New question formats for the TOEIC Listening and Reading test were introduced in May 2016. Due to the complex nature of these formats, learners preparing for the test were concerned that the changes might negatively affect their performance on the test. This study was aimed at determining whether the new and old versions of the test are comparable in terms of scoring and difficulty level. The scores of 141 university students who took both versions were analyzed using descriptive statistics and paired-samples t-tests, and correlations were examined. The average score for the new version was 12 points higher, which is statistically significant but small, and the scores for the two versions correlated at .87. Also, 56 participants responded to a 5-item questionnaire designed to elicit their perception of the difficulty level of the new version. The majority of the respondents indicated that they believed the new version was more difficult.

In November 2015, Educational Testing Service (ETS), the creator of the Test of English for International Communication (TOEIC), announced that question formats in the TOEIC Listening and Reading test (TOEIC L&R) would be changing and that a new version of the exam would be introduced in May 2016 (ETS, 2015a). ETS provided sample questions online (ETS, 2015b), and the Institute for International Business Communication (IIBC), the Japanese administrator of TOEIC programs, posted descriptions of the changes in Japanese on its website along with example questions (IIBC, 2015). When Japanese learners of English who were preparing for the TOEIC L&R reviewed the sample questions, some expressed concern that the changes might negatively affect their performance on the test, resulting in a lower score, as the new question types appeared to be more complex and difficult. In addition, Parts 1, 2, and 5, which are relatively easier than the other sections, would have fewer questions, whereas the harder sections of Parts 3, 6, and 7 would have more.

ETS stated that there would be “no change to the difficulty level” and that TOEIC scores would “continue to be comparable” (ETS, 2015a). However, these assertions cannot be taken at face value, as the organization has a vested interest in maintaining the popularity of the TOEIC L&R. Research was therefore needed to verify their claims.

One way to determine compatibility between the two versions of the test is to have learners take both and then compare their scores. This was possible while the old version continued to be used in the Institutional Program (IP) until the end of March 2017. This is one of two TOEIC programs through which an institution sets the exam time, date, and location, and only those who belong to the institution can take the test. The other program, namely the Public Testing Program (PTP), through which IIBC sets the exam time, date,
and location, administered the new TOEIC L&R in May 2016 for the general public. With two programs simultaneously offering different versions of the test, Japanese learners of English were able to take both up until March 2017, allowing for score comparisons over a short period of time. Taking advantage of the situation, this study was designed to answer the following research questions.

RQ1: How comparable are scores for the new and old versions of the TOEIC L&R?
RQ2: Do test-takers find the new version more difficult than the old version?

**Methods**

The new and old versions of the TOEIC L&R were administered to 141 university students; 81 of these participants took the tests in July 2016 and the other 60 took them in January 2017. Their test scores were analyzed for descriptive statistics, and paired-samples t tests were performed. Correlations between the listening and reading scores for the two versions were also calculated. In addition, 56 of the 60 participants in the January 2017 sessions responded to a 5-item questionnaire, the results of which were compiled and examined.

**Participants**

The participants of this study were 141 university students (121 female and 20 male) who were attending a private Japanese university specializing in foreign languages in the Kanto region. Fourteen of the participants took part in both July 2016 and January 2017, but the related data were treated as if obtained from different individuals on the assumption that their English levels must have been different after 6 months, and therefore they were, in effect, different learners due to the 6-month gap between tests. The participants took part in the study on a voluntary basis and received a 3,000 yen discount on exam fees; the total cost of taking the TOEIC IP test and TOEIC PTP test for each participant was reduced from 8,200 yen to 5,200 yen. Among the participants, 12 were in their first academic year, 26 in their second, 77 in their third, and 26 in their fourth. As for their fields of study, there were 53 English language majors, 52 international communication majors, 12 international business majors, 11 Spanish language majors, four Indonesian language majors, three Vietnamese language majors, three Portuguese language majors, two Chinese language majors, and one Thai language major.

**Materials**

The new and old versions of the TOEIC L&R and a questionnaire were used in this study. The new version was administered in the Public Testing Program (PTP) and the old version in the Institutional Program (IP). IIBC provided the results for both versions, and the scores were analyzed using Microsoft Excel 2013 and IBM SPSS Statistics 23. Those who took part in the January 2017 sessions received the questionnaire online via Google Forms, and 56 of the 60 participants responded. The survey data were then entered into a Microsoft Excel spreadsheet for analysis.

