

Listening Test Methodology for Building Acoustic Issues

Michiel Geluykens^{1,2,3*}, Herbert Müllner², Vojtech Chmelík², Monika Rychtarikova^{2,3}

¹ *Technologisches Gewerbemuseum TGM, Wexstrasse 19, 1200 Vienna, Austria*

² *STU Bratislava, Dep. of Materials Engineering and Physics, Radlinského 2766/11, 810 05 Bratislava, Slovakia*

³ *KU Leuven, Dep. of Architecture, Campus Brussels and Ghent, Paleizenstraat 65/67, 1030 Brussels, Belgium*

* Corresponding author: Michiel.geluykens@tgm.ac.at

Abstract

Sound insulation measurements can accurately characterize the physical attenuation of outdoor sounds by the façade. Nevertheless, to understand the influence of the attenuation on our perception, further investigations are necessary. Listening tests provide a framework for evaluating the perception of sound in a controlled laboratory environment. In a number of studies, listening tests were used to study the subjective assessment of sound insulation. However, among these, few have reported the impact of the methodology on their studies. This work presents the progress and investigations regarding the listening test methodology for evaluating the influence of façade sound insulation on the perception of outdoor noises. In the first experiment, two response collection methods were employed to compare their performance and explore the statistical analysis of the responses. Moreover, in a second experiment, the influence of non-ideal acoustic conditions of the listening test setup is investigated. The results support the scaling method as the preferred method for the proposed context. Moreover, the acoustic conditions of the listening test environment should not be compromised.

Introduction

The response collection method is of crucial importance for the listening test. On the one hand, the ease of the interaction with the participants determines the accuracy and preciseness of the responses. On the other hand, the format of the data will determine which statistical analysis can be applied, and therefore, which information can be extracted. The preferred response collection should consider both aspects. Moreover, the listening test environment affects how the stimuli reach the participants. The background noise, room acoustic conditions, and play-back system determine how the stimuli are reproduced and propagated to the participant.

In the current literature regarding sound insulation, scaling methods have dominated [1]. The participants are presented with a single stimulus and are asked to rate the magnitude of a particular attribute on a scale (e.g. loudness or annoyance). However, the scaling method is subject to various forms of bias, such as avoidance of the use of the endpoints, rubber scale, and sequence effects [2]. These biases complicate the interpretation and analysis of the responses, however, their impact in the context of the intended experiments regarding sound insulation is unclear. Comparison methods have been proposed as a less biased alternative [3]. The participant must only compare the stimuli according to the attribute and is not confronted with having to quantify the magnitude of the perception. However, therefore the perceptual magnitude is not directly represented in the responses. While both methods have their advantages and disadvantages, it ultimately

depends on the context and research goals of the experiment as to which method is most suitable.

Regarding the listening environment, in the current studies, the stimuli are presented in listening rooms or anechoic laboratories. While the background noise in the current literature varies over 10 dB, the influence of the environment on the listening test results has not yet been investigated.

While literature exists investigating the subjective evaluation of the influence of sound insulation on perception, few have investigated the response collection method and listening environment as a part of the experimental design. Therefore, in this study three listening tests were done where the same stimuli were evaluated with the scaling and comparison method, and in two different environments: a listening room as an ‘ideal acoustic environment’, and a seminar room as a ‘non-ideal acoustic environment’. In the following sections, first, a description of the experimental conditions is given. Next, the influence of the response collection method and listening test environment are discussed.

Methodology

Stimuli

4 outdoor environmental sounds considered as relevant to façade sound insulation were selected for the listening tests: pink noise, road traffic, airplane, and railway noise. They were 5s excerpts with limited temporal variation and were extracted from calibrated recordings, however, the average SPL was matched to 91,5 dBA. Their A-weighted spectra are presented in Figure 1. The stimuli for the listening test were created by filtering these sounds with 4 sound insulation spectra shown in Figure 2: two lightweight and two heavyweight constructions, in each case one with and without ETIC system, but all with a weighted sound reduction index R_w of 51 dB. The sound insulation was applied as a 1/3rd octave band filter. The stimuli were presented in different environments, which are described in later sections, however, in both cases, the output of the reproduction system below 50 Hz was negligible. The A-weighted level of the resulting stimuli is shown in Table 1.

