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

Jean-Pierre J. Richard
The University of Nagano

Reference Data:

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*. JALT. https://doi.org/10.37546/JALTPCP2022-02

In 2020, an online version of the Test of English for International Communication Listening and Reading (TOEIC L&R) became publicly available. It was promoted as similar in score interpretation as the paper-based version (IIBC, 2020a, 2020b). However, national cohort results from Japan (IIBC, 2022a) revealed large increases in TOEIC L&R mean scores in 2020. In this paper, two studies from one public university are reported. In Study 1, data from five cohorts (*n* = 1205) who completed either the paper-based or online TOEIC L&R were compared. In Study 2, research groups in 2021 (*n* = 56) and 2022 (*n* = 54) completed both versions of the TOEIC L&R. The online test results were generally higher than the paper-based test ones, and approximately 50% of participants had score differences greater than the *SE Diff*. Implications, related to validity, are briefly discussed.

2020年、国際コミュニケーション英語能力テスト（TOEIC L&R）のオンライン版が公開された。TOEIC L&Rのスコア解釈は、ペーパー版と同様であると宣伝された(IIBC，2020a，2020b)。しかし、日本の全国コホート結果（IIBC，2022a）では、2020年にTOEIC L&Rの平均スコアが大きく上昇したことが明らかになった。本稿では、ある公立大学におけるの2つの研究について報告する。研究1では、ペーパー版またはオンラインTOEIC L&Rを受験した5つのコホート（*n* = 1205）のデータを比較した。研究2では、2021年（*n* = 56）と2022年（*n* = 54）の研究グループがTOEIC L&Rの両バージョンを受験した。その結果、オンライン版の結果は、ペーパー版の結果よりも概して高く、約50%の参加者がSE Diff以上のスコア差を示した。妥当性に関連する含意について簡単に述べる。

Shōzan University (a pseudonym), a public university in rural, central Japan, opened in 2018. The university uses the Test of English for International Communication Listening and Reading (TOEIC L&R) for program evaluation. In 2020, at the beginning of the COVID-19 pandemic, a new online TOEIC L&R was used. Results from the online test revealed unusual data patterns when compared to previous cohorts who had completed the paper-based TOEIC L&R. These unusual results were the impetus for the research reported in this paper. In the following literature review, results from the pre-updated and updated versions of the paper-based TOEIC L&R will be compared. Following this, the new online TOEIC L&R will be introduced. Finally, TOEIC L&R data from the Japanese national cohort will be interpreted. Subsequently, the research presented in this paper will be described. Two studies, one comparing TOEIC L&R results from five cohorts, and another comparing two paired-sample groups who completed two versions of the TOEIC L&R, are reported. After the results, test validity is briefly discussed, along with several questions that follow from the research findings.

Literature Review

The change in format in 2020 is not the first time the TOEIC L&R was updated. For example, in 2016, Educational Testing Services (ETS) introduced an updated paper-based TOEIC L&R. Two studies reported on paired-sample designs in which the same test-takers completed both a pre-updated and an updated paper-based TOEIC L&R. In one study, Japanese university students (*N* = 141) had comparable mean scores for Listening (Δ = 0.96) but larger differences for Reading (Δ = 11.46) (Kanzaki, 2018); however, these differences in Reading scores, while significant, had a negligible effect size (*d* = 0.16). In a second study, the results from test-takers in Japan and Korea (*N* = 3,673) indicated that there were minimal differences in mean scores between pre-updated and updated versions of the paper-based test for both Listening (Δ = 3.11) and Reading (Δ = 1.39) (Cid et al., 2017). Thus, Cid et al., researchers at ETS, concluded that the updated paper-based test "continues to have the same psychometric quality of the pre-updated TOEIC test" (2017, p. 15).

In early 2020, the Institute for International Business Communication (IIBC), which operates the TOEIC L&R in Japan, announced a new online version of this test. In publicity announcements, IIBC indicated that this online version was equivalent to the paper-based version (2020a; 2020b). Indeed, the types of questions on the paper-based and online tests are similar. Nonetheless, these two versions of the TOEIC L&R are structurally different in terms of the number of items, time to complete the test, and mode of delivery. The paper-based test is composed of 100 items for each section of Listening and Reading, compared with 45 items each on the online test. As a result, the paper-based test requires more time to complete, 120 minutes, compared with 62 minutes for the online test. Finally, the paper-based test uses the same questions for all test-takers in each unique setting, whereas the online test includes computer adaptive items, comprising approximately 45% each of Listening and Reading, targeting individual test-takers. Richard (2021) provided a richer description of these two tests.