Two Versions of the TOEIC L&R

Both versions of the TOEIC L&R are composed of two subtests: the listening test and reading test. These subtests each consist of 100 multiple-choice questions, and raw scores of between 0 and 100 are converted to scaled scores of between 5 and 495. The listening test has four parts and the reading test has three. What especially makes the latest version of the TOEIC L&R different from the previous version are its new question types and new formats, which are:

1. Conversations with three speakers in Part 3
2. Questions asking the meaning of a phrase or sentence in a given context in Parts 3, 4, and 7 (e.g., What does the woman mean when she says, "I’ll be using the projector a lot"?)
3. Questions with a graphic, such as a table, floor plan, or coupon, that require the test-taker to connect what they see in the graphic with what they hear, in Parts 3 and 4
4. Questions in which the test-taker chooses a sentence that fits in a blank in a Part 6 passage
5. Questions in which the test-taker chooses the right place to insert a particular sentence in a Part 7 passage
6. Text message chains in Part 7
7. Sets of three related passages with five questions in Part 7
Compared to the questions and formats of the previous version of the test, the new additions appeared to be more complex and difficult.

Another major difference between the test versions are the numbers of questions in each part, except Part 4. Table 1. lists the tasks that test-takers perform in each part and the number of questions for each section in the two versions. Where there are fewer questions in the new version (i.e., Parts 1, 2, and 5), the tasks are relatively simple, whereas in the parts with more questions (i.e., Parts 3, 6, and 7), the tasks are more demanding.

Those who participated in July 2016 took a PTP test (new version) on July 24 and an IP test (old version) on July 29. Those in the January 2017 sessions took an IP test on January 26 and a PTP test on January 29.

<table>
<thead>
<tr>
<th>Table 1.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tasks and Number of Questions in Each Part of the Two Versions of the TOEIC L&amp;R</strong></td>
</tr>
<tr>
<td>Part</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

**5-item Questionnaire**

A 5-item questionnaire was administered to elicit the participants’ views on the two versions. The questions were:
1. Which listening test was more difficult, the new (PTP) or old (IP) test?
2. Which reading test was more difficult, the new (PTP) or old (IP) test?
3. Overall, which version was more difficult?
4. When you prepared for the exam, which version did you focus on?
5. Which version of the TOEIC L&R do you prefer?

The questionnaire in this study was in Japanese (see Appendix) and administered via Google Forms so that participants could fill it out online.
Table 2.

Descriptive Statistics for the Scores for the Two Versions of the TOEIC L&R (N = 141)

<table>
<thead>
<tr>
<th>Test</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old L</td>
<td>370.74</td>
<td>55.94</td>
<td>185</td>
<td>495</td>
<td>-0.35</td>
<td>0.82</td>
</tr>
<tr>
<td>New L</td>
<td>371.70</td>
<td>59.56</td>
<td>150</td>
<td>495</td>
<td>-0.62</td>
<td>1.11</td>
</tr>
<tr>
<td>Old R</td>
<td>295.21</td>
<td>73.76</td>
<td>120</td>
<td>480</td>
<td>-0.23</td>
<td>-0.37</td>
</tr>
<tr>
<td>New R</td>
<td>306.67</td>
<td>71.65</td>
<td>95</td>
<td>440</td>
<td>-0.59</td>
<td>0.14</td>
</tr>
<tr>
<td>Old LR</td>
<td>665.96</td>
<td>120.72</td>
<td>330</td>
<td>970</td>
<td>-0.20</td>
<td>-0.05</td>
</tr>
<tr>
<td>New LR</td>
<td>678.37</td>
<td>122.75</td>
<td>245</td>
<td>930</td>
<td>-0.60</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Note. Old = old version; New = new version; L = listening; R = reading; LR = listening and reading combined; Min = lowest score; Max = highest score.