Experiment 1: Response collection method

To investigate the influence of the response collection method, the stimuli were presented using the comparison and scaling method. In the scaling listening test, an 11-point scale with semantic labels for the endpoints of the scale and linear numeric labels in between was used. The stimulus was played for 5s and repeated after a 2s break. Then, the participant was given sufficient time to rate the loudness in a form. The test consisted of 16 stimuli, which were each presented 5 times in a randomized order, although the order was the same for all participants. To reduce possible sequence effects, it was

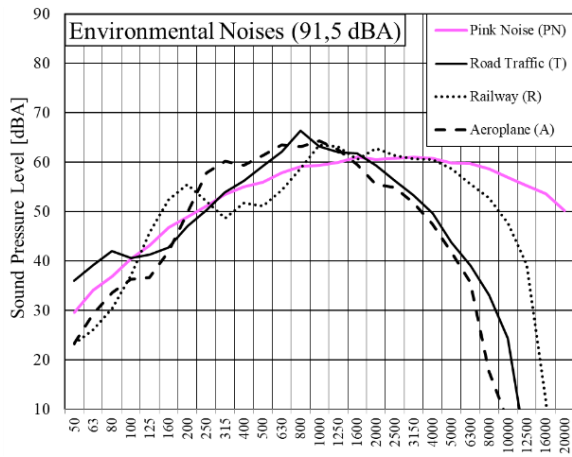


Figure 1: A-weighted spectra of the environmental sounds.

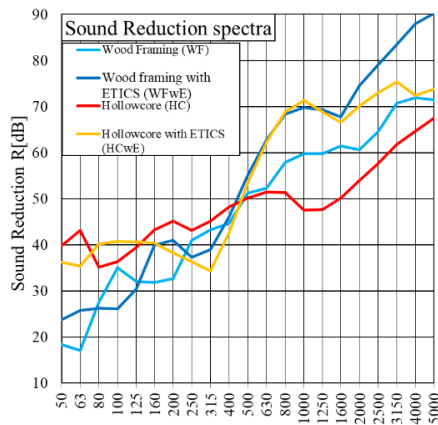


Figure 2: Sound insulation spectra of the considered facades.

Table 1: Sound Pressure level of the stimuli Please put

SPL of stimuli [dBA]	Environmental sound			
	Pink Noise (PN)	Car Traffic (T)	Airplane (A)	Railway (R)
Original level	91,5			
Stimuli Level (Filtered in 1/3rd octave bands)				
Wood Framing	43,1	46,5	43,9	47,0
Wood Framing + ETICS	41,5	43,4	44,7	45,6
Hollowcore	39,9	42,2	43,2	42,3
Hollowcore + ETICS	41,6	42,1	47,5	41,8

guaranteed that subsequent stimuli were not from the same sound source. In total each participant gave 80 ratings, which took 26 minutes. Before the start of the experiment, a selection of stimuli was presented to indicate the loudness range of stimuli to the participants, and they were asked to use the scale entirely. The listening test was preceded by three tryouts to familiarize the participants with the test.

In the comparison method, the participants evaluated the stimuli in a 2-Alternative Choice (2-AC) method wherein two stimuli were presented, and they were asked which of them had more of a certain attribute (in this case: loudness). Two stimuli of 5s were presented with a 2s break in between, after which they were repeated. The participants were asked to respond ‘A’, ‘B’, or ‘Equal/not sure’. The experiment was initiated with two tryouts. In the listening test itself, each comparison was between the same environmental sound with different sound insulations applied. Comparisons between different environmental sources were not included to keep the listening test to a reasonable length (32 min). The order of

comparisons was randomized, and return comparisons were included (S1 – S2 were presented, but also S2 – S1), which led to a total of 48 comparisons.

To leverage the benefit of testing several subjects simultaneously, the experiments comparing the response collection methods took place in the seminar room.

Experiment 2: Listening test environment

To explore the influence of the listening environment, the stimuli were evaluated under two conditions, but using the same scaling method. As the ‘ideal acoustic environment’ the listening room has a low background noise level of 15 dBA, and a reverberation time of 0,3 s. The stimuli were presented to individual participants over a 16-loudspeaker ambisonics system, where a plugin introduced a sense of directivity from the façade of the room. Additionally, as a ‘non-ideal acoustic environment’, the listening test was done in a seminar room with groups of participants simultaneously. The stimuli were reproduced by two Neumann KH-120A loudspeakers connected to an external Roland Octa-Capture soundcard. According to the manufacturer, the loudspeakers have a flat characteristic from approx. 50 Hz. The reverberation time in the room was 0,5s. The background level in the room was relatively stable at 22 dBA. This was measured, however, without the presence of the participants in the room. For both environments, the reproduction system was calibrated by adjusting the gain so that the SPL in the room of the stimulus WF-PN matched the calculated level.