Although IIBC claimed that the online test was equivalent to the paper-based test, unusual results in 2020 bring that claim into question. Publicly available data of Japanese test-takers, of approximately two million per year between 2014 and 2021, from the Secure Program (*i.e.*, IIBC-run TOEIC test centers where test-takers complete the test) and the Institutional Program (*i.e.*, tests administered at individual organizations, including universities) show that for the Secure Program (SP), between 2014 and 2019, the largest year-to-year mean score differences for Reading was 6 points, between 2018 (M = 259) and 2019 (M = 265), and for Listening was 4 points, between 2015 (M = 321) and 2016 (M = 317). See Appendix A. However, in 2020, the year the online test was introduced, mean scores for the national cohort jumped considerably from the previous year, by 17 points for Reading and 14 points for Listening. Importantly, these scores from 2020, and onwards, include data from both the paper-based and online versions of the TOEIC L&R. If the 2020 data included online tests only, the differences between mean scores in 2019 and 2020 might be greater. Note, in publicly available data reviewed by this author, IIBC data does not report the number of paper-based and online test-takers, but rather the combined total only (IIBC, 2022a, p.4 for example). Note also that the mean scores from 2021 are lower when compared with 2020 but not greatly. Appendix A also includes data from the Institutional Program (IP) tests. For both SP and IP, the largest year-to-year differences are between 2019 and 2020 when the online test was introduced.

Statement of Purpose and Research Questions

After being updated in 2016, researchers in Japan (Kanzaki, 2018) and at ETS in the United States (Cid et al., 2017) found the paper-based TOEIC L&R had similar psychometric properties as the pre-updated version. Regarding the new online version, however, while IIBC claimed that it was similar to the paper-based version, structurally the two versions are different, and data from national Japanese cohorts revealed higher scores from 2020 when the online test was introduced compared with pre-pandemic years. As was noted, the impetus for this paper was the unusual data patterns that were observed in the online TOEIC L&R results at Shōzan University in the spring of 2020, which raised questions of test quality. The research reported in this paper examined whether these two versions of the TOEIC L&R, paper-based and online, resulted in equivalent scores for students at Shōzan University. To investigate this, two studies are reported below. The research questions for each study are as follows.

RQ 1: Are results similar across cohorts who used different versions of the TOEIC L&R test?

RQ 2: Do participants who completed both versions of the TOEIC L&R have similar scores?

Methodology and Results

Context

For English class placement, Shōzan University uses the Computerized Assessment for English Communication (CASEC, https://global.casec.com). This test is completed by in-coming first-year students in mid-to-late March before the start of the academic year. For the evaluation of its English language program, students also complete the TOEIC L&R three times over two years, near the beginning of Year 1 (Time 1), at the end of Years 1 (Time 2) and 2 (Time 3). Students are expected to complete these tests, and participation is 96%-99%. Results from this placement test have consistently shown that cohorts are comparatively similar on average. For CASEC details, see Study 1 below for cohort-level comparisons, and Appendix B for results by department. For the TOEIC L&R, two cohorts at Shōzan University, 2018 and 2019, completed the paper-based version; however, in 2020, a new online version was used. Concerns were raised at the university when results from the 2020 cohort differed significantly compared with results from the previous two cohorts. These results from the online TOEIC L&R from the spring of 2020 were the impetus for the two studies reported below. Informed consent and institutional permission was gathered for both studies. In addition, Study 2 was funded by two grants from the president of Shōzan University.

Study 1

In Study 1, data from five cohorts (*n* = 1205) across three departments of non-English majors at Shōzan University were used. Analysis of Variance (ANOVA) tests compared cohort-level mean scores. Assumptions for parametric ANOVA tests were met (*i.e.*, categorical IVs and continuous DVs; cohorts were independent of each other; DVs were normally distributed; no significant outliers were observed; and there was homogeneity of variances between groups). In all, seven ANOVAs were run: CASEC for cohorts 1-5; Time 1 TOEIC Listening and Reading for cohorts 1-5; Time 2 TOEIC Listening and Reading for cohorts 2-4; and Time 3 TOEIC Listening and Reading for cohorts 1-3. For the ANOVA comparing CASEC mean scores, based on the known apparent similarity in CASEC scores for the first three cohorts, it was hypothesized that mean scores would be similar across all five cohorts. For the remaining six ANOVAs, it was hypothesized that these would be significant, and that TOEIC L&R scores would be significantly higher for each cohort that first used the new online TOEIC L&R. Table 1 displays the number of participants for each of the tests between April 2018 and April 2022 for five cohorts, the tests completed, and date.

Table 1

CASEC and TOEIC L&R Testing, per Cohort, n-Size and Date

CASEC	TOEIC Listening & Reading		
Pre Time 1 March	Time 1 (Yr 1 April)	Time 2 (Yr 1 Feb)	Time 3 (Yr 2 Feb)
n (year)	*n* (version, year)	*n* (version, year)	*n* (version, year)
223 (2018)	223 (paper - 2018)	No test	223 (paper - 2020)
238 (2019)	238 (paper - 2019)	242 (paper - 2020)	235 (paper - 2021)
238 (2020)	238 (online - 2020)	242 (online - 2021)	235 (online - 2022)
244 (2021)	244 (online - 2021)	235 (online - 2022)	—
262 (2022)	263 (online - 2022)	—	—

Note. The em dash, —, indicates that that cohort had yet to take this test when these analyses were completed.

