Applied Acoustics 202 (2022) 109093

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Assessment of speech transmission index and reverberation time in standardized English as a foreign language test rooms ☆

Makito Kawata [a,*], Mariko Tsuruta-Hamamura [b], Hiroshi Hasegawa [b]

[a] Liberal and General Education Center, Utsunomiya University, Utsunomiya, Tochigi 321-8505, Japan
[b] Graduate School of Engineering, Utsunomiya University, Utsunomiya, Tochigi 321-8585, Japan

ARTICLE INFO

Article history:
Received 30 December 2021
Received in revised form 8 September 2022
Accepted 24 October 2022
Available online 6 December 2022

Keywords:
Building acoustics
Speech transmission index
Reverberation time
Foreign language
Non-native listeners
Standardized tests

ABSTRACT

Today, millions of standardized English as a foreign language proficiency tests are administered globally each year. A large portion of this is conducted as a paper-based test in which the listening section is commonly delivered through loudspeakers to groups of test takers, a method in which the audio signals are exposed to the acoustic tendencies of each particular venue. As it is well-established in the literature that non-native listeners are more susceptible to adverse listening conditions compared to their native counterparts, there is a need for an objective examination of the acoustic quality of such environments. This study examined the speech transmission index for public address systems (STIPA) for three types of sound sources (wall-mounted speakers, radio cassette player, and amplified speaker) and reverberation time (RT) in 10 unoccupied classrooms commonly used as test rooms at a university in Japan. The results revealed that STI was found to be statistically significantly different for the amplified speaker compared to both or one other sound source in eight out of 10 rooms. The amplified speaker also recorded the highest STI among the three sound sources in eight out of 10 rooms and the most rooms with STI entirely above 0.66, a minimum target value prescribed in IEC 60268–16:2020 as exhibiting high speech intelligibility. Additionally, ≥ 0.66 STI was consistently observed in rooms with $RT_{0.5\text{-}2kHz} \leq 0.7$ s. Further observations are discussed to better understand the current conditions under which these tests are administered

1. Introduction

As the development of globalization continues to establish English as the dominant international language, the number of those who speak English as a second language (L2) has seen incredible growth in recent decades. Today, with an estimated 373 million native language (L1) speakers and over 1 billion L2 speakers, English is the most widely spoken language in the world [1]. This has generated considerable demand for standardized English as a foreign language (EFL) proficiency tests as increasing number of businesses and educational institutions require proof of English proficiency from their applicants.

1.1. Standardized EFL proficiency tests

In particular, IELTS, TOEFL®, and TOEIC® are among the most widely administered and are each recognized by more than 10,000 organizations in over 140 countries [2–4]. Over 3.5 million tests were administered by IELTS in 2018 [2], and, while global figures for TOEFL® and TOEIC® are difficult to pinpoint, published data in Japan reveal that over 2.2 million TOEIC® Listening & Reading Tests (TOEIC L&R) have been administered each year since 2011 [5]. Even amidst the COVID-19 pandemic, TOEIC L&R was reportedly administered to 1.5 million test takers in Japan alone during the 2020 fiscal year [6].

Since English proficiency is commonly evaluated on the four main areas of language (listening, speaking, reading, and writing), a significant portion of these tests cover listening skills and are generally conducted in one of two formats: computer-based and paper-based.

Computer-based tests are administered on computers provided at designated test centers. These tests are individually administered to each test taker and allow the use of headphones during the listening section. This promotes a less distracting listening

☆ Portions of this work appear in a paper entitled "Analysis of standardized foreign language listening test scores and their relationship to speech transmission index and reverberation time in test rooms" presented at the 27th International Congress on Sound and Vibration (ICSV 27), 11–16 July 2021.

* Corresponding author.
E-mail addresses: kawata@cc.utsunomiya-u.ac.jp (M. Kawata), mariko@is.utsunomiya-u.ac.jp (M. Tsuruta-Hamamura), hasegawa@is.utsunomiya-u.ac.jp (H. Hasegawa).

https://doi.org/10.1016/j.apacoust.2022.109093
0003-682X/

environment and grants test takers the ability to adjust the volume according to their preference. IELTS and TOEFL iBT® are examples of computer-based tests. The TOEIC® Speaking & Writing is also a computer-based test but, as the name suggests, does not contain a listening component aside from short instructional cues.

Paper-based tests, on the other hand, are administered simultaneously to groups of test takers in large venues, often across multiple locations. The listening section is generally conducted by playing the audio tracks through loudspeakers equipped at the venue or through portable loudspeakers set up by the examiners. IELTS offers a paper-based version that utilizes this method, but not exclusively as some test centers provide headphones for the listening section. The TOEIC L&R is another example of a paper-based test that typically employs this method. There are two forms of administration for the TOEIC L&R: SP and IP. SP stands for Secure Program and is administered by organizations responsible for the operation of the TOEIC® Program in their designated countries. In Japan, the Institute for International Business Communication (IIBC) takes on this role. IP stands for Institutional Program and is administered by non-IIBC organizations such as companies and universities. Both versions are developed by the Educational Testing Service based in the United States.

Recently, a third online option has emerged in response to the COVID-19 pandemic: IELTS Indicator, TOEFL iBT® Home Edition, and TOEIC® Program IP Test Online. While these tests serve as a welcome solution for millions of students and job applicants by allowing them to participate from the safety of their homes, they will be excluded from further discussion in this study due to their novelty and potentially temporary nature.

1.2. Effect of acoustics on paper-based tests

The paper-based format is advantageous over the use of computers for its ability to be administered to hundreds, even thousands, of test takers simultaneously as well as synchronously across multiple venues. However, conducting the listening test through loudspeakers can be prone to various factors such as the type and condition of the audio source, the acoustic quality of the venue, noise produced by HVAC (heating, ventilation, and air conditioning) systems, and even the seat position of test takers [7]. Some variation in acoustic quality is to be expected as different venues have their own unique physical and acoustic features, but this can become more pronounced in larger venues that are often chosen for their logistical convenience.

This aspect of paper-based tests cannot be overlooked since numerous studies have found that adverse listening conditions can affect speech comprehension and lead to decreased academic performance. Children tend to be more easily affected by unfavorable classroom environments compared to adults (university age or older), as younger learners are still in their developmental stages and require access to clear, unobstructed auditory signals for proper recognition and comprehension of speech [8–17]. This, however, mostly describes listening tasks performed in the listener's L1. When listening in their L2, the performance of both children and adults is often found to suffer under adverse acoustic conditions compared to those listening in their L1 in similar circumstances [18–28]. Given the prominent role of standardized EFL proficiency tests in today's increasingly globalized society, there is a need for an objective examination of the acoustic quality of test rooms as proper administration of such tests is imperative both for the test takers and for ensuring its integrity as a reliable tool for standardized assessment.

1.3. Speech transmission index

A common method for objective prediction and measurement of speech intelligibility is the speech transmission index (STI). It is expressed with a numerical value between 0 and 1 where 0 signifies poor STI and 1 signifies excellent STI [29]. IEC 60268–16:2020 [30] outlines a categorization of STI values in qualification bands from $>$ 0.76 (category A+, recording studios) to $<$ 0.36 (category U, not suitable for PA systems). For example, according to IEC 60268–16:2020, a STI value of 0.66 (category C, speech auditoria and teleconferencing) is considered as exhibiting high speech intelligibility, 0.62 STI (category D, lecture theaters and classrooms) as good speech intelligibility, and 0.58 STI (category E, concert halls and modern churches) as high quality PA systems. The values stated here are the nominal STI values of each designated qualification band.

IEC 60268–16:2020 additionally outlines adjusted STI values for three categories of non-native listeners: advanced, intermediate, and beginner. In these scenarios, to achieve a standard STI equivalent of 0.60 rated as being "fair - good" intelligibility, a transmission system must generate a STI of 0.68 for advanced and 0.86 for intermediate listeners. An adjusted value for beginners is not given as it would require $>$ 1 STI. Similarly, van Wijngaarden et al. [31] proposed a method to predict the intelligibility of non-native listeners by applying a correction function to data from previous studies that explored the performance on intelligibility tasks by native and non-native subjects. They found that achieving a standard STI equivalent of 0.60 required an adjusted STI of 0.79 for non-native subjects who had studied at a university in the United States for at least four years [32]. They also found that 0.68 for early bilinguals (higher proficiency) and $>$ 1 for late bilinguals (lower proficiency) were required for those whose L1 is Mexican-Spanish [33]. Lastly, for Dutch trilinguals with higher proficiency in English and lower proficiency in German, a STI of 0.68 was required for English and $>$ 1 for German [18]. The value indicated by "$>$ 1" signifies that a standard STI equivalent of 0.60 cannot be reached.

While these adjusted STI values reveal the extent of the difficulties non-native listeners experience in various listening conditions, they also highlight the complication involved in employing such values to survey the suitability of a given environment for various proficiency levels of non-native listeners. Thus, for STI, the standard qualification bands would suffice as a framework through which speech perception in standardized EFL proficiency test rooms can objectively be examined.

Of particular interest to the present study are $\geq$ 0.74 STI (category A), $\geq$ 0.70 STI (category B), and $\geq$ 0.66 STI (category C), all noted as exhibiting high speech intelligibility according to the standard STI qualification bands. Since the highest category with $>$ 0.76 STI (category A+) is rated as exhibiting excellent intelligibility but is also mentioned as being "rarely achievable in most environments" [30], high speech intelligibility can be expressed as $\geq$ 0.66 STI.

Also, an intriguing distinction is noted in IEC 60268–16:2020 between the type of messages typically intelligible at or above 0.66 STI (category C, *unfamiliar words*) and below 0.66 STI (category D, *familiar words/contexts*). This may be of some concern for non-native listeners. Not only is the L2 vocabulary knowledge of non-native listeners less developed than that of their L1, but they may also lack skills such as making inferences, the process of guessing the meaning of unknown (*unfamiliar*) words from context. This is an important receptive (listening and reading) skill that L2 learners must acquire if they wish to reach full proficiency. In applying this skill, even when the meaning of the word is unknown to the listener, its phonological cues are still valuable information that can aid comprehension. However, when the phonological cues are unintelligible to the listener, it can lead to greater dependence on cognitive processes such as increased listening effort, which in turn can induce mental fatigue [12,34–37]. In other words, whether or not an unknown word is

intelligible in a given environment may influence the performance of test takers and ultimately the outcome of the test.

Taken together, ≥ 0.66 STI appears to be a reasonable candidate as a minimum recommendation for environments in which standardized EFL proficiency listening tests are administered.

1.4. Reverberation time

One of the major factors contributing to adverse acoustic conditions in an enclosed environment is reverberation. Reverberation is expressed as the measured time in seconds for sound energy to decay by 60 dB in a given space. Since a difference in sound pressure level of 60 dB on top of existing background noise is difficult to achieve in most environments, two alternative methods of measurement, T_{20} and T_{30}, are commonly employed. The time it takes for sound energy to decay by 20 dB from 5 dB to 25 dB below the initial level is labeled T_{20}, and a similar measurement from 5 dB to 35 dB below the initial level is labeled T_{30}. Reverberation time (RT) is typically measured in octave band (125 Hz to 8 kHz) or one-third octave band (100 Hz to 5 kHz) frequencies, but it is also commonly expressed in one averaged value representing the mid-band frequencies of either 500 Hz and 1 kHz ($RT_{0.5\text{-}1kHz}$) or 500 Hz, 1 kHz, and 2 kHz ($RT_{0.5\text{-}2kHz}$).

For listening environments typically categorized as learning spaces (e.g., classrooms, auditoriums, lecture halls), various optimal RT recommendations can be found prescribed in national standards and guidelines around the world [38]. For example, the American National Standards Institute (ANSI) S12.60 [39] recommends $RT_{0.5\text{-}2kHz} \leq 0.6$ s for learning spaces with a volume of ≤ 283 m^3 and $RT_{0.5\text{-}2kHz} \leq 0.7$ s for those between 283 m^3 and 566 m^3. Recommended RT is not specified for learning spaces > 566 m^3. In the UK, Building Bulletin 93 (BB93): Acoustic design of schools—performance standards [40] permits $RT_{0.5\text{-}2kHz} \leq 0.6$ s for primary school classrooms, $RT_{0.5\text{-}2kHz} \leq 0.8$ s for secondary school classrooms, $RT_{0.5\text{-}2kHz} \leq 0.8$ s for small lecture rooms (fewer than 50 people), and $RT_{0.5\text{-}2kHz} \leq 1.0$ s for large lecture rooms (more than 50 people). However, classroom volumes are not specified. In Japan, the AIJES-S001-2020: Academic standards and design guidelines for sound environment in school buildings [41] recommends $RT_{0.5\text{-}1kHz} \leq 0.6$ s for classrooms, meeting rooms, and office spaces with a volume of 200 m^3 and $RT_{0.5\text{-}1kHz} \leq 0.7$ s for those around 300 m^3. The only other recommendations provided for spaces > 300 m^3 are for gymnasiums ($RT_{0.5\text{-}1kHz} \leq 1.6$ s) and auditoriums ($RT_{0.5\text{-}1kHz} \leq 1.3$ s), both around 5,000 m^3.