Participants

For the scaling listening test in the seminar room three groups of students took part for a total of 20 participants (7 + 4 + 9), the age of the students ranged between 18 and 22 years old, with a mean age of 19. Half of the participants were male. For the comparison listening test in the seminar room, two groups of students took part for a total of 12 participants (5 + 7), the age of the students ranged between 17 and 21 years old, with most students 18 or 19. 11 out of 12 participants were male. For the scaling listening tests in the listening room, 3 colleagues from the institute (non-experts in acoustics) were tested in addition to 7 students, therefore the age ranged between 18 and 61, with a mean of 25 and median of 19,5. 9 out of 10 participants were male.

Results & discussion

Experiment 1a: Response collection method - Scaling

The distribution of the responses on the scale may be interpreted as the perceptual magnitude of the attribute under study. The mean ratings for the stimuli are presented in Figure 3. As the scale had 11 options, the responses can be considered as continuous which allows parametric analysis [2]. The Q-Q plots and histograms showed that the data can be assumed to be normally distributed. Firstly, an ANOVA with the participants as the independent variable and responses as the dependent was significant. Therefore, the participants should be incorporated as a factor in the subsequent analysis. Secondly, to investigate the influence of sequence effects, which is that the response on the n^{th} stimulus is dependent on the $(n-1)^{\text{th}}$ stimulus, the correlation between the responses on these stimuli was investigated. The sequence

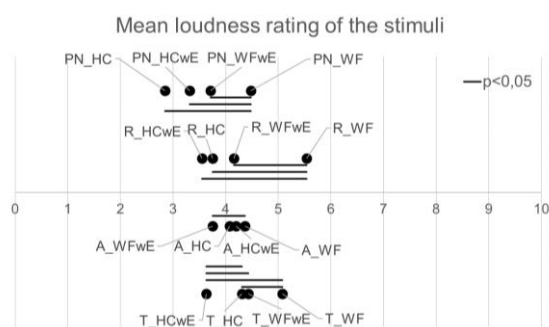


Figure 3: Mean loudness ratings for the stimuli collected with the scaling method in the seminar room

effects are related to the decisional strategy by the participant, which is individual, moreover, the use of the scale was dependent on the participants. Therefore, they were evaluated using Pearson correlation for each participant. The correlation was only significant for 1 of the 20 participants, and therefore does not appear to be systematic.

To incorporate the influence of the sound insulation, environmental sound, and the participants, a Two-Way Repeated Measures ANOVA was constructed. As the use of the scale was dependent on the participant and each participant rated each stimulus, the ratings by one participant can be considered as a repeated or within-subject measure. Each participant rated each stimulus 5 times throughout the listening test, unfortunately, this model does not allow to incorporate these multiple observations. Therefore, these 5 ratings were represented by their mean. Regarding the assumptions for the ANOVA model, normal distribution was given, but Mauchly's test showed that sphericity was violated, this was subsequently accounted for in the ANOVA model. The model showed a significant dependency of the ratings on both sound insulation ($p < 0,0001$) and environmental sound ($p < 0,0001$). Moreover, also their interaction ($p < 0,0001$) was significant, indicating that the influence of sound insulation depends on the source sound. To investigate the significance of the differences between the individual stimuli, pairwise post-hoc dependent sample t-tests were made among stimuli based on the same source sound. The significance of the difference in rating between stimuli based on Bonferroni-corrected p-values in 11 cases and is indicated by the horizontal lines in Figure 3.

The benefit of the scaling method is that the ratings by the participants can be directly interpreted as the perceptual magnitude of the stimuli and can therefore be directly used as the dependent variable in a Pearson correlation analysis. Moreover, a plot of the rating and a regression model can present those stimuli that deviated from the expected behavior, opening the door to investigate the properties that resulted in their rating. An example of the correlation between subjective loudness and calculated loudness (ISO 531-1) is presented in Figure 4. The same strategy can be applied if the ratings correspond to the sound insulation described by single-number quantifiers.

Experiment 1b: Response collection method-Comparison

In the comparison method, the participants reported which of the stimuli they experienced as louder, see Figure 5. It was

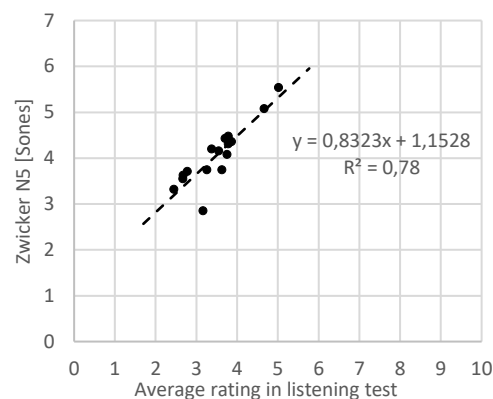


Figure 4: Correlation between calculated and subjective loudness.

expected that the order of the stimuli in the comparison may influence the responses. Therefore, this influence was tested using a Chi-square test which compares the distribution of a categorical variable (the votes for A, B, and Equal/not sure) among conditions (normal vs. return). The test result showed that there was a significant association between stimuli order and votes, $\chi^2(2, N = 575) = 6,4654$, $p = 0,03945$. This indicates that the stimulus presented second is more likely to be selected as louder. To deal with this, the return comparisons should continue to be included in future listening tests.

