Building and Environment xxx (2018) xxx-xxx

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com

Using listening effort assessment in the acoustical design of rooms for speech

Chiara Visentin[a, *], Nicola Prodi[a], Francesca Cappelletti[b], Simone Torresin[c], Andrea Gasparella[c]

[a] University of Ferrara, Department of Engineering, Ferrara, Italy
[b] University Iuav of Venezia, Department of Design and Planning in Complex Environments, Venice, Italy
[c] Free University of Bozen-Bolzano, Faculty of Science and Technology, Bozen-Bolzano, Italy

ARTICLE INFO

Keywords:
Room acoustics
Speech reception
Speech intelligibility
Response time
Listening effort
Auralization

ABSTRACT

This study addresses the issue of an enhanced acoustical design of rooms for speech, which besides targeting high speech intelligibility also ensures minimal effort in speech reception. Speech-in-noise tests in the Italian language were proposed to normal-hearing young adults, both *in situ*, within an existing university classroom, and via headphones, using auralized signals obtained from acoustic simulations of the same environment. Later, auralization was used to investigate the effect of realistic modifications to the room acoustics (acoustical treatment of a wall, change of the room size) by altering the virtual model of the classroom. The speech reception performance was characterized by using both the number of words correctly recognized (speech intelligibility, IS) and two estimates of listening effort: the behavioral measure of response time (RT) and a subjective judgement on a rating scale (LE). Firstly, the correspondence between the IS, RT and LE results *in situ* and in auralized conditions was considered and discussed. Then, the effectiveness of the three metrics in outlining the effect of the acoustic changes of the room was analyzed. The results showed that there were no differences between the compared acoustic conditions in terms of IS. The effects of the characteristics of the room acoustics were instead discriminated when RT and LE were considered, with the greatest number of significant differences observed by using RT. Using RT therefore seems to be an effective and promising strategy to better discern the effects of the room acoustics and to enhance the acoustical design of rooms for speech.

1. Introduction

In the planning stage of the acoustical design of rooms for speech (e.g. classrooms, conference rooms, theatres, etc.) the targets can be assigned either in terms of objective acoustical parameters, as for instance reverberation time [1] and Speech Transmission Index (STI) [2], or in terms of speech reception performance via the percentage of correctly recognized words (i.e. speech intelligibility score, IS) [3]. Intelligibility and the former objective metrics are correlated, and their relationship depends on the type of speech material used for the speech-in-noise tests (logatoms, words in isolation, sentences, etc.) [4]. Compliance to such specifications ensures the correct reception of an appropriate amount of utterances, but does not warrant per se a comfortable speech communication. In fact, even when a good or near ceiling IS is obtained, listeners may experience exertion due to the presence of reverberation [5] and noise [6]. In these cases, the goal of speech reception becomes harder and listeners are required to allocate further cognitive resources to cope with the increased task demands. This process, termed "listening effort" [7], does not just mirror the changes of speech reception accuracy but may also vary independently as it happens in the most favorable listening conditions, when performance accuracy is maintained at the expenses of a more explicit cognitive processing [8]. Owing to the limited availability of personal cognitive capacity [9], a practical consequence is that when increased resources are allocated to word reception, less capacity will be available for higher level processing of speech (e.g. recall of information, understanding of instructions, extraction of discourse meaning, etc.). So, when high levels of effort have to be sustained for long periods (e.g. during lessons), fatigue may arise with negative consequences on learning and cognitive achievements of listeners [10]. Within this context, several categories of listeners are more vulnerable as is the case for non-native listeners, due to their lower language proficiency [11,12].

* Corresponding author.
Email addresses: chiara.visentin@unife.it (C. Visentin); nicola.prodi@unife.it (N. Prodi); francesca.cappelletti@iuav.it (F. Cappelletti); simone.torresin@natec.unibz.it (S. Torresin); andrea.gasparella@unibz.it (A. Gasparella)

https://doi.org/10.1016/j.buildenv.2018.03.020
Received 5 January 2018; Received in revised form 16 March 2018; Accepted 16 March 2018
Available online xxx
0360-1323/

Traditional speech-in-noise tests are rather insensitive to the deployment of processing demands [5], and it follows that an acoustical design based only on speech intelligibility results overlooks information on listening effort that are instead of paramount importance to ensure optimal functionality inside spaces devoted to speech communication. A detailed framework for the understanding of effortful listening was recently elaborated, with a thorough discussion of the many factors that affect this construct [7], pointing out that to date no single measure is available to capture the multifaceted experience of effortful listening. Over the years, several methods have been proposed for the scope, the peculiarities and limits of which are also discussed in the relevant literature, stemming in prevalence from studies on hearing impairment [7,10,13,14]. One family of indices is physiological; they track unconscious reactions of the nervous system, for instance pupil dilation, skin conductance, heart rate and cortisol levels. In some cases, such as pupillometry, they are considered promising [15] but most often cannot be implemented outside a single-user setup in the laboratory or in clinical settings, making measurements in more ecological contexts, such as rooms in working conditions, impracticable. As alternatives, cognitive-behavioral measures and subjective ratings have been proposed to collect information on listening effort. In particular, the subjective rating of "listening difficulty" has been firstly introduced for normal-hearing adult listeners [16] and for children [17], which consists in reporting a personal impression on a categorical scale with opposite anchors. Variations of "listening difficulty" effectively traced a perceived worsening of the reception conditions whilst speech intelligibility had limited variations or was near ceiling [18]. The same effect was monitored in Ref. [19] highlighting that the STI could be a rough predictor of the subjectively rated listening effort, given the correlation of the latter with STI for most of the tested conditions. However, subjective ratings are built upon conscious expressions that are difficult to generalize due to possible individual biases in the scaling adopted [20], caused by choice and interpretation of anchors and of instructions that do not coincide. This was the case for instance in Ref. [17] where participants of different ages interpreted the scale differently. In order to overcome these inherent limitations of self-rated indices, cognitive-behavioral metrics have been used, primarily to monitor the engagement of the cognitive resources underpinning speech reception. This type of metrics can be typically implemented within two alternative paradigms. On the one hand, dual-task experiments are developed where the auditory task is paired with a secondary task of which the variation is the indirect measure of listening effort (for a detailed review see Ref. [21]); on the other, single-task auditory experiments are implemented of which the accuracy results are paired with a suitable additional measure sensitive to the cognitive load. A candidate measure to be collected in single-task paradigms is the response time (RT) to the auditory stimulus, which proved to reflect the amount of resources required to interpret and respond to the incoming signal [22–24]. RT is thought to trace the processing load and thus to be informative on the related effort. Since it carries complementary information to the intelligibility scores, it could be used to improve the means of evaluation of rooms for speech [10]. This quantity was used for instance in monitoring the decrease of speech reception performance resulting from an effortful listening due to prolonged exposure to different types of noises during a 1-h lesson period [25,26].

The present study specifically addresses the issue of an improved acoustical design and evaluation of rooms for speech, based on both performance accuracy (traced by IS) and on feasible estimates of the listening effort; in the latter case, they are achieved by means of both subjective ratings (LE) and the behavioral quantity response time (RT). For the scope, two main issues have been considered.

First, the correspondence between the employed metrics acquired from *in situ* and auralized speech-in-noise tests has been investigated. A university classroom was chosen as a case study, being a room typology for which good environmental comfort was demonstrated to greatly influence the learning capacity of students [27–29]. Auralization techniques based on calibrated acoustical simulations were used to playback sound fields via headphones in a laboratory setting, after the same listening conditions were presented ecologically in the real classroom. Several studies have addressed the issue of the validity of acoustical simulations, finally showing that, once the virtual models are carefully calibrated upon measures, the auralized sound field can almost be equivalent to the real one as concerns acoustical perceptual attributes [30]. A comparison of speech intelligibility data in the framework of auralization was also performed. Literature results showed that a good agreement between real and virtual data can be obtained [31] but consistency in speech intelligibility results seems to decrease for shorter reverberation times and too noisy sound fields [32,33]. Furthermore, the details of the head related transfer functions (HRTF) employed in the simulations might influence the results to a notable extent [34]. Despite several specific studies on speech intelligibility and auralization, the ecological validation of the RT and LE metrics, that is a proof of correspondence between the values retrieved under natural and synthetized conditions, is still lacking. Therefore, this first task of the study is necessary to understand how well the auralization techniques are capable of mimicking everyday realistic communication conditions and is preliminary to further virtual investigations on the effect of changes to the room acoustics on the same metrics.

Second, a proper virtual acoustics design scenario was developed by the simultaneous alteration of the listener's location and room properties (geometry and materials) inside the simulated classroom. The simulated changes to the room acoustics were especially chosen to represent realistic interventions that could be implemented in classrooms for optimization purposes. The auralized outputs were used to prepare comparative speech-in-noise tests of which the intelligibility scores, response times and subjective ratings were collected. By doing so the present study aims at gaining new insights on how the sound in the built environment influences occupants' performance, directly identifying if (and how) changes in the room acoustics affect the outcomes in a speech reception task. As recently pointed out in Ref. [35], to date only a few studies have addressed this topic by investigating on experimental conditions that result from actual acoustic interventions. In the relevant studies reviewed in Ref. [35], speech intelligibility was taken as a direct quantifier of the effects of changes to the room acoustics. The information on the deployment of cognitive resources required during the task would then positively complement the accuracy data, adding valuable knowledge for the design of acoustic environments that best meets the occupants' needs.

Finally, the current study addresses the relationship between different metrics used as proxies of listening effort, which to date is still unclear: in fact, different measures can yield different results [36], as supposedly reflecting underlying constructs that do not entirely match. The results of the experiments will help in identifying to what extent the consideration of the listening effort can enhance the acoustical design of rooms for speech and which of the two viable metrics here selected (RT and LE) is the most suitable for the scope. All experiments in the study involved two groups of participants with different mother tongues, in order to investigate if the method can trace differences in speech reception between native and non-native listeners.

2. Methods

2.1. Participants

Twenty-one young adults participated in the current study, all of them self-reporting normal hearing. The participants were recruited by word of mouth among the students and the academic staff of the Free University of Bozen-Bolzano. They were all Italian citizens born in the

bilingual context of South Tyrol (where the University of Bozen-Bolzano is located) and thus living since birth in an Italian/German speaking environment. Based on their self-declared mother tongue, the participants were divided into two groups: 10 native Italian speakers (5 female, 5 male; mean age: 24.4 years, σ: 1.7 yr) and 11 native German speakers (6 female, 5 male; mean age: 25.9 years, σ: 7.9 yr). In the following, the groups will be named NI and NG respectively.

All NG participants started the acquisition of Italian as a second language before the age of eight, and used the Italian language either for their university studies (with many of their courses being in Italian) or for daily communication. Prior to the experiment, NG participants were asked to self-rate their proficiency in listening of the Italian language on a 7-point category scale, with the highest extreme value labeled as "mother tongue". The resulting median rating was 5.0 (interquartile range: 1.25).

2.2. Speech material

The Diagnostic Rhyme Test (DRT) [37] in the Italian language was used for the experiment. The DRT is a consonant confusion test, which bases on a target word embedded in a carrier phrase (e.g. "La prossima parola che diremo è *riso*", which is Italian for "The next word is *rice*"). The target item is drawn from a corpus of 105 rhyming pairs, all of them meaningful, disyllabic words. Within each pair, the distinctive feature of the initial consonant varies, still keeping the consonant-vowel transition (e.g. /ˈriso/and/ˈlizo/). The speech material is optimized as regards phonemic distribution of the Italian language and word familiarity.

The test sequences were recorded by an adult, native Italian, female speaker; she was instructed to speak at conversational rate, maintaining a natural prosody and avoiding any emphasis on the final, target word. The recordings took place in a silent room, at a sampling frequency of 44.1 kHz. All of the sequences were filtered as to match the long-term spectrum of a female speaker indicated by the IEC60268-16 standard [2], and set to the same root mean square value. The recordings were then organized into five lists of 18 words each. The remaining 15 words were organized into a shorter list, to be used in the initial training phase.

