acoustics

Article

Influence of Test Room Acoustics on Non-Native Listeners' Standardized Test Performance †

Makito Kawata [1], Mariko Tsuruta-Hamamura [2] and Hiroshi Hasegawa [2,*]

[1] Department of English, Kanda University of International Studies, Chiba 261-0014, Chiba, Japan; kawata-m@kanda.kuis.ac.jp

[2] Graduate School of Engineering, Utsunomiya University, Utsunomiya 321-8585, Tochigi, Japan; mariko@is.utsunomiya-u.ac.jp

* Correspondence: hasegawa@is.utsunomiya-u.ac.jp

† Portions of this work appear in the following conference paper: Kawata, M.; Tsuruta-Hamamura, M.; Hasegawa, H. Exploring relationships between acoustic conditions and non-native listener performance in standardized foreign language test rooms. In Proceedings of the 29th International Congress on Sound and Vibration, Prague, Czech Republic, 9–13 July 2023.

Abstract: Understanding the impact of room acoustics on non-native listeners is crucial, particularly in standardized English as a foreign language (EFL) proficiency testing environments. This study aims to elucidate how acoustics influence test scores, considering variables overlooked in prior research such as seat position and baseline language proficiency. In this experiment, 42 Japanese university students' performance on standardized EFL listening tests was assessed in two rooms with distinct acoustic qualities, as determined by the speech transmission index (STI) and reverberation time (RT). The rooms differed significantly in their STI values and RT measurements, with one exhibiting high speech intelligibility qualities of $\geq$0.66 STI and $RT_{0.5\text{–}2kHz}$ $<$ 0.7 s and the other falling below these benchmarks. The findings revealed that listening test scores were consistently higher in the acoustically favorable room across all participants. Notably, the negative effect of poor acoustics was more pronounced for students with lower baseline language proficiency. No significant score differences were observed between front- and rear-seat positions, suggesting that overall room acoustics may be more influential than individual seating locations. The study concludes that acoustics play a significant role in the standardized EFL test performance, particularly for lower-proficiency learners. This highlights the necessity of standardized testing environments to be more carefully selected in order to ensure the fair and reliable assessment of language proficiency.

Keywords: room acoustics; speech transmission index; reverberation time; speech perception; non-native listeners; language assessment; standardized testing

check for updates

Citation: Kawata, M.; Tsuruta-Hamamura, M.; Hasegawa, H. Influence of Test Room Acoustics on Non-Native Listeners' Standardized Test Performance. *Acoustics* **2023**, *5*, 1161–1179. https://doi.org/10.3390/acoustics5040066

Academic Editors: Dadi Zhang and Massimiliano Masullo

Received: 15 November 2023
Revised: 1 December 2023
Accepted: 7 December 2023
Published: 11 December 2023

1. Introduction

Standardized English as a foreign language (EFL) proficiency tests are widely used and relied upon as a measurement of non-native English (L2) users' language abilities. Notable among them are the International English Language Testing System (IELTS), the Test of English as a Foreign Language (TOEFL®), and the Test of English for International Communication (TOEIC®). These assessments serve as trusted benchmarks for millions of yearly test takers as well as over tens of thousands of stakeholders including educational institutions, employers, and various organizations across more than 140 countries [1–3]. The reliance on such standardized tests is intrinsically linked to the global status of English, which is estimated to be spoken by 373 million native (L1) speakers and over a billion second-language (L2) speakers [4]. Attaining a high level of English proficiency, especially for aspiring students and professionals, is not only a tool for communication but also an essential key to unlocking opportunities in an increasingly interconnected world.

Given the ubiquity of standardized EFL proficiency tests, research concerning L2 listening comprehension in various acoustic conditions is paramount. This is because the paper-based versions of these tests are still regularly administered around the world. In contrast to computerized tests that are conducted individually and allow the use of headphones, paper-based tests are designed to be administered simultaneously to large groups of test takers in various types of venues and employ loudspeakers to deliver the audio prompts for the listening section. Examples include IELTS on paper [5], TOEFL ITP® [6], and TOEIC® Listening & Reading (L&R) SP and IP [7]. The TOEFL ITP, short for Institutional Testing Program, and TOEIC IP, short for Institutional Program, are tests for which the administration responsibilities are entrusted to individual organizations such as companies and universities. In addition, online versions were developed in response to the COVID-19 pandemic: IELTS Online [8], TOEFL iBT Home Edition [9], and TOEIC IP Test Online [10]. Although demand has presumably subsided in recent months, information presented on their websites indicates that these online options are currently still operational.

Besides the globally encompassing nature of standardized EFL tests, the pertinence of acoustic research to standardized testing scenarios has also been underscored by numerous anecdotal reports of test takers expressing dissatisfaction with the listening conditions during tests, which range from complaints about the low quality of sound sources to the poor acoustics of the test room itself. Moreover, the first author has encountered these challenges firsthand as a test examiner of the TOEIC L&R IP test administered multiple times a year at the authors' affiliated institution. Despite such warning signs pointing to the need for investigations into the impact of test room acoustics on test taker listening comprehension, studies focusing on this specific issue are scarce. This gap in research should not be ignored as standardized EFL tests serve as critical gateways for non-native speakers pursuing academic and professional opportunities, and any disadvantage posed by poor acoustics, if indeed such distinctions exist between test rooms, could have far-reaching consequences.

1.1. Background

In the current body of the literature, the performance of L2 listeners in indoor spaces has been examined in relation to a wide gamut of acoustic parameters such as the speech transmission index (STI) [11–13], reverberation time (RT) [14–17], background noise level (BNL) [18–21], speech-to-noise ratio (SNR) [22–27], listening effort [28–30], and source-receiver (speaker to listener) distance [31,32]. Research has consistently shown that low STI and high RT, especially in environments with high BNL, can significantly impede L2 listeners' performance in a classroom-like setting. For example, Yang and Mak (2018) [13] investigated speech intelligibility in middle school and university classrooms for second language students aged 12 to 21 and found that not only did intelligibility scores improve with higher STI values across all age groups but also that this correlation became more pronounced as age increased. The study also found that English speech intelligibility scores in Hong Kong were lower than those for native languages under the same STI conditions, which highlights another commonly corroborated idea that L1 listeners consistently outperform L2 listeners in adverse conditions.

MacCutcheon et al. (2019) [16] explored how L2 vocabulary and executive functioning (number updating) abilities of 57 bilingual university students relate to listening comprehension in two RT conditions ($RT_{0.5\text{–}1kHz}$ 0.3 s and 0.9 s (T_{30})) and found that lower RT positively influenced L2 speech comprehension. They discovered that the group with higher L2 vocabulary levels performed 22% better in comprehension in the longer RT condition and 9% better in the shorter RT condition compared to those with lower vocabulary levels. Notably, the advantages of better acoustics for listening comprehension were significant only for the group with higher vocabulary skills, which implies that larger vocabulary knowledge in the L2 can protect against the detrimental effects of reverberation.

A review of the literature by Scharenborg and Os (2019) [21] shed light on the challenges L2 listeners face in background noise, positing that while background noise makes spoken-word recognition harder for L2 compared to L1 listeners, the difference is attributed to varying language exposure rather than the noise itself. Still, the review highlighted that L2 listeners tend to have a disadvantage in noisy environments regardless of proficiency. Examining the effects of different noise types, informational masking, such as competing speech, was found to be more disruptive than energetic masking, like speech-shaped noise. The study also noted that reverberation did not disproportionately affect bilinguals compared to monolinguals and that L2 listeners are less able to use semantic contextual information when listening in noise.

Lastly, Peng and Wang (2019) [30] studied the listening effort required by 115 native and nonnative English speakers in five reverberant ($RT_{0.5\text{–}2kHz}$ 0.37 s, 0.62 s, 0.84 s, 1.05 s, 1.19 s (T_{20})) and three BNL (pink noise calibrated to fit the Room Criteria contours for RC-30 (38 dBA), RC-40 (48 dBA), and RC-50 (58 dBA)) conditions as well as when listening to speech with a foreign accent. The results indicated that nonnative listeners found understanding speech in adverse acoustic conditions more challenging than native listeners, especially at higher BNLs. Interestingly, the performance on the adaptive pursuit rotor (APR) task [33], used to objectively measure listening effort, did not significantly change across different acoustic conditions, but the participants (native and non-native) subjectively reported increased effort when understanding Chinese-accented English speech. Only the non-native listeners who shared the same Chinese accent as the talker reported no change in listening effort, which highlights the influence of talker accent on the listening effort of nonnative listeners.

The studies above represent the wealth of knowledge available in the existing literature that can inform investigations on themes related to L2 listeners' performance in various acoustic conditions. However, an incomplete picture still emerges in addressing the highly specialized experience of L2 listeners during a standardized EFL proficiency test and what specific needs it entails.