The German standard DIN 18041:2016: Acoustic quality in rooms—specifications and instructions for the room acoustic design [42], on the other hand, presents linear RT target values as a function of room volume categorized by usage types as well as a frequency-dependent tolerance range against which the measured RT must be verified. Spaces typical of classrooms are listed in usage types A3 "education/communication" and A4 "education/communication (inclusive)", the latter stipulating stricter conditions for foreign language learners and students with hearing impairments. Target values for usage type A3 are indicated from RT 0.3 s at 30 m^3 to 0.8 s at 1,000 m^3 while those for A4 are presented from RT 2.5 s at 30 m^3 to 5.5 s at 500 m^3. For comparison, the target value at 500 m^3 for A3 is around 7.0 s. These German standards are also referenced in the Italian standard UNI 11532-2: Internal acoustical characteristics of confined spaces—design methods and evaluation techniques—part 2 [43]. The indicated RT values for ANSI S12.60, BB93, and AIJES-S001-2020 refer to unoccupied conditions while those for DIN 18041:2016 and UNI 11532-2 are prescribed for occupied rooms at 80 % capacity.

As the examples illustrate, reverberation times recommended to ensure adequate speech intelligibility are typically prescribed for classrooms, but these spaces are mostly defined as being around 200 to 500 m^3. Larger spaces commonly utilized as venues for standardized tests tend to fall outside of the classification of a classroom, and recommended reverberation values for such spaces are either not given or set at a level unsuitable for speech communication. Moreover, not only do these standards generally apply to L1 speech communication and tend not to reflect the difficulties L2 listening tasks entail, but it has also been repeatedly documented that such standards are rarely enforced and are often found to be insufficiently met [44–51]. Collectively, these factors underscore the complicated process of reliable evaluation and selection of test venues based on current RT standards alone.

The work by Peng and Wang [52] is one of the few studies that has addressed this issue. Citing the growing number of non-native English speakers (both students and teachers) in classrooms across the United States, they examined the effects of five $RT_{0.5\text{-}2kHz}$ scenarios (0.37 s, 0.62 s, 0.84 s, 1.05 s, and 1.19 s) and three background noise level (BNL) conditions (RC-30, RC-40, and RC-50) on three adult listener groups' (NAE = native American English speakers, NNC = non-native English speakers whose L1 is Mandarin Chinese, and NNO = non-native English speakers whose L1 is other than Mandarin Chinese) comprehension of English spoken by native American English speakers and non-native English speakers whose L1 is Mandarin Chinese. While they found that RT, BNL, and non-native accent all had varying degrees of adverse effects on the three listener groups, they concluded by summarizing their findings into four sets of RT and BNL recommendations. For situations involving non-native English listeners and native English speakers, the scenario most relevant to the present study, they proposed a recommendation of $RT_{0.5\text{-}2kHz} \leq 0.6$ s with a BNL of ≤ 48 dB(A). Although their recommendation was based on a simulated classroom with a volume of 260 m^3, it can serve as a valuable point of reference in the assessment of standardized EFL proficiency test rooms.

1.5. The present study

With STI and RT recommendations established for the purposes of this research, the present study aims to investigate the current conditions under which paper-based standardized EFL proficiency tests are conducted by examining the STI and RT in classrooms frequently used as test rooms for the TOEIC L&R IP Test at a university in Japan. In particular, STI for three types of loudspeakers (sound sources) commonly used to administer the listening test will be explored.

It should be noted that one of the authors has been involved in the administration of the TOEIC L&R IP Test at the university as a test examiner since 2013 and has experienced firsthand the difficulty of interpreting the test manuals which make no mention of instructions or procedures to ensure sufficient audio quality during the listening test. Given that the same classrooms have also been used as venues for the TOEIC L&R SP Test several times a year since 2002 according to university records, the findings of this study may serve as a cautionary note for the millions of test takers as well as tens of thousands of organizations that rely on these tests as an objective, standardized assessment of one's language abilities.

Lastly, the focus of the study will solely be on the acoustics of the test rooms. The implications of the test takers' listening skills or the complexity of the test material are beyond the scope of this study and will not be discussed.

2. Materials and methods

2.1. Test rooms

Ten classrooms from a university in Japan were selected for the present study. Rooms 1 to 8 were selected due to their use as test

rooms for the TOEIC L&R IP Test administered two to three times a year to all first- and second-year students at the university since 2012. (In response to the COVID-19 pandemic, the TOEIC® Program IP Test Online was adopted for the 2020 and 2021 academic year.) In addition, two recently renovated classrooms, rooms 9 and 10, were included in the investigation for their prospect as future test rooms. The classrooms will be referred to hereafter as test rooms.

A summary of the test rooms is presented in Table 1. The room numbers are assigned from 1 to 10 according to the order of their room volume (m^3) from largest to smallest. Rooms 1 to 6 have a tiered or sloped floor with an uneven ceiling, so the ceiling height is a mean value of measurements taken at several points around the room. A traditional seating style indicates that every desk is facing forward, while in a stadium seating style the desks on the sides are slightly angled toward the lecture podium.

2.2. Sound sources

Three types of loudspeakers were investigated in each test room: wall-mounted speakers (WMS), radio cassette player (RCP), and amplified speaker (AMP). Wall-mounted speakers refer to the loudspeakers installed in each room. The size, output power, and location of the loudspeakers varied by room, but all rooms were equipped with two loudspeakers positioned around the top right and top left areas at the front of the room. Room 2 was the sole exception with two additional loudspeakers installed on the ceiling at the midway point between the front and the back of the room. Specifications of the power amplifier and loudspeakers by test room are presented in Table 2.

A radio cassette player refers to a portable CD player and loudspeaker unit. Although the cassette player component has long disappeared from the modern version of these devices, in Japan they are still commonly called *rajikase*, an abbreviated version of "radio cassette player." The Sony ZS-RS81BT was used as the instrument for RCP in this study. An amplified speaker refers to a much larger and more powerful portable speaker unit with a built-in CD player. It has the capability to add three wireless microphone channels and a connection to external loudspeakers. The UNI-PEX WA-361A was used as the instrument for AMP in this study. The lecture podium provided in each room was used as the stand for RCP and AMP, a common procedure for standardized EFL proficiency tests. This determined their height which ranged from 0.90 m to 1.24 m to the base of the loudspeakers. The podium was positioned to ensure that the loudspeakers were equidistant from the side walls as well as from the frontmost seats to the front wall. Specifications for RCP and AMP are provided in Table 3.

2.3. Speech transmission index

As the present study concerns listening conditions under the use of loudspeakers, the speech transmission index for public address systems (STIPA) method, a simplified alternative to the full STI method developed specifically for this purpose [30], was employed. STIPA measurements were conducted using the NTi Audio XL2 Audio and Acoustic Analyzer and the M4260 omnidirectional condenser microphone. The M4260 was set on a mic stand at 1.20 m, a typical height of a seated person's ear, and placed over the desk at each measurement (receiver) position. The NTi Audio ASD (Automatic Sensor Detection) cable was used to connect the M4260 and the XL2 Analyzer, allowing the sensitivity and calibration data from the microphone to be read by the XL2 Analyzer. The cable also allowed the researcher to stay away from the acoustic field during the measurement sessions. The accompanying NTi Audio STIPA test signal CD V1.1 was used to produce the pink noise test signal. The test signal was generated by the CD player built into each of the three sound sources. The sound pressure level (SPL) for each loudspeaker was set to L_{AS} 80 dB at a reference point established by placing the M4260 at the center of the front row seats in each test room. This reference SPL was determined to be representative of typical test conditions based on a combination of one of the authors' past experience and anecdotal reports of other test examiners.

The receiver positions for each test room were determined by starting with the four corner seats and distributing the rest as evenly as possible. If the desks were arranged asymmetrically or in an unusual pattern, the measurements in those areas were conducted at the nearest position that a test taker would typically be seated. Between 15 and 30 receiver positions were selected per room depending on the desk arrangement and room size. All source-receiver distances were larger than the critical radius (r_c) calculated for each room in line with the equation reported in Puglisi et al. [53] (see Table 4). The measurement at each position consisted of three 15 s cycles agreeing within 0.03 STI, from which the average was recorded as the STI for that position. If the measurements at each position differed by more than 0.03 STI or if an impulsive noise was detected, three additional cycles were performed at the same position. In every test room, all lights were turned on, all windows were closed, all blinds were shut, the HVAC systems were turned off, and it was made sure beforehand that no other classes or activities would be in progress in nearby classrooms. The measurements were carried out in unoccupied test rooms except for the presence of one researcher.

The STIPA receiver positions, the WMS source positions, the lecture podium (LP) indicating the portable RCP and AMP source positions, and the location of the L_{AS} 80 dB reference point by test room are illustrated in Fig. 1.

2.3.1. Background noise

To account for the effect of background noise typically present in actual test conditions, background noise levels (BNL) for the HVAC systems were subsequently collected for all test rooms. Two noise sources were considered: air conditioner (AC) and

Table 1
Summary of test rooms ordered from largest to smallest volume (m^3). Means and standard deviations (SD) are indicated at the bottom.

Room	Building	Height (m)	Area (m^2)	Volume (m^3)	Seating Capacity	Seating Style	Floor Type
1	D	3.6	271	967	304	Stadium	Tiered
2	A	3.1	283	886	272	Stadium	Sloped
3	C	3.2	210	668	217	Traditional	Tiered
4	C	3.2	210	668	230	Traditional	Tiered
5	C	3.2	210	668	230	Traditional	Tiered
6	D	3.3	182	606	208	Stadium	Tiered
7	B	3.1	130	397	137	Traditional	Flat
8	B	3.0	121	364	146	Traditional	Flat
9	B	2.7	123	333	130	Traditional	Flat
10	B	2.7	111	301	114	Traditional	Flat
	Mean (SD)	3.1 (0.2)	185 (59.4)	586 (220.5)	199 (61.1)		

Table 2
Power amplifier and loudspeaker specifications of the wall-mounted speakers (WMS) by test room.

Room	Power Amplifier		Loudspeaker			
	Brand & Model Number	OutputPower (W)	Brand & Model Number	Speaker Type	Speaker Components	Number of Speaker Units
1	Sony SRP-X500P	90 W	Panasonic WS-AT75-K	2-way	20 cm woofer + SCWG horn tweeter	2
2	Victor PA-604	40 W	TOA F-1000WM	2-way	10 cm woofer + balanced dome tweeter	4
3	JVC PA-806	60 W	JVC PS-S550B	2-way	16 cm woofer + SCWG horn tweeter	2
4	JVC PA-806	60 W	Panasonic WS-AT80	2-way	20 cm woofer + SCWG horn tweeter	2
5	JVC PA-806	60 W	Panasonic WS-AT80	2-way	20 cm woofer + SCWG horn tweeter	2
6	Sony SRP-X500P	90 W	Bose 301 AV Monitor	2-way	20 cm woofer + two 7.6 cm tweeters	2
7	Sony SRP-X350P	50 W	Sony SRP-S320	1-way	12 cm full range	2
8	Panasonic WA-H60	60 W	Sharp AN-SP100	1-way	10 cm full range	2
9	Yamaha MA2030a	30 W	Sony SRP-S320	1-way	12 cm full range	2
10	Sony SRP-X350P	50 W	Sony SRP-S320	1-way	12 cm full range	2

Table 3
Specifications of the portable radio cassette player (RCP) and amplified speaker (AMP).

	Brand & Model Number	Output Power (W)	Speaker Type	Speaker Components
RCP	Sony ZS-RS81BT	2 W + 2 W	1-way	Two 8 cm full range speakers
AMP	UNI-PEX WA-361A	20 W	2-way	20 cm woofer + 2.5 cm balanced dome tweeter

Table 4
Distances (m) from STIPA sources (WMS, RCP/AMP) and RT source (DS3) to their nearest receiver positions in each room. Critical radius (m) for STIPA measurements is denoted as r_c, and the minimum distance (m) for RT measurements is denoted as d_{min}.