2.3. Listening tests in the existing classroom

2.3.1. Outline of the classroom and measurement set up

Speech-in-noise tests were conducted in a university classroom, part of the Classroom Spaces Living Lab of the Free University of Bozen-Bolzano. The room is box-shaped, with floor dimensions 7.29 m × 7.62 m and 3.55 m high, resulting in a volume of 197 m^3. It is characterized by flat surfaces (ceiling: unpainted concrete, floor: linoleum finishing, walls: painted plasterboard); the lateral partition with the adjacent corridor is acoustically treated with Topakustik® 6/2 sound absorbing paneling. The classroom is furnished with wooden desks and chairs and is designed for a maximum of 25 students.

For the experiment, the room was set up as shown in Fig. 1. A B&K type 4720 artificial mouth was placed close to the desk, at a height of 1.5 m, and oriented towards the audience; it was used to deliver the speech signal. Interfering background noise was played back with a B&K type 4292-L omnidirectional source located on the floor, exactly below the speech source. Two measurement positions (R1, R2) were defined within the room, located respectively at 2.5 m and 5.5 m from the loudspeakers. Two omnidirectional, B&K type 4189 1/2 inch microphones were positioned at a height of 1.25 m and used for the objective description of the listening conditions. Also binaural impulse responses were collected *in situ* by means of two head and torso simulators B&K type 4100 placed at R1 and R2 respectively, with ears at 1.15 m in height from the floor.

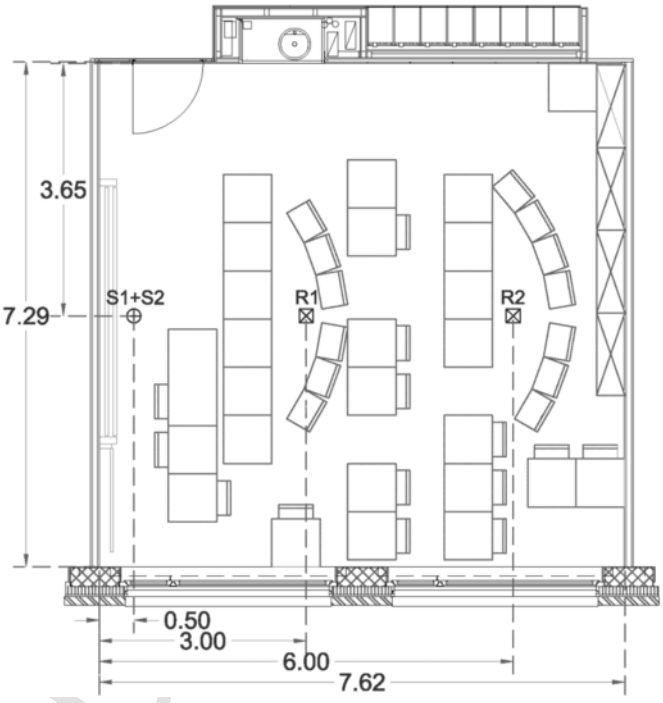


Fig. 1. Classroom plan: relevant dimensions [m], position of the speech and noise sources (S1 and S2), position of the measurements points (R1 and R2). Locations of the participants close to the measurement points are depicted by chair occupancy.

The measurement setup included two B&K type 5935 signal conditioners, a B&K type 4231 calibrator, an RME Fireface© UC full-duplex sound card and a laptop. Test signals and acoustics elaborations were managed by means of Aurora suite in the Adobe Audition® package.

Reverberation time (T_{30}) and speech clarity (C_{50}) values at the two listening positions were derived from monaural impulse responses [38] measured with the swept-sine technique during the experiment (occupancy of the classroom: 50%). The mid-frequency values (average of 500–2000 Hz octave bands) were 0.82 and 0.84 s (T_{30}), and 3.6 and 0.4 dB (C_{50}) respectively at R1 and R2.

2.3.2. Listening conditions and test procedures

For the listening tests, the speech level was fixed at 63 dB(A) at 1 m in front of the source, corresponding to a vocal effort of a speaker between "normal" and "raised" [4]. The resulting levels at R1 and R2 were 61.0 and 57.4 dB(A) respectively. A stationary noise with the same long-term spectrum of speech was played back to partially mask the speech signal. Its level was set so as to achieve the same level of speech as in R1 (signal-to-noise ratio SNR equal to 0 dB); the resulting noise level in R2 was 58.7 dB(A). The choice of the SNR was intended to mimic conditions that may arise during group work, for instance in laboratory assignments or, most often, in the context of open plan group work [28,29]. A comprehensive description of the tested listening conditions within the classrooms was obtained by calculating the Speech Transmission Index (STI), which describes the combined effect of background noise and reverberation on the transmission quality of the speech signal. The STI values were 0.52 in R1 and 0.46 in R2, corresponding to an intelligibility rated as "Fair" [4].

A touchscreen handset was given to each participant, to be used for response selection by means of a soft pen. A wireless test bench was used to manage the experiment [25]; the server application running on a laptop simultaneously controlled the audio rendering, the presenta-

tion of the alternatives on the touchscreen handsets and the collection of words choices and response times. During the experiment, the participants sat around the two receiver positions (Fig. 1). They listened to a target word embedded in the carrier phrase and then selected one of the three options (the rhyming pair and the "*none of the two*" alternative) that were displayed on the touchscreen after the audio playback offset. A training session was firstly proposed, with the aim of familiarizing the participants with the test procedure. During the experiment, the participants completed one test list of 18 words in each position. They were instructed to pay attention, and asked to respond as accurately as possible but they were not urged to provide the quickest possible response. They were also informed that the word played back was always one of the two rhymed alternatives. After each list, the participants were asked to rate their perceived listening effort (LE), answering to the following question: "*How much effort did it take to hear and understand the words?*" The responses were given on a 10-points scale, ranging from *minimum effort* (1) to *maximum effort* (10), which appeared on the handset touchscreen after the last pair of words from each list. The experiment was presented separately to NI and NG listeners, in two subsequent sessions. The same test lists were presented to the two groups of participants.

The data retrieved in the experiments were word scores (correct/incorrect/none of the two), response times defined as the time elapsed between the end of the audio playback and the item selection on the touchscreen, and subjective ratings of listening effort.

2.4. Listening tests in auralized conditions

2.4.1. Set up of the virtual classrooms

A virtual model of the existing classroom was created using the room acoustics software Odeon® v14.01.

Firstly, a geometric model made of 261 surfaces was created in SketchUp® and then imported in the acoustic CAD software. A view of the model is reported in Fig. 2. The geometric model included, besides boundary surfaces, also wooden desks, chairs and all the furnishing elements of the classroom that could be relevant for the acoustic simulation (e.g. lighting fixtures, radiators, shelves). Initial absorption coefficients were assigned to surfaces and objects based on the Odeon® material library and on data available from literature. A mid-frequency scattering coefficient of 0.05 was assigned to all boundary surfaces; for desks and chairs, based on the scale-model measurements of desks and chairs in a row reported in Ref. [39], a value of 0.5 was chosen. A speech source with the directivity pattern of a human talker (*Tlknorm* in Odeon®) and emitting a signal spectrally shaped to match a female talker [2] was defined for the calculation of the room acoustics parameters. The virtual source was located at the same position as in the existing classroom.

A preliminary calibration of the virtual model in unoccupied conditions was initially performed. To the scope, six receiver positions were defined in the audience (Fig. 2) and measures in the real classroom with omnidirectional, B&K type 4189 1/2 inch microphones were achieved at the same locations (height of receivers: 1.25 m) with the speech source in the same position as during the *in situ* experiment. Simulations were performed with a transition order of two, with 2000 early rays and 16000 late rays. During the calibration process, the acoustical material properties were step-by-step adjusted, still keeping physically realistic values, until the differences between measured and simulated values of the selected acoustical parameters were smaller than the Just Noticeable Differences (JND) defined by the ISO 3382-1 standard [38]. In accordance with the literature on the calibration of virtual acoustic models of small classrooms [40], reverberation time (T_{30}, EDT) and speech clarity (C_{50}) were selected as relevant indicators. In Table 1, measured and simulated acoustic parameters (mid-frequency values, averaged over the 500–2000 Hz octave bands) are reported for the six receiver positions; their differences are also indicated. It is relevant that for all positions and acoustic parameters the difference between measured and simulated values was smaller than the corresponding JND (5% for reverberation time, 1 dB for clarity). A

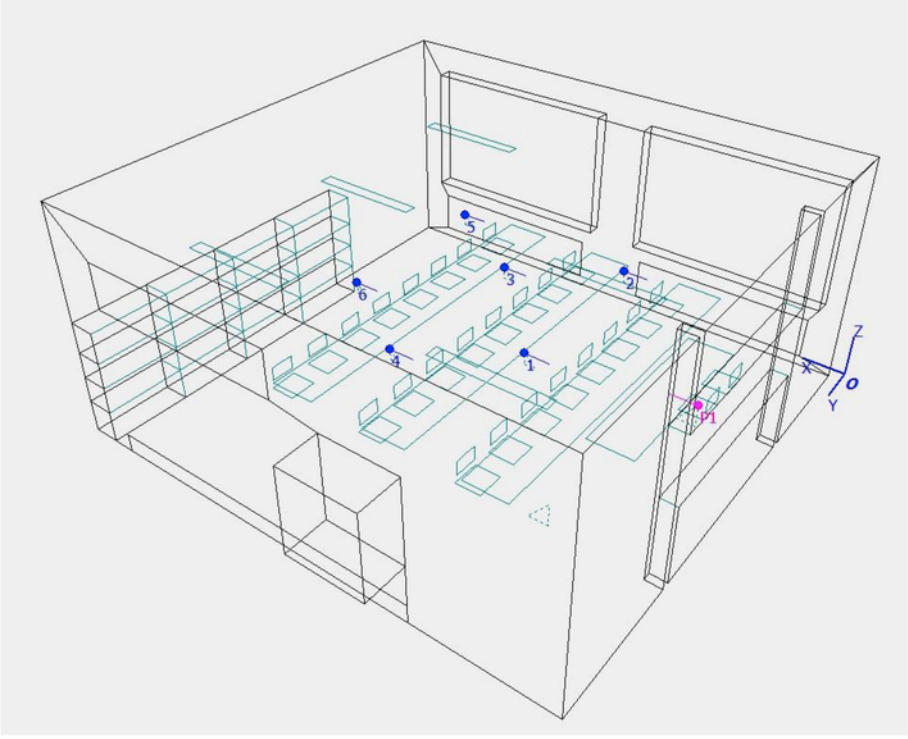


Fig. 2. Geometric model of the classroom. In the model all relevant details of the classroom geometry are represented. The six receiver positions used for the calibration in unoccupied conditions and the speech source are also indicated.

Table 1
Comparison between measured (meas.) and simulated (sim.) acoustic parameters in the unoccupied classroom at six receiver positions (P1–P6) for the source located at the desk position. The values reported are the averages over the 500–2000 Hz octave bands. Differences (diff.) were considered acceptable when smaller than the corresponding JND [38]: 5% for EDT and T_{30}, and 1 dB for C_{50}.

Measurement position	EDT [s]			T_{30} [s]			C_{50} (dB)		
	meas.	sim.	diff. (%)	meas.	sim.	diff. (%)	meas.	sim.	diff. (dB)
P1	1.16	1.13	2.6	1.06	1.04	1.9	−0.6	0	0.6
P2	1.10	1.06	3.6	1.08	1.03	4.6	−2.8	−2.3	0.5
P3	1.11	1.13	1.8	1.04	1.06	1.9	−0.7	−1.1	0.4
P4	1.16	1.12	3.4	1.05	1.05	0.0	−4.1	−3.4	0.7
P5	1.15	1.11	3.5	1.04	1.02	1.9	−1.9	−1.6	0.3
P6	1.08	1.04	3.7	1.04	1.01	2.9	−1.1	−1.7	0.6

similar calibration accuracy was obtained for all listening positions and was deemed appropriate for the scope of the work.