1.2. The Present Study

The present study builds on two previous works by the authors which suggest that acoustic inconsistencies across test rooms may be placing certain test taker populations at an unfair disadvantage, particularly during the listening section of standardized EFL proficiency tests [34,35]. The earlier investigation [34] analyzed a substantial dataset comprising over 20,000 scores from TOEIC IP tests administered over a 12-year period. Although statistically significant differences were found between the scores from the test room exhibiting the worst acoustic condition (0.59 STI, $RT_{0.5\text{–}2kHz}$ 0.87 s (T_{20})) against those from all seven other test rooms, the analysis was limited to comparing mean scores from entire rooms due to the lack of data, specifically information concerning the seat position of test takers and their baseline proficiency levels. Subsequent research [35] aimed to address some of these limitations by examining the speech transmission index for public address systems (STIPA) [36] and RT for different sound sources typically employed in standardized test administration (wall-mounted speakers, portable radio cassette player, and portable amplified speaker) and extending the scope from eight to 10 test rooms. Significant variations were observed not only between test rooms but also between sound sources within the same room, and even across different seat locations. Despite these insights, the question of how these acoustic variations would affect test takers' actual performance remained unanswered. The current work addresses this gap in research by obtaining empirical data from standardized EFL proficiency listening tests administered to participants under controlled conditions in acoustically well-documented test rooms.

Additionally, the present study employs the Common European Framework of Reference for Languages (CEFR) [37], specifically its Japanese adaptation (CEFR-J) [38], to categorize participant proficiency levels. The CEFR is an internationally recognized standard that catalogues a wide range of language abilities expressed through can-do descriptors and organizes them into six levels representing the progressive stages of language development (A1 = beginner, A2 = elementary, B1 = pre-intermediate, B2 = intermediate, C1 = upper-intermediate, C2 = advanced). The CEFR-J refines this further by dividing A1 into three sublevels (A1.1, A1.2, A1.3) and A2, B1, and B2 into two sublevels (A2.1, A2.2, B1.1, B1.2, B2.1, B2.2) to more precisely reflect the English language proficiency of Japanese university students which tend to fall within these ranges. The CEFR-J was adopted for this study to facilitate a more standardized comparison of L2 listening comprehension results across varying acoustic conditions. While existing research provides valuable insights into how acoustic qualities such as STI, RT, and BNL affect L2 speech intelligibility, the lack of a standardized proficiency scale in these studies has often led to conclusions that are not easily generalizable. Typically, categorizations of test subjects into proficiency groups such as "high" and "low" are confined to the context of individual studies, which presents challenges for cross-study comparisons. By aligning our results with the CEFR-J, this study aims to provide a clearer depiction of the influence of acoustic conditions on a defined continuum of language proficiency levels. This alignment not only enhances the comparability of our findings with other research but also contributes to a more cohesive body of knowledge that can inform both future academic inquiry and practical language assessment considerations.

In sum, this study addresses the gaps in current research on L2 listening comprehension in standardized EFL test rooms by specifically investigating the impact of different acoustic conditions on test takers' listening performance. Our approach is novel in several ways: (1) We simulate actual test conditions by conducting the experiments in test rooms routinely used to administer standardized EFL tests, preparing the rooms to reflect conditions typically stipulated in test examiner manuals, and employing official test materials developed by Educational Testing Service, the organization responsible for various standardized tests including TOEFL and TOEIC. (2) Data analysis is performed in light of test takers' baseline proficiency levels and seat positions. (3) The results are aligned with the CEFR-J proficiency scale to promote greater generalizability and comparability with other studies. (4) The findings offer practical insights for improving the selection of test rooms and the equity of standardized EFL assessments.

2. Materials and Methods

2.1. Participants

Forty-three Japanese university students enrolled at the authors' affiliated institution agreed to participate in this study. The participants comprised 38 first year, 4 second year, and 1 third year students, of which 6 were males and 37 were females. All participants were born and raised in Japan with Japanese parents, having started learning English as a foreign language either in the final year of elementary school or the first year of middle school. All self-reported having no hearing impairments, although 4 participants also reported having occasional minor difficulties hearing in certain contexts.

2.1.1. Baseline Proficiency Levels

Score results from the TOEIC L&P IP Test Online administered to the participants by the university English program on two separate occasions (once in July 2022, approximately six months prior to this study; once in January 2023, approximately one week after this study) were obtained, and the mean listening scores were used to determine the participants' baseline English proficiency levels for this study. A boxplot of the mean listening scores (enclosed in orange) along with their corresponding reading and total (listening and reading) scores are shown in Figure 1a.

The listening scores were further analyzed using the CEFR-J [38]. Analysis of the mean listening scores from the two TOEIC L&R IP Online tests revealed that 3 participants were at the B2.1 level, 15 were B1.2, 18 were B1.1, and 7 were A2.2 (which included 1 A2.1) level. These were then combined to create two groups, CEFR-J High (n = 18) and CEFR-J Low (n = 25), which are shown enclosed in orange in Figure 1b.

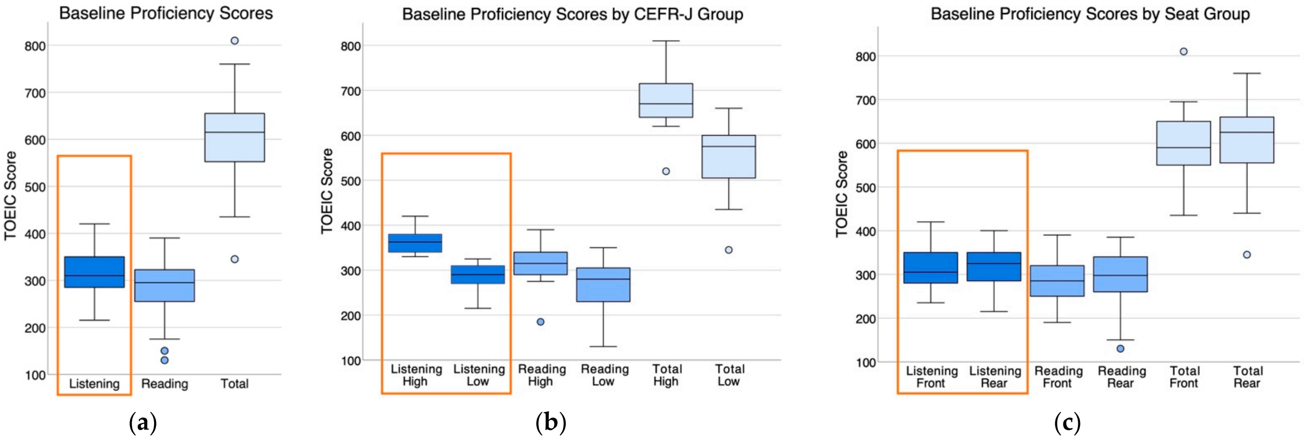


Figure 1. Boxplots of the participants' TOEIC listening, reading, and total scores for the (**a**) Entire Set, (**b**) CEFR-J High and Low, and (**c**) Front and Rear Seat groups. The listening scores enclosed in orange represent the participants' baseline English proficiency levels for each group.

The third set of boxplots shown in Figure 1c presents the participants' baseline proficiency scores according to their seat groups. The seats were assigned, forming a Front Seat group (n = 21) and a Rear Seat group (n = 22) with mean proficiency scores of 319.0 (SD = 51.8) and 320.2 (SD = 44.5), respectively. Details of the seat assignment procedures are presented in Section 2.3.1.

2.2. Test Rooms

The experiment for this study was conducted in two classrooms at the authors' affiliated institution, a national university in Japan regularly employed as a venue for the TOEIC L&R SP and IP tests. Out of the 10 classrooms investigated in [35], Room 5 and Room 9 were selected for this study as representatives of acoustically favorable and unfavorable environments, respectively. A summary of the two rooms is provided in Table 1.

Table 1. Summary of Rooms 5 and 9.

Room	Length (m)	Width (m)	Height (m)	Volume (m^3)	Total Seating Capacity	Exam Seating Capacity	Seating Style	Floor Type	Acoustic Classification
5	14.22	13.80	3.22	631.88	230	128	Front-facing	Tiered	Favorable
9	14.66	8.41	2.72	335.35	130	66	Front-facing	Flat	Unfavorable

2.2.1. Sound Sources

Of the three types of loudspeakers investigated in [35] (wall-mounted speakers, portable radio cassette player, and portable amplified speaker), wall-mounted speakers (WMS) were chosen for this study as the sound source through which the listening test audio was delivered to the participants. WMS were chosen due to their logistical convenience as well as because university records indicated that they have been used to administer the TOEIC L&R IP test at this university repeatedly at least throughout a span of 12 years. Specifications of the audio equipment in Room 5 and Room 9 are presented in Table 2.

2.2.2. Speech Transmission Index

Table 3 shows the STIPA measurement results for WMS at each of the six seat positions assigned to the participants (see Section 2.3.1). Also presented are the mean and standard deviation (SD) for the entire room, for which n denotes the total number of measurement positions in Room 5 and 9. Refer to [35] for details of the STIPA measurement procedures.

