A Comparison of the Online Version and Paper-based Version of TOEIC L&R

著者	Jean-Pierre Joseph Richard
雑誌名	グローバルマネジメント
巻	5
ページ	37-57
発行年	2021-07
URL	http://doi.org/10.32288/00001358

【Articles】

A Comparison of the Online Version and Paper-based Version of TOEIC L&R

Jean-Pierre Joseph RICHARD

Abstract

In Spring 2020, an online version of the Test of English for International Communication for Listening and Reading (TOEIC L&R) became available. The Institute for International Business Communication, IIBC, (2020d) indicated that the paper-based version and the online version are parallel; however, no published studies have discussed the reliability or validity of the online version. In February 2021, volunteer first-year participants (N = 56) at the University of Nagano (UoN) were randomly assigned to complete the paper-based test one day before or after the end-of-year online version. Before combining the paper-based test scores from different days, the data were rigorously checked. Three research questions were investigated. For RQ1, correlational analyses indicated that the two listening tests (r = .742) and the two reading tests (r = .676) had strong correlations. However, for RQ2, paired samples t-tests revealed that the mean scores for the two listening tests were significantly different with a near large sized effect (Cohen's d = .924); whereas the reading tests had a negligible difference (Cohen's d = .308). Finally, for RQ3, independent sample t-tests revealed similarities in scores on the paper-based tests for two cohorts, but differences in scores between the paper-based tests and the online tests. In all, the results raise concerns about the reliability and validity of the online version of the TOEIC L&R. One important limitation is that the online TOEIC L&R was sat by the students without proctors present.

In 2020, the COVID-19 pandemic impacted educational institutions, including disruptions to the academic calendar, a shift to online learning, and changes in

testing programs. The University of Nagano (UoN) was no exception. At UoN, incoming students complete at home the Computerized Assessment System for English Communication (CASEC); then they complete the paper-based Test of English for International Communication for Listening and Reading (TOEIC L&R) a few weeks later on campus once classes have begun. However, due to COVID-19, the paper-based TOEIC L&R was replaced by the new TOEIC L&R Online test which students completed at home.

CASEC test results indicated that the 2020 cohort had comparable scores with previous cohorts; however, results from the online TOIEC L&R indicated that the 2020 cohort had significantly higher scores than cohorts who completed the paper-based TOEIC L&R[1]. Score differences on the TOEIC L&R between the 2020 cohort and previous cohorts might be due to four main possibilities: (1) the late start of the academic year allowed students to prepare; (2) students had higher test-taking motivation while taking the online TOEIC L&R compared with those who sat the paper-based TOEIC L&R; (3) the online TOEIC L&R was taken without proctors and multiple students subverted test-taking procedures; (4) the paper-based and online TOEIC L&R tests are not parallel; or combinations of the above. This study will compare results from a test-retest research program in which first year participants at UoN (N = 56) completed the paper-based and online versions of the TOEIC L&R on consecutive days, and by comparing data from the paper-based TOEIC L&R for the previous cohort (N = 202).

TOEIC L&R

Standardized language tests, such as the TOEIC L&R, are used for admissions, placement, program evaluation, hiring and promotion (Im et al, 2019). At UoN, this test is used for individual and program evaluation, is partially used for class placement in Year 2, and many upper grade students use scores from this test for job hunting. In Japan, more than 2.2 million examinees, including more than 1 million students, completed the TOEIC L&R in 2019 (IIBC, 2020a). Due to the important roles that this test plays, it is imperative that it be consistent and reliable across administrations (*e.g.*, for example at the beginning and end of one academic year) and across test formats (*e.g.*, pre-updated and updated versions of the paper-based TOEIC L&R; the paper-based and the online versions of the TOEIC L&R).

1　Although beyond the scope of this paper, ANOVAs indicated CASEC scores were similar; but TOEIC L&R scores were significantly higher for the 2020 cohort. See Appendix A.

As to the former, Wei and Low (2017) demonstrated that repeater test-taker data could be used to monitor the TOEIC L&R across administrations, concluding that their analyses "support the reliability and validity of the TOEIC scores" (p. 18). As to the latter, the paper-based test format was updated in May 2016. Analyses with examinees in Japan and Korea sitting the updated paper-based test compared with a large reference sample from the pre-updated paper-based test indicated that the updated version performed psychometrically as well as the pre-updated version (Cid et al, 2017). Moreover, mean scale score differences between the pre-updated and updated versions were minimal, 1.39 points for Reading and 3.11 points for Listening (Cid et al). Kanzaki (2017) compared the pre-updated and updated versions of the TOEIC L&R with Japanese students (N = 141), observing strong correlations (Listening: r = .80; Reading: r = .84), and mean scale score differences between versions were 0.96 for Listening and 11.46 for reading. To date, however, there appear to be no studies comparing the paper-based TOEIC L&R and the newer online version.

Limited information is available regarding the online version of the TOEIC L&R. A search using Google in December 2020 for "*TOEIC*® L&R Online", in Japanese and English, resulted in links to university co-ops, cram schools, and press releases. The first news article that was found, from March 2020, reported that the Institute for International Business Communication (IIBC) would begin from April 2020 an online version of the TOEIC L&R (Nikkei Shimbun, 2020 March 10). With the exception of various news aggregator websites, no other news stories were identified in this preliminary search. One press release from the fall of 2020 indicated that more than 1100 Japanese organizations had used the online version of the TOEIC L&R test since April 2020 (IIBC, 2020e). An announcement from IIBC described the online version as:「本物を！ ETS開発の正式なテスト従来のスコアと意味は変わらない」[The real thing! ETS formal test, the interpretation is the same as a traditional score] (IIBC, 2020c). A second announcement from IIBC, indicated that the score interpretation of the two versions were the same,「評価やスコアの意味合いは、公開テストや従来のIPテストと同様で、スコアが同じであれば、英語力も同等です」[the meaning of evaluations and scores is the same as in public tests and conventional IP tests, and if the scores are the same, the English proficiency is also the same] (IIBC, 2020d), although they added with the caution that online tests taken at home may not be controlled.