	STIPA			RT	
Room	WMS	RCP/AMP	r_c	DS3	d_{min}
1	4.22	2.41	1.85	3.49	3.34
2	1.94	2.28	1.87	3.92	3.02
3	1.77	1.95	1.67	3.15	2.70
4	1.72	2.13	1.68	3.25	2.70
5	1.72	2.13	1.66	3.15	2.70
6	3.35	1.89	1.62	2.63	2.53
7	2.68	1.75	1.29	2.67	2.09
8	3.23	1.67	1.24	2.84	2.02
9	2.42	1.20	0.96	2.65	1.97
10	2.69	1.36	0.94	3.91	1.86

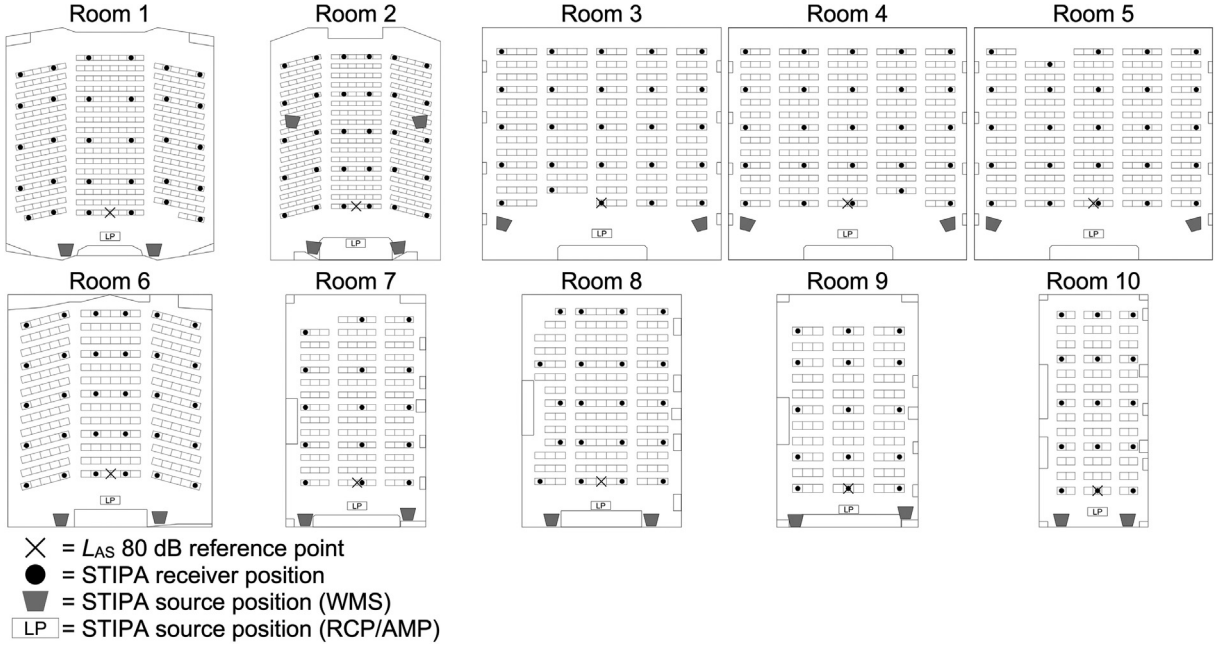

Fig. 1. STIPA sources and receiver positions by test room.

ventilation. Since the type of AC differed by test room, the *middle* setting for the fan strength was selected. For example, the AC in rooms 7, 9, and 10 had five strength settings (1 to 5), so it was set to *3*. The AC in rooms 1, 3, 4, 5, and 6 had three strength settings (weak, strong, powerful), so it was set to *strong*. Room 8 did not have a *powerful* setting, but it was also set to *strong* as the AC system was otherwise identical. Room 2 similarly did not have a *powerful* setting, but it was set to *weak* because the *strong* setting produced an exceedingly louder noise ($>$ 50 dB) compared to that of the other test rooms (see Table 5) and presumably would not be used during the listening portion of a standardized EFL proficiency test. As for the ventilation which simply consisted of an on/off switch, it was turned *on* for rooms 2 to 7. A ventilation system was not present in rooms 1, 8, 9, and 10.

Following the same procedures for STIPA, the measurements were collected using the ambient noise correction setting on the XL2 Analyzer. BNL was measured for 30 s at each STIPA receiver position and recorded in octave band frequency from 125 Hz to 8 kHz. The STIPA and BNL data were then imported into the NTi Audio STI Report Tool (version 4.5), a macro-enabled Excel file, to generate the noise-added STI values for the three sound sources. All measurements were conducted in accordance with IEC 60268-16:2020.

2.4. Reverberation time

Reverberation time (RT) was measured in octave band frequency using the NTi Audio XL2 Audio and Acoustic Analyzer and the omnidirectional DS3 Dodecahedron Speaker Set. The pink noise test signal supplied within the PA3 Power Amplifier was played through the DS3 positioned at a height of 1.50 m from the floor to its base. The M4260 was setup identically to that of STIPA except for its orientation. It was mounted vertically at a height of 1.20 m from the floor to the topmost tip of the microphone.

Three source positions were selected for the DS3 in rooms 1 to 6 whereas two were selected for the smaller rooms 7 to 10. For each DS3 location, the M4260 was placed in three randomly selected receiver positions, and three cycles were performed at each receiver position. All source-receiver distances were beyond the minimum distance (d_{min}) calculated for each room in accordance with the equation prescribed in ISO 3382–2:2008 [54] (see Table 4). The number of positions and measurements also meets the ISO 3382–2:2008 standard for the engineering method requiring $\geq$ 2 source-positions, $\geq$ 2 receiver-positions, and $\geq$ 6 source-receiver combinations. T_{30} measurements were performed for rooms 3 to 5 and 7 to 10. For rooms 1, 2, and 6, T_{20} measurements were performed because the dynamic range necessary for a T_{30} measurement to be evaluated from the decay curve (at least 45 dB above the background noise) could not be reached by the pink noise test signal in all octave band frequencies.

The procedures for preparing the test rooms for the STIPA measurement sessions were also followed for RT. All measurements were conducted in unoccupied conditions in accordance with ISO 3382–2:2008. The RT source and receiver positions for each test room are illustrated in Fig. 2.

3. Results and discussion

3.1. Speech transmission index

Exploratory analysis of the STIPA data was initially performed to determine the appropriate statistical method. With 30 measurements for each of the three sound sources in rooms 1, 2, and 6 and $<$ 30 for those in the remaining seven test rooms, statistical methods appropriate for small sample sizes would be applicable for this study. Since a normal distribution can not be assumed with a small sample size, the Shapiro-Wilk test and a visual inspection of the quantile–quantile (Q-Q) plots were conducted to examine the normality of the data. The analysis suggested that the data were not normally distributed. The test statistic (W) and the significance values (p) from the Shapiro-Wilk tests are presented in Table 6. A significant value ($p < 0.05$) indicates evidence of non-normal distribution.

Based on the evidence of non-normal data, a Friedman test, the non-parametric counterpart for a repeated measures ANOVA, was run in IBM SPSS (version 27) to determine if there were differences in STIPA measurements between the three sound sources. The analysis revealed that a statistically significant difference was found in all test rooms except 7, 9, and 10. The test statistic (χ^2), degrees of freedom (number in parenthesis), and the significance values (p) from the Friedman test results are presented in Table 7.

To examine the differences for each pair of sound sources, the Wilcoxon signed-rank test was used to run the post hoc analysis for all test rooms. A Bonferroni correction for multiple comparisons was applied, which set the statistical significance to be accepted at $p < 0.0167$. The total number of measurements per pairwise comparison (n), standardized test statistic (z), significance values (p), and Pearson's correlation coefficient as effect sizes (r) from the post hoc analysis are presented in Table 8.

The post hoc analysis revealed statistically significant differences in STI between WMS and RCP in test rooms 1, 6, and 8, between WMS and AMP in test rooms 1, 2, 3, 4, and 6, and between RCP and AMP in test rooms 1 to 8. Among the three rooms (rooms 7, 9, and 10) in which the Friedman test yielded no statistical significance, only one pair (RCP and AMP) in room 7 had a statistically

Table 5
Results of BNL measurements (dBA) with (+HVAC) and without HVAC systems by test room. Means and standard deviations (SD) are indicated at the bottom. Selected AC and ventilation settings are bolded and underlined. AC settings W, S, P denote Weak, Strong, Powerful, respectively.

Room	BNL (dBA)	BNL + HVAC (dBA)	AC fan strength setting	Ventilation setting
1	29.8	40.5	W **S** P	n/a
2	31.5	43.3	**W** S	**ON** OFF
3	29.9	39.5	W **S** P	**ON** OFF
4	29.8	43.5	W **S** P	**ON** OFF
5	29.8	40.3	W **S** P	**ON** OFF
6	32.0	41.3	W **S** P	**ON** OFF
7	30.9	41.7	1 2 **3** 4 5	**ON** OFF
8	31.1	44.2	W **S**	n/a
9	30.0	36.8	1 2 **3** 4 5	n/a
10	33.1	35.6	1 2 **3** 4 5	n/a
Mean (SD)	30.8 (1.1)	41.0 (2.8)		

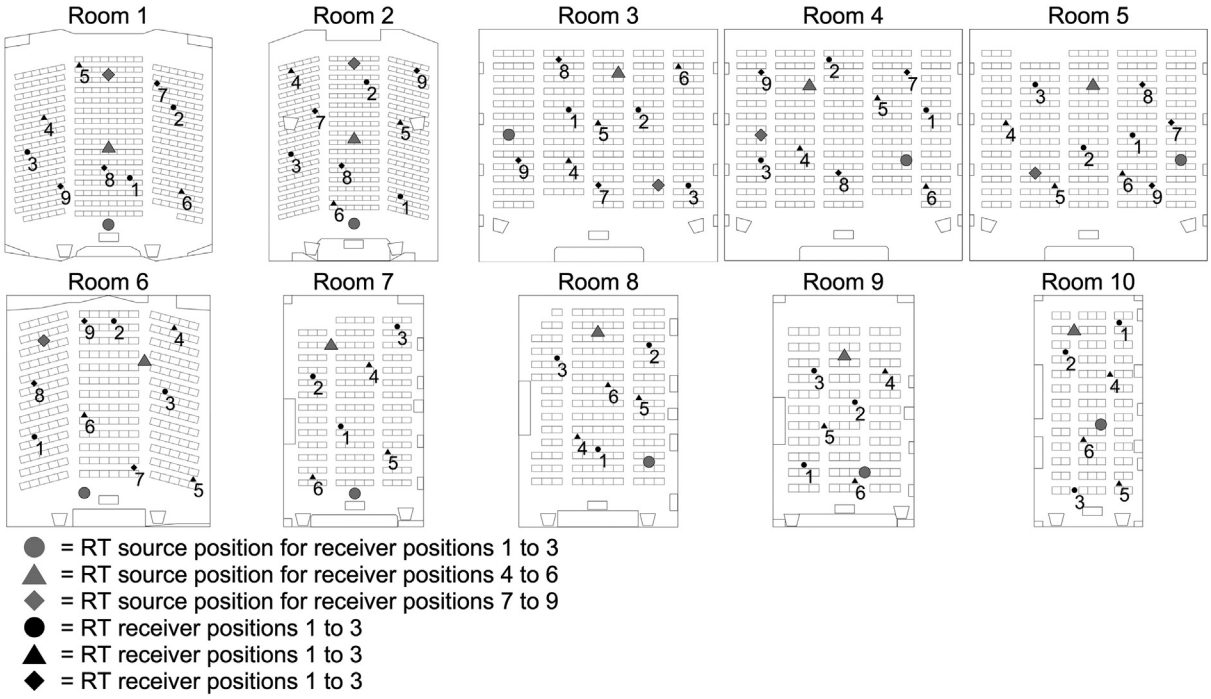

Fig. 2. RT source and receiver positions by test room.

Table 6
Results of the Shapiro-Wilk test of normality for the STIPA measurements by test room and sound source. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Room	WMS (W)	WMS (p)	RCP (W)	RCP (p)	AMP (W)	AMP (p)
1	0.98	0.831	0.98	0.733	0.90	**0.007****
2	0.96	0.355	0.83	< **0.001*****	0.83	< **0.001*****
3	0.95	0.247	0.93	0.107	0.89	**0.014***
4	0.93	0.099	0.93	0.082	0.88	**0.006****
5	0.93	0.076	0.92	0.056	0.90	**0.019***
6	0.95	0.128	0.90	**0.010****	0.95	0.171
7	0.96	0.616	0.82	**0.007****	0.74	< **0.001*****
8	0.96	0.499	0.92	0.081	0.94	0.198
9	0.97	0.802	0.80	**0.004****	0.80	**0.004****
10	0.86	**0.023***	0.71	< **0.001*****	0.65	< **0.001*****

Table 7
Friedman test results for the STIPA measurements by test room. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Room	$\chi^2(2)$	p	Room	$\chi^2(2)$	p
1	56.27	< **0.001*****	6	24.07	< **0.001*****
2	26.60	< **0.001*****	7	5.20	0.074
3	17.84	< **0.001*****	8	18.10	< **0.001*****
4	25.28	< **0.001*****	9	0.40	0.819
5	12.48	**0.002****	10	5.20	0.074

significant difference. Boxplots of the STIPA measurements are presented in Fig. 3. Statistical significance annotation and a solid horizontal line representing a STI value of 0.66, the minimum recommendation proposed in section 1.3, are also displayed.

Additionally, as shown in Fig. 4, the STI values were arranged by their receiver positions, overlayed onto the room schematics, and formatted according to the following conditions. The dark to light grey shades represent the four highest STI qualification categories prescribed in IEC 60268–16:2020 ($\geq$ 0.76 STI = A+, $\geq$ 0.74 STI = A, $\geq$ 0.70 STI = B, $\geq$ 0.66 STI = C). To distinguish these four categories from the rest, areas with STI below 0.66 were left unformatted with a white background. Furthermore, the texts indicating the STI values were displayed down to 0.63 STI while those below it were omitted. Values down to 0.63 were included because Bradley et al. [55] demonstrated that a change in STI up to 0.03 can be classified as a just noticeable difference (JND), a threshold under which a difference is said to be unnoticeable.