Table 2. Specifications of the wall-mounted speakers in Rooms 5 and 9.

	Power Amplifier		Loudspeaker			
Room	**Brand and Model Number**	**Output Power (W)**	**Brand and Model Number**	**Type**	**Components**	**Units**
5	JVC PA-806 [1]	60 W	Panasonic WS-AT80 [3]	2-way	20 cm woofer + SCWG horn tweeter	2
9	Yamaha MA2030a [2]	30 W	Sony SRP-S320 [4]	1-way	12 cm full range	2

[1] Yokohama, Kanagawa, Japan. [2] Hamamatsu, Shizuoka, Japan. [3] Kadoma, Osaka, Japan. [4] Minato City, Tokyo, Japan.

Table 3. STIPA measurement results in Rooms 5 and 9. 'Left' and 'Right' notations represent seat positions as seen from the lecture podium facing towards the rear wall. For results from the entire room, the total number of measurement positions (n), mean, and standard deviation (SD) are presented.

	Front Row			Rear Row			Entire Room		
Room	**Left**	**Center**	**Right**	**Left**	**Center**	**Right**	***n***	**Mean**	**SD**
5	0.75	0.66	0.81	0.70	0.72	0.71	25	0.72	0.04
9	0.61	0.58	0.62	0.56	0.56	0.50	15	0.56	0.03

2.2.3. Reverberation Time

Figure 2 reports the RT measurement results in octave-band frequency from 125 Hz to 8 kHz as well as for the mid-band frequencies $RT_{0.5–1kHz}$ and $RT_{0.5–2kHz}$. Refer to [35] for details of the RT measurement procedures.

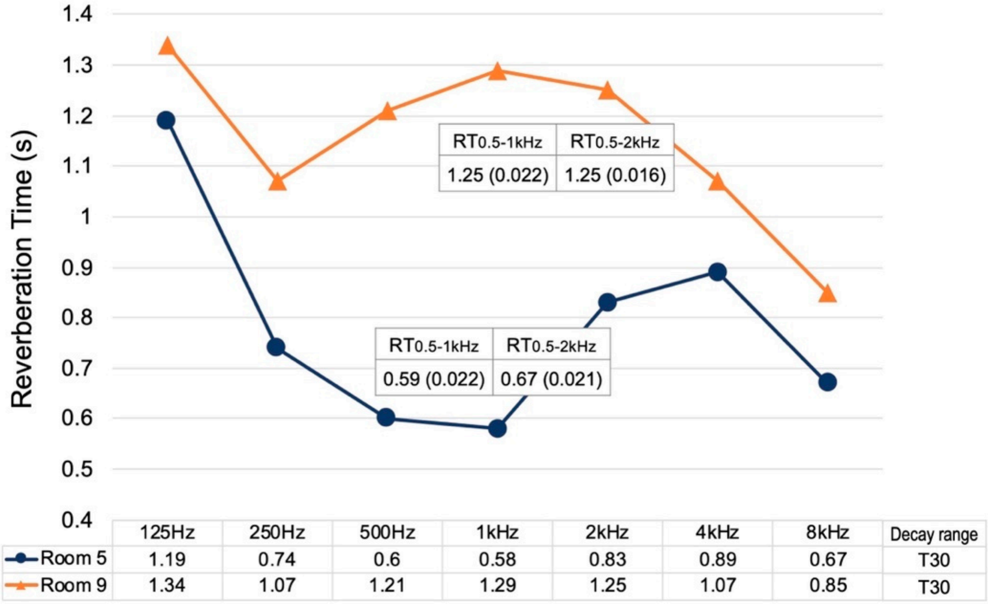


Figure 2. RT measurement results in Rooms 5 and 9. Standard deviations for the mid-band frequencies are given in parenthesis.

2.3. Experiment Procedures

Upon recruitment, the participants were instructed to attend two listening test sessions, one in Room 5 and the other in Room 9. From a list of available time slots for each room, the participants were free to attend any time slot of their choice. The test sessions in the two rooms were administered roughly one week apart and conducted by the first author.

2.3.1. Seat Positions

Each participant was assigned to one of six seat positions (right, center, and left seats in the frontmost and rearmost rows) based on their baseline proficiency score to ensure that two groups similar in size and proficiency level were formed. They were assigned to the same position in both rooms to maintain consistency across the two groups. As mentioned in Section 2.1.1, this resulted in the formation of a Front Seat group (n = 21) and a Rear Seat group (n = 22) with mean proficiency scores of 319.0 (SD = 51.8) and 320.2 (SD = 44.5), respectively. Boxplots of the two groups' baseline proficiency scores (enclosed in orange) are shown in Figure 1c, and the seat positions are indicated in Figure 3.

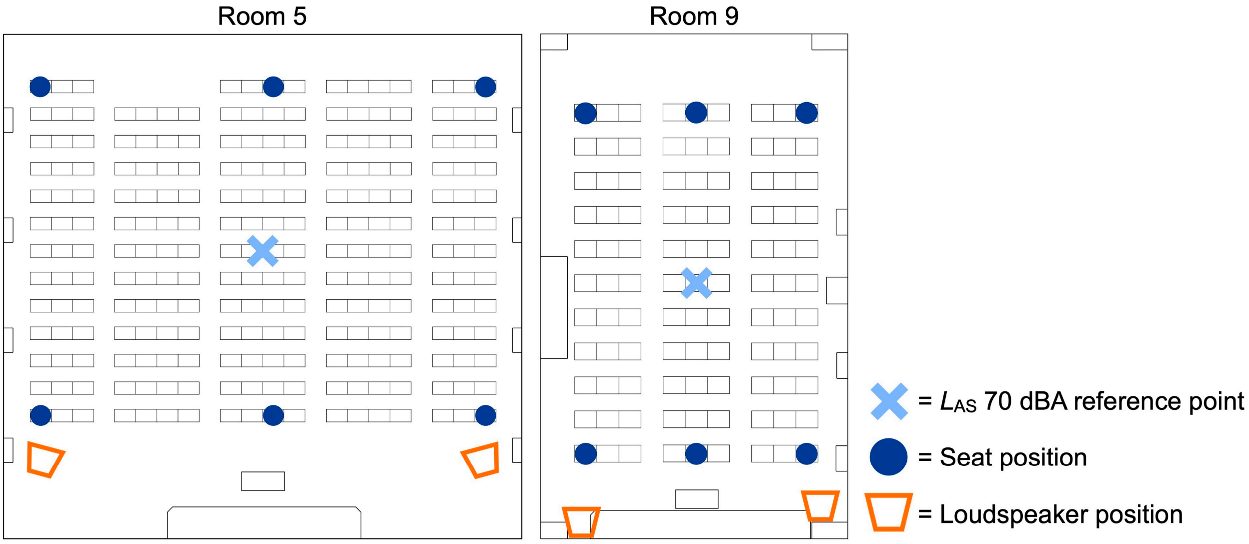


Figure 3. Seat positions and A-weighted sound pressure level set to 70 dBA at a reference point in Rooms 5 and 9.

2.3.2. Test Material

Two practice tests included in the Official TOEIC® Listening & Reading Workbook 9 [39] were employed as test materials for this study. Test 1 was administered in Room 9 while Test 2 was carried out in Room 5. For each test, the full 45 min audio CD containing the listening test prompts for 100 questions was played in its entirety. As this study only concerned the participants' listening performance, the reading section of the practice test was not administered.

2.3.3. Sound Pressure Level

Using the NTi Audio STIPA test signal CD V1.1 (NTi Audio, Schaan, Liechtenstein), the listening test audio signal was set to an A-weighted sound pressure level (SPL) of 70 dBA at a reference point located at the center of the seating area in each room. The reference SPL was measured with the NTi Audio XL2 Audio and Acoustic Analyzer and the M4260 omnidirectional condenser microphone (NTi Audio, Schaan, Liechtenstein) positioned at a height of 1.2 m from the floor. The reference point is indicated in Figure 3.

2.3.4. Room Preparation

The two rooms were prepared to replicate as much as possible the conditions typical of an actual test day. Thus, all lights were turned on, all doors and windows were closed, and all blinds were shut. In addition, the heating, ventilation, and air conditioning (HVAC) system was set up as similarly as possible between the two rooms. The air conditioner (AC) was set to heating at 20 degrees Celsius (68 degrees Fahrenheit) in both rooms, one degree higher than the 19 degrees Celsius (66 degrees Fahrenheit) recommended by the university to promote energy efficiency during the winter season. The AC fan strength was set to the lowest setting (weak 'W' in Room 5, 1 out of 5 in Room 9), and the air flow flap was set to static since its swinging motion seemed to contribute a small amount of additional noise. Finally, any form of ventilation present in both rooms was turned off for the duration of the listening test.