Then, the virtual model of the classroom in occupied conditions was set up and further calibrated. The noise source was added to the model and set as omnidirectional (*Omni* in Odeon®), replicating the directivity pattern of the loudspeaker used in the real classroom. Two omnidirectional receiver points were created, corresponding to R1 and R2. The geometry of the chairs was modified to consider the presence of seated persons; they were modeled as two parallelepipeds: a (0.6×0.5 x 0.4) m seat and a (0.2×0.5 x 0.4) m seatback. Furthermore, in accordance to [40], their scattering coefficient was modified and set to 0.7. Air temperature and relative humidity were set according to average values measured during the *in situ* tests (T = 23 °C, RH = 23%). Simulations were performed with a transition order of two, with 2000 early rays and 16000 late rays. The virtual model in occupied conditions was calibrated with reference to the measured octave-band values of T_{30}, spatially averaged across the two monaural receivers. The choice of calibrating the occupied model over two positions was motivated by the limited spatial variation of the T_{30} values, as resulted by the preliminary calibration. Similarly, in Refs. [25,26] it was found that the spatial deviation of the considered acoustical parameters (i.e. T_{30}, STI) was smaller than the JND values when considering small classroom with size comparable to the present one. Therefore, owing to the spatial distribution of the metric, a calibration over a larger number of positions would not have benefitted the final virtual model. Indeed, the final T_{30} value was 0.81 s to be compared with the measured value of 0.83 s. Then, to further verify the correspondence between the acoustical conditions in the real and the simulated classroom, also the EDT and C_{50} parameters were compared, which are expected to vary with the listening positions. The comparisons are reported in Table 2, where, for the sake of completeness, also the acoustical parameters obtained with the noise source are reported. It has to be noticed that the differences between measured and simulated acoustical parameters at both R1 and R2 are smaller than the corresponding JND also for EDT and C_{50}. A similar calibration accuracy was obtained for the two listening positions and was deemed appropriate for the scope of the work.

Finally, starting from the calibrated virtual model of the existing classroom (C1), two other virtual classrooms were created: C2, replicating the existing classroom except for the acoustical treatment of the lateral wall (replaced by a painted plasterboard finishing), and C3, having the same material properties of C1 but with a doubled volume along the longitudinal direction. The same sources and receiver position as in C1 were set for C2. In C3 a single receiver (R3) located 1.62 m from the end wall of the classroom was defined; the distance from the end wall was chosen to be the same as that of R2 in the C1 and C2 models. The main characteristics of the three classrooms, regarding the receivers, room shape, and material properties are summarized in Table 3.

2.4.2. Listening conditions and procedure

Auralized listening conditions in the three virtual classrooms were created by convolving the anechoic speech signal and noise, used for the *in situ* experiment, with the simulated binaural room impulse responses (BRIRs) at the receiving positions for both speaker and noise sources.

Firstly, the sound power level of the virtual sources was defined, based on the C1 model. Specifically, it was required that the same

Table 2
Measured and simulated acoustic parameters at the listening positions R1 and R2 in the occupied virtual model.

Source	Listening position	T_{30} [s]		EDT [s]		C_{50} (dB)	
		R1	R2	R1	R2	R1	R2
Speech source	real classroom	0.82	0.84	0.81	0.86	3.6	0.4
	model C1	0.80	0.82	0.78	0.83	3.7	0.8
Noise source	real classroom	0.81	0.80	0.74	0.72	3.4	0.8
	model C1	0.78	0.78	0.71	0.69	3.2	1.1

Table 3
Characteristics of the three virtual classrooms: geometry and material properties. The last column reports the α_w values of the treated/untreated lateral wall, according to the standard [41].

Model	Receivers	Dimensions [m]	Volume [m^3]	Acoustical treatment of the lateral wall	α_w of the lateral wall
C1	R1, R2	7.29×7.62 x 3.55	198	yes	0.50
C2	R1, R2	7.29×7.62 x 3.55	198	no	0.05
C3	R3	7.29×15.24 x 3.55	396	yes	0.50

sound pressure levels as measured *in situ*, 1 m away from the sources were also measured in the virtual model. The same sources sound power levels were used for the three virtual models. From a practical point of view, the assumption corresponds to considering the same female speaker, talking with the same vocal effort in the three classrooms. Even though it is known that the vocal output of talkers depends on room acoustics [42,43], the adjustment of diverse sound power levels for each room was deemed not essential for the present purposes. Then, a virtual listener was defined in the acoustical CAD models having the head-related-transfer-functions (HRTFs) of the B&K type 4100 head and torso simulator, which were already available from previous measures. The auralization procedure involved creating separate BRIRs at each selected listening position within the virtual classrooms, for both speech and noise sources; the BRIRs were then convolved with the corresponding anechoic material.

The auralized listening conditions are described in Table 4. A further confirmation of the calibration procedure is obtained by the comparison of the measured binaural values with those simulated in model C1. In fact, all differences between the acoustical parameters obtained by means of binaural *in situ* impulse responses and those output from the simulations (see Sec. 2.3.2) are smaller than the corresponding JNDs (reverberation time: 5% [38]; sound level: 1 dB [38]; STI: 0.04 in Annex F of [2]). The same calibration accuracy was obtained for both listening positions.

The same panel of testers taking part in the *in situ* tests also performed the auralized experiments in a quiet laboratory environment. The testing setup consisted of a laptop with the listening test system, a RME Fireface© UC sound card, a headphones amplifier (Behringer, Powerplay PRO-XL HA4700) and Audio-technica type ATH-m50x headphones. The presentation of the stimuli and the data collection was controlled by the same wireless test bench as described in Sec. 2.3.2. The participants responded by using the touchscreen handset. The experimental set up was calibrated placing the headphones over a B&K type 4100 head and torso simulator.

The experimental session was held almost two months after the *in situ* listening tests, with groups of a maximum of four people at a time, following the same procedure as described in Sec. 2.3.2. Firstly, a training session was proposed; afterwards participants completed five lists of 18 words, each one proposed in a different listening condition. After the completion of each test list, the participants were asked to rate the subjective listening effort over a 10-points scale. Words lists and listening conditions were randomized across the groups of participants.

2.5. Statistical analysis

All statistical analyses were conducted using the software *R* [44]. For the analysis of IS results, the responses were coded using a binary score (0/1, corresponding to incorrect/correct); the selection of "none of the two" was considered an incorrect response. The percentage of correct responses was calculated for each participant in each listening condition (in *situ* and auralized). Individual RT data were examined in detail, to remove excessively large values possibly due to participants' lack of attention [45]. An absolute cut-off of 5000 ms was set, beyond which RT results were discarded and considered as missing data. The procedure yielded the removal of 0.4% of the dataset. As concerns subjective ratings, consistently with literature studies exploring the same subjective metric approach [19,46], the absolute values of LE were considered in the analysis. In fact, adopting a normalization of subjective data aimed at reducing individual variability (e.g. converting individual LE data in the corresponding Z-scores) would have prevented a reliable comparison of the results across the two groups of participants (NI/NG) and the description of the values of the differences across the listening conditions. The effect of individual trends, which is an important aspect of both LE and RT data, was then addressed through the choice of a suitable statistical technique.

To resolve this issue it is useful to recall that several methodological concerns have been raised in using common statistical methods for the analysis of response time data [22,45,47–49] and for elaborating the outcomes of forced-choice tasks [50,51]; these concerns are especially relevant when analyzing data with repeated measurements over the subjects as in the present case. First, the response variable distributions often depart from the normal distribution. As pointed out in Ref. [49], a considerable variation in the shape of the RT distribution is to be expected, both at individual level and for the specific experimental condition. In general, the RT distribution can be considered as positively skewed, rising rapidly on the left and having a long positive tail on the right [45]. Similarly, the normality assumption is not met when considering IS results expressed as the proportion of words correctly identified over those presented. In fact the outcome variable will be bounded to the [0; 1] interval and thus will be better represented by a binomial distribution, not a normal one. This is especially true when considering favorable conditions (as those selected for the present experiment) where the IS distribution will be concentrated on larger values closer to unity due to the increased selection of correct responses. The specific distributions of the data collected in the experiment are reported and discussed in Appendix A. Second, when multiple measures are acquired for each participant in different experimental conditions they will not be statistically independent, even when conditions have been carefully controlled. For instance, each participant potentially has a slightly biased response time, and this characteristic will affect all the responses from that participant. Then, the added variability in responses related to individualities need to be addressed together with the variability explained by the fixed factors of the experiment (e.g. listening conditions).

Therefore, the generalized mixed-effects model GLMM (*lme4* package [52]) was chosen for data analysis, being a statistical method showing the twofold advantage of dealing with not-normal data distributions and with non-independent individual responses. In particular, the latter issue is handled through the definition of the participants as a random factor within the model, i.e. a subset of subjects randomly sampled from a larger population. Instead of analyzing aggregated data based on the participants mean, GLMM predicts the individual responses to the fixed factors of the experiment. For each participant a

Table 4
Auralized listening conditions in the three virtual classrooms (C1, C2 and C3) at the listening positions (R1: front, R2 and R3: back). Data refer to the average of left and right channels. The data measured in the real classroom are also included for comparison with the auralized classroom C1.

Classroom	T_{30} [s]	Receiver	Speech level dB(A)	Noise level dB(A)	SNR	STI
C1	0.81	R1	61.0	60.9	0.1	0.52
		R2	57.4	58.7	−1.3	0.45
C2	1.21	R1	62.6	62.0	0.6	0.49
		R2	60.7	60.8	−0.1	0.42
C3	0.88	R3	55.2	54.7	0.5	0.48
real classroom (ref. C1)	0.83	R1	61.0	61.0	0.0	0.52
		R2	57.4	58.7	−1.3	0.46

different baseline value is assumed, upon which the other subject's responses will depend; furthermore, a different degree of variation across a fixed factor can be assumed for each participant. Including in the model the random effect of participants allows the estimation of how much of the variance in the outcome variable is due to the different individuals rather than to the effect of the experimental conditions. Furthermore, this type of analysis will reduce the type I error rate [53]. Indeed, if not partialed out, the variability associated to participants could mask patterns in the analysis of the fixed effects. A further advantage of GLMM is that it does not rely on the assumption of a normal distribution of the data, allowing for the selection of the distribution most appropriate to the response variable. Consequently, raw RT data can be directly analyzed without the need for transformation prior to analysis: it has been shown that data transformation does not affect type I errors [46]48 and that analysis on the transformed RT might be uninformative as to whether the same significant effects exist on the original, untransformed metric [53].

In this study a GLMM with a binomial distribution was used to analyze IS data, whereas RT results were analyzed using a Gamma distribution with a log-link function [24,49,53]. For the analysis of the RT results, both correct and incorrect responses were considered; using only correct responses did not change the statistical results. The analysis of LE data was instead accomplished with a cumulative link mixed model (*ordinal* package [54]), which describes the relationship between a categorical variable with a clear ordering of the levels (ordinal response variable) and the explanatory variables, still considering the subject variability as a random effect.

For all statistical analyses, model selection was based on a forward procedure using the likelihood ratio test. The statistical assumptions of the final model have been verified by checking the normality of the random effect terms and the residuals. In case of statistically significant effects of the main factors or of the interactions, pairwise comparisons based on the difference of the means were performed using the *lsmeans* package [55]; in order to account for planned multiple comparison, a Benjamini-Hochberg procedure was used.

Prior to analysis, RT data corresponding to the "none of the two" responses were considered. This alternative was made available in order to minimize the occurrence of false positives in the IS results, and, depending on the listening condition, represented a percentage of the overall responses ranging between 1.5% and 4.1%. A preliminary observation of the dataset suggested that this choice was always associated with RTs larger than the RTs of correct/incorrect responses. To test this occurrence, a dedicated statistical model was setup, with RT as the response variable and response type (correct/wrong/none of the two) as the fixed factor. Following the significant effect ($\chi^2(2) = 116.2$, $p < 0.001$), the pairwise comparisons showed that RT data associated to the "none of the two" alternative were significantly larger than RT associated with correct or wrong responses ($p < 0.001$ and $p = 0.006$ respectively). The finding indicates that, even though participants were aware of having always been presented with one of the rhymed alternatives, in some occasions more processing time was needed, but in the end a choice could not be made. In this sense, these RT data were not representative of a successful decision process and were thus removed from the analysis. Overall, 20 RTs were removed from the *in situ* experiment (2.8% of the dataset) and 47 RTs were removed from the auralized experiment (2.7% of the dataset).