3.1.1. STI differences by sound source

One pattern that emerged from this investigation is the clear difference in STIPA measurements between WMS and the other two loudspeakers. The analysis revealed that RCP and AMP exhibited higher STI compared to WMS in every test room except room 8. This finding sharply contrasts with the authors' initial belief that WMS would yield higher and more consistent STIPA measurements. We assumed that WMS would be the best method of distributing audio signals evenly and directly to each receiver position compared to a portable loudspeaker unit simply placed

Table 8
Pairwise comparison results of the Wilcoxon signed-rank test for sound sources by test room. Bonferroni corrections applied at $p < 0.0167$. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Room	Sample 1	Sample 2	n	z	p	r
1	WMS	RCP	60	4.78	$<$ **0.001*****	0.62
	WMS	AMP	60	4.78	$<$ **0.001*****	0.62
	RCP	AMP	60	4.62	$<$ **0.001*****	0.60
2	WMS	RCP	60	−0.13	2.681	−0.02
	WMS	AMP	60	3.28	$<$ **0.003****	0.42
	RCP	AMP	60	4.72	$<$ **0.001*****	0.61
3	WMS	RCP	50	−0.79	1.282	−0.11
	WMS	AMP	50	2.87	**0.013***	0.41
	RCP	AMP	50	3.67	$<$ **0.001*****	0.52
4	WMS	RCP	50	−0.71	1.428	−0.10
	WMS	AMP	50	2.62	**0.026***	0.37
	RCP	AMP	50	4.37	$<$ **0.001*****	0.62
5	WMS	RCP	50	0.09	2.775	0.01
	WMS	AMP	50	2.01	0.135	0.28
	RCP	AMP	50	4.18	$<$ **0.001*****	0.59
6	WMS	RCP	60	3.16	**0.005****	0.41
	WMS	AMP	60	3.67	$<$ **0.001*****	0.47
	RCP	AMP	60	3.24	**0.004****	0.42
7	WMS	RCP	30	−0.80	1.280	−0.15
	WMS	AMP	30	1.02	0.920	0.19
	RCP	AMP	30	2.44	**0.044***	0.45
8	WMS	RCP	40	−3.25	**0.004****	−0.51
	WMS	AMP	40	−1.01	0.940	−0.16
	RCP	AMP	40	3.14	**0.005****	0.50
9	WMS	RCP	30	1.14	0.768	0.21
	WMS	AMP	30	0.91	1.091	0.17
	RCP	AMP	30	−0.68	1.487	−0.12
10	WMS	RCP	30	−0.11	2.729	−0.02
	WMS	AMP	30	1.48	0.419	0.27
	RCP	AMP	30	2.33	0.060	0.43

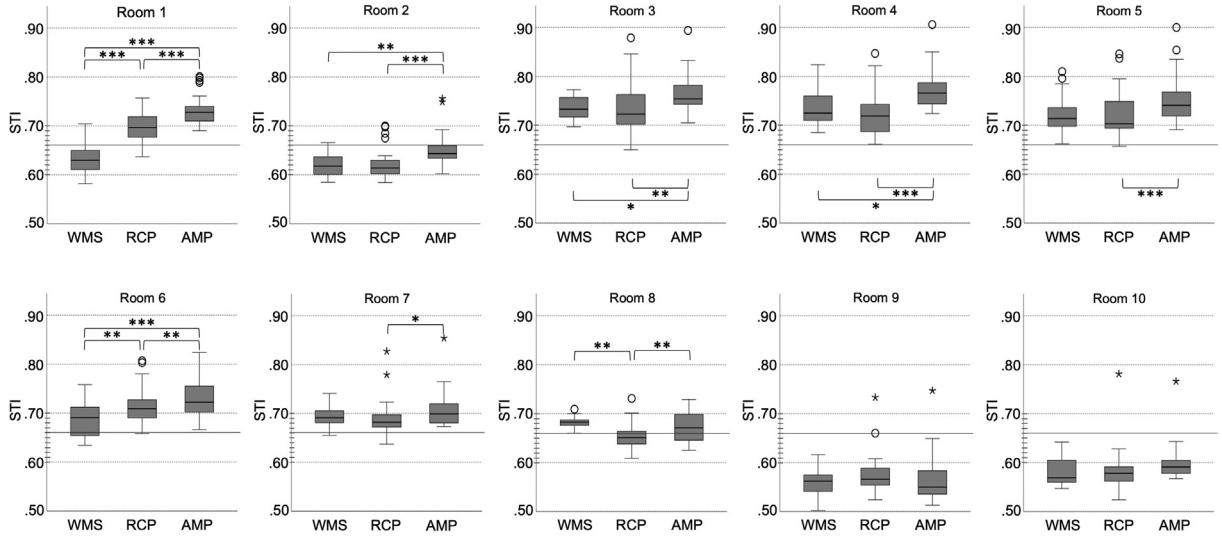

Fig. 3. Boxplots of STIPA measurement results by test room and sound source. The median is depicted by the bold lines in the boxes, the boxes represent the first and the third quartiles, and the whiskers extend to 1.5 interquartile range. Outliers are denoted by circles and asterisks beyond the whiskers. The solid horizontal line across the graph indicates 0.66 STI. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

on the lecture podium. We also expected WMS to have a greater advantage over the portable loudspeaker units as the test room increased in size since the power amplifier and loudspeaker specifications of WMS seemingly complemented the acoustic demands of each room while those of RCP and AMP remained the same. In fact, the RCP failed to reach the baseline SPL of L_{AS} 80 dB at the reference point even at the maximum volume setting in rooms 1 to 7. Nevertheless, RCP and AMP performed equally well or better than WMS in almost all test rooms.

One possible explanation is that our initial assumptions were predicated on the idea that PA systems are selected, installed, and calibrated upon careful consideration of the acoustic characteristics unique to each room. However, this may have been a case of naïve overestimation, and such ideal scenarios may only be reserved for acoustically demanding spaces such as recording studios and concert halls. In her review of research on classroom sound field amplification, Millett [56] highlighted numerous advantages of adopting such technology but also raised concerns regarding potential issues that may arise from factors such as improper installation, inadequate teacher/user training, and lack of system maintenance. Lafargue & Lafargue [57] even suggested that inattention to these matters can lead to a listening experience

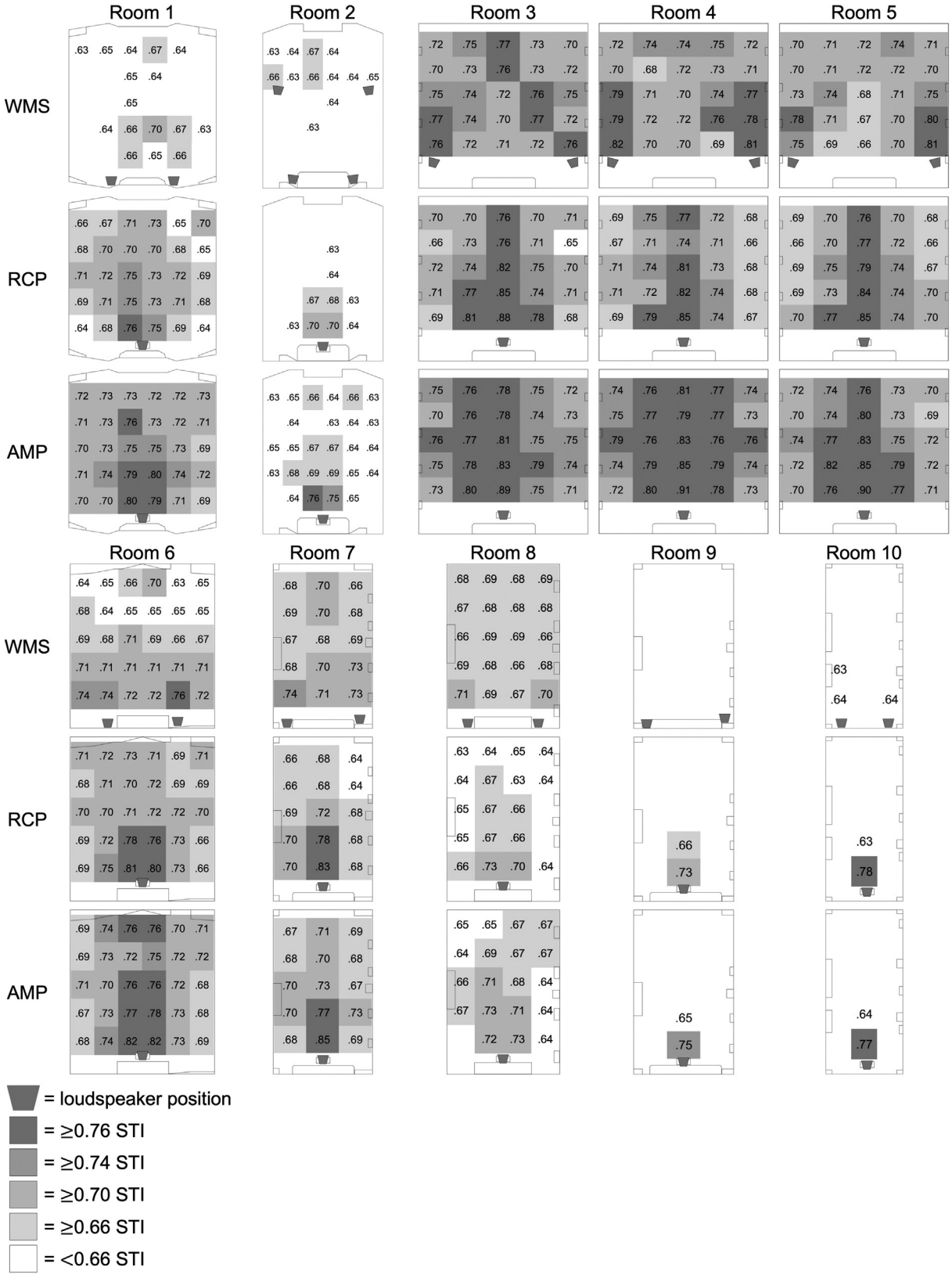

Fig. 4. STIPA measurement results by test room, sound source, and receiver position. Dark to light grey shades represent the four highest STI qualification categories ($\geq$ 0.76 STI, $\geq$ 0.74 STI, $\geq$ 0.70 STI, $\geq$ 0.66 STI). White areas denote $<$ 0.66 STI. Numbers indicating the STI values $\geq$ 0.66 are displayed over their corresponding shades, and those between 0.66 and 0.63 STI (representing a just noticeable difference of 0.03 STI) are displayed over the white background. Plain white areas with omitted STI values represent measurement results exhibiting $<$ 0.63 STI.

more detrimental than the one prior to its installation. As the primary purpose of university classrooms generally entails delivering lectures in the listeners' L1, and L1 speech comprehension is consistently found to be more robust to various levels of adverse acoustic conditions compared to listening tasks in L2 [24–28], the PA systems in the test rooms investigated in this study may have been installed without careful consideration of the issues above but remained adequate enough to facilitate L1 speech intelligibility. Moreover, lack of system maintenance may have led to equipment deterioration in one of the test rooms in this study. One of the WMS loudspeakers in room 1 produced a noticeable hum throughout the measurement session which may have interfered with the pink noise test signal, which in turn may have caused the unfavorable STIPA measurement results (mostly $<$ 0.66 STI) obtained in this room.

Another possible explanation concerns the location and configuration of the loudspeakers. As stated in section 2.2., RCP and AMP were placed on the lecture podium and positioned equidistant from the side walls as well as from the frontmost seats to the front wall. This means that there was only one central loudspeaker source from which the test signal was generated for the STIPA measurement sessions. On the other hand, WMS consisted of loudspeakers positioned around the top right and top left areas at the front of the room, meaning that the test signal was always generated simultaneously by two loudspeakers from the right and left sides of the room. While such stereo configuration would promote an enjoyable listening experience for music, studies suggest that the same may not be true for speech comprehension. This is because playing audio in this configuration creates an effect called acoustical crosstalk where the sound signals reach the listener's ears simultaneously but slightly out of phase. This in turn produces a phenomenon known as central stereo image, or phantom center image, where the sound signal appears to be coming from a point between the loudspeakers [58]. Under this condition, there is evidence suggesting that cancellations may occur in certain frequencies crucial to the intelligibility of speech [59]. In fact, Shirley et al. [58] showed that a word recognition task administered using a central source yielded significantly higher speech intelligibility compared to the same task performed with a phantom center image. The German standard DIN 18041:2016 [42] also advises against the use of decentralized sound sources in rooms where comprehension of speech is of critical importance. Although an in-depth comparison of these acoustical effects was not conducted in this investigation, the studies above suggest that the location and configuration of the loudspeakers may also play a role in the strength and distribution of STI observed in test rooms, and that the central loudspeaker position of RCP and AMP may be more suitable for standardized EFL proficiency listening tests.

3.2. Reverberation time

The RT measurement results are presented in Table 9 and Fig. 5. Table 9 shows the measured RT decay range (T_{30} or T_{20}), the mean values for octave band frequencies from 125 Hz to 8 kHz (9 measurements in rooms 1 to 6; 6 measurements in rooms 7 to 10), the combined mean values for mid-band frequencies 500 Hz to 1 kHz and 500 Hz to 2 kHz, and their standard deviations in parenthesis. Boxplots for the mid-band frequencies 500 Hz to 2 kHz generated for visual analysis are displayed in Fig. 5.

Comparing the RT results with the recommendations mentioned in section 1.4, only room 7 met the ANSI S12.60 recommendation of $RT_{0.5\text{-}2kHz} \leq 0.7$ s for learning spaces between 283 m^3 and 566 m^3, while rooms 1 to 8 met the BB93 recommendation, which does not specify any classroom volume, of $RT_{0.5\text{-}2kHz} \leq 1.0$ s for large lecture rooms designed for more than 50 people. For AIJES-S001-2020, the standard most relevant to the Japanese university classrooms investigated in this study, only rooms 9 and 10 fit the description of classrooms around 300 m^3, both of which far exceeded the recommendation of $RT_{0.5\text{-}1kHz} \leq 0.7$ s with $RT_{0.5\text{-}1kHz}$ over 1.25 s and 1.17 s, respectively. None of the test rooms met Peng and Wang's [52] recommendation of $RT_{0.5\text{-}2kHz} \leq 0.6$ s for situations involving non-native English listeners and native English speakers, but it should also be noted that all the rooms in this study were much larger than their simulated classroom of 260 m^3.