2.3.5. Background Noise Level

With the room conditions established, the BNL was measured for 30 s at each seat position using the NTi Audio XL2 Audio and Acoustic Analyzer and the M4260 omnidirectional condenser microphone. Six measurements, one at each seat position, were made in each room. The mean (dBA) and standard deviation (SD) of the BNL measurements along with the HVAC settings mentioned above are presented in Table 4.

Table 4. Heating, ventilation, and air conditioning (HVAC) settings, background noise levels (BNL), test audio sound pressure levels (SPL), and speech-to-noise ratios (SNR) in Rooms 5 and 9.

Room	AC Fan Strength	Ventilation	BNL + HVAC Mean (dBA)	BNL + HVAC SD	Test Audio SPL (dBA)	Test Audio SPL SD	SNR (dB)	SNR SD
5	[W] S P	ON [OFF]	34.1	0.08	73.1	5.07	+39.0	5.09
9	[1] 2 3 4 5	ON [OFF]	36.4	0.96	72.6	2.06	+36.2	1.30

2.3.6. Compensation

All participants were shown appreciation for their cooperation with a monetary compensation equivalent to a two-hour minimum wage.

3. Results

For initial exploratory analysis, the test results were examined by generating a boxplot (see Figure 4) of the score difference values between Room 5 and 9 (Room 9 scores subtracted from Room 5 scores). The boxplot, with a median of 5 and the interquartile range from 1 (lower) to 11 (upper), indicated that the participants generally scored higher in Room 5. However, it also revealed an outlier with a value of −18, which signifies that one participant scored 18 points higher in Room 9, the acoustically unfavorable environment. For further examination, the raw score results (1–100 points; 1 point increments) were converted to their corresponding TOEIC listening scores (5–495; 5 point increments), which revealed that this participant's converted score (220) in Room 5 was 120 points lower than the baseline proficiency score (340), the largest score difference among all participants. Such a considerable difference in performance suggests that there may have been an external factor that caused this participant to greatly underperform in the acoustically favorable environment. Since (1) this was the sole outlier, (2) the data for this study consisted of small sample size, which may be greatly affected by the presence of an outlier, and (3) the raw score difference value (−18) was twice as large as the next largest value (−9), it was decided that the outlier be omitted from the dataset prior to conducting further statistical analysis. The removal of the outlier resulted in the change in total sample size from n = 43 to n = 42, CEFR-J High group from n = 18 to n = 17, and Front Seat group from n = 21 to n = 20.

A summary of the test results following removal of the outlier is presented in Table 5. As mentioned above, the raw scores range from 1 to 100 in 1-point increments, but in the official TOEIC tests the results are issued in 5-point increments between 5 and 495. The official score, however, is not simply generated by converting the raw scores in multiples of five, but rather is "calculated using a statistical processing method called 'equating' [40]." A conversion table is provided in the Official TOEIC Listening & Reading Workbook 9 [39] to approximate the score that one would receive on an official TOEIC test, but it merely serves as a rough estimate as it only indicates the range within which the official score may fall. For example, according to the conversion table, a score between 41 and 45 on the practice test listening section would equate to an official score between 160 and 230. Thus, as direct comparison of the converted scores to the baseline proficiency scores would be difficult to carry out reliably, the converted scores in Table 5 (generated by multiplying the raw scores by five) are presented purely for reference purposes. For this study, only the raw scores were considered for statistical analysis.

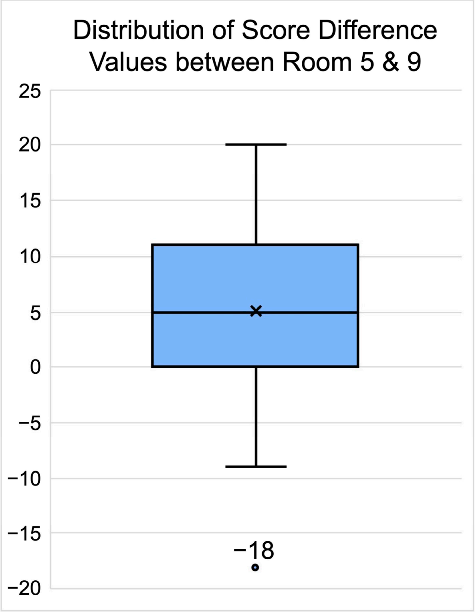


Figure 4. Boxplot of score difference values between Room 5 and 9.

Table 5. Raw and converted test results with corresponding baseline proficiency scores by group.

		Raw Scores				Converted TOEIC Scores				Baseline Proficiency Scores	
		Room 5		Room 9		Room 5		Room 9			
	n	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Entire Set	42	62.90	9.44	57.29	10.95	314.52	47.20	286.43	54.76	319.17	48.07
CEFR-J High	17	66.71	8.75	63.47	9.85	333.53	43.76	317.35	49.25	366.18	28.37
CEFR-J Low	25	60.32	9.16	53.08	9.73	301.60	45.82	265.40	48.67	287.20	27.95
Front Seat	20	63.90	7.48	58.10	10.69	319.50	37.41	290.50	53.46	318.00	52.88
Rear Seat	22	62.00	11.02	56.55	11.38	310.00	55.12	282.73	56.92	320.23	44.49

To test for statistical significance of differences between the mean score results from Room 5 and 9, paired-samples *t*-tests were performed for the (1) Entire Set, (2) CEFR-J High, (3) CEFR-J Low, (4) Front Seat, and (5) Rear Seat groups. Table 6 shows the results of the paired-samples *t*-tests and their corresponding normality tests. The scores for each group were normally distributed, as assessed by Shapiro–Wilk's test ($p > 0.05$). The results indicate that scores from Room 5 were found to be statistically significantly higher compared to those from Room 9 for all groups, all with medium to large effect sizes. In particular, the

Entire Set, CEFR-J Low, and Front Seat groups were significant at the p = 0.001 level, while the CEFR-J High group yielded a barely significant p value of 0.049.

Table 6. Paired-samples *t*-test results. Sample size (n), test statistic t-value (t), degrees of freedom (df), 95% confidence interval (CI), statistical significance (p), and Cohen's d effect size (d) are presented. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

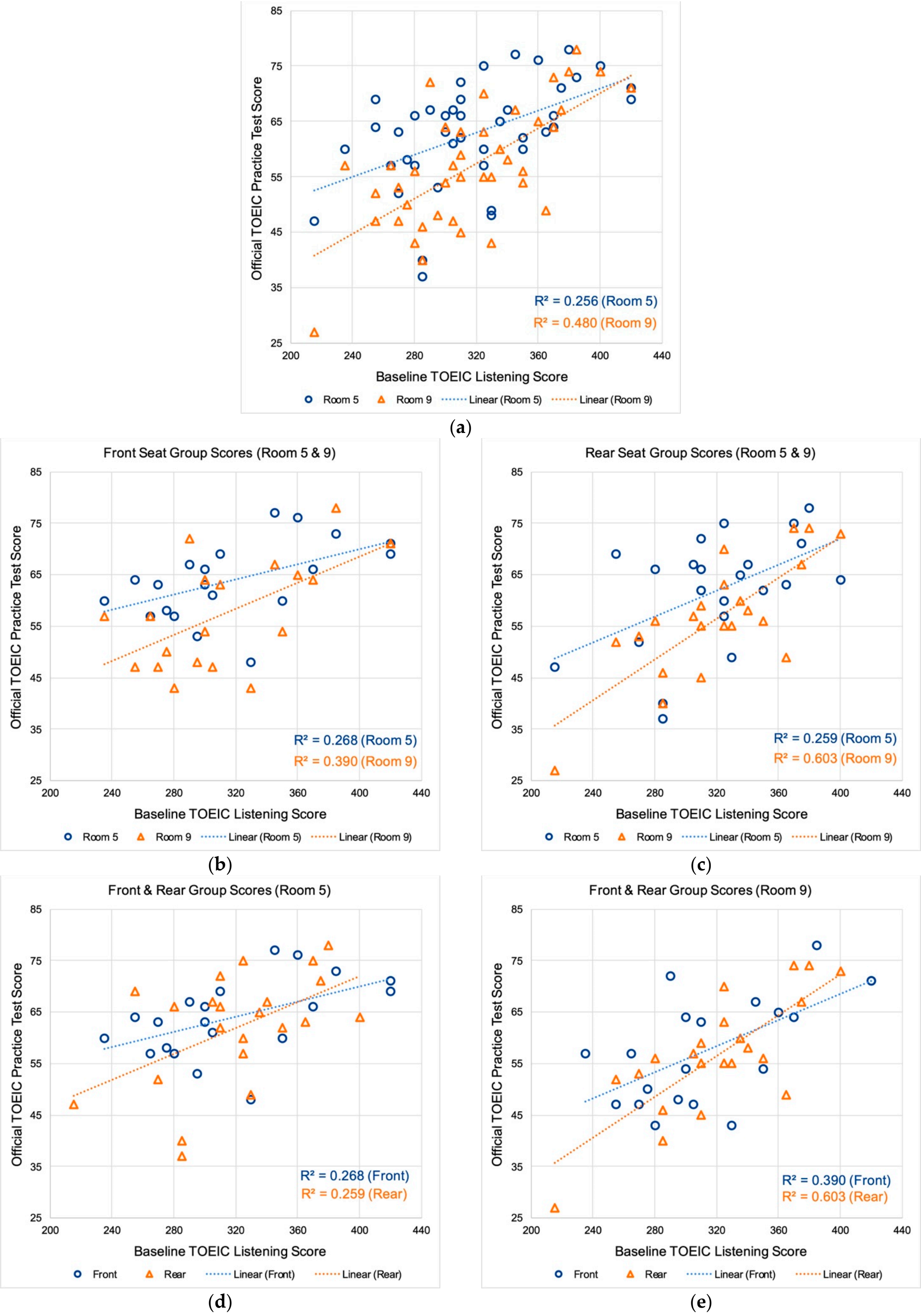
		Shapiro–Wilk Normality Test	**Paired Samples *t*-Test**				
	n	***p***	***t***	***df***	**95% CI [Lower, Upper]**	***p***	***d***
Entire Set	42	0.502	4.81	41	[3.258, 7.980]	<0.001 ***	0.74
CEFR-J High	17	0.367	2.13	16	[0.022, 6.449]	0.049 *	0.52
CEFR-J Low	25	0.354	4.48	24	[3.905, 10.575]	<0.001 ***	0.90
Front Seat	20	0.354	3.80	19	[2.609, 8.991]	0.001 ***	0.85
Rear Seat	22	0.584	3.06	21	[1.746, 9.163]	0.006 **	0.65