Table 2 displays the results for the seven ANOVAs. As hypothesized, the first ANOVA comparing mean scores by cohort for CASEC was non-significant, indicating that each cohort was similar on average as they entered the university, as measured by CASEC. Moreover, within each cohort, the three departments at the university were generally different from each other, but similar across cohorts. For example, the Economics Department had on average the highest mean scores, followed by the Health Department, and the Education Department, and mean score differences generally held across all cohorts. See Appendix B for department-level descriptives for CASEC.

Table 2

ANOVA Results by Cohort

		Cohort *M* (*SD*)							
		2018	2019	2020	2021	2022			
Time	Test	(paper)	(paper)	(online)	(online)	(online)	F[1]	η^2	η_p^2
Pre 1	CASEC	566 (78)	574 (75)	577 (69)	571 (80)	566 (83)	(4, 1200) = 0.94	0.00	0.00
1	Listening	241 (63)	234 (59)	**280 (59)**	268 (62)	241 (65)	(4, 1201) = 25.95	0.08	0.08
	Reading	182 (59)	186 (54)	**228 (55)**	201 (45)	194 (71)	(4, 1201) = 20.42	0.06	0.06
2	Listening	(no test)	277 (66)	**319 (59)**	298 (66)	—	(2, 716) = 25.83	0.07	0.07
	Reading	(no test)	222 (60)	**247 (66)**	236 (72)	—	(2, 716) = 8.42	0.02	0.02
3	Listening	313 (77)	304 (67)	**329 (67)**	—	—	(2, 690) = 8.13	0.02	0.02
	Reading	248 (72)	242 (70)	**282 (72)**	—	—	(2, 690) = 21.67	0.06	0.06

Notes. The em dash, —, indicates that that cohort had yet to take this test when these analyses were completed.

[1] For CASEC, the *F*-test was non-significant ($p = .439$). The remaining six *F*-tests were significant ($p < .001$).

Table 2 also shows the results for the remaining six ANOVAs, three each for TOEIC Listening and Reading. As hypothesized, each of these were significant, with eta squared (η^2) and partial eta squared (η_p^2) effect sizes ranging from small (.02) to medium (.08). Mean score TOEIC L&R data from 2020 (bolded in Table 2), when the online test was introduced, were highest for each test; thus, 2020 can be seen as an inflection point. Previous to 2020, when the paper-based test was used, results were statistically significantly lower, and after 2020, when the online test continues to be used, results have decreased. Furthermore, as hypothesized, for each of Listening and Reading, ad hoc comparisons showed that the cohort which first used the online test had significantly higher mean scores than other cohorts. See Appendix C for ad hoc test results. These results appear to be similar to data from the national cohort.

Study 2

In Study 2, two groups from Shōzan University voluntarily completed both versions, paper-based and online, of the TOEIC L&R: Group A[1], in February 2021 (*n* = 56, 100% 1st-year students); and Group B, in February 2022 (*n* = 54, 76% 2nd-year students, 24% 1st-year students), in the days after their end-of-year final examinations. The participants received no compensation, monetary, academic or other, for participating. However, these were motivated learners, interested in experiencing both tests and knowing their scores. While the test-taking motivations of the participants as they completed the tests is unknown, it was observed by the test proctors that the participants completed the tests seriously.

To avoid a test fatigue effect, in 2021 participants were randomly assigned to complete the online test a day or two before or after all participants completed the paper-based test (*i.e.*, half of the participants took the online test, all wrote the paper-based test, the remaining half completed the online test). In 2022, this procedure was reversed (*i.e.*, half took the paper-based test, all wrote the online test, half completed the paper-based test). In both 2021 and 2022, for both Listening and Reading, before combining data from the randomly assigned groups, each half was checked for normality, and then independent sample *t*-tests were run to ensure that the results from the two halves were similar. No problems were identified, and these halves were combined into one group for the paired-sample *t*-tests.

Four paired-sample parametric *t*-tests were used to compare mean scores from the two research groups who completed both the paper-based and online versions of the TOEIC L&R. The Bonferroni corrected alpha was set to *p* = .0125 (.05/4) The assumptions to run these *t*-tests were met (*i.e.*, DVs were on a continuous scale; IVs were matched-participant groups; differences between matched pairs were normally distributed; no significant outliers were observed). The null hypotheses were that there were no differences between mean test scores.

Table 3 displays the means and standard deviations for both research groups, for the online and the paper-based TOEIC L&R. Differences in means for Listening in 2021 (Δ = 37) and Reading in 2022 (Δ = 33) were particularly large. The participants completed both TOEIC L&R tests a few days apart; thus, new learning cannot account for large differences in scores. Table 3 also displays the paired-sample *t*-test results. The means for the Listening tests in 2021 were significantly different, with a medium effect size (*d* = .92). In 2022, the means for the Reading tests were significantly different, with a small effect size (*d* = .59). See Plonsky and Oswald (2014) for within-group Cohen's *d* labels in L2-language research.