As for the German standard DIN 18041:2016, verification of the measurement results between 125 Hz and 4 kHz was performed as outlined in Annex A [42]. Data conversion was carried out to generate the frequency-dependent reverberation time for the occupied state at 50 % as it is common procedure to assign test takers to every other seat when administering standardized tests. The sound absorption area per person (ΔA1Person) values for "person sitting on lightly upholstered seating" [42] were applied to the conversion. The adjusted results for usage types A3 and A4 mentioned in section 1.4 along with the recommended optimal RT range as a function of frequency are presented in Fig. 6.

It can be observed in Fig. 6(a) that, with the exception of room 8, all test rooms exhibited RT outside of the optimal range in at least one octave band frequency. However, DIN 18041:2016 states that "a moderate increase in reverberation time at low frequencies does not compromise acoustic quality," and "in rooms for spoken information and speech communication, the reverberation times should be shorter rather than longer" (p.12) [42]. Therefore, it can be argued that rooms 1 to 8 all met the required RT for their respective volumes for usage type A3, that is, for normal hearing persons in spaces intended for "education/communication." In the case of usage type A4 shown in Fig. 6(b), all test rooms exceeded the upper tolerance range in at least one octave band frequency due to the stricter requirements stipulated for foreign language learners and students with hearing impairments. However, given that a 5 % difference in perceived reverberance is listed as a JND in ISO 3382–1:2009 [60], it can be said that rooms 3 to 8 for the most part also met the stricter requirements whereas rooms 1 and 2 exhibited slightly less favorable conditions. In both cases, rooms 9 and 10 resulted in RT completely beyond the optimal range even after the conversion.

3.3. STI & RT results by test room

3.3.1. Rooms 1 and 6

Rooms 1 and 6 are located in the same department and, aside from room volumes of 967 m^3 and 606 m^3, share a similar architectural style including being the only two test rooms that feature stadium seating and a tiered floor. They are also the only test rooms in which statistically significant differences in STI were found between all three sound sources as well as a clear advantage for RCP and AMP compared to WMS. A closer examination of the results reveals that rooms 1 and 6 had the two longest reverberation times for mid-band frequencies $RT_{0.5\text{-}1kHz}$ (0.73 s and 0.65 s) and $RT_{0.5\text{-}2kHz}$ (0.77 s and 0.69 s) among the six test rooms (1, 3, 4, 5, 6, and 7) in which $\geq$ 0.66 STI was recorded in every receiver position (including those below 0.66 but within the JND range) for RCP and AMP. In other words, RT in rooms 1 and 6 represent the upper boundaries for an optimal range that would be conducive to exhibiting $\geq$ 0.66 STI throughout the entire room.

Comparison with other studies that report the STI and RT measurements of university classrooms with similar architectural styles seem to corroborate this point. Analyzing the STI, RT, and BNL of 17 university rooms of varying purposes and sizes, Escobar and Morillas [61] found that rooms LR1 (815 m^3) and AU4 (1,300 m^3) had $RT_{0.5\text{-}2kHz}$ 0.73 s with 0.71 STI and $RT_{0.5\text{-}2kHz}$ 0.64 s with 0.76 STI, respectively, while rooms CL1 (840 m^3) and CL4 (610 m^3) had $RT_{0.5\text{-}2kHz}$ 2.41 s with 0.50 STI and $RT_{0.5\text{-}2kHz}$ 1.94 s

Table 9
RT measurement results (s) by test room and octave band frequency from 125 Hz to 8 kHz. Measured decay range is noted as T_{20} or T_{30}. Mid-band frequencies are denoted by $RT_{0.5\text{-}1kHz}$ and $RT_{0.5\text{-}2kHz}$ and their standard deviations in parenthesis. Lowest standard deviations are bolded and underlined.

Room	Decay range	RT_{125Hz}	RT_{250Hz}	RT_{500Hz}	RT_{1kHz}	RT_{2kHz}	RT_{4kHz}	RT_{8kHz}	$RT_{0.5\text{-}1kHz}$	$RT_{0.5\text{-}2kHz}$
1	T_{20}	1.35	1.05	0.80	0.67	0.85	0.94	0.67	0.73 (0.037)	0.77 (0.046)
2	T_{20}	0.59	0.62	0.67	0.86	1.07	1.00	0.75	0.77 (0.052)	0.87 (0.045)
3	T_{30}	1.17	0.78	0.58	0.58	0.84	0.89	0.62	0.58 (0.023)	0.67 (0.023)
4	T_{30}	1.08	0.78	0.58	0.56	0.81	0.92	0.66	0.57 (0.023)	0.65 (0.018)
5	T_{30}	1.19	0.74	0.60	0.58	0.83	0.89	0.67	0.59 (0.022)	0.67 (0.021)
6	T_{20}	1.02	0.80	0.65	0.65	0.78	0.84	0.57	0.65 (0.018)	0.69 (0.020)
7	T_{30}	1.02	0.86	0.59	0.58	0.79	0.84	0.65	0.59 (0.023)	0.65 (0.017)
8	T_{30}	1.06	0.77	0.75	0.72	0.76	0.73	0.51	0.73 (**0.008**)	0.74 (**0.005**)
9	T_{30}	1.34	1.07	1.21	1.29	1.25	1.07	0.85	1.25 (0.022)	1.25 (0.016)
10	T_{30}	1.28	1.01	1.13	1.20	1.18	1.03	0.87	1.17 (0.014)	1.17 (0.011)

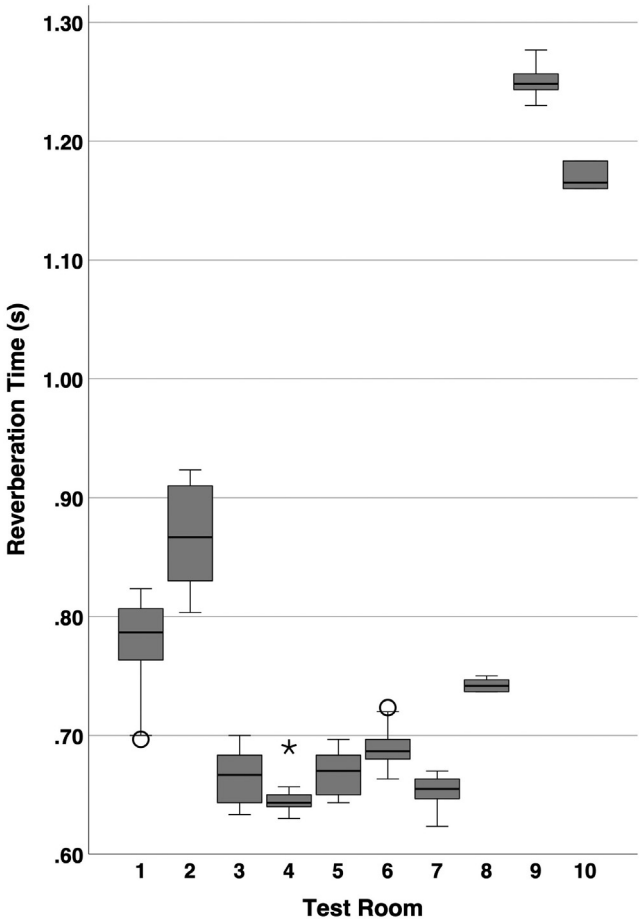

Fig. 5. Boxplots of $RT_{0.5\text{-}2kHz}$ measurements by test room. The median is depicted by the bold lines, the boxes represent the first and the third quartiles, and the whiskers extend to 1.5 interquartile range. Outliers are denoted by circles and asterisks beyond the whiskers.

with 0.52 STI, respectively. Nestoras and Dance [62] examined 10 representative university classrooms and found that their lecture theater (753 m^3) with a short RT (frequency unspecified) of 0.56 s yielded a STI of 0.78. Ricciardi and Buratti [63] documented a host of thermal, acoustic, and lighting data gathered from seven university classrooms, reporting that $RT_{0.125\text{-}8kHz}$ 0.95 s with 0.62 STI and $RT_{0.125\text{-}8kHz}$ 1.40 s with 0.52 STI were recorded for stadium style rooms 1 (808 m^3) and 4 (680 m^3), respectively. Finally, Fratoni et al. [64] collected measurements from six university lecture halls, of which an "amphitheater" shaped hall *a* (1000 m^3) and hall *b* (900 m^3) exhibited $RT_{0.5\text{-}1kHz,occ}$ 0.99 s with 0.49 STI and $RT_{0.5\text{-}1kHz,occ}$ 0.90 s with 0.48 STI, respectively. The reverberation times stated in Fratoni et al. are values converted to 80 % occupancy according to UNI 11532-2 [43]. A visual analysis of the unoccupied measurements provided in their figures indicates $RT_{0.5\text{-}1kHz}$ of around 1.7 s for both rooms.

Although these RT results are reported in different frequency ranges, only one STI value is reported for each room, and the STI measurements are not compared across different sound sources, it can still be deduced from the studies above that RT values comparable to those in rooms 1 and 6 may serve as a reference in determining if similar lecture halls would ensure $\geq$ 0.66 STI for all occupants, provided that a central loudspeaker source is used.

3.3.2. Room 2

Room 2, the second largest room in this study (886 m^3) with the third highest $RT_{0.5\text{-}2kHz}$ of 0.87 s, met the BB93 recommendation of $RT_{0.5\text{-}2kHz} \leq 1.0$ s for large lecture rooms designed for more than 50 people, and the same can be argued for the optimal RT range specified in DIN 18041:2016 for usage type A3 "education/communication" (see section 3.2). Nevertheless, all three sound sources in room 2 exhibited STI values mostly below 0.66, the only other room besides rooms 9 and 10, both with $RT_{0.5\text{-}2kHz}$ well over 1.0 s, to yield such unfavorable results. A closer examination of the results by frequency in room 2 reveals that 125 Hz and 250 Hz, which tend to exhibit longer RT compared to the mid-band frequencies, had the shortest RT out of all the test rooms. They also reveal that 500 Hz to 2 kHz had the steepest rise in RT from 0.67 s to 1.07 s, resulting in a RT boxplot (Fig. 5) with a large interquartile range that spans almost an entire second.

Short reverberation times in the lower frequencies are not necessarily detrimental, as DIN 18041:2016 even suggests the use of a high-pass filter from 150 Hz to 200 Hz for PA systems intended for spoken presentations [42]. Compared to room 1, however, room 2 had a 0.10 s longer $RT_{0.5\text{-}2kHz}$ while also being 81 m^3 smaller. This difference of $RT_{0.5\text{-}2kHz}$ 0.10 s between rooms 1 and 2 seems trivial, but Prodi and Visentin [65] reported that an increase of just $RT_{0.5\text{-}2kHz}$ 0.11 s from 0.57 s to 0.68 s resulted in a statistically significant effect on listening effort (reaction time) during speech perception and sentence comprehension tasks. Although only young L1 learners were considered in their study, it is entirely possible for adult L2 learners to perform similarly in adverse acoustic conditions (see section 1.2). Unfortunately, with $RT_{0.5\text{-}2kHz}$ 0.87 s and STI mostly below 0.66 for all three sound sources, this would indeed likely be the case in room 2.

3.3.3. Rooms 3, 4, and 5

Rooms 3, 4, and 5 were found to exhibit among the lowest RT and the highest STI across all three loudspeakers. In fact, they were the only rooms in which $\geq$ 0.66 STI was recorded in every receiver position (except one with 0.65 STI) for all three sound sources. These rooms are located in the same building and are practically identical in size, shape, furnishing, seating orientation, and

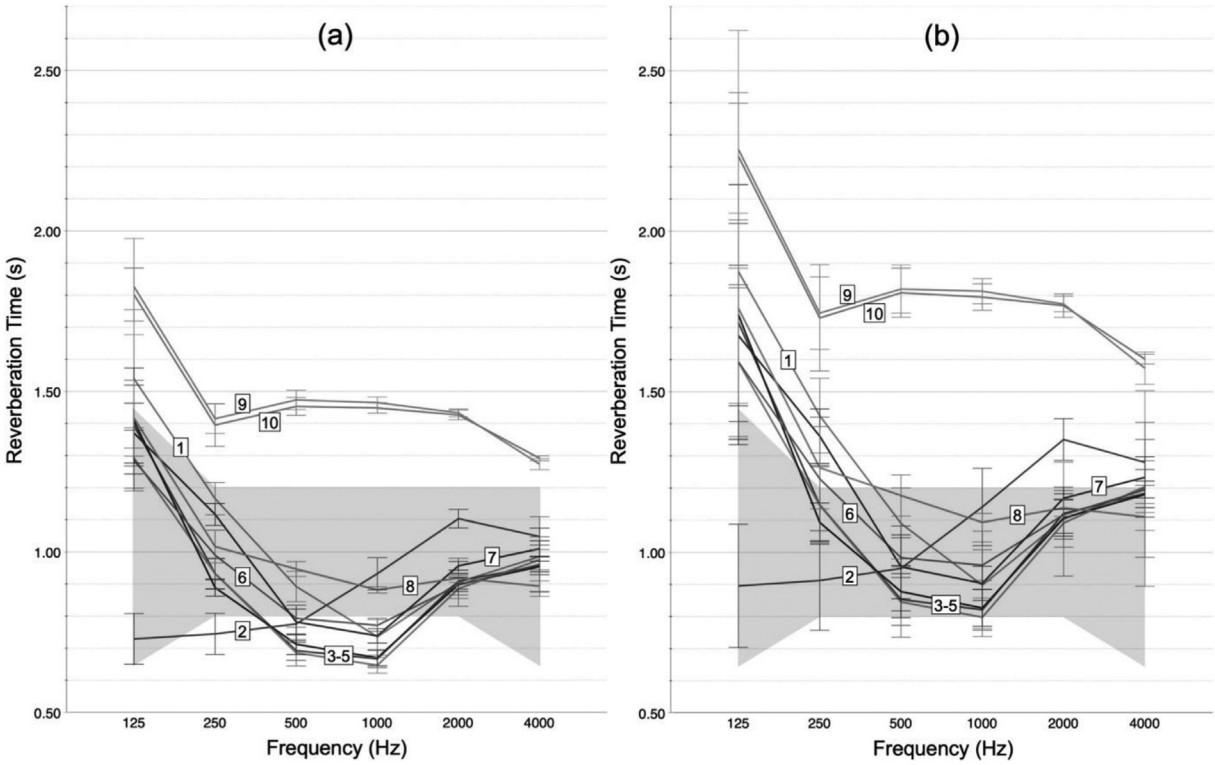

Fig. 6. RT measurement results converted to 50 % occupancy for usage types A3 (a) and A4 (b) in accordance with DIN 18041:2016. Optimal RT tolerance range as a function of frequency is represented by the grey shaded area. Numbers labeled 1–10 refer to their corresponding test rooms. Error bars represent the standard deviation for all measurements performed in each test room.