Additionally, independent-samples *t*-tests were performed to determine if there were differences in the mean listening scores between the Front and Rear Seat groups in each room (see Table 7). There were 20 and 22 participants in the Front and Rear seat groups, respectively, in both rooms. Listening scores for each seat group were normally distributed, as assessed by Shapiro–Wilk's test ($p > 0.05$), and there was homogeneity of variances, as assessed by Levene's test for equality of variances. In Room 5, the listening scores were higher for the Front Seat group (M = 63.90, SD = 7.48) than the Rear Seat group (M = 62.00, SD = 11.02); however, the difference was not statistically significant, M = 1.90, 95% CI [−4.04, 7.84], t(40) = 0.647, p = 0.521, d = 0.20. Likewise, in Room 9, the listening scores were higher for the Front Seat group (M = 58.10, SD = 10.69) than the Rear Seat group (M = 56.55, SD = 11.38), but the difference was not statistically significant, M = 1.56, 95% CI [−5.35, 8.46], t(40) = 0.455, p = 0.652, d = 0.14. It should be noted that inspection of the listening score boxplots revealed one outlier in the Rear Seat group in both rooms. The outliers were included in the results presented here since the independent samples *t*-test analysis yielded a non-significant result for datasets with and without the outlier.

Table 7. Independent samples *t*-test results.

			Shapiro–Wilk Normality Test	**Levene's Test**	**Independent Samples *t*-Test**				
		n	***p***	***p***	***t***	***df***	**95% CI [Lower, Upper]**	***p***	***d***
Room 5	Front	20	0.990	0.182	0.647	40	[−4.04, 7.84]	0.521	0.20
	Rear	22	0.111						
Room 9	Front	20	0.244	0.602	0.445	40	[−5.35, 8.46]	0.652	0.14
	Rear	22	0.292						

For visual analysis, scatter plots were generated for the following combinations: (1) Entire Set from Rooms 5 and 9, (2) Front Seat from Rooms 5 and 9, (3) Rear Seat from Rooms 5 and 9, (4) Front and Rear Seat from Room 5, and (5) Front and Rear Seat from Room 9. Since the CEFR-J High and Low groups are represented by the right and left halves of Figure 5a, respectively, individual scatter plots for these two groups were not generated.



(**a**)

(**b**) (**c**)

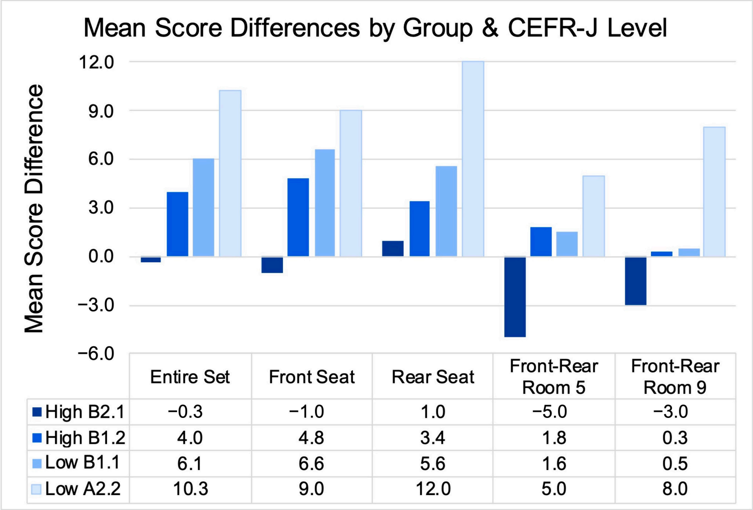
(**d**) (**e**)

Figure 5. Scatter plots of the test results plotted against baseline proficiency scores. (**a**) Entire Set group scores from Rooms 5 and 9, (**b**) Front Seat group scores from Rooms 5 and 9, (**c**) Rear Seat group scores from Rooms 5 and 9, (**d**) Front and Rear Seat group scores from Room 5, and (**e**) Front and Rear Seat group scores from Room 9.

It can be observed that similar patterns emerged in Figure 5a–c. As the paired-samples *t*-tests indicated, the scatter plots show that the participants, not only for the entire set but also independently within the CEFR-J High, CEFR-J Low, Front Seat, and Rear Seat groups, consistently scored higher in Room 5, the acoustically favorable environment, compared to Room 9, the acoustically unfavorable environment. In addition, the scatter plots reveal a linear relationship in which the score differences between Room 5 and 9 are at their greatest in the lowest baseline proficiency range and diminishes as the proficiency level increases. Ultimately, the best-fit lines can be seen converging among data points at the highest proficiency range, which for this study was at around the baseline TOEIC Listening scores of 400 to 420. The converging best-fit lines in the highest proficiency range aptly explains the paired-samples *t*-test result for the CEFR-J High group, which was barely statistically significant at $p = 0.049$, suggesting that there may be a threshold beyond which the acoustic differences between test rooms have no meaningful effect on the test takers' performance. More critically, however, the diverging best-fit lines seem to indicate that the acoustic quality of a given test room may considerably influence the L2 listening performance of lower-proficiency test takers.

The scatter plots presented in Figure 5d–e, on the other hand, are not as conclusive. Although it can be observed that Room 5 yielded higher scores for the low-baseline-proficiency participants, the best fit lines converge at a lower baseline proficiency range compared to those from Figure 5a–c. This suggests that the acoustic differences between seat locations (sound source to test taker distances) within a given room have less of an effect on L2 listening performance than those between test rooms with dissimilar acoustic characteristics, at least to the extent observed in the two rooms in this study.

Finally, given that (1) the paired-samples *t*-test results for the CEFR-J High and Low groups yielded contrasting degrees of statistical significance (*p*) values, and (2) the scatter plots seemed to indicate a greater difference in scores among the lower-proficiency participants, a bar chart of the mean score differences between Room 5 and 9 sorted by individual CEFR-J levels (B2.1, B1.2, B1.1, A2.2) in each group were generated to examine the extent to which the compared conditions influence L2 listening performance at each level (see Figure 6).



	Entire Set	Front Seat	Rear Seat	Front-Rear Room 5	Front-Rear Room 9
■ High B2.1	−0.3	−1.0	1.0	−5.0	−3.0
■ High B1.2	4.0	4.8	3.4	1.8	0.3
■ Low B1.1	6.1	6.6	5.6	1.6	0.5
■ Low A2.2	10.3	9.0	12.0	5.0	8.0

Figure 6. Mean score differences of the Entire Set, Front Seat, Rear Seat, Front-Rear Room 5, and Front-Rear Room 9 groups between Room 5 and 9 sorted by Common European Framework of Reference for Languages—Japan (CEFR-J) levels. A positive value indicates higher mean scores in Room 5 for Entire Set, Front Seat, and Rear Seat groups and higher mean scores in the Front seats in Room 5 and 9. Conversely, a negative value indicates that the participants overall performed better in Room 9 for Entire Set, Front Seat, and Rear Seat groups and likewise in the Rear seats in Room 5 and 9.

A clear trend of the mean score differences intensifying from higher to lower proficiency levels can be observed in the Entire Set, Front Seat, and Rear Seat groups. This pattern could denote a potential correlation between lower language proficiency and increased susceptibility to acoustic conditions during L2 listening tests. A similar overall trend can be identified for the Front–Rear Room 5 and Front–Rear Room 9 groups, although the score differences are less pronounced, which may imply that the impact of seat location within the same acoustic environment is less substantial than the effect of the overall acoustic quality of the room.