In addition to Google, three newspaper websites (Yomiuri, Asahi, Japan Times) were searched in December 2020 for TOEIC-related stories. None were found which referred to the new online test. Next, an online web search in December 2020 using

Google Scholar with variations of "*TOEIC*® L&R Online" in Japanese and English found no related articles. A second search in March 2021 found one (Suzuki, 2021) that merely summarized the operating procedures for TOEFL, Eiken, and TOEIC since the beginning of the pandemic, but which did not report on the performance of the online tests. Lastly, a search of the TOEIC research database in December 2020 and March 2021 at ETS (https://www.ets.org/toeic/organizations/research/topics/) found no papers related to the TOEIC Online L&R. In short, there is limited information available related to the online version of the TOEIC L&R, and the claim that the online test scores are equivalent to the paper-based test scores appears to be untested.

Research Questions

The first research question is interested in the correlations of scores from both skills across the online and paper-based TOEIC L&R tests. Although Kanzaki (2017) identified strong correlations between the pre-updated and updated forms for the paper-based TOEIC test for listening and reading, the correlations between the scores from the online and paper-based TOEIC L&R are unknown. The second research question investigates whether the results from the paper-based TOEIC L&R are the same as those from the online TOEIC L&R, per skill of listening and reading. Kanzaki observed a small difference in scale points between the two listening tests, but a much larger difference between the reading tests. However, whether the scores from the paper-based and online versions of the TOEIC L&R are similar or not are unknown. The final research question compares the 2019[2] and 2020 cohorts on the end-of-Year 1 TOEIC L&R. These results could provide more support for any claims about the reliability and validity of the online TOEIC L&R.

Methodology

UoN and its English Program

UoN, a small, regional, public university located in the northern part of central Japan, opened in 2018. The required English program is semi-intensive over two years. Year 1 students have four 100-minute English lessons per week, and Year 2 students have two-to-four 100-minute lessons per week, depending on the faculty. Electives for students in Years 3 and 4 are available. In the weeks before entering, incoming students complete at home CASEC for class placement. This computer

2 The 2018 cohort did not complete the TOEIC L&R at the end of their first year.

adaptive test, developed by the Japan Institute for Educational Measurement, takes approximately 40–50 minutes. It includes four sections: vocabulary (k = 16), phrasal expressions and usage (k = 16), listening for the main idea (k = 17), and listening for specific information (k = 11) (CASEC, n.d., a). Official score reports, received upon test completion, include a chart of the examinee's performance on each section, approximate TOEIC and STEP Eiken comparison scores, and estimated can-do abilities (CASEC, n.d., b).

In addition to CASEC, UoN students complete the TOEIC L&R at the beginning and end of Year 1 and at the end of Year 2. The 2018 and 2019 cohorts completed the paper-based TOEIC L&R supervised at the university within approximately one week of entering the university. This paper-based test (Table 1) has 200 questions: listening (k = 100, 45 minutes) and reading (k = 100, 75 minutes) (IIBC, 2020b). In 2020, the start of the academic year was postponed by approximately six weeks and the online TOEIC L&R replaced the paper-based TOEIC L&R. The online computer adaptive TOEIC L&R, was given in May 2020. For this test (Table 2), for each of listening and reading, examinees have the same sets of 25 questions in Unit 1 and depending on the degree of correctness each candidate receives a different set of 20 questions in Unit 2 (IIBC, 2020c).

Table 1. *TOEIC L&R Paper-Based Test Format*

Section	Part	Questions (Type)	Questions (k)	Minutes
Listening	1	Photographs	6	45
	2	Question-Response	25	
	3	Conversations	39	
	4	Short talks	30	
Reading	5	Incomplete sentences	30	75
	6	Text completion	16	
	7	Single passages	29	
		Multiple passages	25	

Table 2. *TOEIC L&R Online Test Format*

Section	Unit	Computer Adaptive	Questions (Type)	Questions (k)	Minutes
Listening	1	No	Photographs	3	25
			Question-Response	4	
			Conversations	9	
			Short talks	9	
	2	Yes	Question-Response	5	
			Conversations	9	
			Talks	6	

Reading	1	No	Incomplete sentences	5	37
			Text completion	4	
			Reading comprehension	16	
	2	Yes	Incomplete sentences	7	
			Text completion	4	
			Reading comprehension	9	

Participants

Approximately 25% of the Year 1 population (N = 56) at UoN participated. Recruitment was done via an online form distributed in December 2020 through email. The research program provided for up to 100 Year 1 students to participate. The email indicated that participants would not be renumerated. In all, 67 students were recruited, and 58 completed both versions of the TOEIC L&R of whom 57 consented for their data to be used. Of these 57, 93% were from the largest faculty at the UoN, although this faculty represents 70% of the student population at UoN. Following the detailed analyses described in the following section, one participant was deleted from the data set, resulting in a pool of 56 participants. In addition to this group of participants, to answer RQ3, data from 202 participants from the 2019 cohort were also used.

To avoid a test fatigue effect, half of the participants were each randomly assigned to complete the paper-based test one day before or after the online test. See Table 3.