loudspeaker specifications. The best explanation for the RT and STI results in these rooms can be credited to the perforated, fluted wooden acoustic panels installed across the entire front and rear walls (excluding doors, blackboards, and electrical enclosures). It can be observed in Fig. 5 that $RT_{0.5\text{-}2kHz}$ in room 4 is noticeably lower than the other two. This may be attributable to the fact that the rear wall in room 4 is entirely covered with the acoustic panels except for two small electrical enclosures. In rooms 3 and 5, two sets of steel double doors are installed along the rear wall, and flat wooden panels surround the area around the doors from floor to ceiling. However, this seemed to have no detrimental effect on the STIPA measurements which resulted in $\geq$ 0.66 STI even in the vicinity of the steel double doors.

The benefits of such acoustic treatment, especially on the rear wall, are supported in the literature. In all combinations of the placement of sound absorptive materials deemed favorable in DIN 18041:2016 [42], installations are always present across the rear wall. Addressing the issue of background noise in classrooms, Smirnowa and Ossowski [66] advocated for a minimum of 30 % of the ceiling and rear wall to be covered with absorptive materials to meet the targets of 0.80 STI and RT 0.50 to 0.60 s. Sala and Viljanen [67] examined the effectiveness of 20 combinations of mineral wool panels and found that highest speech intelligibility measurement results were achieved in cases where the entire rear wall and at least 35 % of the ceiling were covered with the absorptive material. Finally, Minelli et al. [68] investigated several acoustic parameters of two similar primary school classrooms, one with and the other without acoustic treatment. While the untreated classroom had a $RT_{0.25\text{-}2kHz}$ of 1.40 s, the other classroom treated with glass fiber absorption panels on the left and rear walls and the ceiling showed a much-improved $RT_{0.25\text{-}2kHz}$ of 0.45 s.

Combined with the fact that even WMS with its decentralized configuration performed comparatively to the center loudspeaker sources of RCP and AMP, it appears that rooms 3, 4, and 5 may be exemplar cases in which both the design of the room as well as the installation of the PA system have been carried out upon careful consideration of their acoustic characteristics and relevant building standards.

3.3.4. Rooms 7 and 8

Rooms 7 and 8 present a puzzling case. Located in the same building, they have similar features. Both have a rectangular shape and feature a flat floor with wooden desks and seats bolted onto the floor. The entire rear wall is covered with perforated wooden acoustic panels with a blackboard attached at the center, which is similar to what was found in rooms 3, 4, and 5, except in this case the panels are not fluted and blackboards are installed on both the front and the rear walls. Nevertheless, there was a difference of nearly 1.0 s in $RT_{0.5\text{-}2kHz}$ between rooms 7 and 8, the longer RT belonging to the smaller room 8. The STIPA results were similarly distinct. Room 7 was the only flat, rectangular test room in which $\geq$ 0.66 STI was recorded in every receiver position (except two with 0.64 STI) for all three sound sources, while room 8 was the only test room in which WMS was found to exhibit the highest STI out of the three loudspeakers. Upon closer inspection, it can be observed from the RT boxplots (Fig. 5) that room 8 had the smallest interquartile range as well as one of the shortest whiskers in the entire data set. A similar observation can be made from the STI boxplots (Fig. 3) for WMS. This indicates that both RT and STIPA measurements for WMS were consistent throughout room 8. The bolded and underlined standard deviation values in Table 9 and Table 10 validate this point. The only surmisable explanation is that the particular physical features of room 8 better acoustically suited WMS over RCP and AMP.

One additional point is worth mentioning here. Although it was found that WMS performed the best out of the three sound sources

Table 10
Standard deviations (SD) of STI values by test room and sound source. Lowest value is bolded and underlined.

Room	WMS STI_{SD}	RCP STI_{SD}	AMP STI_{SD}
1	0.027	0.028	0.028
2	0.022	0.027	0.032
3	0.024	0.054	0.039
4	0.038	0.046	0.040
5	0.036	0.047	0.051
6	0.034	0.034	0.037
7	0.023	0.047	0.045
8	**0.012**	0.027	0.031
9	0.031	0.054	0.061
10	0.033	0.058	0.049

in room 8, the data analysis revealed no statistical significance between WMS and AMP. Also, AMP was only one STIPA measurement result shy of reaching the 0.66 STI (including the JND range) in every receiver position. Thus, perhaps an inexpensive solution such as a temporary use of absorptive materials or strategically assigning seats to avoid less favorable areas during standardized testing may be enough to justify allowing the use of AMP in room 8.

3.3.5. Rooms 9 and 10

Rooms 9 and 10 were particularly extreme cases with $RT_{0.5\text{-}2kHz}$ 1.27 s and 1.17 s, respectively. STI exceeded 0.66 in receiver positions only directly in front of the sound sources and the others reached as low as 0.50 STI, which according to IEC 60268–16:2020 is comparable to the intelligibility in shopping malls and public offices. Both rooms have a rectangular shape with a slightly low ceiling height of 2.71 m and are the two smallest rooms in this study with a volume of 333 m^3 and 301 m^3. Walls on all sides mostly consist of textured drywall with protruded concrete columns and pipe shafts. Whiteboards are installed on both front and rear walls, and windows are located on the right side. Further exploration revealed that, while the ceiling tiles in rooms 7 and 8 are composed of 12 mm rock wool with a sound absorption coefficient of 0.50, rooms 9 and 10 are fitted with 9.5 mm decorative gypsum boards with no sound absorptive qualities according to the product catalogue [69]. Such abundance of hard, flat surfaces was most likely responsible for the unusually high RT. Interestingly, the BNLs recorded in rooms 9 (36.8 dBA) and 10 (35.6 dBA), including the noise from HVAC systems, were not particularly detrimental as the standards ANSI S12.60, BB93, AIJES-S001-2020, and DIN 18041:2016 all prescribe $\leq$ 35 dB for learning spaces comparable to these rooms. In fact, DIN 18041:2016 notes that its requirement can be adjusted to $\leq$ 40 dB when speech is delivered through a sound system, and the Italian standard UNI 11532–2 specifically prescribes $\leq$ 38 dB as a reference value for continually operating equipment such as HVAC systems, both with which rooms 9 and 10 already comply. Nevertheless, this does not mitigate the fact that the least favorable STI and RT results in this study were observed in the two smallest and the most recently renovated test rooms. The rooms themselves are clean, well-furnished, and well-equipped. It is unfortunate that the acoustics of the room was not given equal consideration.

3.4. Implications for standardized tests

Some important implications concerning standardized EFL proficiency listening tests can be gleaned from the above findings. First, the investigation revealed large variations in both STI and RT across different sound sources and test rooms. These findings alone are not new as numerous studies have reported similar results [17,49,61,70–72]. However, considering that standardized EFL proficiency listening tests have been administered in these conditions for years, such inconsistency in acoustic environments may very well have potentially influenced the performance of past test takers. Indeed, there have been tens of thousands of university students to whom WMS in rooms 1 to 8 have been used to administer the TOEIC L&R IP Test over a span of more than 10 years. Similarly, while information regarding the specific type of loudspeaker utilized for the TOEIC L&R SP Test is not made public, anecdotal reports suggest that portable loudspeaker devices similar to RCP and AMP have also been used in rooms 1 to 8 several times a year over the past 20 years. According to university records, rooms 9 and 10 have likewise been in use since the completion of their renovation two years ago. With millions of standardized tests having been administered over the same time period [6], it would not be surprising for there to have been similar inconsistencies in other venues.

The results also underscore the difficulty involved in the assessment and selection of test rooms. This was notably apparent in room 2. On the surface, room 2 is a clean, well-equipped, typical university lecture hall with ample seating, a solid candidate for a test room. Nevertheless, it was found to exhibit inadequate STI across all three sound sources. A similar observation can be made for WMS in room 1. This stresses the point that acoustic features which may have detrimental effects on L2 speech intelligibility are difficult to detect, and given that L1 speech intelligibility in the same room is likely to be perceived as being satisfactory by normal hearing adults [19,21,24,28], it can easily be overlooked in the test room selection process.

As particularly relevant to the present investigation, two studies raised concerns regarding the challenging nature of administering standardized foreign language listening tests in non-standardized environments, suggesting that current conditions may in fact be promoting an unfair disadvantage to some learners. Sörqvist et al. [23] explored the effect of classroom reverberation on L2 (English) listening comprehension by native adult Swedish speakers. Citing previous studies reporting that a large percentage of classrooms in Sweden did not meet the RT recommendation of 0.8 s "in the low frequency range," they simulated three $RT_{125\text{-}8kHz}$ conditions of 0.26 s, 0.92 s, and 1.77 s in which a listening task from a standardized English test for high school students was administered. The corresponding STI for the three RT conditions were 0.87, 0.62, and 0.49, respectively. They found that the participants' performance decreased as the RT increased and that their baseline L2 proficiency (assessed with a reading comprehension task from the same standardized English test) was a stronger predictor of the effects of RT compared to working memory capacity.

A related study by Hurtig et al. [7] examined the effect of listening positions and reverberation on high school (15-year-old) students' performance on the same standardized English listening comprehension test in Sweden. They simulated two $RT_{125\text{-}4kHz}$ conditions of 0.33 s and 1.07 s and two distances of 1.05 m and 6.13 m from the sound source. The corresponding STI for the short and long distances for RT 0.33 s were 0.95 and 0.84, respectively, and those for RT 1.07 s were 0.71 and 0.62, respectively. They found that increases in both RT and distance negatively influenced the participants' performance, concluding that not only the room acoustics but also the seat position can have a detrimental effect on L2 listening comprehension. Similar findings were reported in a recent work by Puglisi et al. [53], though they only examined listening in L1, in which source-receiver distances of 1.5 m and 4 m in an acoustically compliant room (171 m^3, $RT_{0.25\text{-}2kHz}$ 0.4 s) and 1.5 m, 4 m, and 6.3 m in an acoustically non-compliant room (282 m^3, $RT_{0.25\text{-}2kHz}$ 3.1 s) all resulted in significantly worse speech intelligibility scores by adult listeners as the source-receiver distances increased. This effect was observed in the presence of both informational and energetic masking noise. These studies corroborate the measurement results obtained in the present study which revealed that STI can not only differ significantly but also span sev-

eral qualification categories (particularly above and below 0.66 STI) depending on the receiver location.

To the authors' knowledge, Sörqvist et al. [23] and Hurtig et al. [7] are the only studies that have covered the combined topics of room acoustics with respect to STI and RT, L2 listening comprehension, use of loudspeakers as sound sources, young adult subjects, and their implication on standardized EFL proficiency tests.

3.5. Recommendations for standardized foreign language listening tests

3.5.1. Sound source

This investigation revealed that utilizing identical loudspeakers as sound sources in every test room yielded more consistent STIPA results. It also found that STIPA results for AMP were statistically significantly higher in the largest number of test rooms compared to WMS and RCP. Moreover, given that AMP was able to produce well over 0.66 STI in all receiver positions in room 1, the largest test room in this study with a volume of 967 m^3, it can be deduced that the unfavorable STIPA results for AMP observed in rooms 2, 8, 9, and 10 were not due to power deficiency on the part of the sound source but were rather consequences of the acoustics and environmental factors particular to these rooms. Thus, we recommend a portable amplified loudspeaker comparable to the instrument employed in this study (UNI-PEX WA-361A) to be used as the sound source when administering standardized foreign language proficiency listening tests. Matching the height of the loudspeakers (perhaps 1.5 m or 1.7 m) as well as the approximate distances from the loudspeaker to the frontmost seats and to the front wall in every test room are also encouraged.

3.5.2. Speech intelligibility index

STI has been shown to be a reliable metric of speech intelligibility [61], and the amount of adjustment necessary for different proficiency levels of non-native listeners has been well documented (see section 1.3). However, these adjusted STI values are often unrealistic, even unachievable, in most environments. Therefore, taking into account the types of messages typically intelligible above (*unfamiliar words*) and below (*familiar words/contexts*) 0.66 STI according to IEC 60268–16:2020 [30], the standard STI value of $\geq$ 0.66 was offered as a general recommendation in section 1.3. The STIPA measurement results revealed that $\geq$ 0.66 can be attained at every receiver position in a wide range of test room sizes. AMP, the amplified loudspeaker recommended above, accomplished this in rooms 1, 3, 4, 5, 6, and 7, a difference of nearly 600 m^3. Thus, we believe that $\geq$ 0.66 STI is a reasonable and attainable recommendation for standardized EFL proficiency test rooms.