Interestingly, the data revealed that B2.1 was the sole level at which the mean score differences resulted in a negative number. This is especially evident in the Front–Rear comparisons where the negative values indicate that front seat scores were lower than those of the rear seats. This could reflect a complex interaction between proficiency level and acoustic perception that may be less affected by environmental factors at higher proficiency levels. On the other hand, the participants at the A2.2 level across all groups exhibited the greatest mean score differences between the compared conditions by a considerable margin. This observation held true even when comparing Front and Rear Seat groups within the same acoustic settings, which is particularly striking given that the independent samples *t*-tests did not reveal significant differences. This is compounded by the fact that Room 9 yielded a larger mean score difference than Room 5, suggesting that test takers at lower proficiency levels are more vulnerable to suboptimal acoustic environments.

Lastly, the results for participants at the B1.1 and B1.2 levels consistently demonstrate a preference for the acoustically favorable environment of Room 5. However, these groups exhibited only a slight tendency towards improved performance in the front seats of each room. This could indicate that while the overall room acoustics play a critical role in L2 listening comprehension, the relative advantage of front seating is not as strongly felt within the confines of a well-designed acoustic space.

4. Discussion

4.1. Reverberation Time

The current study's findings corroborate the established notion in the literature that adverse acoustic conditions, particularly high reverberation time, pose a significant challenge to L2 listening comprehension. Examining the effect of three simulated $RT_{125\text{–}8kHz}$ (T_{30}) conditions of 0.26 s, 0.92 s, and 1.77 s with a corresponding STI of 0.87, 062, and 0.49, respectively, on native adult Swedish participants' performance on an English listening task, Sörqvist et al. (2014) [14] observed that the performance scores decreased as RT increased. Similarly, investigating a combination of two $RT_{125\text{–}4kHz}$ (T_{30}) conditions of 0.33 s and 1.07 s and two source–receiver distances of 1.05 m and 6.13 m, for which the STI were 0.95, 0.84, 0.71, and 0.62, Hurtig et al. (2016) [31] found that both the higher RT and the farther distance negatively affected the participants' performance on a L2 listening comprehension test.

The influence of noise on comprehension has also been substantiated in the literature. Kilman et al. (2014) [20] reported how speech perception performance degraded in the presence of four types of background noise: stationary noise, fluctuating noise, two-talker babble in Swedish (participants' L1), and two-talker babble in English (participants' L2), particularly for non-native speakers, highlighting that non-native listeners' ability to understand speech in noisy environments is significantly more compromised than that of native listeners. Tabri et al. (2011) [19] took a novel approach by comparing the speech perception among monolingual, bilingual, and trilingual listeners in noise levels of 50, 55, 60, 65, and 70 dB SPL. They found that, while all groups performed similarly at the lowest noise signal, speech perception scores of bilinguals and trilinguals degraded more rapidly than those of monolinguals at higher noise levels. Scharenborg and van Os (2019) [21] comprehensively elucidated the struggle non-native listeners encounter in noisy environments, suggesting that listening in a non-native language is inherently more demanding, a factor exacerbated by poor acoustic conditions. Lastly, Visentin et al. (2019) [32] and Peng and Wang (2019) [30] both observed that non-native listeners must exert more effort to achieve comprehension amidst noise, which becomes especially pronounced when dealing with reverberation and talker foreign accent.

The insights from the studies above highlight the intricate ways in which acoustic conditions interact with language comprehension abilities, and collectively underscore the vulnerability of L2 listeners in acoustically challenging environments, a concern that is pertinent in standardized testing scenarios. The current study extended this narrative by quantitatively demonstrating the degradation of listening performance in environments with poor acoustic characteristics, offering empirical evidence for advocating for acoustically optimized testing conditions.

4.2. Speech-to-Noise Ratio

Another key component often included in the discussion of speech intelligibility in classroom-like spaces is the speech-to-noise ratio (SNR). In educational settings, the consensus has been that an SNR of +15 dB is ideal for learning, particularly for children. Bistafa and Bradley (2000) [22] and Crandell and Smaldino (2000) [23] were among the first to highlight the importance of this SNR threshold in classrooms. However, Bradley and Sato (2008) [25] found that even a +15 dB SNR may be insufficient for the youngest learners. The implications of these findings suggest that the heightened sensitivity of young learners must be accounted for when considering acoustic design and noise management in educational settings. For adults, the tolerance for lower SNRs seems to be more robust. Choi's (2020) [27] study with Korean university students demonstrated that an SNR of merely +3 dB can facilitate a high level of speech intelligibility. This is in stark contrast to the requirements for younger learners and suggests that adult learners have coping mechanisms or linguistic competencies that may mitigate the need for higher SNRs. Such resilience is only reserved for native adult listeners though, as a completely different picture emerges for non-native listeners. The study by van Wijngaarden et al. (2002) [24] concluded that non-native listeners require more favorable SNR conditions, anywhere from +1 to +7 dB, to match the intelligibility scores of their native counterparts. Likewise, Warzybok et al. (2015) [26] found that participants, whose non-native language in this study was German, required +3 dB and +6 dB for closed-set and open-set speech tests, respectively, to achieve 50% speech recognition comparable to that of native listeners.

The studies above present critical considerations and implications aimed at providing optimal listening conditions in educational spaces. In the context of the present study, the SNRs in Room 5 and 9 far exceeded the thresholds identified for both children and adult learners, including non-native listeners. This indicates that other factors, potentially related to the specific acoustic profiles of the rooms, may have had a more significant influence on the L2 listening comprehension of the test subjects.

4.3. Standardized Assessment Methods

One of the main findings from the present study was the trend of increasing mean score differences from high to low proficiency levels observable across all comparison groups, but particularly in the Entire Set, Front Seat, and Rear Seat groups (see Figure 6). The proficiency levels were determined by applying the CEFR-J categorization to the subjects' mean listening scores derived from their official TOEIC IP test results. Although the subdivided categorizations (i.e., B1 into B1.1 and B1.2) are unique to the CEFR-J, it can be argued that the four levels represented in this study (A2.2, B1.1, B1.2, B2.1) can also be expressed in the original CEFR as high-A2, B1, and low-B2.

A few related studies have taken a similar approach of allocating their subjects into groups by their baseline proficiency levels. Both Sörqvist et al. (2014) [14] and Hurtig et al. (2016) [31] utilized the listening section of the National Tests of English for senior Swedish high school students as the experiment test material and the reading section from the same test to determine the baseline proficiency of their participants. No grouping was applied to the baseline proficiency scores in [14], but [31] assigned their subjects into High and Low groups based on their reading scores. In both studies, the higher-proficiency individuals or groups performed better than their lower-proficiency counterparts in all conditions. According to the authors, the National Tests of English are a set of standardized tests administered to high school seniors in Sweden, implying that they may be comparable to the CEFR in their design and intent; however, the data were not explicitly aligned with the CEFR in their studies.

Similarly, MacCutcheon et al. (2019) [16] used a picture naming test to gauge their subjects' vocabulary knowledge and categorized them into High Vocab and Low Vocab groups. The groups listened to monaural audio files in virtual spaces representing the actual measurements of two classrooms, one with $RT_{0.5–1kHz}$ 0.3 s and the other with $RT_{0.5–1kHz}$ 0.9 s (both T_{30}), at a source–receiver distance of 6 m. They found that while the Low Vocab group showed a mere $<$ 1% increase in their listening comprehension scores between the two RT conditions, the High Vocab group boasted a 14% increase between the short and long RTs. The authors discussed that their research design using vocabulary knowledge to gauge the subjects' proficiency allowed them to consider that perhaps top-down language processing such as reliance on contextual cues may have benefited the High Vocab group, and that this takes effect beyond "a certain level of L2 vocabulary" [16] (p. 182). This finding seems to contradict the results of the present study in which the greatest increase in scores between Room 5 and 9 belonged not to the highest-proficiency group (B2.1) but rather to the lowest-proficiency group (A2.2). However, without reference to a benchmark like the CEFR, it is difficult to pinpoint the exact proficiency levels at which the findings in [16] were observed, rendering direct comparison of the two studies impractical.

Other examples of various assessment methods include the Swedish and American English versions of the Hearing in Noise Test (HINT) [20], the Diagnostic Rhyme Test in Italian [32], and a dual-task approach to measure listening effort using the NASA Task Load Index [30], as well as interviews and questionnaires [19]. As illustrated above, these varied methodologies, although they are valuable instruments in their own right designed for specific purposes, highlight the need for a standardized approach. The CEFR-J utilized in the present study serves as a framework through which the results can be interpreted, ensuring comparability of results and better understanding of the nuanced effects of acoustic conditions on L2 listening comprehension across different proficiency levels.

