 fourth-year students. Among them, 81 were enrolled in programs where English was the primary foreign language, while 46 were enrolled in programs focused on other languages such as Indonesian, Spanish, and Chinese.

Analysis Procedure

Before analyzing the comparability of scores across the four tests, the normality of data distributions was assessed using multiple methods. Shapiro–Wilk tests for all variables indicated no significant departures from normality (all $ps > .05$), although several p-values approached the .05 threshold. Skewness and kurtosis values were all within ±1, suggesting only minor deviations from normality. Visual inspection of histograms revealed approximately symmetric, unimodal distributions for each variable, with some nonnormality observed in several cases. Taken together, these results indicate that the assumption of normality was reasonably satisfied, supporting the use of parametric analyses.

To examine mean score differences, repeated-measures ANOVAs were conducted separately for each of the three score categories (listening, reading, and total). Pairwise comparisons were performed for six score pairs within each category, with p-values adjusted using the Bonferroni correction for multiple comparisons. Additionally, the effect sizes (Cohen's d) were calculated for all pairs using paired-samples t-tests in IBM SPSS, with correction for the correlation between measures (Cohen, 1988).

To evaluate participant-level score differences, the standard error of difference (SEdiff) was used as a reference, following Richard's (2023) approach. The SEdiff is 35 points for both the listening and reading sections (Educational Testing Service, 2022a, 2022b); the ±35-point band corresponds to a 68% confidence interval, meaning that if a test taker's English proficiency remains stable, score differences between two test administrations are expected to fall within this range 68% of the time.

Although the SEdiff for total scores is not publicly available, it can be derived from the SEdiff values for the listening and reading sections using the following formula:

$$\text{SEdiff}_{\text{total}} = \sqrt{(\text{SEdiff}^2_{\text{listening}} + \text{SEdiff}^2_{\text{reading}})}$$

Substituting the known values:

$$\text{SEdiff}_{\text{total}} = \sqrt{(35^2 + 35^2)} = \sqrt{(1225 + 1225)} = \sqrt{2450} \approx 50$$

Therefore, a ±50-point band was used as the reference for interpreting total score differences.

Results

For clarity and conciseness, the following abbreviations are used hereinafter: Online = online version, Paper = paper-based version, 1 = first test, 2 = second test, L = listening scores, R = reading scores, T = total scores. For example, Online1L refers to the listening scores from the first online test, and Paper2T refers to the total scores from the second paper-based test.

Descriptive Statistics

Table 2 presents descriptive statistics for the 12 sets of scores.

Table 2

Descriptive Statistics for 12 Sets of Scores

Test	Score category / Possible score	Score range	Mean	*SD*
Online1L	Listening 5–495	220–495	369.8	64.0
Online2L		200–495	375.4	65.5
Paper1L		205–495	347.5	62.7
Paper2L		180–495	342.2	61.0
Online1R	Reading 5–495	105–485	292.9	85.2
Online2R		55–495	305.4	90.1
Paper1R		115–465	288.3	82.6
Paper2R		85–465	296.7	82.5
Online1T	Total 10–990	380–980	662.6	138.4
Online2T		315–965	680.9	146.3
Paper1T		355–910	635.8	137.2
Paper2T		340–935	638.9	135.2

Note. $N = 127$.

Correlations

Tables 3 through 5 display Pearson's *r* correlations between the six pairs of the four score sets in each of the three score categories.

Table 3

Correlations Between the Four Sets of Listening Scores

	Online1L	Online2L	Paper1L	Paper2L
Online1L	—			
Online2L	.82**	—		
Paper1L	.73**	.79**	—	
Paper2L	.76**	.81**	.74**	—

Note. $^{**}p < .001$ (two-tailed).

Table 4

Correlations Between the Four Sets of Reading Scores

	Online1R	Online2R	Paper1R	Paper2R
Online1R	—			
Online2R	.81**	—		
Paper1R	.86**	.85**	—	
Paper2R	.81**	.83**	.88**	—

Note. $^{**}p < .001$ (two-tailed).

Table 5

Correlations Between the Four Sets of Total Scores

	Online1T	Online2T	Paper1T	Paper2T
Online1T	—			
Online2T	.87**	—		
Paper1T	.88**	.87**	—	
Paper2T	.86**	.87**	.87**	—

Note. $^{**}p < .001$ (two-tailed).