Table 3. *Test Dates and n-sizes.*

	Day 1 (Feb 8, 2021)	Day 2 (Feb 9, 2021)	Day 3 (Feb 10, 2021)
Test *n*	Paper-based TOEIC L&R *n* = 29	Online TOEIC L&R *n* = 57	Paper-based TOEIC L&R *n* =28

Descriptives and Analyses

Before combining the scores from the paper-based test from Day 1 and Day 3, the data underwent multiple inspections. First, two independent sample Student's t-tests were run to investigate whether mean scores were similar. The data were analyzed using JASP, a free and open-sourced program for statistical analyses (JASP Team, 2020). The data met assumptions for parametric testing. The t-tests were nonsignificant and effect sizes were negligible (following Plonsky & Oswald, 2014, where d = .40 is small, d =. 70 is medium, and d = 1.00 is large) with the confidence intervals crossing the zero: Listening [t(55) = 0.86, p = .39, Cohen's d = .23 (95% CIs = −.29, .75)]; and Reading [t(55) = 0.71, p = .48, Cohen's d = .19 (95% CIs = −.33, .71)]. This

indicated that the mean scores for the paper-based TOEIC L&R, for each section of Listening and Reading, from Day 1 were likely similar to the mean scores for the respective tests from Day 3, possibly allowing for the data to be combined.

Data inspection continued. Table 4 displays the number of participants whose scores between the two versions (*i.e.* paper-based and online) differed by ±35 scale points. This value, ±35, was chosen because it represents the Standard Error of Difference (SE*diff*) between two administrations of the paper-based TOEIC L&R (ETS, 2019). In the current study, 55% of the participants had score differences within the SE*diff* range (*i.e.*, between-35 and 35) possibly indicating no difference in scores. However, approximately 45% of the participants had differences in scores outside the SE*diff*. For Listening and Reading, 24 and 18 participants respectively had higher scores on the online test; compared with two and nine who had higher scores on the paper-based test.

Table 4. *Number (%) of Participants with Scale Score Differences of* ±35 *Points between the Online and Paper-based TOEIC Tests (n* = 57*)*

Difference	TOEIC Listening	TOEIC Reading
>35	24 (42.11%)	18 (31.58%)
-35 to 35	31 (54.39%)	30 (52.63%)
<-35	2 (3.51%)	9 (15.79%)

Note. >35 = participants scored higher on the online test; <35 = they scored lower on the paper-based test.

The TOEIC L&R tests were administered on consecutive days; thus, differences beyond ±35 points are likely due to either test or within-subject variability (*e.g.*, differences in motivation). Table 5 shows differences in scale scores for Listening and Reading for each participant. Day 1 and Day 3 indicate which day the participants completed the paper-based TOEIC L&R. A range of scale score differences can be seen; however, the value of 225 stands out. This value indicates that one participant's online listening test score from Day 2 was 225 points higher than their paper-based listening score from Day 3. This difference, 6.4 times greater than the SE*diff*, and 2.1 times larger than the nearest value for listening tests of 110, over a 24-hour period might be accounted for by differing motivational levels.

Finally, 3x2 chi-square tests of independence were run to investigate test-day bias, by comparing the number of participants in the three scale score difference categories (*i.e.*, >35, -35 to 35, and <35) by Day 1 or Day 3. No significant associations for Listening [X^2 (2, N = 57) = 0.44, p = .81, Cramer's V = .09] or Reading [X^2 (2, N = 57) = 0.85, p = .65, Cramer's V = .12] were observed. Removing the participant with the

scale score difference of 225 does not change the non-association [Listening, X^2 (2, N = 56) = 0.26, p = .88, Cramer's V = .07].

Table 5. *Individual Participant's Score Differences between the Online and Paper Versions of the TOEIC L&R (n = 57)*

	TOEIC Listening		TOEIC Reading	
Difference	Day 1	Day 3	Day 1	Day 3
>35		225		
			140	
			120	
	110 110 100	110 110 110		
				105
				100
		95		95
	85			
		80		
	75 75 75 75	75	75	75 75 65
	70		70	
		65 65 65		
			60 60	
			55	
	50	50 50		50
	45	45		45
			40	40 40
−35 to 35	35 35 35	35		
		30	30 30	30 30
	25 25 25 25	25	25 25 25 25	25 25
	20	20 20	20	
	15 15	15 15 15		15 15 15
	10		10 10	10
		5 5 5	5	
	0 0 0			
	−5		−5	
		−10	−10 −10	−10
	−15		−15 −15	−15 −15
	−20	−20	−20	−20
		−25	−25	
				−30
<−35	−40	−40		−40
				−50
			−60 −60	−60
			−65	−65
			−75	
				−125

Note. Positive values indicate participants scored higher on the TOEIC online test.

Figures 1 and 2 allow for a visual inspection of each participant's paired scores (*i.e.*, paper-based and online) for Listening and Reading. To create these figures, scores

from the paper-based tests were organized in ascending order (green line), hence the appearance of a near linear line for these scores. These scores were used as a baseline on which to map the scores from the online test (blue line) because the former is thought to be known. For both listening and reading, it appears that the paired scores frequently varied.