3. Results

3.1. Comparison between in situ and auralized listening tests

In the setup of the statistical models, listening position (R1 *vs.* R2), mother tongue (NI *vs.* NG), mode of presentation (*in situ vs.* auralized), and their two- and three-way interactions were considered fixed factors. Participants were considered a random factor; a random slope was also specified, supposing that the effect of the mode of presentation might be different for each participant. The descriptive statistics of the measured IS, RT and LE data averaged over participants for the two experimental conditions are shown in Fig. 3. In the following, the outcomes of the statistical analysis are presented not with reference to the descriptive statistics but rather to the estimates of the metrics as an output from the statistical GLMM model where, as discussed above, the random factors have been controlled for.

The statistical analysis of IS results revealed that the effect of mother tongue alone was significant ($\chi^2(1) = 29.16$, $p < 0.001$). When averaged over mode and position, the estimated probability of a correct response was higher for NI than for NG, with an IS difference equal to 10.5% (NI: 94.1%, NG: 83.6%). The effects of mode of presentation and listening position were not significant.

In the analysis of RT data only a significant effect of position ($\chi^2(1) = 13.03$, $p < 0.001$) was found, indicating that, when results were collapsed across mother tongue and mode of presentation, participants always showed significantly smaller RTs in position R1 versus position R2 (R1: 1342 ms, R2: 1426 ms; mean ratio: 0.94). No main effects of mode of presentation and mother tongue were found. It should be noted that a visual inspection of the descriptive statistics results (Fig. 2Fig. 3b) might suggest an increase in the mean values in RTs of NG listeners and a decrease in RTs in auralized conditions for position R1 alone. The GLMM statistical model showed that neither the effect of mother tongue nor the interaction between mode of presentation and position reached the level of significance ($p = 0.068$ and $p = 0.057$). It has to be noticed that the p value resulting for the interaction between mode of presentation and position was close to the significance threshold. In order to understand if the absence of a significant interaction was prompted by the high variance of the NG group (as showed for instance by the IS results), the statistical analysis was repeated separately for the two groups. In the case of NI a significant effect of position was found ($\chi^2(1) = 13.28$, $p < 0.001$) but no effect of mode of presentation ($\chi^2(1) = 1.58$, $p = 0.21$) nor of the interaction ($\chi^2(1) = 1.29$, $p = 0.26$). Then, including NG participants alone, the results indicated the absence of significant effects for listening position ($\chi^2(1) = 2.96$, $p = 0.075$), mode of presentation ($\chi^2(1) = 0.60$, $p = 0.44$) and their interaction ($\chi^2(1) = 2.59$, $p = 0.11$). Overall, these results further confirmed the absence of an interaction between listening condition and mode of presentation, but this aspect devises further researches. A detailed discussion is taken on in Sec. 4.1.

Finally, the statistical analysis of the LE ratings showed the presence of a significant interaction between both position and mode ($p = 0.007$), and position and mother tongue ($p = 0.01$). The interaction between position and mode indicated that, for position R2 alone, LE ratings in higher (more effortful) categories were more likely for the auralized tests than for the *in situ* tests ($z = 2.68$, $p = 0.007$). No significant difference in the LE ratings was found at position R1. When examining the pairwise comparisons between positions within each mode of presentation, only the comparison for the auralized condition was significant ($z = -5.25$, $p < 0.001$), showing that LE ratings in position R2 were higher than in R1. Then, as concerns the interaction between position and mother tongue, of which the descriptive statistics is shown in Fig. 3c, the GLMM model outlined that LE ratings in position R1 were lower than in position R2 for both groups of participants (NI: $z = -4.84$, $p < 0.001$; NG: $z = -2.40$, $p = 0.016$). No significant difference between NG and NI was found in the LE results in position R1, whereas in R2 the LE ratings of the NI participants were found to be higher than the LE ratings of the NG participants ($z = 2.67$, $p = 0.008$).

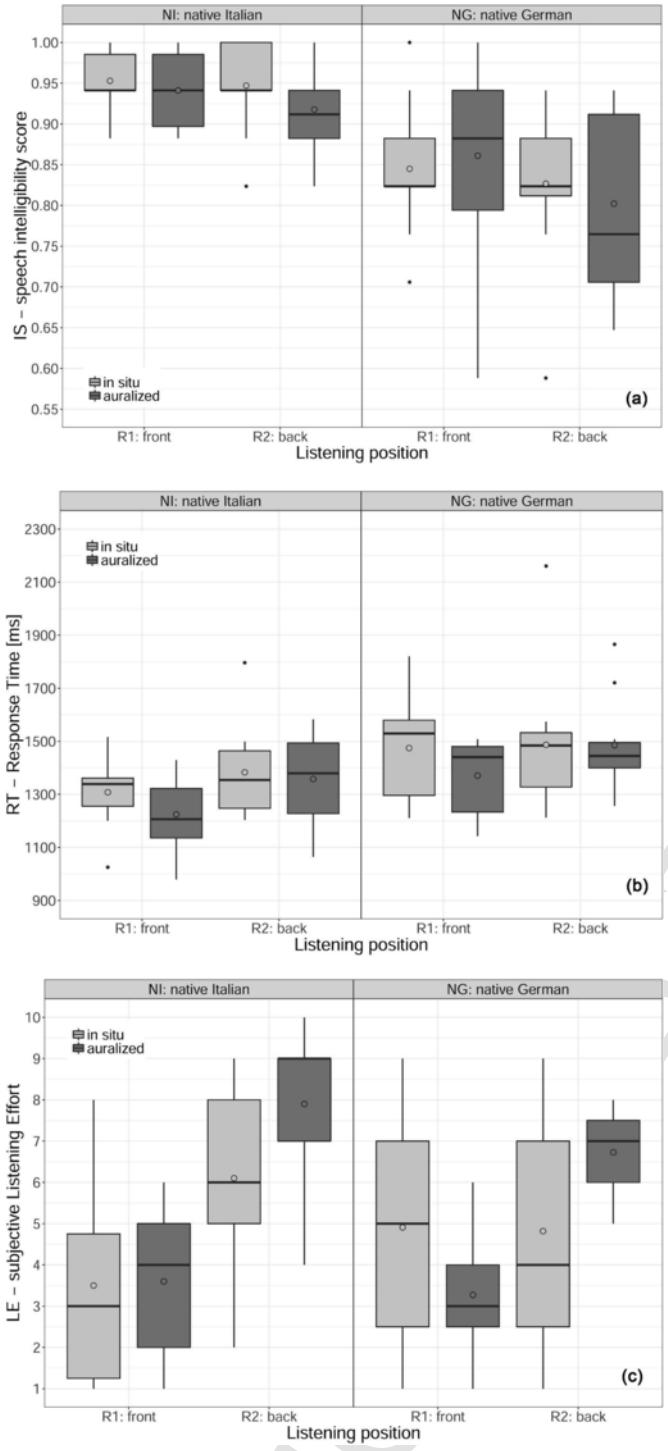


Fig. 3. Boxplots of the (a) speech intelligibility IS, (b) response time RT, and (c) subjective effort LE results averaged across participants. The results are divided according to the participants' mother tongue (NI, NG), the listening position (R1, R2) and the mode of presentation of the tests (*in situ*, auralized). The auralized condition refers to model C1, as defined in Table 2Table 3. The bottom and the top of the boxes are the first and the third quartiles of the data distributions, the central, bold line is the median value, and the white circle is the mean value; 99% of the data fall within the whiskers. The outliers are shown as points outside the whiskers.

3.2. Auralized listening tests: effects of the acoustical treatment of a lateral wall

In this case, the GLMM model included mother tongue (NI *vs.* NG), listening position (R1 *vs.* R2), finishing of the lateral surface (with *vs.* without acoustical treatment, corresponding to the virtual models C1 and C2) and their interactions as fixed factors; participants were considered a random factor.

Fig. 4a displays the descriptive statistics of IS data across the two listening positions for the two virtual classrooms (C1 and C2). The statistical analysis revealed that the main effect of mother tongue was significant ($\chi^2(1) = 10.95$, $p < 0.001$): the NI participants showed significantly higher IS results than the NG participants, with a predicted increase of 7.2%. The analysis showed that the interaction between position and wall finishing was also significant for the IS results ($\chi^2(1) = 4.47$, $p = 0.035$), and the pairwise comparisons indicated the presence of a significant decrease of the quantity between R1 and R2 only for classroom C2, without the acoustical treatment ($z = 4.39$, $p < 0.001$). The IS gap between positions was 9.9%.

As regards RT, the descriptive statistics across the two listening positions for the two virtual classrooms (C1 and C2) are presented in Fig. 4b. The GLMM analysis showed a significant main effect of both listening position ($\chi^2(1) = 41.49$, $p < 0.001$) and wall finishing ($\chi^2(1) = 12.53$, $p < 0.001$), whereas the effect of mother tongue and the interactions were not significant. Specifically, the pairwise comparisons revealed that the participants responded significantly faster in R1 than in R2 (mean ratio: 0.90). Furthermore, greater RT results were found in the virtual classroom without the acoustically treated wall in comparison to the acoustically treated classroom (mean ratio: 1.06).

Finally, the descriptive statistics of LE subjective ratings are shown in Fig. 4c. From the statistical model a significant effect of mother tongue ($p = 0.035$) was found, and the pairwise comparisons showed that the NI participants gave on average higher LE ratings than the NG listeners ($z = 2.13$, $p = 0.033$), thus indicating greater perceived subjective effort. Then, there appeared an interaction between position and wall finishing ($p = 0.009$) with the LE values increasing in the rear position of the classrooms (R1 *vs.* R2 – with treatment: $z = -5.68$, $p < 0.001$; without treatment: $z = -3.07$, $p = 0.002$). However, whereas in R1 significantly higher LE results were found for the classrooms without acoustic treatment ($z = -3.97$, $p < 0.001$), no significant difference between the two room preparations was observed in R2.

3.3. Auralized listening tests: effects of room shape

The effect of room shape was analyzed with reference to the rear position alone (R2 for C1 and C2 models, R3 for C3). Fig. 5 reports the descriptive boxplots of the IS, RT and LE measured data for this analysis. Data were then modeled using GLMMs with mother tongue (NI *vs.* NG), classroom typology (C1, C2, C3) and their interaction as fixed factors; again, participants were considered as a random factor.