Results

Descriptive Statistics for the Scores for the Two Versions

Scores for the two versions of the TOEIC L&R were examined in three categories: listening scores, reading scores, and total scores (i.e., the listening and reading scores combined). Table 2 shows the descriptive statistics for the three categories. The average scores for the new version were higher in all the categories; the average listening score was 0.96 points higher, the average reading score was 11.46 points higher, and the average total score was 12.41 points higher, compared to the old version. Histograms, normal Q-Q plots, and box plots for the six sets of scores were created to gauge the normality of score distributions, and a visual inspection of these confirmed that the scores were approximately normally distributed. However, the box plots indicated several outliers: For the old version there were two in the listening scores and one in the total scores, and for the new version there were two in the listening scores, two in the reading scores, and one in the total scores, and all were below the bottom boundary. These outliers were not removed from the study so as not to reduce the number of participants.

Paired-Samples T Tests

Paired-samples t tests were conducted to compare the scores for the old and new versions in terms of listening, reading, and total scores. On average, the listening scores for the new version were higher (M = 371.70, SD = 59.56) than those for the old version (M = 370.74, SD = 55.94). This difference, -0.96, BCa 95% CI [-7.82, 5.50], was not significant, t(140) = -0.31, p = .759, and represented a very small effect size, d = 0.02. In other words, there was no significant difference between the listening scores for the two versions. The reading scores for the new version were also higher (M = 306.67, SD = 71.65) than those for the old version (M = 295.21, SD = 73.76). This difference, -11.45, BCa 95% CI [-18.37, -4.82], was significant, t(140) = -3.30, p = .001, but represented a small effect size, d = 0.16, which was below Cohen’s (1988) convention for a small effect (d = .20). This means that there was a statistically significant difference between the reading scores for the two versions, but its effect was small. Similarly, the total scores for the new version were higher (M = 678.37, SD = 122.75) than those for the old version (M = 665.96, SD = 120.72). This difference, -12.41, BCa 95% CI [-22.39, -2.74], was also significant, t(140) = -2.42, p = .017, but represented a small effect size, d = 0.10. Again, between the total scores for the two versions, a statistically significant difference was found with a small effect size.

Correlations

The scores for the two versions were significantly correlated in all three categories; the listening scores, r = .80, the reading scores, r = .84, and the total scores, r = .87 (all ps < .001). Figures 1, 2, and 3 are scatterplots of the listening, reading, and total scores, respectively, for both versions and show how closely the two sets of scores correlated.
Questionnaire
The questionnaire was used only for the January 2017 sessions, and 56 participants responded. Table 3. shows the descriptive statistics for their scores, for both versions and the three categories. Again the average scores for the new version were higher in all categories: The average listening score was 5.98 points higher; the average reading score was 7.76 points higher; and the average total score was 13.78 points higher, compared to the old version.

The first item on the questionnaire asked the participants which listening test they found more difficult (see Table 4. for responses). Nearly three quarters of the respondents indicated that the new version was more difficult.

The second item on the questionnaire asked the participants which reading test they found more difficult (see Table 5. for responses). Nearly two thirds of the respondents indicated that the new version was more difficult.

The third item on the questionnaire asked the participants which version they found more difficult overall (see Table 6. for responses). More than three quarters of the respondents indicated that the new version was more difficult.

The fourth item on the questionnaire asked the participants which version they focused on while preparing for the test, and about three times more of them had focused on the old version, compared to the new version (see Table 7. for responses).

The fifth item on the questionnaire asked the participants which version they preferred, and there were almost twice as many participants who preferred the old version as those who preferred the new version (see Table 8. for responses).

Discussion
First, according to the data obtained in this study, the average listening, reading, and total scores for the new version were higher than those for the old version by 0.96, 11.45, and 12.41 points, respectively. The results of the paired-samples t tests indicated that there was not any significant difference between the two sets of listening scores, but the differences between the reading scores and between the total scores were statistically significant. However, these statistically significant differences are almost negligible in practice for two reasons. One is that their effect sizes were small; Cohen’s ds for the reading score comparison and the total score comparison were 0.16 and 0.10, respectively, which means the differences were a 0.16 standard deviation for the reading scores and a 0.10 standard deviation for the total scores. Their effects are therefore limited. The other is that, as a measurement instrument, the TOEIC L&R is not very accurate when measuring such small differences. ETS (2015c) reports that the standard error of measurement for the TOEIC L&R is about 25 points for each of the listening and reading sections, which means a test-taker’s true total score could be ±50 points of the reported score they receive. Since measurement errors cause a difference of ±50 points to a total score, a difference of around 10 points is not so important by comparison.