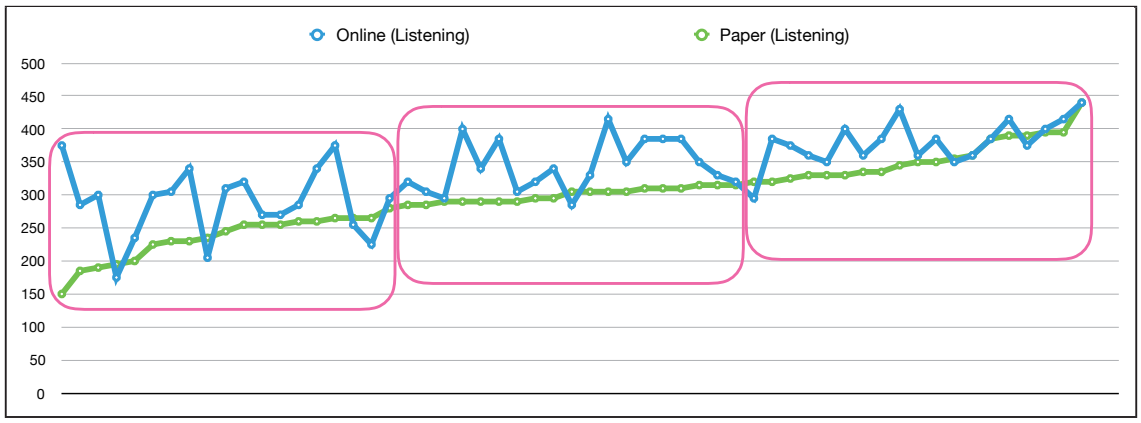

Figure 1. Participants' paired scores for the online and paper-based TOEIC Listening test (N = 57). Mean differences: Lower (left) m = 64.21 (95% CIs = 40.46, 87.96), sd = 52.82, n = 19 [after removing the most left pair difference of 225 points: m = 55.28 (95% CIs = 42.94, 67.62), sd = 26.72, n = 18], Middle (center) m = 46.32 (95% CIs = 31.00, 61.64), sd = 34.07, n = 19, Upper (right) m = 28.16 (95% CIs = 16.82, 39.51), sd = 25.23, n = 19.

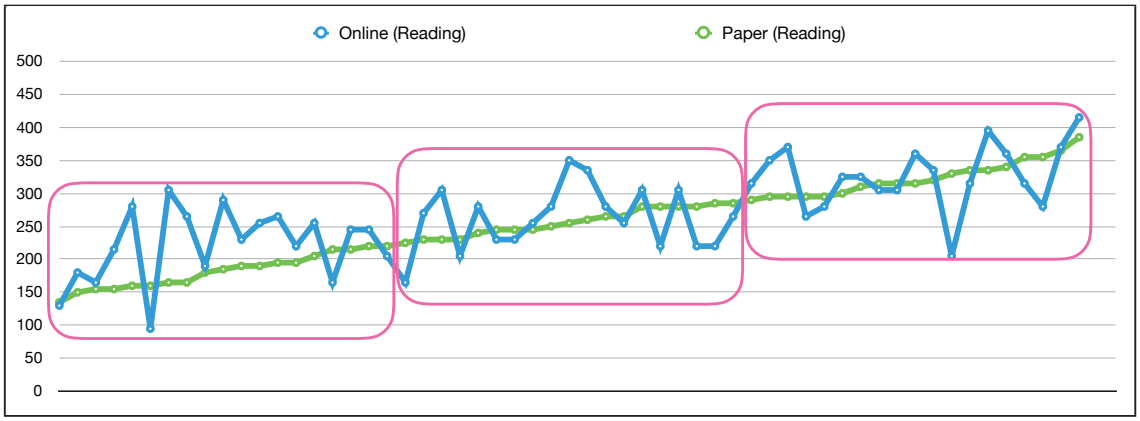

Figure 2. Participants' paired scores for the online and paper-based TOEIC Reading test (N = 57). Mean differences: Lower (left) m = 53.42 (95% CIs = 35.73, 71.12), sd = 39.34, n = 19, Middle (center) m = 40.00 (95% CIs = 28.32, 51.68), sd = 25.98, n = 19, Upper (right) m = 36.58 (95% CIs = 22.99, 50.17), sd = 30.23, n = 19.

By dividing the participants into groups of 19[3] (rounded rectangles in Figures 1

3 Participants were divided into three groups because groups were equal in size and participants with the same paper-based scores were not in different groups. Comparing three groups might have created artificial group differences. Two and four groups were also investigated. See Appendix B.

and 2), listening scores appear to stabilize in the upper right one-third of Figure 1. This would indicate that students scoring higher on the paper-based TOEIC listening test also generally scored higher on the online test, and that differences between these two TOEIC listening tests were narrower for higher scoring participants. A one-way ANOVA tested whether there were statistical differences between the scale score differences for these three groups (*i.e.*, higher, middle, lower). Assumptions of normality were met. For Listening, the ANOVA was significant with a near large effect [$F(2, 54) = 4.04$, $p = .02$, $\omega^2 = .096$], with the Lower and Upper groups being significantly different from each other. One participant, in the most left of Figure 1, has a difference of 225 scale points between these two listening tests. Temporarily removing this participant also resulted in a significant ANOVA with a medium-sized effect [$F(2, 53) = 3.41$, $p = .04$, $\omega^2 = .079$], with once again the Lower and Upper groups being different. For Reading, the ANOVA, however, indicated that there were no statistical differences between differences in test scores between these three groups (*i.e.*, Lower, Middle, Upper) [$F(2, 54) = 1.44$, $p = .25$, $\omega^2 = .051$); however, the effect size was small-to-medium.

Finally, before combining the data from participants who completed the paper-based TOEIC L&R on Day 1 with those who completed this test on Day 3, the participant who scored 225 points higher on the online listening test than on the paper-based test was permanently removed. Descriptive statistics for the participants ($N = 56$) are displayed in Table 6. The online Listening test data were somewhat negatively skewed. The remaining variables had acceptable values for skewness, kurtosis, and the Shapiro-Wilk test. Histograms, density plots and Q-Q plots were visually inspected, and no unexpected observations were made, with no outliers. Thus, the variables were assumed to be normally distributed; however, taking into consideration the performance of the listening tests on the analyses described above, caution might be warranted. An earlier draft of this paper ran all further analyses with and without the participant with the gap of 225 points on the listening tests. Descriptive statistics for the participants including this participant can be seen in Appendix C.