As concerns speech intelligibility, a significant main effect of mother tongue alone was found ($\chi^2(1) = 4.78$, $p = 0.029$); when the results were collapsed over position and room type, the NI participants had higher IS results than the NG participants (mean difference: 5.9%). No effects were significant for classroom type ($\chi^2(2) = 1.35$, $p = 0.51$) or for the interaction ($\chi^2(2) = 4.29$, $p = 0.12$). Visual inspection of Fig. 5a suggest a high IQR in the results of NG, which could be motivated by a different linguistic proficiency of the participants and potentially mask the presence of a significant effect of listening condition for the NI participants (having a much smaller IQR). Then, the statistical

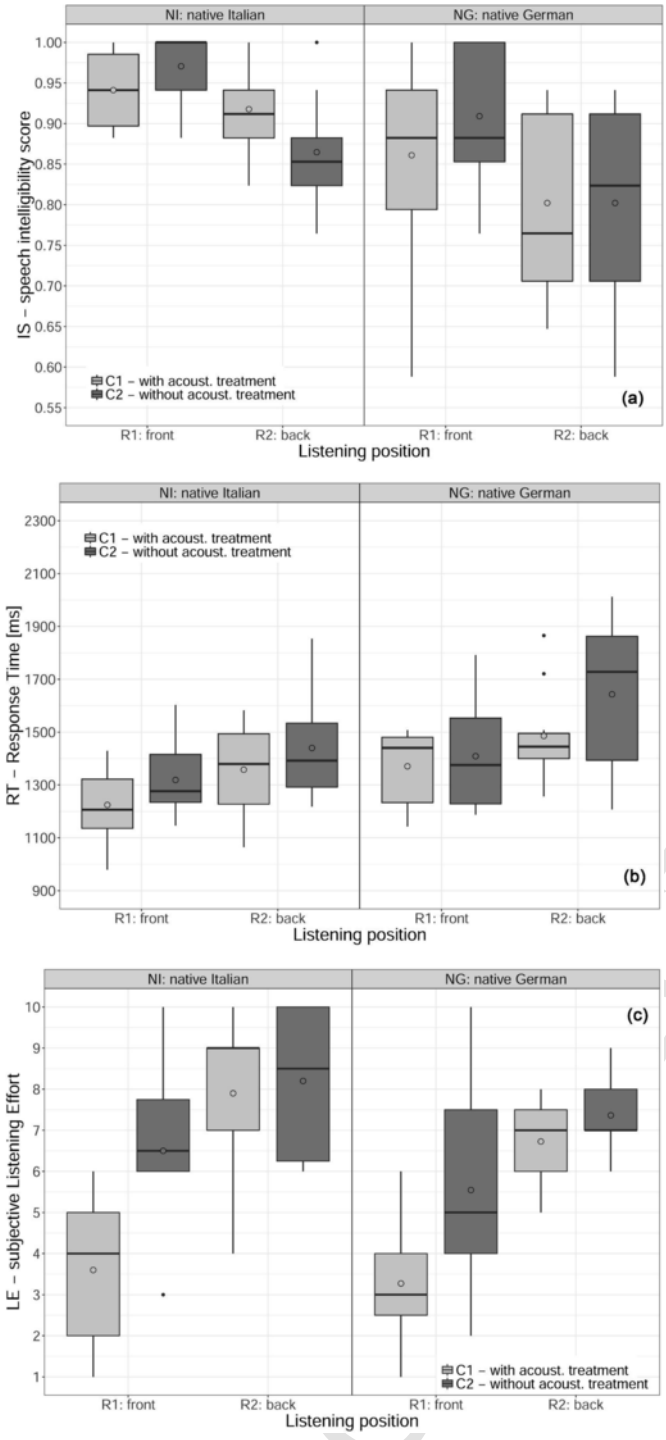


Fig. 4. Boxplots of the (a) speech intelligibility IS, (b) response time RT, and (c) subjective effort LE results averaged across participants. The results are divided according to the participants' mother tongue (NI, NG), the listening position (R1, R2) and the finishing of the lateral wall (with or without acoustical treatment, corresponding to models C1 and C2). The bottom and the top of the boxes are the first and the third quartiles of the data distributions, the central, bold line is the median value, and the white circle is the mean value; 99% of the data fall within the whiskers. The outliers are shown as points outside the whiskers.

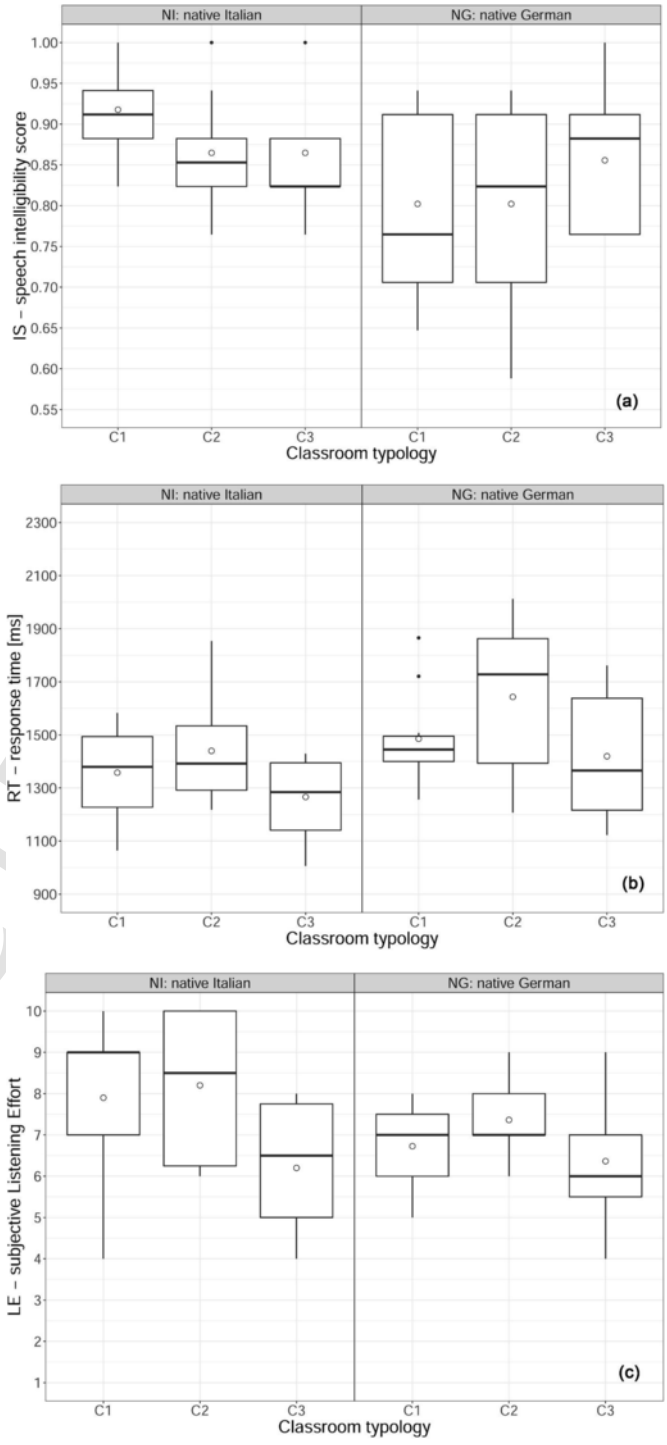


Fig. 5. Boxplots of the (a) speech intelligibility IS, (b) response time RT, and (c) subjective effort LE results averaged across participants. The results are divided according to the participants' mother tongue (NI, NG) and the classroom typology: rectangular shape with acoustical treatment (C1), rectangular shape without acoustical treatment (C2), elongated shape with acoustical treatment (C3). The bottom and the top of the boxes are the first and the third quartiles of the data distributions, the central, bold line is the median value, and the white circle is the mean value; 99% of the data fall within the whiskers. The outliers are shown as points outside the whiskers.

analysis was repeated including NI listeners alone, to check if their higher intra-group consistency would provide different outcomes. In this case listening condition was considered as a fixed factor and the results indicated that the effect of listening condition on the IS metric was still not significant ($\chi^2(2) = 3.15$, $p = 0.21$).

Differently, for the RT results the statistical analysis revealed a significant main effect of classrooms type ($\chi^2(1) = 32.64$, $p < 0.001$) whereas the effects of mother tongue and the interaction were not significant. All pairwise comparisons were significant at $p < 0.01$; the RT

results were the quickest in the elongated classroom C3 (mean RT: 1330 ms), slightly increased in the real-size, acoustically treated classroom C1 (mean RT: 1418 ms), and were the slowest in the classroom with reflective boundaries C2 (mean RT: 1520 ms).

Similarly to RT, the statistical analysis of LE showed a significant effect of classroom typology ($p < 0.001$) alone. The pairwise comparisons indicated that no difference was present between the ratings of C1 and C2. The LE results were instead significantly lower in configuration C3 (C1 *vs.* C3: $z = 2.65$, $p = 0.008$; C2 *vs.* C3: $z = 3.61$, $p < 0.001$), indicating that the elongated classroom was subjectively perceived as less demanding.

4. Discussion

4.1. Comparison between in situ and auralized tests

Based on the comparison of the outputs of the statistical model of the speech-in-noise tests in auralized and *in situ* conditions, insights on the potentials of the auralization techniques in recreating perceptually equivalent environments can be drawn, specifically referring to the relevant metrics of RT and LE.

The IS estimates in auralized conditions matched the corresponding ones obtained with *in situ* testing, confirming that when properly calibrated virtual models are used, where the differences between auralized and measured relevant acoustic parameters are smaller than the JND, there is the same speech intelligibility as in real settings. It is interesting to notice that, for NG group alone, the variance in IS results greatly increased between real and auralized conditions. The finding could be explained by differences in the states of arousal or motivation between the two experimental settings. In fact, within the real classroom, the participants are prompted to a more focused attention due to the presence of the other participants, carrying out the same task simultaneously. During the auralized tests, that were self-paced and without the constraints of community testing, a decrease in participants' arousal emphasized the presence of differences in the individual performance.

As concerns the measure of RT, no main effect of mode of presentation was found in the statistical analysis, suggesting that, upon proper calibration of the virtual model the same absolute values as measured *in situ* could be replicated in the laboratory tests. The finding points toward the ecological validity of the measure of RT in auralized conditions. The equivalence of two modes of presentation as regards RT was ensured also by the absence of significant interactions, even though that with listening position was close to the threshold of statistical significance ($p = 0.057$). The result might be influenced by the high variance of the results of the NG group, as suggested by the results of the analysis including NI listeners alone, where the effect of mode of presentation ($p = 0.21$) and the interaction with listening position ($p = 0.26$) were clearly not significant. However, this critical point prompts further investigation into the relationship between mode of presentation, acoustical conditions and language proficiency to strengthen and generalize the present validation. In particular, a wider range of listening conditions would be helpful and an extension of the panel of participants would be also appropriate, either including only native listeners, or objectively assessing the language proficiency of non-native participants (see Sec. 4.3). Here, only a significant main effect of the listening position on RT was found. Whilst IS did not show differences, it was possible to distinguish between the front and the rear position of the classroom using the RT metric, with the latter position associated to slower RTs. The same RT increase was found in both *in situ* and auralized conditions (ΔRT: 85 ms and 82 ms respectively, averaged over the listening positions), showing that the increase in the deployed cognitive resources tracked by RT was independent of the mode of presentation.

As concerns the LE results, a significant interaction between mode of presentation and position was found. When the results were averaged across the participants' mother tongue, it was found that the listening positions were discriminated only when presented in the laboratory setting, thus suggesting that other factors, beside the response to the auditory stimulus, affected the subjective response *in situ* and decreased its potentials. It should be noted that the experimental setup and auditory-visual interaction are all factors that could influence the LE results, especially in the more complex conditions of field experiments. In particular a great dispersion of the results was observed for the NG participant in the real classroom, whereas the IQR of the LE results reduced (and became comparable with the IQR of NI participant) when the tests were proposed in auralized conditions. It is believed that the result stemmed from the subjective nature of the LE rating itself. Indeed, listening effort depends not only on the task demands but also on an individual cost/benefit evaluation, involving an appraisal of listening demands in relation to one's capacity [56]. Then, the LE rating will reflect the listening demands in relation to the participant auditory and cognitive abilities, but also the more subjective aspect of the participant appraisal of its capacity to meet the demands. The high IQR of NG participants might then indicate a difficulty in the appraisal of their abilities in performing the speech perception task in the non-native language. The same difficulty was not observed during the test in auralized conditions, where the participants possibly built upon their previous experience to better assess their capacity in relation to the task.

Finally, since no interactions were found between mother tongue and mode of presentation for any of the response variables, it follows that the results held independently of the participants' mother tongue, since they showed the same trends across the two modes of presentation.

In summary, as regards the comparison between tests in real and auralized conditions, the present study shows that, upon proper calibration of the virtual model:

- the IS results obtained in the real setting could be replicated in auralized condition;
- the RT results obtained in the real setting could be replicated in auralized conditions. Further investigations are needed to better explore the relationship between mode of presentation, listening condition and language proficiency on this outcome.
- the LE results could not be replicated in the two modes of presentation, and a significant interaction was found between mode of presentation and listening position. The result was driven by the subjective nature of the LE ratings that beside the effect of the listening condition also reflect individual, extra-acoustic factors.

Therefore, an ecological validation was achieved with reference to IS and RT alone, on which the assessment builds upon. In the following, the LE results will be anyhow discussed, to explore their sensitivity with reference to the RT results (see sec. 4.4).

4.2. Effects of room acoustics on the speech reception performance

Table 5 summarizes the results of the statistical analyses output from the GLMM models reported in Sec. 3.2 and 3.3, relevant for the present discussion.