Mean Differences

Repeated-measures ANOVAs were conducted to examine mean score differences across the four sets of scores in each of the three categories. Mauchly's tests indicated that the assumption of sphericity was met for the listening scores, $\chi^2(5) = 7.74$, $p = .171$, and the total scores, $\chi^2(5) = 1.71$, $p = .888$. However, it was violated for the reading scores, $\chi^2(5) = 16.51$, $p = .006$; therefore, degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = .93$). The analyses revealed significant differences across the four score sets of the listening, $F(3, 378) = 37.49$, $p < .001$, $\eta^2_p = .229$, the reading, $F(2.79, 351.91) = 5.67$, $p = .001$, $\eta^2_p = .043$, and the total, $F(3, 378) = 22.66$, $p < .001$, $\eta^2_p = .152$.

In addition to overall comparisons in each category, scores were compared pairwise as well. Tables 6 through 8 display pairwise comparisons of the six pairs in each of the three categories. *P*-values are adjusted for multiple comparisons using the Bonferroni correction. Cohen's *d* values, obtained by running paired-samples *t*-tests with IBM SPSS, are corrected for the correlation between measures.

Table 6

Pairwise Comparisons of Listening Scores

Comparison	Mean difference	*SE*	*p*	Cohen's *d*
Online1L vs Online2L	−5.67	3.42	.598	−0.09
Online1L vs Paper1L	22.28*	4.13	<.001	0.35
Online1L vs Paper2L	27.52*	3.87	<.001	0.44
Online2L vs Paper1L	27.95*	3.72	<.001	0.44
Online2L vs Paper2L	33.19*	3.46	<.001	0.52
Paper1L vs Paper2L	5.24	3.98	1.000	0.09

Note. *$p < .05$. Bonferroni adjustment applied.

Table 7

Pairwise Comparisons of Reading Scores

Comparison	Mean difference	*SE*	*p*	Cohen's *d*
Online1R vs Online2R	−12.56	4.77	.057	−0.14
Online1R vs Paper1R	4.57	3.99	1.000	0.05
Online1R vs Paper2R	−3.82	4.65	1.000	−0.05
Online2R vs Paper1R	17.13*	4.29	<.001	0.20
Online2R vs Paper2R	8.74	4.50	.326	0.10
Paper1R vs Paper2R	−8.39	3.57	.121	−0.10

Note. *$p < .05$. Bonferroni adjustment applied.

Table 8

Pairwise Comparisons of Total Scores

Comparison	Mean difference	*SE*	*p*	Cohen's *d*
Online1T vs Online2T	−18.23*	6.44	.032	−0.13
Online1T vs Paper1T	26.85*	5.96	<.001	0.20
Online1T vs Paper2T	23.70*	6.33	.002	0.17
Online2T vs Paper1T	45.08*	6.50	<.001	0.32
Online2T vs Paper2T	41.93*	6.49	<.001	0.30
Paper1T vs Paper2T	−3.15	6.14	1.000	−0.02

Note. *$p < .05$. Bonferroni adjustment applied.

Participant-Level Score Differences

Individual score differences in the six listening and six reading pairs were examined to determine whether they fell within or outside the ±35-point band. Table 9 summarizes the results.

Table 9

Numbers of Score Differences Within and Outside the ±35-point Band

Comparison	Within ±35	Outside ±35
Online1L vs Online2L	83 (65.4%)	44 (34.6%)
Online1L vs Paper1L	71 (55.9%)	56 (44.1%)
Online1L vs Paper2L	68 (53.5%)	59 (46.5%)
Online2L vs Paper1L	75 (59.1%)	52 (40.9%)
Online2L vs Paper2L	64 (50.4%)	63 (49.6%)
Paper1L vs Paper2L	77 (60.6%)	50 (39.4%)
Online1R vs Online2R	71 (55.9%)	56 (44.1%)
Online1R vs Paper1R	81 (63.8%)	46 (36.2%)
Online1R vs Paper2R	79 (62.2%)	48 (37.8%)
Online2R vs Paper1R	69 (54.3%)	58 (45.7%)
Online2R vs Paper2R	63 (49.6%)	64 (50.4%)
Paper1R vs Paper2R	86 (67.7%)	41 (32.3%)

Individual score differences in the six total score pairs were examined to determine whether they fell within or outside the ±50-point band. Table 10 summarizes the results.