Table 6. *Descriptive Statistics for TOEIC L&R per Skill per Format (N = 56)*

	Listening		Reading	
	Online	Paper	Online	Paper
M	336.79	299.64	270.27	254.11
Lower 95% *CI* for *M*	321.87	255.27	252.84	237.39
Upper 95% *CI* for *M*	351.71	314.02	287.70	270.83
5% Trimmed *M*	339.40	299.60	271.30	253.70
SD	56.97	54.90	66.54	63.83
Median	340	305	275	235
Variance	3245.84	3013.51	4427.65	3106.49
Min (Max)	175 (440)	185 (440)	95 (415)	135 (385)
Range	265	255	320	250
IQR	85.00	66.25	95.00	93.75
Skewness (SE)	−0.58 (.32)	0.06 (.32)	−0.03 (.32)	−0.02 (.32)
Kurtosis (SE)	0.24 (.63)	.062 (.63)	.92 (.63)	−0.04 (.63)
Shapiro-Wilk (*P*)	.97 (.17)	.99 (.70)	.97 (.25)	.99 (.95)

Results

Research Question 1 – TOEIC L&R Online and Paper-based Score Correlations

Correlation analyses investigated the strength of the relationships between the online and paper-based TOEIC tests for listening and reading. The data met the assumptions required for parametric testing; thus, Pearson's *r* was used. There was a significant correlation between the online and paper-based TOEIC Listening tests [r = .742 (95% CIs = .596, .841), p = <.001], and there was a significant correlation between the online and paper-based TOEIC Reading tests [r = .676 (95% CIs = .503, .797), p = <.001].

Research Question 2 – TOEIC L&R Online and Paper-based Score Comparisons

Paired samples t-tests were run comparing mean scores for each of Listening and Reading (*i.e.*, paper-based and online). The data met the assumptions for parametric tests; thus, Student's t-tests were run. Results are shown in Table 7. Following the Bonferroni correction, the *p*-value for Reading (α_{altered} = .05/2 = 0.025) remained statistically significant. Table 7 also includes effect sizes with confidence intervals. For Listening, the effect size was near large, and for Reading, the effect size was near small, with wide boundaries for both.

Table 7. *Paired Samples Student's t-Tests for TOEIC L&R Online vs Paper-Based*

					95% CI for Effect Size	
TOEIC	Statistic	df	*p*	Effect Size	Lower	Upper
Listening	6.916	55	<.001	0.924	0.608	1.235
Reading	2.303	55	0.025	0.308	0.038	0.575

Note. The effect size for Student's t-test is Cohen's *d*.

Research Question 3 – 2019 and 2020 Cohort TOEIC L&R Data Comparisons

This question was answered using independent samples t-tests comparing scores from the end-of-Year 1 paper-based TOEIC L&R for 2019 (N = 202) with the scores from the paper-based and online version of the TOEIC L&R from the main group of participants in this study (N = 56). For descriptives for the 2019 cohort, see Appendix C. The Shapiro-Wilk Test of Normality and Levene's Test of Equality of Variances were violated for the comparison of the two groups' paper-based scores for Reading; thus, the Mann-Whitney test was used because this test does not require the assumption of normality nor homogeneity of variance (Goss-Sampson, 2020). For the paper-based and online comparison of scores for Reading, the Shapiro-Wilk Test of Normality was violated; however, equality of variance was met. For this comparison, both the Student's t-test and the Mann-Whitney test were used. As shown in Table 8, for Listening and Reading, the differences in mean scores for the paper-based tests were nonsignificant (initially for Listening, and for Reading following the Bonferroni correction). Also, the effect sizes were negligible for Listening and Reading. However, the differences in mean scores between the paper-based and online tests were significant with a medium-sized effect size for Listening, and small-to-medium for Reading, with wide confidence intervals.

Table 8. *Independent Samples t-Test for TOEIC Listening and Reading Scores*

						95% CI for Effect Size	
TOEIC	Comparison	Statistic	df	*p*	Effect Size	Lower	Upper
Listening	Paper-Paper	1.530	256	<.127	0.231	-0.066	0.528
	Paper-Online	5.559	256	<.001	0.840	0.533	1.144
Reading	Paper-Paper	4547.500		<.025	0.196	0.027	0.354
	Paper-Online	4.180	256	<.001	0.631	0.330	0.932
	Paper-Online	3669.500		<.001	0.351	0.193	0.492

Note. *The effect sizes for the Student's t-tests are Cohen's *d;* except for Mann-Whitney test for Reading Paper-Paper which is given by the rank biserial correlation.

Discussion

This study reported on a test-retest design in which approximately 25% of first-year students at UoN completed the paper-based and online TOEIC L&R tests on consecutive days at the end of one academic year. The participants were randomly assigned to complete the paper-based TOEIC L&R either on Day 1 of the research program or Day 3, and the online TOEIC L&R on Day 2. Before combining the data from Day 1 and Day 3, the data underwent close inspection. Mean scores on these two days were similar. It was observed that a large percentage of participants, for both listening and reading, scored ±35 scale points different on the two tests (*i.e.*, online and paper). It was also observed that higher performing participants on the paper-based TOEIC listening test generally performed higher on the online TOEIC listening test. The same phenomenon was not observed for the two reading tests. Before combining the data from test days, one participant was permanently deleted. The correlations (RQ1) between the two listening tests (r = .742) and two reading tests (r = .676) were large; however, these correlations were smaller than those observed by Kanzaki (2017) for two versions of the paper-based TOEIC L&R (r = .80 and r = .84, respectively). The paired sample Student's t-tests results (RQ2) indicated that the mean scores for the two listening tests significantly differed, but after applying the Bonferroni correction, the mean scores for the two reading tests did not, with a medium-sized effect size for the former. Comparing the scores of the 2019 cohort with the participants in this study (RQ3), the results from the two paper-based TOEIC L&R were similar; however the results from the paper-based TOEIC L&R (2019) were significantly different from the online version of the TOEIC L&R (2020).