Table 3

Study 2 TOEIC L&R Mean Scores and Paired Sample t-test Results

Group: Year (*n*-size)	Online *M* (*SD*)	Paper *M* (*SD*)	*M* Δ	*t* (df)	*p*	*d*	95% CIs for *d* Lower	Upper
A: 2021 (*n* = 56)								
Listening	336.8 (57.0)	299.6 (54.9)	37.2	6.92 (55)	<.001	0.92	0.61	1.24
Reading	270.3 (66.5)	254.1 (63.8)	16.2	2.30 (55)	.025	0.31	0.04	0.58
B: 2022 (*n* = 54)								
Listening	354.5 (47.4)	348.2 (44.4)	6.3	0.92 (53)	.362	0.13	-0.14	0.88
Reading	314.4 (53.6)	280.7 (51.1)	33.7	4.33 (53)	<.001	0.59	0.30	0.88

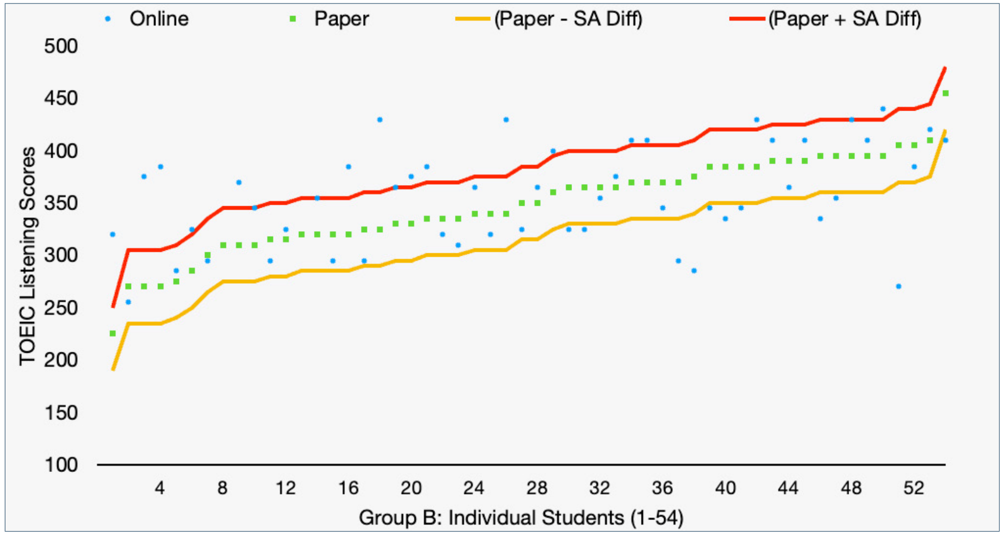
In addition to *t*-tests, the standard error of differences (*SE Diff*) was used to investigate participant-level variation in scores. The *SE Diff* is the error of measurement associated with the difference between scores from two test administrations; and ETS estimates that, for both Listening and Reading, the SE *Diff* is ±35 scaled score points (ETS, 2022, p. 14). This useful statistic allows us to compare results for the same section of the test from two different administrations of that test, to investigate whether an individual's scores have changed.

In 2021, 50% of Group A participants had differences in scores between both Listening tests that were greater than the *SE Diff*, and for Reading this figure was 48%.[2] In 2022, for Group B, these figures were 56% and 57%. Moreover, the percentage of participants whose scores were two times greater than the *SE Diff* (*i.e.*, differences in scores greater than 70 points on the same section between tests) were large in 2021 (Listening: 27%; Reading: 20%) and 2022 (Listening: 15%; Reading: 39%). Thus, while the *t*-tests compare mean scores for the groups, the *SE Diff* provides a participant-level inspection of the data, with different results. For example, mean scores for Listening in 2022 were not statistically different; however, more than half of the students (56%) had differences in scores that were greater than the *SE Diff*.

Figure 1 shows the paired-sample results for TOEIC Listening from February 2022. The green squares represent each of the 54 participants, from lowest-to-highest-scoring, on the paper-based test. The corresponding blue circles on the vertical axis above or below each green square represent the scores from the same participant on the same section on the online test. The orange and red lines represent the lower and upper scores of the *SE Diff* for that participant. Blue circles below the orange line or above the red line represent participants whose online test score was greater than the *SE Diff*. See Appendix D for figures representing the results from TOEIC Listening and Reading from February 2021, and TOEIC Reading from February 2022.

Figure 1

Paired-Sample Results from TOEIC Listening, February 2022 (n =54)



Discussion and Conclusion

In recent years in Japan, TOEIC L&R is taken by approximately 2 million individuals annually (IIBC, 2022a). TOEIC L&R results are used by many Japanese companies during the annual hiring of new recruits and or for company-wide internal promotion (IIBC, 2019, p.14). Recruiting websites advise job-seekers that a combined TOEIC L&R score of 600 is a minimum requirement when job-hunting in Japan, but that higher scores are necessary for particular industries or employers (JWS, 2022; Leverages, 2022). In addition, TOEIC L&R has been adopted by Japanese academic institutions for a variety of purposes, including placement, awarding of credits, and program evaluation (Kanzaki, 2020; Nishitani, 2022). See also “School Case Studies” (IIBC, 2022b). In short, TOEIC L&R serves multiple important functions for employers and academic institutions in Japan. For these reasons, valid and reliable tests are essential.