Table 10

Numbers of Score Differences Within and Outside the ±50-point Band

Pair	Within ±50	Outside ±50
Online1T vs Online2T	76 (59.8%)	51 (40.2%)
Online1T vs Paper1T	67 (52.8%)	60 (47.2%)
Online1T vs Paper2T	62 (48.8%)	65 (51.2%)
Online2T vs Paper1T	62 (48.8%)	65 (51.2%)
Online2T vs Paper2T	56 (44.1%)	71 (55.9%)
Paper1T vs Paper2T	80 (63.0%)	47 (37.0%)

Discussion

The descriptive statistics in Table 2 show that both online tests yielded higher mean scores than the paper-based tests across all sections. The differences were larger in listening than in reading, and the mean total scores for the online tests (M = 662.6 and 680.9) notably exceeded those of the paper-based tests (M = 635.8 and 638.9). In addition, standard deviations were consistently larger in the online version across all sections, suggesting greater variability in test-taker performance. The adaptive design of the online test may have contributed to increased score dispersion, as performance-based branching could have widened the performance gap between higher- and lower-ability students.

The correlations presented in Tables 3 through 5 were in the high range, from .73 to .88. In the listening section, the highest correlation was between the two online tests (r = .82), while the correlation between the two paper-based tests was the second lowest (r = .74). This lower correlation may have been influenced by differences in audio quality between the two paper-based administrations. In the open-ended comments from the post-test questionnaire after the first test, 37 participants reported difficulty following the listening section during the test due to excessive volume or echoing. To address this issue, the volume was carefully adjusted prior to the second test. Interestingly, despite this adjustment, the average listening score on the second paper-based test was 5.3 points lower than that of the first, suggesting that while some participants found the louder audio problematic, it may have been beneficial for others.

In the reading section, the highest correlation was between the two paper-based tests (r = .88), while the lowest correlations were observed between the two online tests and between Online1R and Paper2R (both rs = .81). The correlations for the six pairs of total scores ranged from .86 to .88, indicating very strong relationships between all four tests.

The results of the repeated-measures ANOVAs revealed statistically significant differences across the four tests in all three score categories. The effect was strongest in the listening section, $F(3, 378) = 37.49$, $p < .001$, $\eta^2_p = .229$, and weakest in the reading section, $F(2.79, 351.91) = 5.67$, $p = .001$, $\eta^2_p = .043$. The total scores also showed significant differences, $F(3, 378) = 22.66$, $p < .001$, $\eta^2_p = .152$. Score differences were most pronounced in listening and least evident in reading.

The pairwise comparisons presented in Tables 6 through 8 highlight mean score differences within and across test formats. For listening, all cross-format comparisons were statistically significant, while within-format comparisons were not. In reading, only one comparison reached statistical significance: Online2R vs. Paper1R (M =

17.13, $p < .001$). All other comparisons were non-significant, indicating that reading scores remained relatively stable both within and across formats. Total scores showed statistically significant differences in all cross-format comparisons, with mean differences ranging from 23.70 to 45.08. No significant difference was observed between the two paper-based tests, but the difference between the two online tests was significant, suggesting that scores were more stable in the paper-based version than in the online version.

For all statistically significant differences, the effect sizes expressed as Cohen's *d* ranged from 0.13 to 0.52 in absolute value. Plonsky and Oswald (2014) proposed a benchmark of $d = 0.6$ as a small effect for repeated-measures designs in second language research. According to this criterion, the effect sizes for all significant differences are considered small.

Each participant's score differences between test pairs were examined against standard error of difference bands: ±35 points for listening and reading, and ±50 points for total scores. For listening, the two within-format comparisons had a higher proportion of participants within the ±35-point band (65.4% for the online pair and 60.6% for the paper-based pair) than cross-format comparisons, which ranged from 50.4% to 59.1%. This pattern suggests greater score stability within the same format. For reading, the highest proportion within the ±35-point band was between the two paper-based tests (67.7%), while the proportion for the two online tests was 55.9%, which was lower than two of the four cross-format pairs. For total scores, the highest stability was again between the two paper-based tests, with 63.0% of participants within the ±50-point band. The second highest was between the two online tests (59.8%), and proportions for cross-format pairs ranged from 44.1% to 52.8%. Taken together, these results suggest that score consistency was generally higher within the same format than across formats, with the only exception being the reading scores of the two online tests.