3.5.3. Background noise

The adverse effect of BNL on speech intelligibility for both L1 and L2 listeners is well established in the literature [12,22,26,27,28,33,35]. In fact, RT and BNL together constitute the biggest obstacle to achieving high speech intelligibility. Thus, in addition to prescribing RT recommendations, national standards and guidelines commonly list acceptable levels of BNL for different scenarios. For classrooms and educational spaces, $\leq$ 35 dB is generally prescribed [39–43]. This would ensure adequate speech comprehension by allowing a teacher's (or any talker's) typical speech level to exceed a signal-to-noise ratio (SNR) of +15 dB [73,74]. Naturally, these recommendations for educational spaces assume the teacher's speaking voice as the primary sound signal and the students' class activities such as their movements and chatter (along with HVAC systems and outside sources such as traffic or adjacent classrooms) as the main contributors of noise. In the case of standardized proficiency listening tests, however, loudspeakers are always employed, allowing continuous production of the sound signal at higher SPL, and test takers generally remain silent in order to focus on the listening task. Therefore, HVAC systems would likely constitute much of the noise in test environments.

In the present study, the air conditioner (AC) and ventilation were measured to account for background noise. The fan strength for AC (which accounts for most of the noise produced by the system) was set to the *middle* setting, while ventilation was simply switched on. This resulted in a mean BNL of 41.0 dB (SD = 2.8) (see Table 5). As the reference SPL determined to be representative of typical test conditions was established at L_{AS} 80 dB for all three sound sources in this study, it can be reasonably assumed that a SNR of at least +15 dB would be achievable in similar scenarios. Of course, in real test situations, there is no guarantee that these HVAC settings would be replicated since at the moment there seem to be no guidelines specifying such matters. Thus, we recommend selecting a combination of settings for the HVAC system (e.g., the lowest fan strength setting) that would generate the least amount of noise but would also maintain a sufficient level of thermal comfort to ensure that the test takers' health is not jeopardized. Since this procedure is only required during the actual listening test, conditioning the test rooms in advance as well as readjusting the settings afterwards, such as when the test transitions to the reading or writing section, are also advised.

3.5.4. Reverberation time

Optimal RT in primary and secondary school classrooms has been a focus of research for some time, but those concerning university classrooms and lecture halls are a relatively new development. Prodeus and Diskovska [75] called attention to this gap in the literature by postulating that, while the acoustics of small school classrooms and large concert halls have been studied extensively, those of university lecture rooms have been largely neglected due to their "intermediate" size. Though not explicitly an investigation on optimal RT, they examined three university lecture rooms labeled small (177 m^3), medium (270 m^3), and large (370 m^3) to analyze different methods of STI assessment. Leccese et al. [76], with the aim of formulating a predictive equation to estimate STI from RT, surveyed 11 university classrooms and also categorized them into three size groups: small (less than 350 m^3), medium (350 to 650 m^3), and large (more than 650 m^3). Nestoras and Dance [62] similarly categorized 10 university lecture rooms into small (less than 230 m^3), medium (230 to 350 m^3), and large (more than 350 m^3) to explore alternative methods of assessing room acoustics. Other studies involving a collection of university learning spaces include the investigation of L1 speech intelligibility of Korean lectures in 12 university classrooms (188 to 343 m^3) [72], effect of occupants on RT in 12 Korean university classrooms (193 to 2535 m^3) [77], L2 (English) speech intelligibility in 11 university classrooms (109 to 229 m^3) in Hong Kong [25], and L1 speech intelligibility in 17 auditoria and conference rooms (190 to 2000 m^3) at a university in Spain [61]. Although it is difficult to identify any sort of definitive pattern from the studies above, a slight adjustment of the size categories by Leccese et al. [76] seems most fitting to the present study. Thus, the test rooms investigated in this study can be categorized into small rooms less than 350 m^3 (rooms 9 and 10), medium rooms 350 to 700 m^3 (rooms 3 to 8), and large rooms 700 to 1050 m^3 (rooms 1 and 2).

Based on the size categories established here, recommendations for optimal RT shall be explored. A research review that has specifically examined the topic of optimal RT found 0.7 s to be applicable for students over the age of 12, although anywhere from 0.3 to 0.9 s is also reported to be adequate for a similar demographic [17]. Analyzing speech intelligibility test scores and acoustic measurements, Bradley [78] found RT_{1kHz} 0.4 to 0.5 s to be an optimal range for normal hearing L1 listeners, provided that BNL be around

30 dBA. Similarly, Bistafa and Bradley [44] solved for a range of reverberation times in classrooms with volumes of 100 m^3, 300 m^3, and 500 m^3 and found that RT_{1kHz} of 0.4 s would yield a STI equivalent of $\geq$ 0.75 in all three classroom sizes. For reference, none of the test rooms investigated in the present study exhibited RT_{1kHz} 0.4 s although a mean STI of $\geq$ 0.75 was achieved for AMP in rooms 3, 4, and 5 with RT_{1kHz} 0.58 s, 0.56 s, and 0.58 s, respectively. Hodgson and Nosal [79] calculated the optimal RT for unoccupied classrooms with volumes of 50 m^3, 100 m^3, 300 m^3, 500 m^3, 1000 m^3, and 4000 m^3, and, in the case of a speech-to-noise difference of 20 dB, presented optimal RT (and acceptable ranges) of 0.2 s (0–0.8 s), 0.2 s (0–0.8 s), 0.3 s (0–0.8 s), 0.4 s (0.1–0.8 s), 0.6 s (0.2–0.8 s), and 0.9 s (unachievable), respectively. Lastly, Bottalico and Astolfi [80] proposed a range of $RT_{0.5\text{-}2kHz}$ from 0.75 to 0.85 s; however, this was intended for talkers (teachers) in primary school classrooms and not adult L2 listeners.

The studies above present a broad spectrum of optimal RT values. While they generally fall under RT 1.0 s, the lower values around 0.0 to 0.4 s seem impractical considering that RT in educational spaces are routinely found to exceed national standards and guidelines (see section 1.4). Also, one detail that must be pointed out is that the studies reviewed above do not take L2 listeners into account. Furthermore, Prodi and Visentin [65] raised a valid point that the effect of optimal RT may vary depending on the complexity of the listening task. These factors suggest that optimal RT is a fluid concept rather than one that is fixed, and assigning one RT value to represent an array of environments used as standardized proficiency test rooms is not only challenging but may also result in an incomplete portrayal of the acoustic tendencies of each room. On the other hand, however, identifying and recommending precise optimal RT for every conceivable type of test room is not only beyond the scope of this study but would also not be generalizable outside of the rooms investigated here.

Thus, with this in mind, the following recommendations shall be presented for the proposed size categories based on the RT measurement results obtained in the present study. Starting with the medium rooms, the RT boxplots (Fig. 5) indicate that $RT_{0.5\text{-}2kHz} \leq$ 0.7 s seems appropriate as a general recommendation. Indeed, test rooms 3 to 7 had $RT_{0.5\text{-}2kHz}$ under 0.7 s and consistently recorded high STI values across all three sound sources, while room 8 exhibited strange acoustic patterns in terms of STI as discussed in section 3.3.4. Expanding on this idea, $RT_{0.5\text{-}2kHz} \leq$ 0.8 s seems applicable for the large rooms as room 1 with $RT_{0.5\text{-}2kHz}$ 0.77 s (SD = 0.046) yielded STI $\geq$ 0.66 in almost all receiver positions for RCP and AMP, while room 2 with $RT_{0.5\text{-}2kHz}$ 0.87 s (SD = 0.045) yielded only a handful of STI $\geq$ 0.66 for all three sound sources. Lastly, the excessive RT in the two small rooms make it difficult to formulate a recommendation, but Peng and Wang's [52] $RT_{0.5\text{-}2kHz} \leq$ 0.6 s for a 260 m^3 simulated classroom appears to be a suitable substitute in this size category.

4. Conclusions

This study has investigated the STIPA and RT of test rooms frequently used for standardized EFL proficiency tests at a university in Japan with particular interest in the differences between three types of sound sources (WMS, RCP, and AMP). Overall, AMP yielded the most favorable STIPA results in the highest number of test rooms, although RCP produced comparable results and performed well despite the power output difference between the two. Together, AMP and RCP, the two sound sources that were constant in all test rooms, yielded more consistent STIPA results compared to WMS which varied by test room. As for RT, test rooms with $RT_{0.5\text{-}2kHz} \leq$ 0.7 s recorded STI $\geq$ 0.66 in nearly all receiver positions for all three sound sources. The test rooms were assigned three size categories (small, medium, and large) and recommended $RT_{0.5\text{-}2kHz}$ ranges were proposed for each ($\leq$ 0.6 s, $\leq$ 0.7 s, and $\leq$ 0.8 s). These values may serve as a reliable criterion in the test room selection process.

It should be reiterated that the measurements for STIPA and RT in this study were conducted in unoccupied test rooms, and the results would likely be affected in the presence of test takers. Further research is needed to corroborate these findings before any generalizations about test rooms can be made. Also, as the results for WMS were specific to the test rooms selected for this study, loudspeakers installed in other venues may yield different results. A more thorough investigation of WMS in other settings could shed additional light on the efficacy of their application in standardized EFL proficiency tests. Additionally, the results for RCP were comparable to those for AMP in many respects despite their specification differences. Other radio cassette players with a higher output power should also be considered.

Finally, this study was not meant to cast negative light on current practices but rather to inform future decisions that would allow standardized EFL proficiency tests to be administered under more acoustically consistent environments. For example, instead of one large test room, the use of multiple medium to small test rooms may be better suited for this purpose. Replicating the methods from this study to measure STIPA and RT in every potential test room, however, would be unrealistic. Estimation of STI from RT and occupied RT from unoccupied RT data have been proposed in recent years [81–86], but more research and development are needed for these methods to become accessible to the test organizers. Instead, reviewing and, if necessary, updating the protocol for selection and operation of test rooms may be a good place to start. The recommendations presented in section 3.5 may be referenced for this purpose. For now, it would be in the best interest of the test takers for the organizers to gain greater awareness of these acoustic characteristics in test rooms, take STIPA, RT, or BNL measurements if possible, and utilize a consistent sound source across all test rooms.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Makito Kawata: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Mariko Tsuruta-Hamamura:** Methodology, Resources, Supervision, Project administration, Funding acquisition. **Hiroshi Hasegawa:** Methodology, Resources, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Kohei Sato, Kodai Motono, Komei Endo, Hinako Toda, Ayaka Kurita, and Haruki Akagi for their assistance in the measurement of test rooms. The authors are also indebted to the anonymous reviewers for their insightful comments and suggestions on earlier versions of this manuscript.