According to IIBC, the paper-based and online versions of TOEIC L&R are parallel. However, data from Japanese national cohorts (>2 million test-takers per cohort), and results from the two studies reported in this paper revealed that the two versions of the test resulted in different scores, at least when considering data from the inflection point of 2020 when the new online test was introduced. Importantly, data from two paired-sample groups (*i.e.*, study two) indicated that approximately 50% of the test-takers had statistically different scores on the same sections but different versions of the test. This variation in scores cannot be explained by new learning, as these participants completed the two tests within days of each other. This variation in scores likely reduces test validity for these test-takers, and anecdotally several students at Shōzan University who participated in Study 2 reported being confused and disappointed by the large differences in their scores between the two versions of the test. In addition, the large change in scores between versions of the test, and from year-to-year, complicate program evaluations. Academic institutions, including Shōzan University, assume that their students are being evaluated fairly, accurately, and reliably, and that the test can be used as a reliable reference point for the evaluation of the English-program at the university. However, results discussed in this paper lead to many questions.

Why were the results on the online TOEIC L&R from 2020 so much higher compared with the paper-based test? Is this due to differences in test structure? As was noted, the paper-based test has more than twice as many questions as the online test (k = 200 *vs* k = 90) and it takes twice as long to finish the former (2 hours *vs* 1 hour). In addition, the latter test has computer adaptive items. These structural differences might account for some score differences. For example, at Shōzan University, more students finish the online test compared with the paper-based test. In this way, the online test might be a better estimate of the participants' ability compared with the paper-based test which might seem as an exercise in test motivation. That is, more students at Shōzan University are able to maintain their motivation to complete the one-hour online test compared with the two-hour paper-based test. Another possibility for the higher scores in 2020 both nationally and at Shōzan University is cheating occurring in poorly proctored settings, such as test-takers' homes. At the start of the 2020 academic year when there were many unknowns related to the COVID-19 pandemic, Shōzan University opted for the online test in lieu of the paper-based test, and allowed test takers to complete the test at their homes. This was likely the case at many other organizations who used the institutional program (IP). However, is it possible for cheating to go unnoticed on a national scale? Moreover, results from secure program (SP) test sites also have shown that from 2020, scores increased greatly compared with previous years when paper-based test results only were reported. While all instances of possible cheating cannot be ruled out, either at Shōzan University or nationally, it is highly unlikely that cheating on a national scale has taken place at these secure test venues.

One anonymous reviewer pointed out that the large drop in test-takers between 2019 and 2020 might account for the change in scores observed in the national cohort. That is, compared with 2019, in 2020 there were 275,000 fewer SP test-takers and 300,000 fewer IP test-takers, a decrease of 33% and 26% respectively. Then in 2021, the number of test-takers returned to pre-pandemic levels. Perhaps many of these missing test-takers represent unmotivated and or weaker-skilled individuals, and without these individuals the test scores rose in 2020. As the number of test-takers returned to pre-pandemic levels in 2021, test scores regressed towards the mean. This conjecture might partially explain results, especially those from IP test sites, where tests are frequently used for quality assurance, but students might not be obligated to complete. Yet, similar large gains were observed in the national cohort among SP test-takers, and these test-takers likely include individuals who are more invested, because they are likely paying for the test. Moreover, this conjecture of the missing unmotivated and unprepared would not explain the results observed at Shōzan University that were described above.

This leads to further questions. Was the online test in 2020 of "the same psychometric quality " (Cid et al., 2017, p .15) as the paper-based test? Is the online TOEIC L&R "a fair, valid and reliable assessment of everyday and workplace English" (Cid et al., 2017, p .i)? Why does 2020 appear to be an inflection point for scores, but now scores appear to be falling again, reverting to pre-online test levels? Were there problems with the algorithm when the online test was introduced? Is the online test being purposefully made more difficult? Or is the algorithm recalibrating test items or scores in some way? Unfortunately, ETS has yet to publish detailed analytical reports about the online test as Cid et al. (2017) did when the updated paper-based version was introduced. Without transparent reports from ETS or IIBC, these questions will remain unanswered and open to speculation.

Shōzan University has, for the time being, decided to continue to use the online TOEIC L&R for program evaluation. Considering the results reported above, this might seem counterintuitive or inappropriate. However, the shorter test time for the online version allows more students to complete the test, and it is believed by faculty members that this allows more students to perform to the best of their abilities. This might be especially true for most students when taking the test for the first time in April of their first year at the university. In addition to these perceived benefits for students, the online test is easier to administer. However, faculty and administration at the university also recognize that students and future employers might want to know results from the paper-based TOEIC L&R. Therefore, the university also organizes and arranges opportunities for students to take the paper-based tests throughout the year. This, however, increases costs and diverts limited resources, which is particularly burdensome at smaller institutions, such as Shōzan University.