When examining the effects of the acoustics treatment of the room on the speech reception performance, it was found that little information was returned by the accuracy results. The IS metric was sensitive to the participants' mother tongue but not to the modifications to the room acoustics. Inserting the acoustic treatment on the lateral wall was

Table 5
Summary of the statistical analysis outputs as concerns the effects of room acoustics on intelligibility results (IS), response times (RT) and subjective ratings of listening effort (LE). The results refer to the two groups of participants (NI: native Italian, NG: native German), the three virtual classrooms (C1: regular room with acoustical treatment, C2: regular room without treatment, C3: elongated room with acoustical treatment), and the two listening positions (R1: front, R2: back). The dash within a cell indicates that the corresponding effect (or interaction) is not statistically significant.

		IS	RT	LE
effects of room finishing	wall treatment (C1 *vs.* C2)	–	$RT_{C1} < RT_{C2}$	–
	listening position (R1 *vs.* R2)	–	$RT_{R1} < RT_{R2}$	–
	mother tongue (NI *vs.* NG)	$IS_{NI} > IS_{NG}$	–	$LE_{NI} > LE_{NG}$
	treatment X position	C2: $IS_{R1} > IS_{R2}$	–	$LE_{R1} < LE_{R2}$
				R1: $LE_{C1} > LE_{C2}$
effects of room shape	room shape (C1 *vs.* C2 *vs.* C3)	–	$RT_{C2} > RT_{C1} > RT_{C3}$	$LE_{C1} > LE_{C3}$
				$LE_{C2} > LE_{C3}$
	mother tongue (NI *vs.* NG)	$IS_{NI} > IS_{NG}$	–	–

not picked out as a change by IS, as pointed out by the absence of statistically significant differences between C1 and C2. Solely based on this finding, one could argue that given the high IS results already scored in the untreated classroom (higher than 80%, thus corresponding to intelligibility rated as "Fair" [4]) the addition of acoustical treatment had a marginal benefit on the listeners' speech reception. Indeed, the increase of 0.03 in the objective measure of STI accomplished in both listening positions by inserting the acoustical treatment was lower than the associated JND of 0.04. As concerns the STI gap between the listening positions, it was the same for both C1 and C2 ($\Delta STI = 0.07$), so that one could expect an almost equivalent measurable reduction in IS in both cases. Instead, a statistically significant difference between the front and rear position was found for the untreated room alone (10% decrease). However, it has to be considered that the STI variations were realized for different absolute values of the objective metric (C1: 0.52–0.45; C2: 0.49–0.42). Relying on the psychometric curve (i.e. the sigmoid curve relating STI and IS results [2,4]) it is reasonable to expect that fewer differences will be obtained in the IS results when moving towards higher STI values. The metric will become less informative for the highest STI values, where it undergoes a ceiling effect.

As also pointed out in Ref. [5], the fact that the number of correctly recognized words was not affected by the changes in acoustic conditions, does not necessarily imply that the task was not cognitively more demanding. Specifically, RTs were significantly greater in the classroom without acoustic treatment (average increase: 82 ms), indicating that more time was spent to process the auditory information. Even though the task presented in the experiment is not directly representative of the memorization and recall processes, which are critical for listening environments [57], it can be inferred that the increase of RTs already in the speech reception stage will negatively affect speech communication. Prolonged speech processing will limit the amount of information that can be held in memory, impairing the subsequent storing and recall. When the listening effort was tracked using the perceived effort LE, only a partial difference could be found between the two classroom configurations, with the presence of acoustical treatment yielding lower (i.e. better) ratings for the anterior listening position alone. In R2, despite the significant increase in the RTs results, the two classroom configurations were rated as similarly effortful. As listening conditions in R2 were those with the lowest STI (and SNR) values, it could be hypothesized that the lack of differences was driven by the minor sensitivity of the LE results in the unfavorable listening condition range. For instance, in Ref. [19] it was argued that the LE measure is more sensitive at an SNR higher than -2 dB (no reverberation) and in Ref. [16] it was found that as the listening conditions worsened (either increased reverberation or lowered SNR) the "listening difficulty" results showed fewer variations between conditions and consequently lesser discriminating potential.

A similar scenario was outlined when the effect of a change in room shape was considered. No differences were found in the accuracy results, whereas both the RT and LE data indicated that less demanding listening was achieved in the treated long classroom, with smaller RTs and lower LE ratings in comparison with the treated normal-sized environment. Again, a change in the acoustic configuration of the room that yielded a STI difference lower than the JND between the tested sound fields, was not tracked by the accuracy results, while it was found to affect the listening experience as concerns the reported effort devoted to the task and the response time. The result was caused by the more favorable listening condition realized in the long classroom. In fact, despite an increase in the reverberation time of the room, a positive SNR was observed at the back of the classroom thanks to the position and the omni-directivity of the background noise source as compared to the directional source of the speaker. From the point of view of room acoustic design, it would be of interest to examine the effects of a similar change in room size in the presence of spatially distributed, not punctual, background noise sources mimicking, for instance, unintelligible student babble. The results would allow the specific pro and cons of a similar design strategy to be assessed.

4.3. Effects of mother tongue on the speech reception performance

The issue of second-language listeners is especially relevant in school settings where the effect of the sound environment sums up with the incomplete linguistic knowledge, thus making the speech reception task harder, and, in turn, recall and memorization of information. Based on previous literature studies [11,12,58] a discrepancy in the accuracy performance of the two groups of participants was expected: the NI participants scored significantly higher than the NG listeners (with a 6% average difference), irrespective of the listening conditions. The finding shows a disadvantage of the NG listeners based on inaccurate perceptual processing of non-native words, which could be explained by not entirely equivalent language proficiency [12].

Interestingly, the RT metric did not disclose the difference of the two groups. Visual inspection of the descriptive statistics data (see Figs. 4b and 5b) might suggest an increase of RTs for the NG participants, which could be explained by interferences from their native language on the lexical or phonetic level [59,60]. However, the GLMM analysis of the RT data did not reveal any statistically significant differences between the two groups. It has to be recalled that the present experiment collected only self-report assessments of language proficiency. The NG listeners participating in the experiment self-rated their Italian proficiency as quite high; they have lived since birth in a bilingual environment and, according to the definition in Ref. [2], they are considered as "experienced, daily second language users". It has to be noticed that differences might still exist in the individual proficiency in the non-native language, which can only been disclosed by using an objective as-

sessment, such as testing their linguistic abilities [12,61]. Controlling individual abilities within the statistical analysis using the test results as a covariate, or testing native and non-native listeners at the same IS level would help in better outlining the effects of room acoustics on the RT results. For instance, in Ref. [62], the performance of native and non-native listeners was compared in quiet and noisy acoustical conditions where 100% intelligibility was scored by both groups of participants. In this case, non-native listeners, though highly proficient, always showed greater RTs suggesting that when the same level of accuracy is reached, longer processing times are needed to cope with the task. No interaction was found between listening condition and mother tongue, leaving open the hypothesis that the increase of RT could have been carried over from the quiet to the noisy condition, without an additional effect of the latter. Non-native listeners are required to deploy greater cognitive resources (traced by longer RTs) already in quiet conditions due to the effect of native language interference. As in Ref. [62], the specific hypothesis under investigation in the present study was that unfavorable listening conditions would add more demands on the processing time of non-natives compared to native listeners, but this occurrence could not be confirmed by the results. Therefore, dedicated experiments are needed to understand better how listening in reverberant or noisy conditions affects the RT results of non-native listeners; in the experiments, a careful control of the language proficiency shall be implemented.

Concerning the self-reported measure of listening effort (LE) it is noteworthy that NG participants reported a lower degree of perceived effort than the NI participants despite poorer accuracy of results and no differences in the RTs. It can be speculated that the two groups either interpreted the concept of "listening effort" differently or, similarly to [17], scaled the judgments according to peculiar anchors differently (e.g. the same categorical value was associated with a different degree of effort, depending on the participant). Possibly both aspects played a role in the result which raises the question of consistency between the IS and LE results across the groups.

4.4. Advantages of an acoustical design based on behavioral indexes

In the present experiment acoustic conditions representative of realistic classroom activities were modeled, and the modifications of the virtual classroom model reflected actual acoustic intervention plans or design strategies. The obtained results yield a direct and immediate insight on how passive acoustic interventions influence the task performance and on the deployment of the cognitive resources needed to achieve it.

Monitoring metrics that are considered informative on listening effort is especially important when intelligibility is satisfactory or even near ceiling, as in many everyday environmental situations. In the high-intelligibility region, the metrics of LE and RT show a remarkable resolution and clarify conditions characterized by similar STI values (e.g. within one JND) or considered potentially equivalent if traced by IS. From an applied perspective, this approach would allow for the design of environments were comfortable listening conditions can be achieved, minimizing the risks due to suboptimal room acoustics. RT and LE are respectively a behavioral measure (tracing the processing time) and a measure of subjectively perceived listening effort; they reflect different perceptual mechanisms [7,63] and might not output matching results. In fact, the present experiment highlighted that the two metrics varied peculiarly as a function of the task demands, with the changes in one measure not necessarily reflected by the other. Relying on RT, the effect of modifications to the room acoustics on speech reception was always successfully traced within each group of participants with the same mother tongue. As also found in Ref. [46], self-report ratings of LE discriminated fewer conditions than the RT, especially when the listening conditions worsened. The individual interpretation of the "listening effort" concept influenced the LE results, even more so when comparing the listeners' group of different mother tongue. The finding suggests that the response time might be a more sensitive metric for detecting the effects of changes in the listening conditions, which using LE are not discriminated because they either do not reach the listeners' conscious awareness or undergo ceiling effects.

A strategy for integrating RT in the acoustical design process of public spaces where the requirements of communication are paramount (e.g. in compliance with [1] schools, conference halls, lecture rooms for medium and long range communication, social/senior centers, canteens and restaurants for short range communication) can be depicted based on the present results. Specifically, the assessment of RT should follow a preliminary design based on the STI metric, the target values of which ensure the required IS results based on the conventional psychometric functions. Unfortunately, since statistically different RTs can be obtained for the same intelligibility score (see Sec. 4.2 and 4.3) a unique relationship between STI and RT is not viable. Furthermore, RT will strongly depend on the nature of the specific background noise, the details of which (e.g. stationary, fluctuating, impulsive, with informational content, etc.) are known to affect the speech processing time to a notable extent [64]. For this reasons much work is needed to discern the link that the various acoustical factors have with the behavioral quantity, pursuing the ultimate aim of providing more widely accessible design tools other than listening tests. At present, an acoustical design also taking over RT necessarily involves the setup and presentation of case-specific speech-in-noise tests. This added complexity to the acoustical design is not at hand for the most common types of acoustical projects but can be taken on for the most elaborated or technically demanding ones.

4.5. Study limitations

Despite the present relevant findings, further research has to be devoted to better understand the link between sound environment (as composed by room acoustics, source, and noise characteristics [35]), and the listening effort in the speech reception task. For instance, the impact of the type of noise on listening effort in realistic scenarios needs to be clarified.

Moreover, the present work is limited to the speech reception task, making the inference that more cognitive demanding tasks, which imply speech understanding (i.e. interpreting the meaning of the message) or speech recall, will be more affected by a speech reception worsening. This assumption underlies the current speech evaluation methods [57] and warrants generality to the present approach. However, employing more specific and demanding tasks (e.g. word recall, speech comprehension, etc.) would better outline the role of unfavorable acoustics on memory and learning performance in rooms for speech [65–67]. Systematic information on the effort on the specific tasks could then be used in the context of acoustical design as well.

Finally, it should be noted that, when using the RT metric, it is hardly possible to rely on an absolute scale because the RT course reflects changes other than just the response to the stimulus. Whereas the relative differences in RTs between conditions (e.g. with respect to a baseline in quiet conditions [46]) in a given experiment are driven by the complexity of the stimulus, the absolute values of the metric might be affected by inter-experiment variability due to the specific measurement layout and to the test material used. In this work, the ecological validity of the RT metric was demonstrated by the consistency of the results obtained in field and in auralized conditions using the same experimental layout.