Conclusion

This study evaluated the comparability of scores between the online and paper-based versions of the TOEIC L&R, relative to the comparability within the same format. Correlational analyses suggested that scores were comparable across formats, as the correlations for cross-format pairs were as high as those within the same format. However, pairwise comparisons revealed that listening and total scores across formats were not as comparable, with significant mean score differences observed between cross-format pairs, while within-format differences were mostly non-significant. Furthermore, when individual score differences between two tests were evaluated against standard error of difference bands, listening and total score differences were generally larger in cross-format comparisons than in within-format comparisons. In sum, the listening and total scores between the two formats were less comparable than those within the same format, whereas the reading scores showed higher comparability across formats.

The most noteworthy finding is that the total scores on the online version were higher than those on the paper-based version. The mean differences were 26.58 points between Online1T and Paper1T, 23.70 between Online1T and Paper2T, 45.08 between Online2T and Paper1T, and 41.93 between Online2T and Paper2T (an average of 34.39 points). These differences were statistically significant, but their practical significance is debatable, as the effect sizes were small, with Cohen's *d*s between 0.17 and 0.32.

As Richard (2023) speculated, one possible factor contributing to the higher online scores is the difference in test length. While previous studies suggest that test length does not significantly affect performance in high-stakes settings (Ackerman & Kanfer, 2009; Jensen et al., 2013; Laitusis et al., 2007; Liu et al., 2004), this study was conducted in a low-stakes setting. Maintaining focus for the 2-hour paper-based test may be more challenging, whereas the shorter 1-hour online test may be more manageable. In this regard, the online version might provide a fairer assessment, particularly for test takers with shorter attention spans who may be disadvantaged by the longer paper-based test.

Beyond stamina and concentration, other human factors may influence TOEIC scores, including conscientiousness, motivation, the desire for high performance, and test-takers' physical or mental conditions. Administration conditions and individual testing environments may also affect performance. A limitation of this study is its inability to distinguish between score discrepancies due to measurement issues and those caused by these other factors.

In conclusion, although scores on the paper-based test are more stable, this paper supports the use of the online version, as it offers several advantages. For test takers, it provides greater convenience and requires less time and effort due to its shorter length. For institutions, it reduces logistical burdens by eliminating the need for physical test rooms and proctors. Although the online version may yield slightly higher scores, this difference could be viewed positively, as it may reflect a fairer assessment of proficiency for those who struggle to sustain concentration over longer periods.

Data Sharing

The score data, SPSS output, and other materials related to this study are available at http://bit.ly/3TSwkpZ.

Acknowledgments

This study was supported by a research grant from Kanda University of International Studies.

Bio Data

Masaya Kanzaki has been affiliated with Kanda University of International Studies since 2010. His research interests include language testing, vocabulary acquisition, and corpus linguistics. <kanzaki-m@kanda.kuis.ac.jp>

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163–181. https://doi.org/10.1037/a0015719

Boo, J., & Vispoel, W. (2012). Computer versus paper-and-pencil assessment of educational development: A comparison of psychometric features and examinee preferences. *Psychological Reports*, *111*(2), 443–460. https://doi.org/10.2466/10.03.11.PR0.111.5.443-460

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Educational Testing Service. (2022a). *TOEIC listening and reading test examinee handbook*. Retrieved from https://www.ets.org/pdfs/toeic/toeic-listening-reading-test-examinee-handbook.pdf

Educational Testing Service. (2022b). *TOEIC listening and reading test score user guide*. Retrieved from https://www.ets.org/pdfs/toeic/toeic-listening-reading-score-user-guide.pdf

Educational Testing Service. (2024). *TOEIC Institutional and Public Programs*. Retrieved July 26, 2025, from https://www.etsglobal.org/fr/en/content/institutional-and-public-programmes

Hochlehnert, A., Brass, K., Moeltner, A., & Juenger, J. (2011). Does medical students' preference of test format (computer-based vs. paper-based) have an influence on performance?. *BMC Medical Education*, *11*, 89. https://doi.org/10.1186/1472-6920-11-89

IIBC. (n.d.-a). 公式データ・資料 [Official data and resources]. https://www.iibc-global.org/toeic/official_data.html