Second, the scores for the two versions significantly correlated for the listening subtest, $r = .80$, for the reading subtest, $r = .84$, and for the whole test, $r = .87$ (all $p$s < .001). Usually, this level of correlations is considered to be strong; however, one could argue that the figures should be higher because two versions of the same test were compared, and ETS claims that the two versions are equivalent. It must be remembered, though, that the TOEIC L&R is not a highly accurate measurement instrument, and measurement errors of ±50 points out of 990 points are reported (ETS, 2015c). Spearman (1904) suggested that raw correlations are lower than true correlations because of measurement errors and, therefore, in order to estimate the real correlations, the raw figures need to be corrected on the basis of reliability estimates. ETS does not disclose reliability estimates for the scores of a particular test form, but it reports that the KR-20 reliability index for the listening and reading scores across all forms from their norming samples “has been approximately 0.90 and up” (ETS, 2015c, p. 22). If reliability estimates for the scores of the two forms used in this study were .09, then the raw correlations of .80, .84, and .87 could be corrected to .89, .93, and .97, respectively, by using the formula for correction for attenuation proposed by Spearman (1904). Therefore, considering measurement errors, the figures obtained in this study could be considered high enough to support ETS’s claim that the two versions are equivalent.
Finally, 44 of the 56 respondents to the questionnaire survey indicated that the new version was more difficult. One could argue that the new version was more difficult due to the fact that nearly 80% of the respondents believed it to be so. However, people use their subjective judgment to determine which version is more difficult, and therefore responses can vary from person to person, even among those who have taken the same tests. In addition, two respondents indicated that the old version was more difficult, and 10 respondents were unable to say which version was more difficult.

Even though nearly 80% of the respondents considered the new version to be more difficult, their average score for the new version was higher than that for the old version, which may seem counterintuitive. One reason for this is that ETS made some adjustments to offset the difference in the overall level of difficulty when converting raw scores to scaled scores, as ETS (2015c) explained:

Scaled scores are transformed and derived from test takers’ raw scores through a proven statistical procedure called “equating.” This procedure adjusts for test-form difficulty and establishes the relationship between test takers’ raw and scaled scores so that the scaled scores from different test administrations are comparable. (p. 4)

Thus, even if the new version was more difficult and the participants performed poorly, this was not necessarily reflected in their scores. For example, on an easier test form, test-takers need to get 53% of the questions correct in order to score 500, but on a more difficult form they can get the same score with a lower percentage of correct answers. Considering the results of the score comparisons in this study, it seems that ETS is capable of making these adjustments with a high degree of accuracy.

In conclusion, scores for the two versions could be considered comparable because the score differences were almost negligible, despite their statistical significance, and the correlations between the scores seem reasonably strong. Also, as the majority of the participants thought the new version was more difficult, it is likely that the difficulty level of the TOEIC L&R has increased. However, because ETS makes sufficient adjustments based on test-form difficulty when converting raw scores to scaled scores, TOEIC-takers should not worry that their scores will be negatively affected by the question and format changes.

Acknowledgments
This study was supported by the Center for Foreign Language Proficiency (FLP) at Kanda University of International Studies. The author wishes to express his gratitude to Professor Nobuko Hasegawa, the Director of the FLP, for providing financial aid and useful advice, without which this study would not have been possible.

References


Figure 1.
Scatterplot of the listening scores for the two versions of the TOEIC L&R (N = 141).

Figure 2.
Scatterplot of the reading scores for the two versions of the TOEIC L&R (N = 141).

Figure 3.
Scatterplot of the total scores for the two versions of the TOEIC L&R (N = 141).
Table 3.