4.4. Sound Source to Receiver Distance

The current study's examination of the source–receiver distance offers insights into the spatial dynamics of listening comprehension. Previous research has identified significant differences in speech intelligibility performance at varying distances. Hurtig et al. (2016) [31] set the source–receiver distance at 1.05 m and 6.13 m, and Visentin et al. (2019) [32] examined distances of 2.5 m and 5.5 m. Both studies found differences in L2 listening performance based on these proximities, underscoring the essential role of spatial acoustics particularly for non-native listeners. More critically, it is posited that such variability of speech intelligibility within a given educational setting where foreign language instruction and learning takes place may even lead to disparities in the learners' grades [31]. This is a serious claim that is highly pertinent to investigations that concern standardized test performance outcomes. In the current study, however, with source–receiver distances of 3.25 m (SD = 1.73) and 10.15 m (SD = 1.18) in Room 5 and 3.34 m (SD = 0.86) and 12.16 m (SD = 0.78) in Room 9, no significant differences were observed. This suggests that the relationship between distance and listening performance may be more complex than previously understood and that the overall acoustic quality of the room might play a more decisive role than the distance from the sound source.

4.5. Listening Effort

The concept of listening effort is another factor in L2 listening comprehension worth briefly exploring here. Lam et al. (2018) [29], using response times as metric of listening effort, examined L1 and L2 listeners' accuracy and response times across three conditions:

anechoic with no noise, with reverberation only, and with reverberant noise. They observed longer response times for L2 listeners in all three conditions. Interestingly, however, the accuracy scores were found to be comparable between L1 and L2 listeners across all conditions. The authors postulated that, even when L2 accuracy matches that of L1 listeners, the cognitive load is significantly higher for L2 individuals in the presence of reverberation and noise. Their findings suggest that improved acoustic conditions can enhance not only the accuracy of test takers but also the efficiency with which they respond to each question. Moreover, the L2 listeners in [29] are defined as those exposed to the English language after the age of 3, which would be considered near-native level in a Japanese setting, as English learners in Japan typically begin learning English much later at around ages 10–12. This implies that Japanese L2 learners' response times would be much longer than those of L2 listeners in [29]. Given the timed nature of standardized tests, these considerations are critical as each correctly answered question within the time limit contributes to the final score.

4.6. University Facilities

The selection of university facilities for administering standardized English-language proficiency tests is an important logistical consideration, as it is common practice to utilize university classrooms and lecture halls as venues for standardized tests. In fact, among the list of possible test venues listed by prefecture in the official TOEIC Website in Japan, the overwhelming majority are universities [41]. This is likely due to the fact that universities have the capacity to rent out multiple classrooms and lecture halls to outside organizations on days when classes are not in session such as weekends and holidays. According to the latest report by TOEIC [42], in Japan alone, more than two million tests are administered each year at these venues. Given the unique combination of specific areas of research, this issue has received limited attention in the literature, which poses another challenge in drawing direct comparisons with this study. However, studies such as those by Nestoras and Dance (2013) [43], Escobar and Morillas (2015) [12], Choi (2020) [27], Prodeus and Didkovska (2021) [44], Minelli et al. (2022) [45] and Mealings (2023) [46], although they do not explicitly focus on standardized testing conditions, provide insights into the acoustic conditions of university classrooms and lecture halls. These studies collectively suggest that the acoustics of these venues can significantly influence listening comprehension outcomes, thus emphasizing the need for careful acoustic consideration in the selection of test venues.

4.7. Recommendations for Standardized Tests

The general recommendations presented in the authors' previous research [35] for STI ($\geq$0.66) and $\text{RT}_{0.5\text{–}2\text{kHz}}$ ($<$0.7 s) remain relevant. However, the current study's findings suggest that room sizes, which were proposed in [35] in conjunction with RT as small (less than 350 m^3 with $\text{RT}_{0.5\text{–}2\text{kHz}} < 0.6$ s), medium (350 to 700 m^3 with $\text{RT}_{0.5\text{–}2\text{kHz}} < 0.7$ s), and large (700 to 1050 m^3 with $\text{RT}_{0.5\text{–}2\text{kHz}} < 0.8$ s), may be a secondary concern to the overall acoustic quality of the room. It is noteworthy that Room 5, despite being twice the volume of Room 9, exhibited better acoustic conditions, which can most likely be attributed to the presence of perforated, fluted wooden acoustic panels installed across the entire front and rear walls. Such use of acoustic treatment or any other consideration promoting a favorable acoustic environment was not apparent in Room 9. Consequently, the non-significant differences between the Front and Rear Seat groups imply that as long as acoustic conditions are within the recommended STI and RT ranges, test administrators may be able to take advantage of larger rooms without compromising listening comprehension. To this end, utilizing smartphone apps [47] for a preliminary assessment of room acoustics may offer a practical and economical alternative to more sophisticated equipment. Indeed, with careful selection and simple assessment tools, environments conducive to optimal L2 listening comprehension can certainly be identified and utilized for testing.

4.8. Limitations and Future Studies

Several limitations must be noted. First, only 43 subjects participated. The study should be expanded with more participants, preferably in each CEFR-J level, in future research. Second, the listening tests were conducted in empty classrooms except for the participants sitting in their designated seat positions. A fully occupied test room, usually around 50% to 70% occupancy, which is a common procedure for standardized tests, may affect the listening experience of test takers. Third, the experiment was conducted in only two classrooms with particularly contrasting acoustic conditions. Further investigations including other acoustic environments is needed before any generalizations regarding the potential positive or negative effect of room acoustics on the performance of standardized EFL proficiency listening tests can be made. Fourth, only wall-mounted speakers were employed as sound sources for this study. Since the investigation in [35] revealed notable differences in STI (especially those above and below 0.66 STI) between different sound sources even within the same test room in some cases, score results from tests conducted using other sound sources should also be compared. Fifth, the relatively small sample sizes of CEFR-J High B2.2 (n = 3) and A2.2 (n = 7) groups compared to those of B1.2 (n = 14) and B1.1 (n = 18) groups may have disproportionately exaggerated the results illustrated in Figure 6. Sixth, L2 listening performance may also depend on the complexity of the test material. Identical to the TOEIC SP and IP tests, the practice tests included in Official TOEIC Listening & Reading Workbook 9 is composed of four sections of varying format and difficulty. However, examining the results at this level of detail was beyond the scope of this study. Finally, there are various other factors that could have had an effect on the participants' listening performance. Perhaps most notably, due to logistical reasons with regard to classroom availability, the listening test was conducted in Room 9 first, then in Room 5 approximately one week later. Some participants commented that they felt more comfortable and prepared in Room 5 because they had already taken the test once in Room 9. This may be attributable to the fact that this was the first time that the participants had taken the full 45 min listening test. Although they had already taken the TOEIC L&R IP Test Online twice within a span of a year at the time of this study, the online test is a consolidated version which only takes one hour to complete rather than the two hours required for the standard SP and IP tests. In the future, the order of the test rooms should be randomized to see if the statistically significant results still hold true.

5. Conclusions

This study examined the relationship between classroom acoustics and L2 listening comprehension in the context of standardized EFL testing environments. It found statistically significant differences in test outcomes between two rooms previously used for the TOEIC L&R SP and IP tests. The analysis also revealed that participants with lower proficiency levels, specifically at the CEFR-J A2.2 level, experienced more acoustical challenges, while those at the B2.1 level were least affected. This finding demonstrates the practicality of considering the CEFR(-J) framework in future studies, especially when assessing the impact of acoustic variables on non-native populations. The empirical evidence presented in this study bears important implications for the administration of standardized tests and invites test organizers to consider the acoustics of testing environments and reassess their criteria for test room selection. The research affirms that acoustic quality in test rooms is a critical factor in language assessment and calls for heightened acoustical awareness in the standardized testing process to promote equitable language evaluation opportunities.

Author Contributions: Conceptualization, M.K.; methodology, M.K., M.T.-H. and H.H.; software, M.K.; validation, M.K.; formal analysis, M.K.; investigation, M.K.; resources, M.T.-H. and H.H.; data curation, M.K.; writing—original draft preparation, M.K.; writing—review and editing, M.K.; visualization, M.K.; supervision, M.T.-H. and H.H.; project administration, M.T.-H. and H.H.; funding acquisition, M.T.-H. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Utsunomiya University (Project ID: H22-0102).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available in the article.