Mean differences for Listening (37.15) were greater than for Reading (16.16), and these values were much greater than those observed by Cid et al (2017) (1.39 and 3.11 respectively) The value for Listening was also much greater than that observed by Kanzaki (2017 (0.96); however, the value for Reading was comparable (11.46). However, Cid et al and Kanzaki compared two versions of the paper-based TOEIC L&R, whereas this study compared the paper-based and online versions. Score differences between the two versions in this study, in particular for Listening, along with the noted phenomenon that the Listening tests had narrower differences for higher-performing participants, might challenge the claim from IIBC that the two tests, online and paper-based, produce comparable results. Unfortunately, one important limitation with the current study relates to test security. The online version of the TOEIC L&R was completed offsite without proctors overseeing test

security. The possibility of students subverting standard test-taking procedures in order to gain unfair advantages is not zero. However, while this possibility is not zero, no advantages are gained for doing so.

Assuming that both the online and paper-based TOEIC L&R are reliable, producing comparable scores, other explanations are needed, in particular for the differences in listening scores. One possible factor is test-taking motivation (*i.e.*, participants might have been more motivated to complete the online test, for which they had higher scores). A second possible factor is academic dishonesty when completing the online TOEIC L&R (*e.g.*, sharing answers while sitting the test together, reporting questions to those who sat the test later, sitting the test for a another). But why were mean score differences for the listening tests so much greater than those for the reading tests? Might we assume that dishonest participants were those with the greatest differences between the online (higher) and paper-based (lower) scores? Appendix D displays Student's t-test simulation data for paired samples. The rows, with diminishing *n*-sizes, display the t-test results after removing at each new step the top 5% of scores with the greatest differences between the online and paper-based TOEIC Listening tests. In all, 45% of the participants would need to be eliminated before the t-test is nonsignificant and the effect size crosses the zero. For Reading, removing less than 4% of the top values with the greatest differences between the online and paper-based tests resulted in a nonsignificant t-test. In other words, assuming that score difference is accounted for by academic dishonesty, we need to be convinced that approximately half of the participants successfully engaged in academic dishonesty on the online listening test, but not on the online reading test. Although unprovable, it is difficult to believe that nearly half of the participants completed the online test under these conditions with these results.

Assuming that some participants were dishonest when sitting the online TOEIC L&R, what might UoN or other institutions do in the future? To ensure that students follow proper procedures during the online test, institutions might benefit from adopting the following procedures. First, require students to complete an academic integrity pledge before the online exam. Remind students of the academic integrity codes and the consequences for breaking those codes, and require students to agree to or sign, even electronically, the pledge. Second, restrict the test window by starting the test at the same time for all examinees. This would remove the possibility of one examinee reporting questions to others, or one examinee taking the online test for multiple participants. Third, sit the test within an online meeting with cameras turned on, such as during an online synchronous class via Zoom or Microsoft Teams.

This eliminates the possibility that people other than the examinees complete the exam. If this is not possible due to hardware or software limitations, an alternative could be to track IP addresses that reveal the locations of the examinees. Alternatively, safe exam browser software that locks down browser windows and applications that can open during an online test could be used. Safe exam browsers, however, would increase the cost of the test.

However, assuming that participants sat the online test honestly and seriously, the results of this study raise a number of points. Specifically, the results would fail to meet several underlying claims of language assessment validity (Chapelle, 2021). For example, if the online tests are overestimating each participant's scores, compared with the paper-based test, and if the different sections are performing differently, the test scores do not accurately reflect or explain the tested construct of English listening and reading for international communication. Consequently, the participants are likely unable to accept the meaning of the scores. This would be especially true for students whose two scores, on consecutive days vary greatly. These results would also impact generalizability as they do not appear to reflect consistent, or reliable, performance. Therefore, the scores would fail to accurately summarize test-taking performances. However, as has been stated, the possibility of some students having engaged in academic dishonesty is not zero; therefore, challenges to validity need to be accepted cautiously.

Conclusion

This paper compared the mean scores from two TOEIC L&R tests (*i.e.*, the paper-based version and the newer online version). Participants were randomly assigned to complete the paper-based TOEIC L&R test either one day before or one day after the online TOEIC L&R test. The paper-based test results were closely inspected before being combined into one data set. This inspection revealed that (a) a large percentage of participants, for both listening and reading, scored outside the SEdiff of ±35 scale points on the two tests (*i.e.*, online and paper); and higher performing participants on the paper-based TOEIC listening test were more likely to also score higher on the online TOEIC listening test, but the same was not observed for the two reading tests. While paired-skill correlations were large, the paired-sample Student's t-test results indicated that the two listening tests were significantly different with the online test being approximately 37 scale points higher, and the reading tests were not significantly different after applying the Bonferroni correction. This paper is important because there are heretofore no published papers which have compared