Descriptive Statistics for the Scores of the 56 Participants Who Responded to the Questionnaire

<table>
<thead>
<tr>
<th>Test</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old L</td>
<td>374.11</td>
<td>52.56</td>
<td>250</td>
<td>495</td>
<td>0.25</td>
<td>0.35</td>
</tr>
<tr>
<td>New L</td>
<td>380.09</td>
<td>61.04</td>
<td>170</td>
<td>495</td>
<td>−0.85</td>
<td>1.59</td>
</tr>
<tr>
<td>Old R</td>
<td>309.29</td>
<td>65.85</td>
<td>150</td>
<td>480</td>
<td>0.00</td>
<td>0.52</td>
</tr>
<tr>
<td>New R</td>
<td>317.05</td>
<td>67.10</td>
<td>110</td>
<td>440</td>
<td>−0.78</td>
<td>1.01</td>
</tr>
<tr>
<td>Old LR</td>
<td>683.39</td>
<td>109.18</td>
<td>400</td>
<td>970</td>
<td>0.24</td>
<td>0.77</td>
</tr>
<tr>
<td>New LR</td>
<td>697.17</td>
<td>117.72</td>
<td>350</td>
<td>930</td>
<td>−0.63</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Note. Old = old version; New = new version; L = listening; R = reading; LR = listening and reading combined; Min = lowest score; Max = highest score.

Table 4.

Responses to Question 1 (n = 56)

<table>
<thead>
<tr>
<th>New (PTP)</th>
<th>Old (IP)</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>41 (73.3%)</td>
<td>3 (5.4%)</td>
<td>12 (21.4%)</td>
</tr>
</tbody>
</table>

Note. Question 1 = Which listening test was more difficult, the new (PTP) or old (IP) test?

Table 5.

Responses to Question 2 (n = 56)

<table>
<thead>
<tr>
<th>New (PTP)</th>
<th>Old (IP)</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>36 (64.3%)</td>
<td>6 (10.7%)</td>
<td>14 (25.0%)</td>
</tr>
</tbody>
</table>

Note. Question 2 = Which reading test was more difficult, the new (PTP) or old (IP) test?

Table 6.

Responses to Question 3 (n = 56)

<table>
<thead>
<tr>
<th>New (PTP)</th>
<th>Old (IP)</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>44 (78.6%)</td>
<td>2 (3.6%)</td>
<td>10 (17.9%)</td>
</tr>
</tbody>
</table>

Note. Question 3 = Overall, which version was more difficult?

Table 7.

Responses to Question 4 (n = 56)

<table>
<thead>
<tr>
<th>New (PTP)</th>
<th>Old (IP)</th>
<th>Both equally</th>
<th>Didn’t do much preparation</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 (12.5%)</td>
<td>23 (41.1%)</td>
<td>6 (10.7%)</td>
<td>20 (35.7%)</td>
</tr>
</tbody>
</table>

Note. Question 4 = When you prepared for the exam, which version did you focus on?

Table 8.

Responses to Question 5 (n = 56)

<table>
<thead>
<tr>
<th>New (PTP)</th>
<th>Old (IP)</th>
<th>Hard to say</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 (23.2%)</td>
<td>25 (44.6%)</td>
<td>18 (32.1%)</td>
</tr>
</tbody>
</table>

Note. Question 5 = Which version of the TOEIC do you prefer?
Appendix

Original Questionnaire in Japanese
Q1. リスニングはどちらが難しく感じましたか。
   1) 公開テストの方が難しい。 2) 学内 IP の方が難しい。 3) どちらとも言えない。
Q2. リーディングはどちらが難しく感じましたか。
   1) 公開テストの方が難しい。 2) 学内 IP の方が難しい。 3) どちらとも言えない。
Q3. 総合するとどちらが難しく感じましたか。
   1) 公開テストの方が難しい。 2) 学内 IP の方が難しい。 3) どちらとも言えない。
Q4. テストに向けた学習は新旧どちらの形式がメインでしたか。
   1) 新形式がメイン。 2) 旧形式がメイン。 3) 両方同じくらい。 4) テスト対策は特にしなかった。
Q5. 新形式と旧形式ではどちらが好きですか。
   1) 新形式の方が好き。 2) 旧形式の方が好き。 3) どちらとも言えない。

Author’s Biography

Masaya Kanzaki teaches at Kanda University of International Studies. His research interests include language testing, vocabulary acquisition, and corpus linguistics. He can be reached at kanzaki-m@kanda.kuis.ac.jp.