Acknowledgments: The authors would like to thank the subjects for their participation in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. IELTS. *IELTS Guide for Teachers 2019*; IELTS: Manchester, UK, 2019.
2. Educational Testing Service. *TOEFL iBT Test and Score Data Summary 2022*; ETS: Princeton, NJ, USA, 2022.
3. Educational Testing Service. *TOEIC Listening & Reading Test Examinee Handbook 2022*; ETS: Princeton, NJ, USA, 2022.
4. Eberhard, D.M.; Simons, G.F.; Fennig, C.D. (Eds.) *Ethnologue: Languages of the World*, 26th ed.; SIL International: Dallas, TX, USA, 2023; Available online: http://www.ethnologue.com (accessed on 1 November 2023).
5. Should I Take IELTS on Computer or Paper? Available online: https://ieltsjp.com/japan/about/article-computer-delivered-paper-based-ielts-comparison/en-gb (accessed on 1 November 2023).
6. The TOEFL Assessment Series. Available online: https://www.ets.org/toefl/itp.html (accessed on 1 November 2023).
7. TOEIC Listening & Reading Test. Available online: https://www.iibc-global.org/english/toeic/corpo/toeic.html (accessed on 1 November 2023).
8. IELTS Online. Available online: https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-online (accessed on 1 November 2023).
9. At Home Testing for the TOEFL iBT Test. Available online: https://www.ets.org/toefl/test-takers/ibt/test-day/at-home-test-day.html (accessed on 1 November 2023).
10. TOEIC Program IP Test Online. Available online: https://www.iibc-global.org/toeic/corpo/guide/online_program.html (accessed on 1 November 2023).
11. van Wijngaarden, S.J.; Bronkhorst, A.W.; Houtgast, T.; Steeneken, H.J.M. Using the Speech Transmission Index for predicting non-native speech intelligibility. *J. Acoust. Soc. Am.* **2004**, *115*, 1281–1291. [CrossRef]
12. Escobar, V.G.; Morillas, J.M.B. Analysis of intelligibility and reverberation time recommendations in educational rooms. *Appl. Acoust.* **2015**, *96*, 1–10. [CrossRef]
13. Yang, D.; Mak, C.M. An investigation of speech intelligibility for second language students in classrooms. *Appl. Acoust.* **2018**, *134*, 54–59. [CrossRef]
14. Sörqvist, P.; Hurtig, A.; Ljung, R.; Rönnberg, J. High second-language proficiency protects against the effects of reverberation on listening comprehension. *Scand. J. Psychol.* **2014**, *55*, 91–96. [CrossRef]
15. Mealings, K. Classroom acoustic conditions: Understanding what is suitable through a review of national and international standards, recommendations, and live classroom measurements. In Proceedings of the Acoustics 2016: The Second Australasian Acoustical Societies Conference, Brisbane, Australia, 9–11 November 2016.
16. MacCutcheon, D.; Hurtig, A.; Pausch, F.; Hygge, S.; Fels, J.; Ljung, R. Second language vocabulary level is related to benefits for second language listening comprehension under lower reverberation time conditions. *J. Cogn. Psychol.* **2019**, *31*, 175–185. [CrossRef]
17. Puglisi, G.E.; Warzybok, A.; Astolfi, A.; Kollmeier, B. Effect of reverberation and noise type on speech intelligibility in real complex acoustic scenarios. *Build. Environ.* **2021**, *204*, 108137. [CrossRef]
18. Mayo, L.H.; Florentine, M.; Buus, S. Age of second-language acquisition and perception of speech in noise. *J. Speech Lang. Hear. Res.* **1997**, *40*, 686–693. [CrossRef]
19. Tabri, D.; Chacra, K.M.S.A.; Pring, T. Speech perception in noise by monolingual, bilingual and trilingual listeners. *Int. J. Lang. Commun. Disord.* **2011**, *46*, 411–422. [CrossRef]
20. Kilman, L.; Zekveld, A.; Hällgren, M.; Rönnberg, J. The influence of non-native language proficiency on speech perception performance. *Front. Psychol.* **2014**, *5*, 651. [CrossRef] [PubMed]
21. Scharenborg, O.; van Os, M. Why listening in background noise is harder in a non-native language than in a native language: A review. *Speech Commun.* **2019**, *108*, 53–64. [CrossRef]
22. Bistafa, S.R.; Bradley, J.S. Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics. *J. Acoust. Soc. Am.* **2000**, *107*, 861–875. [CrossRef]
23. Crandell, C.C.; Smaldino, J.J. Classroom acoustics for children with normal hearing and with hearing impairment. *Lang. Speech Hear Serv. Sch.* **2000**, *31*, 362–370. [CrossRef] [PubMed]
24. van Wijngaarden, S.J.; Steeneken, H.J.M.; Houtgast, T. Quantifying the intelligibility of speech in noise for non-native listeners. *J. Acoust. Soc. Am.* **2002**, *111*, 1906–1916. [CrossRef] [PubMed]

25. Bradley, J.S.; Sato, H. The intelligibility of speech in elementary school classrooms. *J. Acoust. Soc. Am.* **2008**, *123*, 2078–2086. [CrossRef] [PubMed]
26. Warzybok, A.; Brand, T.; Wagener, K.C.; Kollmeier, B. How much does language proficiency by non-native listeners influence speech audiometric tests in noise? *Int. J. Audiol.* **2015**, *54*, 88–99. [CrossRef] [PubMed]
27. Choi, J.Y. The intelligibility of speech in university classrooms during lectures. *Appl. Acoust.* **2020**, *162*, 107211. [CrossRef]
28. Picou, E.M.; Gordon, J.; Ricketts, T.A. The effects of noise and reverberation on listening effort for adults with normal hearing. *Ear Hear.* **2016**, *37*, 1–13. [CrossRef]
29. Lam, A.; Hodgson, M.; Prodi, N.; Visentin, C. Effects of classroom acoustics on speech intelligibility and response time: A comparison between native and non-native listeners. *Build. Acoust.* **2018**, *25*, 35–42. [CrossRef]
30. Peng, Z.E.; Wang, L.M. Listening effort by native and nonnative listeners due to noise, reverberation, and talker foreign accent during English speech perception. *J. Speech Lang. Hear. Res.* **2019**, *62*, 1068–1081. [CrossRef]
31. Hurtig, A.; Sörqvist, P.; Ljung, R.; Hygge, S.; Rönnberg, J. Student's second-language grade may depend on classroom listening position. *PLoS ONE* **2016**, *11*, e0156533. [CrossRef]
32. Visentin, C.; Prodi, N.; Cappelletti, F.; Torresin, S.; Gasparella, A. Speech intelligibility and listening effort in university classrooms for native and non-native Italian listeners. *Build. Acoust.* **2019**, *26*, 275–291. [CrossRef]
33. Srinivasan, N.K. The Perception of Natural, Cell Phone, and Computer-Synthesized Speech during the Performance of Simultaneous Visual–Motor Tasks. Ph.D. Dissertation, University of Nebraska, Lincoln, NE, USA, 2010.
34. Kawata, M.; Sato, K.; Tsuruta-Hamamura, M.; Hasegawa, H. Analysis of standardized foreign language listening test scores and their relationship to speech transmission index and reverberation time in test rooms. In Proceedings of the 27th International Congress on Sound and Vibration, ICSV27, Virtual, 11–16 July 2021.
35. Kawata, M.; Tsuruta-Hamamura, M.; Hasegawa, H. Assessment of speech transmission index and reverberation time in standardized English as a foreign language test rooms. *Appl. Acoust.* **2023**, *202*, 109093. [CrossRef]
36. Steeneken, H.J.M.; Houtgast, T. A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.* **1980**, *67*, 318–326. [CrossRef]
37. Council of Europe. Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Companion volume, Council of Europe Publishing, Strasbourg. 2020. Available online: www.coe.int/lang-cefr (accessed on 1 November 2023).
38. CEFR-J, European Language Portfolio CAN-DO Descriptors. 2012. Available online: http://www.cefr-j.org/index.html (accessed on 1 November 2023).
39. Educational Testing Service. *Official TOEIC Listening & Reading Workbook 9*, 1st ed.; The Institute for International Business Communication: Tokyo, Japan, 2023.
40. TOEIC Test Results. Available online: https://www.iibc-global.org/english/toeic/test/lr/guide05.html (accessed on 1 November 2023).
41. TOEIC SP Schedule & Test Locations by Prefecture. Available online: https://www.iibc-global.org/toeic/test/lr/guide01/schedule/area.html (accessed on 1 November 2023).
42. Educational Testing Service. *TOEIC Program Data Analysis 2022*; ETS: Princeton, NJ, USA, 2022.
43. Nestoras, C.; Dance, S. The interrelationship between room acoustics parameters as measured in university classrooms using four source configurations. *Build. Acoust.* **2013**, *20*, 43–54. [CrossRef]
44. Prodeus, A.; Didkovska, M. Assessment of speech intelligibility in university lecture rooms of different sizes using objective and subjective methods. *East-Eur. J. Enterp. Technol.* **2021**, *3*, 47–56. [CrossRef]
45. Minelli, G.; Puglisi, G.E.; Astolfi, A. Acoustical parameters for learning in classroom: A review. *Build. Environ.* **2022**, *208*, 108582. [CrossRef]
46. Mealings, K. A scoping review of the effect of classroom acoustic conditions on university students' listening, learning, and well-being. *J. Speech Lang. Hear. Res.* **2023**, *66*, 4653–4672. [CrossRef]
47. Mealings, K. Validation of the SoundOut room acoustics analyzer app for classrooms: A new method for self-assessment of noise levels and reverberation time in schools. *Acoust. Aust.* **2019**, *47*, 277–283. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.