5. Conclusions

The present study investigated the advantages of introducing the listening effort concept beside IS in the acoustical design of rooms for speech, and tested two metrics that are considered informative of the construct, one behavioral (RT) and one self-rating (LE). Four major observations were made:

(1) Speech-in-noise tests performed via headphones using auralized material were found in this work perceptually equivalent to *in situ* experiments, not only as regards the intelligibility results but also for the RT metric. The LE results were found to depend on the mode of presentation.
(2) Using a feasible metric to depict the complex construct of listening effort to complement traditional intelligibility results is a valuable strategy, which allows the discrimination of listening conditions equivalent as regards speech intelligibility. In fact, realistic modifications to the room acoustics yielding similar accuracy in word identification were found to change the amount of processing resources involved in the speech reception task, which can be monitored by using the two proposed metrics.
(3) Using the measure of response time it was possible to discriminate more listening conditions than using the self-reported effort. The finding suggests that RT might be a more sensitive metric than LE, providing information that positively contributes to the optimization of rooms for speech; its integration in the process of the acoustical design has been discussed.
(4) Concerning the effects of mother tongue on the speech reception task, it was found that non-native (but still highly proficient) listeners had a disadvantage at the perceptual level, which was not paired by changes in RT or LE results. An inconsistency between IS and LE results was found since the group which achieved the worst accuracy rated the task as being less effortful. More dedicated experiments are required to investigate the comparison of RT results of native and non-native listeners, with an effective objective qualification of language proficiency in order to better discern the effects of room acoustics.

Acknowledgements

This study was funded by the internal project "Human-Centered Design of the Built Environment: definition of a methodology for the experimental assessment of the overall Indoor Environmental Quality" of the Free University of Bozen-Bolzano.

Appendix A.

The distributions of the raw IS, RT and LE data are reported in the following Figures A1–A3. For each metric, three distributions were considered, one for each of the three statistical analyses performed and described in text in Sec. 3.1, Sec. 3.2 and Sec. 3.3. As the participants responses were coded as 0/1 (corresponding to wrong/correct word), the distribution of IS data is reported with reference to the average result over a listening condition (for each participant, ratio between the number of correct words and those presented in a given condition). Histograms and kernel density plots were obtained with the *R* software using the following commands: *hist*() and *density*().

It can be observed that, in all cases, the distribution of IS data departed from normality due to a prevalence of results close to the unity. As expected, the RT distributions were skewed with a long tail on the right. The distributions of LE data showed instead a high variability, with data similarly spread over all the levels of the response scale. Based on the observed distribution of the raw data and on literature evidences, the most appropriate GLMM models were selected for data analysis (e.g. a Gamma distribution with a log-link for RT data).

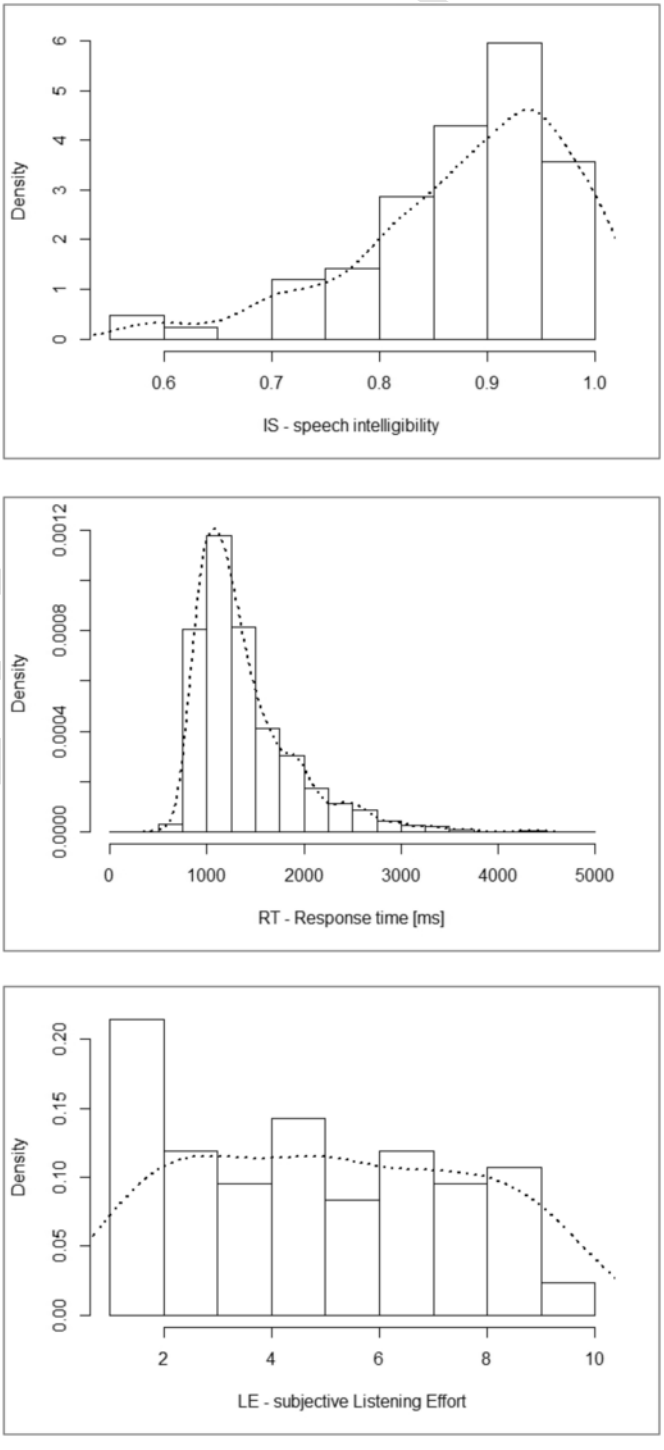
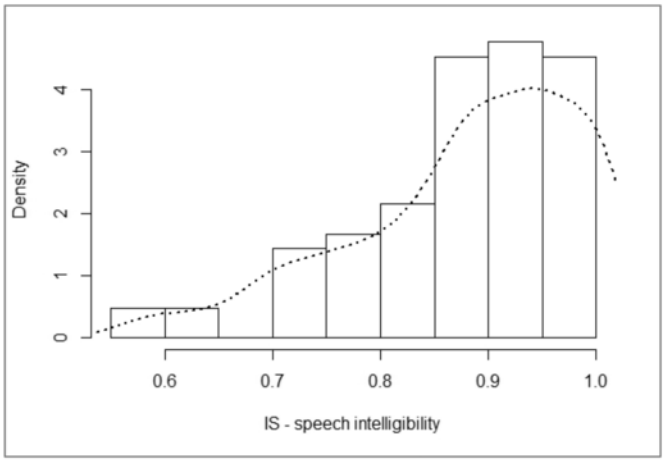
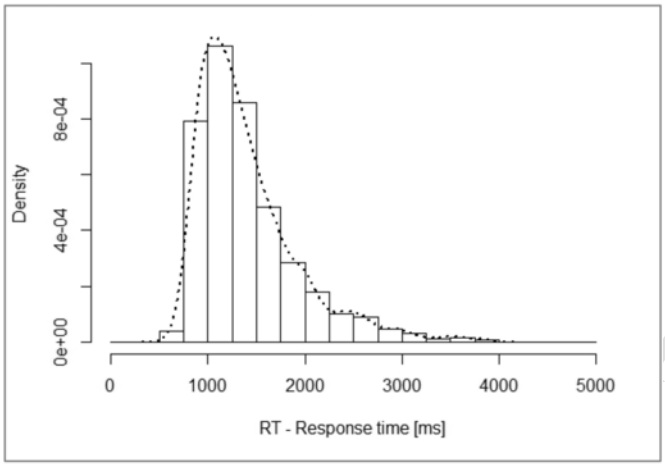


Fig. A1. Histograms and kernel density plot for the raw IS, RT and LE data. The distributions refer to the data used for the statistical analysis reported in Sec. 3.1.





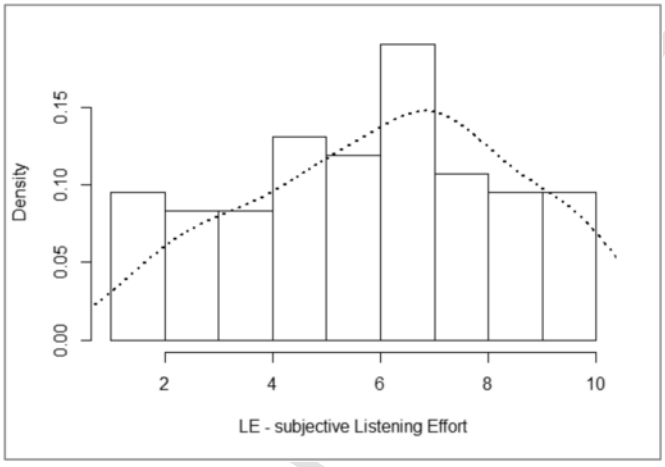
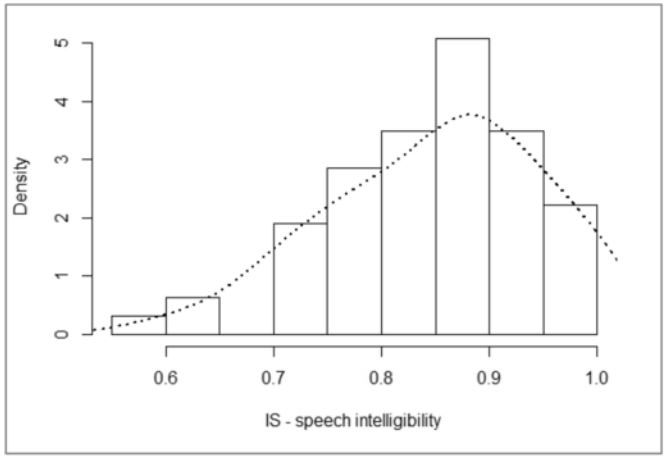
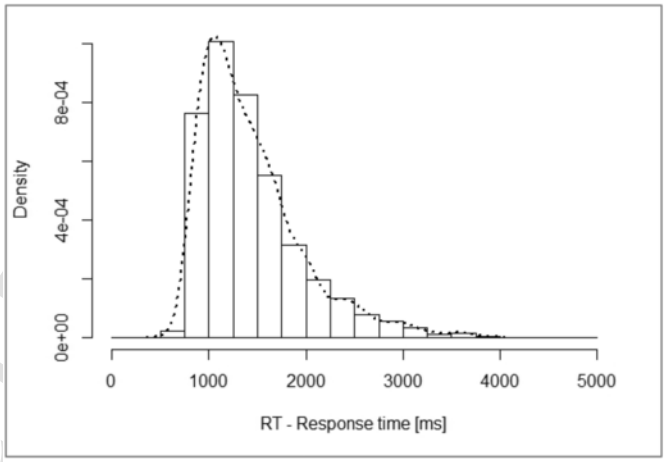


Fig. A2. Histograms and kernel density plot for the raw IS, RT and LE data. The distributions refer to the data used for the statistical analysis reported in Sec. 3.2.





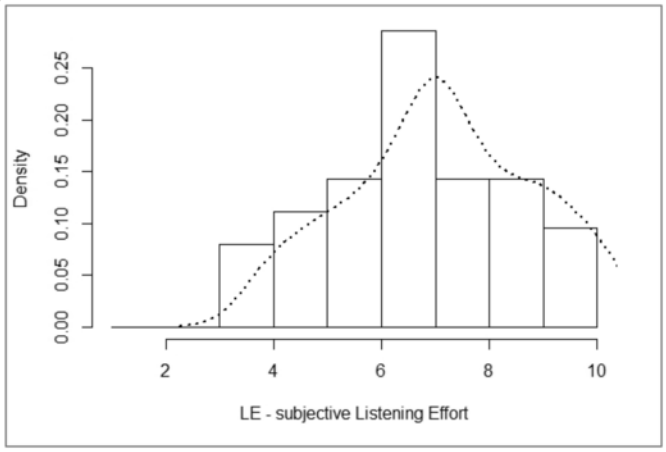


Fig. A3. Histograms and kernel density plot for the raw IS, RT and LE data. The distributions refer to the data used for the statistical analysis reported in Sec. 3.3.