As shown in this paper, cohorts at Shōzan University who completed a new online version of the TOEIC L&R had significantly higher results than those cohorts which completed the paper-based version of this test. This was particularly true for the first cohort, 2020, that completed the new online test. Moreover, for paired-sample groups, there were large differences in individual scores between the two versions of the test. Importantly, with the many questions raised, the results reported here should not be used to generalize to other institutions. At Shōzan University, these results have decreased test validity in the eye of the test-takers, and have caused difficulties regarding program evaluation. While IIBC initially indicated that the new online test was similar to the paper-based test, neither IIBC nor ETS have publicly released detailed reports on the psychometric properties of this new online test. Without a publicly available detailed report similar to Cid et al. (2017), speculation remains about the quality, validity, and reliability of the online TOEIC L&R.

Notes

1. Richard (2021) previously reported data from Group A.
2. Koizumi et al. (2015) used the following formula to calculate the *SE Diff*, <SE Diff = (Time 1 *SD*) * (√[2 – (Reliability at Time 1) – (Reliability at Time 2)])>. Using this

formula results in much smaller *SE Diffs*: *2021,* Listening = 19.7 and Reading = 23.0; 2022, Listening = 16.4 and Reading =18.8. If these calculated *SE Diffs* were used instead of the ±35 scaled score points (ETS, 2022, p. 14), the percentage of students whose scores were different between the two versions would be much greater.

Biodata

Jean-Pierre J. Richard, EdD, is an associate professor in the Faculty of Global Management at the University of Nagano. His research interests include testing and individual differences. <richard.jean-pierre@u-nagano.ac.jp>

Acknowledgements

Study 2 was funded by two research grants from the President of Shōzan University.

References

Cid, J., Wei, Y., Kim, S., & Hauck, C. (2017). *Statistical analysis 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

ETS. (2022). *TOEIC score user guide: TOEIC Listening and Reading Test*. https://www.ets.org/content/dam/ets-org/pdfs/toeic/toeic-listening-reading-score-user-guide.pdf

IIBC. (2019). *Eigo katsuyō jittai chōsa, kigyō dantai/ bijinesupāson 2019* [*Survey on the use of English, companies, businesspersons 2019*]. https://www.iibc-global.org/library/default/toeic/official_data/lr/katsuyo_2019/pdf/katsuyo_2019_corpo.pdf

IIBC. (2020a). *TOEIC program IP tesuto (onrain)* [*TOEIC program IP test (online)*]. https://www.iibc-global.org/toeic/corpo/guide/toeic/online_program.html

IIBC. (2020b). *Tokushū. Basho to jikan o awazu ni katsuyō dekiru IIBC no onrain puroguramu.* [*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. (2022a). *TOEIC Program Data & Analysis 2022, IIBC, 2021-Nendo jukenshasū to heikin sukoa* [*Number of test takers and average scores in 2021*]. https://www.iibc-global.org/library/default/toeic/official_data/pdf/DAA.pdf

IIBC. (2022b). *Gakkō no jirei* [*School case studies*]. https://www.iibc-global.org/toeic/corpo/case.html#anchor02

JWS. (2022 October 14). *Shūkatsu o yūri ni susumeru tōikku sukoa no meyasu ya gyōkai o kaisetsu* [Commentary on TOEIC score guidelines and industry to advance job hunting]. White Career. https://jws-japan.or.jp/whitecareer/blog/4394

Kanzaki, M. (2018). New and old TOEIC L&R: Score comparison and test-taker views on difficulty level. *PanSIG Journal 2017*, 104–112.

Kanzaki, M. (2020). TOEIC Listening and Reading Test and overall English ability. In P. Clements, A. Krause, & R. Gentry (Eds.), *Teacher efficacy, learner agency*. Tokyo: JALT (pp. 559–567). https://doi.org/10.37546/JALTPCP2019-63

Koizumi, R., In'nami, Y., Azuma, J, Asano, K., Agawa, T., and Eberl, D. (2015). Assessing L2 proficiency growth: Considering regression to the mean and the standard error of difference. *Shiken*, *19*(1), 3–15. https://hosted.jalt.org/teval/node/22

Leverages. (2022, September 05). *Tōikku wa shūshoku ni yūri? Shutoku shite okitai tensū ya apīru hōhō o kaisetsu* [Is TOEIC useful for job-hunting? Explanation of the score you should get and how to appeal]. Hataractive. https://hataractive.jp/useful/3039/

Nishitani, A. (2022). Curricular innovation impact analysis: Parallel process growth curve models. In S. J. Ross and M. C. Masters (Eds.), *Longitudinal Studies of Second Language Learning* (pp. 150–170). Routledge. DOI: 10.4324/9781003087939-9

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

Richard, J. - P. J. (2021). A comparison of the online version and paper-based version of TOEIC L&R. *The Global Management* (the University of Nagano), 5, 37–57. http://doi.org/ 10.32288/00001358

Appendix A

Summary of Japan National Cohort Data for TOEIC Listening and Reading, 2014-2021

	2014	2015	2016	2017	2018	2019	2020	2021	Δ^1	Δ^2	Δ^3
Secure Program (SP)											
Reading *M*	262	264	262	261	259	265	282	279	6	17	3
Listening *M*	320	321	317	320	321	323	337	331	4	14	6
N-size (million)	0.91	1.00	0.96	0.96	0.98	0.83	0.55	0.90			
Institutional Program (IP)											
Reading *M*	202	203	203	205	205	206	221	219	2	15	2
Listening *M*	259	260	263	262	266	264	282	279	4	18	3
N-size (million)	1.29	1.32	1.32	1.29	1.24	1.15	0.85	1.0			

Notes. Δ^1 refers to the maximum year-to-year differences between 2013-2019. These are underlined. Δ^2 refers to the difference between 2019 and 2020. Δ^3 refers to the difference between 2020 and 2021.