the paper-based TOEIC L&R test with its online version. For listening, the differences between the higher online scores and lower paper-based scores challenge the claim from IIBC that the two tests result in parallel scores with parallel interpretations. Test-taking motivation variation might be one factor to explain the differences in scores as shown by the comparisons between scores from the CASEC test at the beginning of the year and the paper-based TOEIC L&R test at the end of the year. Unfortunately, this paper had one important limitation—test security. While the participants sat the paper-based TOEIC L&R test with proctors present, they sat the online test at their homes without proctors. Thus, the probability that some participants used nonstandard test-taking procedures is not zero, and this is a potential factor in the differences in scores. However differences between the results from the pairs of listening tests and the pairs of reading tests lessen this possibility. Finally, one additional possibility was recently suggested. An audiophile has suggested that audio quality differences between the paper-based listening test where the participants sat in a university classroom using built-in ceiling speakers and the online listening test where many participants might have sat the test while using higher quality earphones or headphones that they are used to might account for some of the difference, in particular might account for higher scores on the online listening test.

Acknowledgements

This research was supported by a grant from UoN President Kindaichi Masumi. Prof. Saka Junichi and other English faculty members provided material support. Two research assistants, Oshima Shiori and Sakuyama Rinka assisted with preliminary analyses.

References

CASEC. (n.d., a). *About CASEC*. https://global.casec.com/about/

CASEC. (n.d., b). *CASEC official score report*. https://casec.evidus.com/about/feedback.pdf

ETS. (2019). Score User Guide: TOEIC Listening and Reading Test. https://www.ets.org/s/toeic/pdf/toeic-listening-reading-test-user-guide.pdf

Chappelle, C. (2021). Validity in language assessment. In P. Winke & T. Brunfaut, *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 11-20). New York: Routledge.

Cid, J., Wei, Y., Kim, S., & Hauck, C. (2017). Statistical Analyses for the Updated

TOEIC® Listening and Reading Test. Research Memorandum: *ETS RM*-17-05. ETS. https://www.ets.org/Media/Research/pdf/RM-17-05.pdf

Field, A. (2020). *Discovering Statistics Using SPSS* (5th ed.). Sage Publications.

Goss-Sampson, M. A. (2020). *Statistical analysis in JASP: A guide for students. JASP v0.14.* https://doi.org/0.6084/m9.figshare.9980744

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Duxbury Press.

IIBC. (2020a). TOEIC Program Data & Analysis: Number of examinees and average scores in FY2019. https://www.iibc-global.org/library/default/english/toeic/official_data/pdf/DAA_english.pdf

IIBC. (2020b). About the TOEIC® Listening & Reading Test. https://www.iibc-global.org/english/toeic/test/lr/about.html

IIBC. (2020c). *TOEIC Program IP* テスト(オンライン). https://www.iibc-global.org/toeic/corpo/guide/toeic/online_program.html

IIBC. (2020d). 特集。場所と時間を合わずに活用できる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

IIBC (2020e). TOEIC®リスニング&リーディングIPテスト(オンライン)、AIを活用した試験冠詞サービスの開発に関するお知らせ [Notice regarding the development of TOEIC (R) Listening & Reading IP test (online) and AI-based test monitoring service]. https://www.iibc-global.org/iibc/press/2020/p165.html

Im, G. - H., Shin, D., Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9,* 14. https://doi.org/10.1186/s40468-019-0089-4

JASP Team (2020). *JASP* (Version 0.13.1) [Computer software].

Kanzaki, M. (2017). New and old TOEIC L&R: Score comparison and test-taker views on difficulty level. *PanSIG Journal 2017,* 104-112.

Nikkei (2020 March 10). IIBC、TOEIC Programの団体特別受験制度（IPテスト）にオンライン方式を追加 [IIBC adds online method to TOEIC Program group special examination system (IP test)]. https://www.nikkei.com/article/DGXLRSP530559_Q0A310C2000000/

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

Suzuki, Y. (2021). An overview of new operating practices for private English tests during and after the COVID-19 outbreak. *Tokyo University of Marine Science and Technology Bulletin, 17.* 72-77. http://id.nii.ac.jp/1342/00002042/

Wei, Y. & Low, A. C. (2017). Monitoring Score Change Patterns to Support TOEIC

Listening and Reading Test Quality. *ETS Research Report RR*-17-54. https://www.ets.org/research/policy_research_reports/publications/report/2017/jyez

Appendix A

CASEC

The three cohorts had similar mean CASEC scores: 2018 (m = 566, sd = 78), 2019 (m = 574, sd = 75), and 2019 (m = 577, sd = 69). For each cohort, there were several outliers, kurtosis was high and significant with significant values for the Shapiro-Wilk's test. A Kruskal-Wallis one-way non-parametric ANOVA was run to test for group differences by year. CASEC scores were not significantly different by year $H(2) = 2.43$, $p = .297$, $\varepsilon^2 = .003$.

TOEIC Listening

The 2020 cohort seemed to have higher TOEIC listening mean scores: 2018 (m = 241, sd = 63), 2019 (m = 234, sd = 59), and 2020 (m = 281, sd = 59). Skewness, kurtosis, and the Shapiro-Wilk p-values were significant. A Kruskal-Wallis one-way non-parametric ANOVA was run to test for group differences by year, for TOEIC Listening scores. TOEIC listening scores were significantly different by year $H(2) = 84.30$, $p < .001$, $\varepsilon^2 = .121$. Pairwise comparisons showed that Cohort 2018 and 2019 were not significantly different ($z = 1.31$, $p = .10$); however, Cohort 2020 was significantly different from both Cohort 2018 and Cohort 2019 ($z = 7.11$, $p < .001$, and $z = 8.55$, $p < .001$, respectively).