IIBC. (n.d.-b). 企業・団体の主なTOEIC Program活用目的 [Main uses of the TOEIC Program by companies and organizations]. https://group.iibc-global.org/organizations/corporate/study

IIBC. (n.d.-c). 短大・大学・大学院・専門学校の主なTOEIC Program活用目的 [Main uses of the TOEIC Program by junior colleges, universities, graduate schools, and vocational schools]. https://group.iibc-global.org/organizations/university/study

IIBC. (n.d.-d). TOEIC Program IPテスト（オンライン） [TOEIC program IP test online]. https://www.iibc-global.org/toeic/corpo/guide/online_program.html

IIBC. (2020). 特集：場所と時間を問わずに活用できるIIBCのオンラインプログラム [*Special feature: IIBC's online program that can be used at any time and place*]. https://www.iibc-global.org/iibc/activity/iibc_newsletter/nl141_feature_01.html

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PloS ONE*, *8*(8), e70270. https://doi.org/10.1371/journal.pone.0070270

Karay, Y., Schauber, S. K., Stosch, C., & Schüttpelz-Brauns, K. (2015). Computer versus paper—Does it make any difference in test performance? *Teaching and Learning in Medicine*, *27*(1), 57–62. https://doi.org/10.1080/10401334.2014.979175

Laitusis, C. C., Morgan, D. L., Bridgeman, B., Zanna, J., & Stone, E. (2007). Examination of fatigue effects from extended-time accommodations on the SAT reasoning test. *ETS Research Report Series*, *2007*(2), i–13. https://doi.org/10.1002/j.2333-8504.2007.tb02073.x

Li, D., Yi, Q., & Harris, D. (2016). Evidence for paper and online ACT comparability: Spring 2014 and 2015 mode comparability studies. *ACT Working Paper 2016-02*. ACT, Inc. https://www.act.org/content/dam/act/unsecured/documents/Working-Paper-2016-02-Evidence-for-Paper-and-Online-ACT-Comparability.pdf

Liu, J., Allspach, J. R., Feigenbaum, M., Oh, H. J., & Burton, N. (2004). A study of fatigue effects from the new SAT. *ETS Research Report Series*, *2004*(2), i–13. https://doi.org/10.1002/j.2333-8504.2004.tb01973.x

Mohammadi, M., & Barzgaran, M. (2012). Comparability of computer-based and paper-based versions of writing section of PET in Iranian EFL context. *Journal of Foreign Language Teaching and Translation Studies*, *1*(2), 1–20. https://efl.shbu.ac.ir/article_79172_b760dabe26e418d65627673b0914f266.pdf

Pengelley, J., Whipp, P. R., & Malpique, A. (2024). A testing load: a review of cognitive load in computer and paper-based learning and assessment. *Technology, Pedagogy and Education*, *34*(1), 1–17. https://doi.org/10.1080/1475939X.2024.2367517

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Powers, D. E., & Schmidgall, J. E. (2018). The TOEIC test: A brief history. In D. E. Powers, & J. E. Schmidgall (Eds.), *The research foundation for the TOEIC tests: A compendium of studies: Vol. 3* (pp. 1.1–1.5). https://www.ets.org/s/toeic/pdf/research-compendium.pdf

Richard, J. - P. J. (2021). A comparison of the online version and paper-based version of TOEIC L&R. *The Global Management of Nagano*, *5*, 37–57. https://shinshu.repo.nii.ac.jp/records/46430

Richard, J. - P. J. (2023). Score differences between the paper-based and online TOEIC L&R. In P. Ferguson, B. Lacy, & R. Derrah (Eds.), *Learning from Students, Educating Teachers–Research and Practice* (pp. 10–20). JALT. https://doi.org/10.37546/JALTPCP2022-02

Steedle, J., Pashley, P., & Cho, Y. (2020). *Three studies of comparability between paper-based and computer-based testing for the ACT.* ACT, Inc. https://www.act.org/content/dam/act/unsecured/documents/R1847-three-comparability-studies-2020-12.pdf

Yu, W., & Iwashita, N. (2021). Comparison of test performance on paper-based testing (PBT) and computer-based testing (CBT) by English-majored undergraduate students in China. *Language Testing in Asia*, *11*, 32. https://doi.org/10.1186/s40468-021-00147-0