Appendix B

Descriptives for Department-Level CASEC Results by Cohort

Department	Cohort	Mean	*SD*	*N*
Economics	2018	581.3	74.0	154
	2019	587.4	67.0	168
	2020	587.0	68.3	170
	2021	578.5	81.7	173
	2022	574.5	84.8	188
Health	2018	557.2	66.9	29
	2019	570.5	69.5	30
	2020	572.1	79.1	28
	2021	585.6	67.3	30
	2022	556.4	83.6	32
Education	2018	513.9	78.4	40
	2019	517.9	88.0	40
	2020	539.1	50.8	40
	2021	530.2	68.1	41
	2022	535.8	64.1	42

Note. Department names are pseudonyms.

Appendix C

Ad Hoc Comparisons for TOEIC Listening and Reading by Cohort

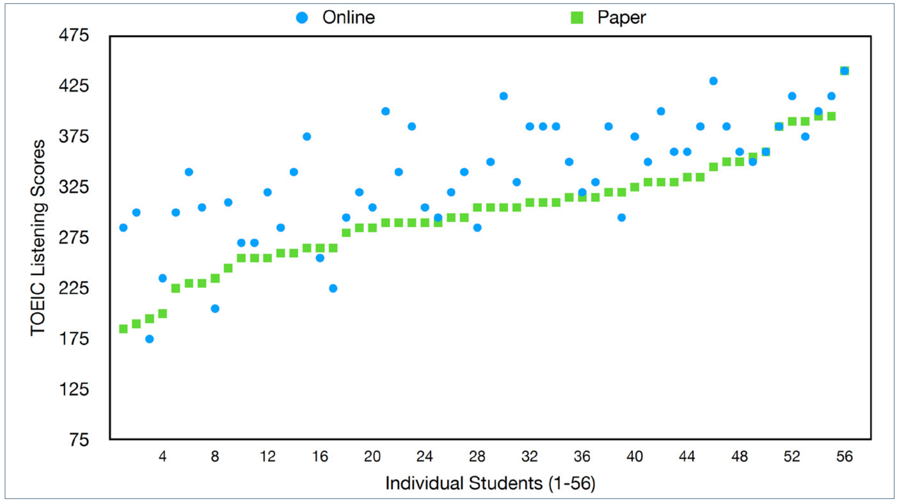
					95% CI for Mean Δ					
Test (Time)	Cohorts		Tests	Mean Δ	Lower	Upper	SE	*t*	*d*	p^{tukey}
L (1)	2018	2019	PP	7.2	-8.4	22.9	5.73	1.26	0.12	0.717
		2020	PO	-39.5	-55.1	-23.8	5.73	-6.89	-0.65	< .001***
		2021	PO	-27.0	-42.6	-11.4	5.70	-4.74	-0.43	< .001***
		2022	PO	0.0	-15.2	15.3	5.60	0.01	0.00	1.000
	2019	2020	PO	-46.7	-62.1	-31.3	5.64	-8.28	-0.79	< .001***
		2021	PO	-34.2	-49.5	-18.9	5.60	-6.11	-0.57	< .001***
		2022	PO	-7.2	-22.2	7.9	5.50	-1.30	-0.12	0.690
	2020	2021	OO	12.5	-2.9	27.8	5.60	2.22	0.21	0.172
		2022	OO	39.5	24.5	54.5	5.50	7.18	0.64	< .001***
	2021	2022	OO	27.0	12.1	42.0	5.47	4.95	0.43	< .001***
R (1)	2018	2019	PP	-4.4	-20.2	11.5	5.79	-0.75	-0.08	0.944
		2020	PO	-46.4	-62.2	-30.6	5.79	-8.02	-0.82	< .001***
		2021	PO	-19.3	-35.0	-3.5	5.75	-3.35	-0.30	0.008**
		2022	PO	-12.5	-28.0	2.9	5.65	-2.22	-0.19	0.175
	2019	2020	PO	-42.1	-57.6	-26.5	5.69	-7.39	-0.77	< .001***
		2021	PO	-14.9	-30.4	0.6	5.66	-2.63	-0.24	0.065
		2022	PO	-8.2	-23.3	7.0	5.56	-1.47	-0.13	0.583
	2020	2021	OO	27.2	11.7	42.6	5.66	4.81	0.44	< .001***
		2022	OO	33.9	18.7	49.1	5.56	6.10	0.53	< .001***
	2021	2022	OO	6.7	-8.4	21.8	5.52	1.22	0.10	0.740