TOEIC Reading

The 2020 cohort seemed to have higher TOEIC listening mean scores: 2018 (m = 182, sd = 59), 2019 (m = 186, sd = 54), and 2020 (m = 228, sd = 55). Skewness, kurtosis, and the Shapiro-Wilk p-values were significant. A Kruskal-Wallis one-way non-parametric ANOVA was run to test for group differences by year, for TOEIC Reading scores. TOEIC reading scores were significantly different by year $H(2) = 100.22$, $p < .001$, $\varepsilon^2 = .144$. Pairwise comparisons showed that Cohort 2018 and 2019 were not significantly different ($z = 1.35$, $p = .09$); however, Cohort 2020 was significantly different from both Cohort 2018 and Cohort 2019 ($z = 9.19$, $p < .001$, and $z = 7.97$, $p < .001$, respectively).

Appendix B

Independent Samples T-Test for Listening (n = 57)

The data were nonparametric (Levene's test of equality of variance p = .038). A Mann-Whitney test showed that the mean differences for the lower half (Median = 42.50, n = 28) were similar to the upper half group (Median = 25.00, n = 29), U = 526.00, p = .06, with a negligible effect size, r = .30.

Independent Samples T-Test for Listening (n = 56)

Removing the participant with the gap of 225 scale points between the two tests resulted in the data meeting the assumptions of normality. The Student's t-test was nonsignificant, $t(55)$ = 0.91, p = .37, and Cohen's d was small (0.51) but its 95% CIs were wide and crossed zero (-0.03, 1.04).

Independent Samples T-Test for Reading (n = 57)

The data met the assumptions of normality. The Student's t-test was non-significant, $t(54)$ = 1.89, p = .06, and Cohen's d was negligible (0.24), with its 95% CIs being wide and crossing zero (-0.28, 0.76).

ANOVAs for Online and Paper-Based TOEIC Listening with Four Groups (n = 57)

					Post Hoc Comparisons			
	F	Df	p	ω^2	*Groups*	*t*	*Cohen's d*	p_{Tukey}
Listening (n = 57)	4.50	3, 53	0.007	0.16	1 vs 4	3.65	1.23	0.00
Listening (n = 56)[a]	3.99	3, 52	0.012	0.14	1 vs 4	3.34	1.36	0.01
Reading (n = 57)	2.19	3, 53	0.108	0.06	NA			

a. The t-test was rerun after temporarily removing the participant with the largest gap between the online and paper-based scores.

Appendix C

Descriptive Statistics for TOEIC L&R per Skill per Format (N = 57)

	Listening		Reading	
	Online	Paper	Online	Paper
M	337.46	297.02	269.56	253.95
Lower 95% *CI* for *M*	322.74	281.99	252.39	237.52
Upper 95% *CI* for *M*	352.17	312.05	286.74	270.37
5% Trimmed *M*	335.73	297.55	270.49	253.53
SD	56.69	57.90	66.16	63.27
Median	340	305	270	255
Variance	3213.50	3352.55	4377.04	4002.44
Min (Max)	175 (440)	150 (440)	95 (415)	135 (250)
Range	265	290	320	250
IQR	85.00	70.00	95.00	90.00
Skewness (SE)	−0.62 (.32)	−0.13 (.32)	−0.21 (.32)	−0.02 (.32)
Kurtosis (SE)	0.28 (.62)	0.24 (.62)	.01 (.62)	−0.89 (.62)
Shapiro-Wilk (*P*)	.97 (.13)	.99 (.81)	.99 (.97)	.98 (.30)

Descriptive Statistics for Cohort 2019 *(N = 202) for the Paper-Based TOEIC L&R*

	Listening	Reading
M	285.67	233.52
Lower 95% *CI* for *M*	277.13	225.83
Upper 95% *CI* for *M*	294.21	241.21
5% Trimmed *M*	285.00	232.34
SD	61.91	55.74
Median	285	235
Variance	3833.25	3106.49
Min (Max)	125 (475)	110 (415)
Range	350	305
IQR	80.00	75.00
Skewness (SE)	.16 (.17)	.32 (.17)
Kurtosis (SE)	.44 (.34)	.48 (.34)
Shapiro-Wilk (*P*)	.99 (.26)	.99 (.03)

Appendix D

Paired Sample T-Test Results for Online and Paper-Based TOEIC Listening and Reading

Listening						95% CI for Effect Size	
N	Cumulative% Deleted	Statistic	df	p	Effect Size	Lower	Upper
54	5.26	6.61	53	<.001	0.90	0.58	1.21
51	10.53	6.20	50	<.001	0.87	0.54	1.19
48	15.79	5.72	47	<.001	0.83	0.49	1.15
45	21.05	5.14	44	<.001	0.77	0.43	1.10
42	26.32	4.55	41	<.001	0.70	0.36	1.04
39	31.58	3.92	38	<.001	0.63	0.28	0.97
36	36.84	3.24	35	<.001	0.54	0.19	0.89
33	42.11	2.49	32	0.018	0.43	0.07	0.79
32	43.86	2.25	31	0.032	0.40	0.04	0.76
31	45.61	2.00	30	0.054	0.36	−0.01	0.72
30	47.37	1.74	29	0.090	0.32	−0.05	0.68
Reading							
57	0.00	2.26	56	0.028	0.30	0.03	0.56
56	1.75	2.01	55	0.049	0.27	−0.01	0.53
55	3.51	1.76	54	0.083	0.24	−0.03	0.50
54	5.26	1.53	53	0.129	0.13	−0.63	0.48

Note. Student's t-test effect sizes are given by Cohen's *d*.
a Wilcoxon paired sample t-test for nonparametric data was used, the effect size is given by matched rank biserial correlation