References

[1] DIN 18041:2016–03, Hörsamkeit in Räumen – Anforderungen, Empfehlungen und Hinweise für die Planung (Acoustic quality in rooms – Specifications and instructions for the room acoustic design), 2016.
[2] IEC 60268-16, Sound System Equipment – Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index, 2011.
[3] ANSI/ASA S3.2:2009, Method for Measuring the Intelligibility of Speech over Communication Systems, 2009.
[4] ISO 9921:2003, Ergonomics – Assessment of Speech Communication, 2003.

[5] A. Kjellberg, Effects of reverberation time on the cognitive load in speech communication: theoretical considerations, Noise Health 7 (2004) 11–21.
[6] P.M. Rabbitt, Channel-capacity, intelligibility and immediate memory, Q. J. Exp. Psychol. 20 (1968) 241–248.
[7] M.K. Pichora-Fuller, S.E. Kramer, M.A. Eckert, B. Edwards, B.W. Hornsby, L.E. Humes, et al., Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL), Ear Hear. 37 (2016) 5S–27S.
[8] A.M. Surprenant, The effect of noise on memory for spoken syllables, Int. J. Psychol. 34 (1999) 328–333.
[9] D. Kahneman, Attention and Effort, vol. 1063, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[10] R. McGarrigle, K.J. Munro, P. Dawes, A.J. Stewart, D.R. Moore, J.G. Barry, S. Amitay, Listening effort and fatigue: what exactly are we measuring? A british society of audiology cognition in hearing special interest group 'white paper', Int. J. Audiol. (2014).
[11] S.J. Van Wijngaarden, H.J. Steeneken, T. Houtgast, Quantifying the intelligibility of speech in noise for non-native listeners, J. Acoust. Soc. Am. 111 (2002) 1906–1916.
[12] L. Kilman, A. Zekveld, M. Hällgren, J. Rönnberg, The influence of non-native language proficiency on speech perception performance, Front. Psychol. 5 (2014), article 651.
[13] K.B. Klink, M. Schulte, M. Meis, Measuring listening effort in the field of audiology – a literature review of methods (part 1), Z. Audiol. 51 (2012) 60–67.
[14] K.B. Klink, M. Schulte, M. Meis, Measuring listening effort in the field of audiology – a literature review of methods (part 2), Z. Audiol. 51 (2012) 96–105.
[15] A. Zekveld, S.E. Kramer, Cognitive processing load across a wide range of listening conditions: insight from pupillometry, Psychophysiology 51 (2014) 277–284.
[16] M. Morimoto, H. Sato, M. Kobayashi, Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces, J. Acoust. Soc. Am. 116 (2004) 1607–1613.
[17] N. Prodi, C. Visentin, A. Farnetani, Intelligibility, listening difficulty and listening efficiency in auralized classrooms, J. Acoust. Soc. Am. 128 (2010) 172–181.
[18] H. Sato, H. Sato, M. Morimoto, Effects of aging on word intelligibility and listening difficulty in various reverberant fields, J. Acoust. Soc. Am. 121 (2007) 2915–2922.
[19] J. Rennies, H. Schepker, I. Holube, B. Kollmeier, Listening effort and speech intelligibility in listening situations affected by noise and reverberation, J. Acoust. Soc. Am. 136 (2014) 2642–2653.
[20] F.N. Jones, M.J. Marcus, The subject effect in judgments of subjective magnitude, J. Exp. Psychol. 61 (1961) 40.
[21] J.-P. Gagné, J. Besser, U. Lemke, Behavioral assessment of listening effort using a dual-task paradigm: a review, Trends Hear. 21 (2017) 1–25.
[22] R. Houben, M. van Doorn-Bierman, W.A. Dreschler, Using response time to speech as a measure for listening effort, Int. J. Audiol. 52 (2013) 753–761.
[23] C. Pals, A. Sarampalis, H. van Rijn, D. Başkent, Validation of a simple response-time measure of listening effort, J. Acoust. Soc. Am. 138 (2015) EL187–EL192.
[24] D. Lewis, K. Schmid, S. O'Leary, J. Spalding, E. Heinrichs-Graham, R. High, Effects of noise on speech recognition and listening effort in children with normal hearing and children with mild bilateral or unilateral hearing loss, J. Speech Lang. Hear. 59 (2016) 1218–1232.
[25] N. Prodi, C. Visentin, A. Feletti, On the perception of speech in primary school classrooms: ranking of noise interference and of age influence, J. Acoust. Soc. Am. 133 (2013) 255–268.
[26] N. Prodi, C. Visentin, Listening efficiency during lessons under various types of noise, J. Acoust. Soc. Am. 138 (2015) 2438–2448.
[27] M. Klatte, T. Lachmann, M. Meis, Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting, Noise Health 12 (2010) 270–282.
[28] M.R. Hodgson, Experimental investigation of the acoustical characteristics of university classrooms, J. Acoust. Soc. Am. 106 (1999) 1810–1819.
[29] P. Ricciardi, C. Buratti, Environmental quality of university classrooms: subjective and objective evaluation of the thermal, acoustic, and lighting comfort conditions, Build. Environ. 127 (2018) 23–36.
[30] B.N. Postma, B.F. Katz, Perceptive and objective evaluation of calibrated room acoustic simulation auralizations, J. Acoust. Soc. Am. 140 (2016) 4326–4337.
[31] M. Rychtáriková, T. Van den Bogaert, G. Vermeir, J. Wouters, Perceptual validation of virtual room acoustics: sound localization and speech understanding, Appl. Acoust. 72 (2011) 196–204.
[32] W. Yang, M. Hodgson, Validation of the auralization technique: comparative speech-intelligibility tests in real and virtual classrooms, Acta Acustica united Acustica 93 (2007) 991–999.
[33] M. Hodgson, N. York, W. Yang, M. Bliss, Comparison of predicted, measured and auralized sound fields with respect to speech intelligibility in classrooms using CATT-Acoustic and ODEON, Acta Acustica united Acustica 94 (2008) 883–890.
[34] P. Zhu, F. Mo, J. Kang, G. Zhu, Comparisons between simulated and in-situ measured speech intelligibility based on (binaural) room impulse responses, Appl. Acoust. 97 (2015) 65–77.
[35] J. Reinten, P.E. Braat-Eggen, M. Hornikx, H.S. Kort, A. Kohlrausch, The indoor sound environment and human task performance: a literature review on the role of room acoustics, Build. Environ. 123 (2017) 315–332.
[36] R. McGarrigle, P. Dawes, A.J. Stewart, S.E. Kuchinsky, K.J. Munro, Pupillometry reveals changes in physiological arousal during a sustained listening task, Psychophysiology 54 (2017) 193–203.
[37] P. Bonaventura, F. Paoloni, F. Canavesio, P. Usai, Realizzazione di un test diagnostico di intelligibilità per la lingua italiana (Development of a diagnostic intelligibility test for the Italian language), International Technical Report No. 3C1286 Fondazione Ugo Bordoni, Rome, 1986.
[38] ISO 3382, Acoustics – Measurement of Room Acoustic Parameters – Part 1: Performance Spaces (2009); Part 2: Reverberation Time in Ordinary Rooms, 2008.
[39] R. Vitale, Perceptual Aspects of Sound Scattering in Concert Halls, vol. 21, Logos Verlag Berlin GmbH, 2015.
[40] A. Astolfi, V. Corrado, A. Griginis, Comparison between measured and calculated parameters for the acoustical characterization of small classrooms, Appl. Acoust. 69 (2008) 966–976.
[41] ISO 11654, Acoustics – Sound Absorbers for Use in Buildings – Rating of Sound Absorption, 1997.
[42] D. Pelegrín-García, J. Brunskog, Speakers' comfort and voice level variation in classrooms: laboratory research, J. Acoust. Soc. Am. 132 (2012) 249–260.
[43] J. Brunskog, A.C. Gade, G.P. Bellester, L.R. Calbo, Increase in voice level and speaker comfort in lecture rooms, J. Acoust. Soc. Am. 125 (2009) 2072–2082.
[44] R. Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017 https://www.R-project.org/.
[45] R. Whelan, Effective analysis of reaction time data, Psychol. Rec. 58 (2008) 475–482.
[46] M. van den Tillaart-Haverkate, I. de Ronde-Brons, W.A. Dreschler, R. Houben, The influence of noise reduction on speech intelligibility, response times to speech, and perceived listening effort in normal-hearing listeners, Trends Hear. 21 (2017) 1–13.
[47] S. Gatehouse, J. Gordon, Response times to speech stimuli as measures of benefit from amplification, Br. J. Audiol. 24 (1990) (1990) 63–68.
[48] R. Ratcliff, Methods for dealing with reaction time outliers, Psychol. Bull. 114 (1993) 510–532.
[49] R.H. Baayen, P. Milin, Analyzing reaction times, Int. J. Psychol. Res. 3 (2010) 12–28.
[50] T.F. Jaeger, Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models, J. Mem. Lang. 59 (2008) 434–446.
[51] H. Quené, H. Van den Bergh, Examples of mixed-effects modeling with crossed random effects and with binomial data, J. Mem. Lang. 59 (2008) 413–425.
[52] D. Bates, M. Maechler, B. Bolker, S. Walker, Fitting linear mixed effects models using lme4, J. Stat. Software 67 (2015) 1–48.
[53] S. Lo, S. Andrews, To transform or not to transform: using generalized linear mixed models to analyze reaction time data, Front. Psychol. 6 (2015), article 1171.
[54] R.H.B. Christensen, Ordinal – Regression Models for Ordinal Data, R package version, 20156–28 http://www.cran.r-project.org/package=ordinal/.
[55] R.V. Lenth, Least-squares means: the R package, J. Stat. Software 69 (2016) 1–33.
[56] M.K. Pichora-Fuller, How social psychological factors may modulate auditory and cognitive functioning during listening, Ear Hear. 37 (2016) 92S–100S.
[57] S. Hygge, Classroom noise and its effect on learning, 11th International Congress on Noise as a Public Health Problem (ICBEN), Nara, JAPAN, vol. 4, June 2014.
[58] M.L.G. Lecumberri, M. Cooke, A. Cutler, Non-native speech perception in adverse conditions: a review, Speech Commun. 52 (2010) 864–886.
[59] A. Wagner, M. Ernestus, Identification of phonemes: differences between phoneme classes and the effect of class size, Phonetica 65 (2008) 106–127.
[60] M. Broersma, A. Cutler, Phantom word activation in L2, System 36 (2008) 22–34.
[61] Z.E. Peng, L.M. Wang, Effects of noise, reverberation and foreign accent on native and non-native listeners' performance of English speech comprehension, J. Acoust. Soc. Am. 139 (2016) 2772–2783.
[62] A. Lam, M. Hodgson, N. Prodi, C. Visentin, Effect of classroom acoustics on speech intelligibility and response time: a comparison between native and non-native listeners, Build. Acoust. 25 (2018) 35–42.
[63] T. Neher, G. Grimm, V. Hohmann, Perceptual consequences of different signal changes due to binaural noise reduction: do hearing loss and working memory capacity play a role?, Ear Hear. 35 (2014) e213–e227.
[64] B. Larsby, M. Hällgren, B. Lyxell, S. Arlinger, Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects, Int. J. Audiol. 44 (2005) 131–143.
[65] D.L. Valente, H.M. Plevinsky, J.M. Franco, E.C. Heinrichs-Graham, D.E. Lewis, Experimental investigation of the effects of the acoustical conditions in a simulated classroom on speech recognition and learning in children, J. Acoust. Soc. Am. 131 (2012) (2012) 232–246.
[66] L. Fontan, J. Tardieu, P. Gaillard, V. Woisard, R. Ruiz, Relationship between speech intelligibility and speech comprehension in babble noise, J. Speech Lang. Hear. 58 (2015) 977–986.
[67] A. Kjellberg, R. Ljung, D. Hallman, Recall of words heard in noise, Appl. Cognit. Psychol. 22 (2008) 1088–1098.