					95% CI for Mean Δ					
Test (Time)	Cohorts		Tests	Mean Δ	Lower	Upper	SE	*t*	*d*	p^{tukey}
L (2)	2019	2020	PO	-42.3	-56.1	-28.5	5.89	-7.19	-0.66	< .001***
		2021	PO	-21.4	-35.3	-7.5	5.93	-3.61	-0.32	< .001***
	2020	2021	OO	20.9	7.0	34.9	5.93	3.53	0.33	0.001**
R (2)	2019	2020	PO	-24.6	-38.8	-10.5	6.02	-4.10	-0.39	< .001***
		2021	PO	-13.6	-27.8	0.6	6.06	-2.24	-0.21	0.065
	2020	2021	OO	11.0	-3.2	25.3	6.06	1.82	0.16	0.163
L (3)	2018	2019	PP	9.6	-5.9	25.0	6.57	1.46	0.13	0.313
		2020	PO	-16.3	-31.7	-0.9	6.57	-2.48	-0.23	0.036*
	2019	2020	PO	-25.9	-41.1	-10.6	6.49	-3.99	-0.39	< .001***
R (3)	2018	2019	PP	6.8	-8.9	22.4	6.67	1.02	0.10	0.568
		2020	PO	-33.8	-49.4	-18.1	6.67	-5.06	-0.47	< .001***
	2019	2020	PO	-40.5	-56.0	-25.1	6.58	-6.16	-0.57	< .001***

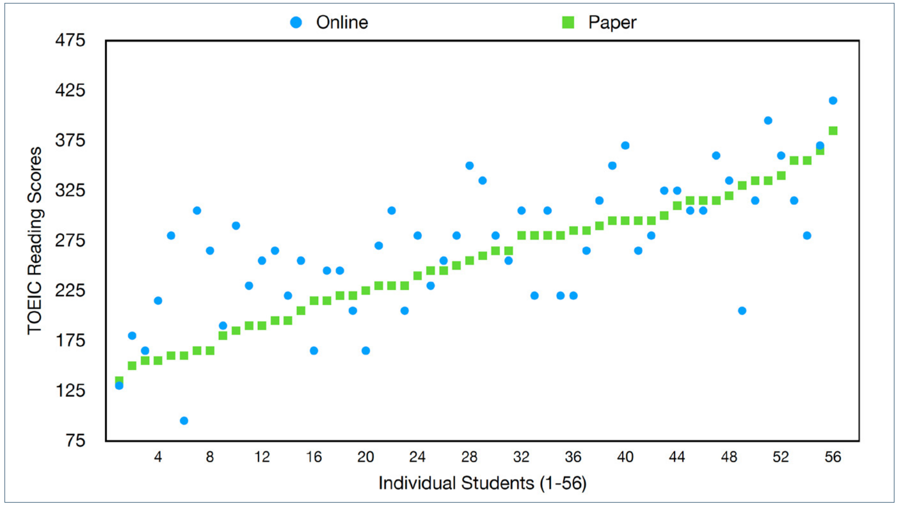
Notes. L and R refer to Listening and Reading tests. Times 1, 2, and 3 refer to Year 1 April, Year 1 February, and Year 2 February respectively. In column 4, “Tests”, PP compares 2 paper-based tests, PO compares 1 paper-based and 1 online test, OO compares 2 online tests.

Appendix D

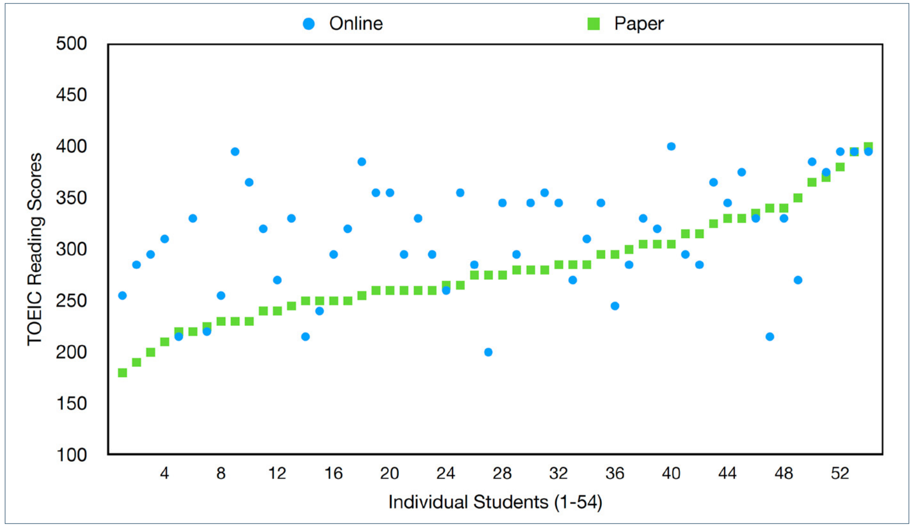
Paired-Sample Results: TOEIC Listening, February 2021



Paired-Sample Results: TOEIC Reading, February 2021



Paired-Sample Results: TOEIC Reading, February 2022


● Online
🟩 Paper
TOEIC Reading Scores
Individual Students (1-54)