References

[1] Eberhard DM, Simons GF, Ethnologue FCD. Languages of the World. SIL Int. http://www.ethnologue.com; 2022.
[2] IELTS. IELTS Guide for Teachers 2019.
[3] Educational Testing Service. TOEIC Listening & Reading Test Examinee Handbook 2019.
[4] Educational Testing Service. TOEFL iBT Test and Score Data Summary 2020.
[5] Educational Testing Service. TOEIC Program Data & Analysis 2020.
[6] Educational Testing Service. TOEIC Program Data & Analysis 2021.
[7] Hurtig A, Sörqvist P, Ljung R, Hygge S, Rönnberg J. Student's second-language grade may depend on classroom listening position. PLoS One 2016;11:1–8. https://doi.org/10.1371/journal.pone.0156533.
[8] Mendell MJ, Heath GA. Do indoor pollutants and thermal conditions in schools influence student performance? A critical review of the literature. Indoor Air 2005;15:27–52. https://doi.org/10.1111/j.1600-0668.2004.00320.x.
[9] Dockrell JE, Shield BM. Acoustical barriers in classrooms: The impact of noise on performance in the classroom. Br Educ Res J 2006;32:509–25. https://doi.org/10.1080/01411920600635494.
[10] Ljung R, Sörqvist P, Kjellberg A, Green AM. Poor listening conditions impair memory for intelligible lectures: Implications for acoustic classroom standards. Build Acoust 2009;10:39–46. https://doi.org/10.1260/135101009789877031.
[11] Yang W, Bradley JS. Effects of room acoustics on the intelligibility of speech in classrooms for young children. J Acoust Soc Am 2009;125:922–33. https://doi.org/10.1121/1.3058900.
[12] Howard CS, Munro KJ, Plack CJ. Listening effort at signal-to-noise ratios that are typical of the school classroom. Int J Audiol 2010;49:928–32. https://doi.org/10.3109/14992027.2010.520036.
[13] Klatte M, Hellbrück J, Seidel J, Leistner P. Effects of classroom acoustics on performance and well-being in elementary school children: A field study. Environ Behav 2010;42:659–92. https://doi.org/10.1177/0013916509336813.
[14] Rabelo ATV, Santos JN, Oliveira RC, de Magalhães M. Effect of classroom acoustics on the speech intelligibility of students. CoDAS 2014;26:360–6. https://doi.org/10.1590/2317-1782/20142014026.
[15] Prodi N, Visentin C, Borella E, Mammarella IC, di Domenico A. Noise, age, and gender effects on speech intelligibility and sentence comprehension for 11- to 13-year-old children in real classrooms. Front Psychol 2019;10:1–18. https://doi.org/10.3389/fpsyg.2019.02166.
[16] Prodi N, Visentin C, Peretti A, Griguolo J, Bartolucci GB. Investigating listening effort in classrooms for 5- to 7-year-old children. Lang Speech Hear Serv Sch 2019;50:196–210. https://doi.org/10.1044/2018_LSHSS-18-0039.
[17] Minelli G, Puglisi GE, Astolfi A. Acoustical parameters for learning in classroom: A review. Build Environ 2022;208:108582. https://doi.org/10.1016/j.buildenv.2021.108582.
[18] van Wijngaarden SJ, Steeneken HJM, Houtgast T. Quantifying the intelligibility of speech in noise for non-native listeners. J Acoust Soc Am 2002;111:1906–16. https://doi.org/10.1121/1.1456928.
[19] Rogers CL, Lister JJ, Febo DM, Besing JM. Abrams HB. Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. Appl Psycholinguist 2006;27:465–85. https://doi.org/10.1017/S014271640606036X.
[20] Lecumberri MLG, Cooke M, Cutler A. Non-native speech perception in adverse conditions: A review. Speech Commun 2010;52:864–86. https://doi.org/10.1016/j.specom.2010.08.014.
[21] Tabri D, Chacra KMSA, Pring T. Speech perception in noise by monolingual, bilingual and trilingual listeners. Int J Lang Commun Disord 2011;46:411–22. https://doi.org/10.3109/13682822.2010.519372.
[22] Kilman L, Zekveld A, Hällgren M, Rönnberg J. The influence of non-native language proficiency on speech perception performance. Front Psychol 2014;5:1–9. https://doi.org/10.3389/fpsyg.2014.00651.
[23] 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;55:91–6. https://doi.org/10.1111/sjop.12115.
[24] 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. https://doi.org/10.1177/1351010X18758477.
[25] Yang D, Mak CM. An investigation of speech intelligibility for second language students in classrooms. Appl Acoust 2018;134:54–9. https://doi.org/10.1016/j.apacoust.2018.01.003.
[26] 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. https://doi.org/10.1016/j.specom.2019.03.001.
[27] Peng ZE, Wang LM. 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;62:1068–81. https://doi.org/10.1044/2018_JSLHR-H-17-0423.
[28] 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;26:275–91. https://doi.org/10.1177/1351010X19882314.
[29] Steeneken HJM, Houtgast T. A physical method for measuring speech-transmission quality. J Acoust Soc Am 1980;67:318–26. https://doi.org/10.1121/1.384464.
[30] IEC 60268-16:2020 Edition 5.0 Sound system equipment—Part 16: Objective rating of speech intelligibility by speech transmission index; 2020.
[31] van Wijngaarden SJ, Bronkhorst AW, Houtgast T, Steeneken HJM. Using the Speech Transmission Index for predicting non-native speech intelligibility. J Acoust Soc Am 2004;115:1281–91. https://doi.org/10.1121/1.1647145.
[32] Florentine M. Speech perception in noise by fluent, non-native listeners. J Acoust Soc Am 1985;77:S106. https://doi.org/10.1121/1.2022152.
[33] Mayo LH, Florentine M, Buus S. Age of second-language acquisition and perception of speech in noise. J Speech Lang Hear Res 1997;40:686–93. https://doi.org/10.1044/jslhr.4003.686.
[34] Hornsby BWY. The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. Ear Hear 2013;34:523–34. https://doi.org/10.1097/AUD.0b013e31828003d8.
[35] Picou EM, Gordon J, Ricketts TA. The effects of noise and reverberation on listening effort for adults with normal hearing. Ear Hear 2016;37:1–13. https://doi.org/10.1097/AUD.0000000000000222.
[36] Borghini G, Hazan V. Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. Front Neurosci 2018;12:1–13. https://doi.org/10.3389/fnins.2018.00152.
[37] Peelle JE. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. Ear Hear 2018;39:204–14. https://doi.org/10.1097/AUD.0000000000000494.
[38] 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 Acoustics 2016: Second Australas Acoust Soc Conf, Brisbane, Australia; 2016.
[39] American National Standards Institute. ANSI S12.60: Acoustical performance criteria, design requirements, and guidelines for schools. Part 1: Permanent schools. 2010.
[40] Building Bulletin 93, Acoustic design of schools—A design guide. London, 2003.
[41] AIJES-S0001-2020: Standard and Design Guidelines for Sound Environment in School Buildings. 2020.
[42] DIN 18041:2016-03 Acoustic quality in rooms—Specifications and instructions for the room acoustic design. Berlin, Germany; 2016.
[43] Astolfi A, Parati L, D'orazio D, Garai M. The new Italian standard UNI 11532 on acoustics for schools. In: Proceedings of the 23rd international congress on acoustics; 2019.
[44] Bistafa SR, Bradley JS. Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics. J Acoust Soc Am 2000;107:861–75. https://doi.org/10.1121/1.428268.
[45] Crandell CC, Smaldino JJ. Classroom acoustics for children with normal hearing and with hearing impairment. Lang Speech Hear Serv Sch 2000;31:362–70. https://doi.org/10.1044/0161-1461.3104.362.
[46] Knecht HA, Nelson PB, Whitelaw GM, Feth LL. Background noise levels and reverberation times in unoccupied classrooms: Predictions and measurements. Am J Audiol 2002;11:65–71. https://doi.org/10.1044/1059-0889(2002/009).
[47] Nelson P, Kohnert K, Sabur S, Shaw D. Classroom noise and children learning through a second language: Double jeopardy? Lang Speech Hear Serv Sch 2005;36:219–29. https://doi.org/10.1044/0161-1461(2005/022).
[48] Larsen JB, Vega A, Ribera JE. The effect of room acoustics and sound-field amplification on word recognition performance in young adult listeners in suboptimal listening conditions. Am J Audiol 2008;17:50–9. https://doi.org/10.1044/1059-0889(2008/006).
[49] Mikulski W, Radosz J. Acoustics of classrooms in primary schools – Results of the reverberation time and the speech transmission index assessments in selected buildings. Arch Acoust 2011;36:777–93. https://doi.org/10.2478/V10168-011-0052-6.
[50] Brill LC, Wang LM. Building a sound future for students: Considering the acoustics in occupied active classrooms. Acoust 2018;14:14–22.
[51] Wilson WJ, Downing C, Perrykkad K, Armstrong R, Arnott WL, Ashburner J, et al. The 'acoustic health' of primary school classrooms in Brisbane. Australia Speech Lang Hear 2020;23:189–96. https://doi.org/10.1080/2050571X.2019.1637042.
[52] Peng Z, Wang LM. Effects of noise, reverberation and foreign accent on native and non-native listeners' performance of English speech comprehension. J Acoust Soc Am 2016:139. https://doi.org/10.1121/1.4800071.
[53] Puglisi GE, Warzybok A, Astolfi A, Kollmeier B. Effect of reverberation and noise type on speech intelligibility in real complex acoustic scenarios. Build Environ 2021:204. https://doi.org/10.1016/j.buildenv.2021.108137.
[54] ISO 3382-2:2008(E). Acoustics—Measurement of room acoustic parameters—Part 2: Reverberation time in ordinary rooms; 2008.
[55] Bradley JS, Reich R, Norcross SG. A just noticeable difference in C 50 for speech. Appl Acoust 1999;58:99–108. https://doi.org/10.1016/S0003-682X(98)00075-9.
[56] Millett P. Sound field amplification research summary. https://simeoncanada.com/wp-content/uploads/2018/12/Soundfield-Amplification-Research-Summary.pdf; 2015.
[57] Lafargue C, Lafargue A. Sound field systems in New Brunswick classrooms: Let's enhance their use! Antistasis 2012. https://journals.lib.unb.ca/index.php/antistasis/article/view/19591/21123.
[58] Shirley B, Kendrick P, Churchill C. The effect of stereo crosstalk on intelligibility: Comparison of a phantom stereo image and a central loudspeaker source. J Audio Eng Soc 2007;55:852–63. https://www.aes.org/e-lib/browse.cfm?elib=14174.

[59] Vickers E. Fixing the phantom center: Diffusing acoustical crosstalk. In: Proceedings of the 127th Convention of Audio Engineering Society, New York; 2009.
[60] ISO 3382-1. Acoustics—Measurement of room acoustic parameters—Part 1: Performance spaces. 2009.
[61] Escobar VG, Morillas JMB. Analysis of intelligibility and reverberation time recommendations in educational rooms. Appl Acoust 2015;96:1–10. https://doi.org/10.1016/j.apacoust.2015.03.001.
[62] Nestoras C, Dance S. The interrelationship between room acoustics parameters as measured in university classrooms using four source configurations. Build Acoust 2013:43–54. https://doi.org/10.1260/1351-010X.20.1.43.
[63] Ricciardi P, Buratti C. Environmental quality of university classrooms: Subjective and objective evaluation of the thermal, acoustic, and lighting comfort conditions. Build Environ 2018;127:23–36. https://doi.org/10.1016/j.buildenv.2017.10.030.
[64] Fratoni G, D'Orazio D, de Salvio D, Garai M. Predicting speech intelligibility in university classrooms using geometrical acoustic simulations. In: Proceedings of the 16th IBPSA Conference, Rome, Italy; 2019.
[65] Prodi N, Visentin C. A slight increase in reverberation time in the classroom affects performance and behavioral listening effort. Ear Hear 2021;43:460–76. https://doi.org/10.1097/AUD.0000000000001110.
[66] Smirnowa J, Ossowski A. A method for optimising absorptive configurations in classrooms. Acta Acust United Acust 2005;91:103–9.
[67] Sala E, Viljanen V. Improvement of acoustic conditions for communication in classrooms. Appl Acoust 1995;45:81–91. https://doi.org/10.1016/0003-682X(94)00035-T.
[68] Minelli G, Puglisi GE, Astolfi A, Hauth C, Warzybok A. Binaural speech intelligibility in a real elementary classroom. J Phys Conf Ser 2021;2069. https://doi.org/10.1088/1742-6596/2069/1/012165.
[69] Yoshino Gypsum Co., Ltd. Tiger Gyptone Light: Direct-overlay decorative ceiling board. https://yoshino-gypsum.com/en/prdt/Tiger%20Gyptone%20Light; 2020.
[70] Zannin PHT, Zwirtes DPZ. Evaluation of the acoustic performance of classrooms in public schools. Appl Acoust 2009;70:626–35. https://doi.org/10.1016/j.apacoust.2008.06.007.
[71] Schmidtke J. The bilingual disadvantage in speech understanding in noise is likely a frequency effect related to reduced language exposure. Front Psychol 2016:7. https://doi.org/10.3389/fpsyg.2016.00678.
[72] Choi JY. The intelligibility of speech in university classrooms during lectures. Appl Acoust 2020;162:107211. https://doi.org/10.1016/j.apacoust.2020. 107211.
[73] Bradley JS. Optimising sound quality for classrooms. In: Proceedings of XX Meeting of SOBRAC (Brazilian Acoustical Association), Rio de Janeiro, Brazil; 2002.
[74] American Speech-Language-Hearing Association. Acoustics in educational settings: Technical report. www.asha.org/policy; 2005.
[75] 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 2021:47–56. https://doi.org/10.15587/1729-4061.2021.228405.
[76] Leccese F, Rocca M, Salvadori G. Fast estimation of Speech Transmission Index using the Reverberation Time: Comparison between predictive equations for educational rooms of different sizes. Appl Acoust 2018;140:143–9. https://doi.org/10.1016/j.apacoust.2018.05.019.
[77] Choi YJ. Effect of occupancy on acoustical conditions in university classrooms. Appl Acoust 2016;114:36–43. https://doi.org/10.1016/j.apacoust.2016.07.010.
[78] Bradley JS. Speech intelligibility studies in classrooms. J Acoust Soc Am 1986;80:846–54. https://doi.org/10.1121/1.393908.
[79] Hodgson M, Nosal E. Effect of noise and occupancy on optimal reverberation times. J Acoust Soc Am 2002;111:931–9. https://doi.org/10.1121/1.1428264.
[80] Bottalico P, Astolfi A. Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms. J Acoust Soc Am 2012;131:2817–27. https://doi.org/10.1121/1.3689549.
[81] Galbrun L, Kitapci K. Accuracy of speech transmission index predictions based on the reverberation time and signal-to-noise ratio. Appl Acoust 2014;81:1–14. https://doi.org/10.1016/j.apacoust.2014.02.001.
[82] Nowoświat A, Olechowska M. Fast estimation of speech transmission index using the reverberation time. Appl Acoust 2016;102:55–61. https://doi.org/10.1016/j.apacoust.2015.09.001.
[83] Choi YJ. Predicting classroom acoustical parameters for occupied conditions from unoccupied data. Appl Acoust 2017;127:89–94. https://doi.org/10.1016/j.apacoust.2017.05.036.
[84] 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;47:277–83. https://doi.org/10.1007/s40857-019-00166-1.
[85] Feng Y, Chen F. Nonintrusive objective measurement of speech intelligibility: A review of methodology. Biomed Signal Process Control 2022;71:103204. https://doi.org/10.1016/j.bspc.2021.103204.
[86] Duangpummet S, Karnjana J, Kongprawechnon W, Unoki M. Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response. Appl Acoust 2022;185:108372. https://doi.org/10.1016/j.apacoust.2021.108372.