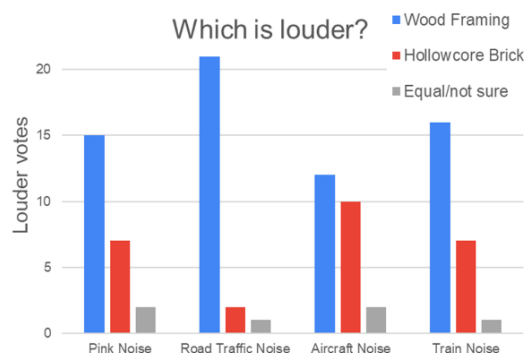


Figure 5: Responses in the comparison method: 1 set of results of the comparison of different sounds of timber frame wall (WF) and a wall of hollow core bricks (HC)

While in some comparisons one stimulus clearly dominated, in other comparisons, the distribution of votes for both stimuli is closer to equal. Here it is necessary to test if the difference in votes is significant. The current experiment was not a strict 'forced choice' experiment because it included an 'equal/not sure' answer option. This option only accounted for less than 4% of the responses. Moreover, at most, it made up for 3 out of 12 responses in one comparison. To facilitate the statistical analysis, these responses were ignored, and the statistical significance of the vote distribution was analyzed using a binomial test with a probability parameter of 0.5 (the aim is to reject the null hypothesis of no loudness difference, under which an equal distribution of votes for both stimuli would be expected). The results indicated that there was a significant difference in loudness perception in 13 out of 24 comparisons of different sound insulation applied to the same environmental sound, see Table 2. Therefore, sound insulation characteristic can be expected as an influencing factor.

Table 2: Sign. of binomial test (x) in the comparison method

Comparison	Environmental sound			
	PN	R	A	T
WF-WFwE	x	x	x	x
WF-HC		x		
WF-HCwE	x	x	x	x
WFwE - HC			x	
WFwE - HCwE				x
HC - HCwE		x		x

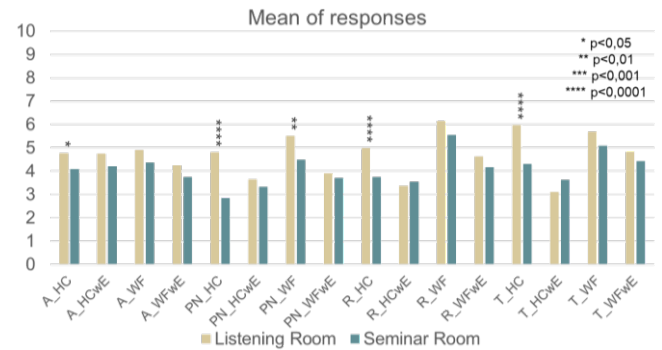
Unlike for scaling, the perceptual magnitude is not represented in the responses itself, which complicates the next step of the analysis (quantification of the relationship between variables). In a first attempt, the correlation of the 5 percentile loudness values in Sones according to ISO 532-1 with the 'loudness votes' for each stimulus is evaluated. It was opted to use Spearman correlation since it applies ordinal values: The votes are ordinal as they present the ranks of the wall structures but do not present the perceptual distance. The result was significant and showed a moderate Spearman's Rho of 0.65.

Models have been developed to derive the perceptual distance between stimuli based on comparison methods. de la Prida et al. applied Thurstonian models in a listening test on sound insulation [3]. These models are based on Signal Detection Theory, in which the magnitude of the perception of a stimulus is normally distributed. In a comparison of stimuli, the distributions of two stimuli will overlap depending on the perceptual distance between the stimuli. In Thurstonian models, it is assumed that the distribution of responses is related to the overlap of the perceptual distributions, and therefore to the perceptual distance between the presented stimuli. This method applies to stimuli with small perceptual differences. However, when the difference between the stimuli is large, there is no more overlap of the perceptual distributions and the Thurstonian models get no usable results. As shown in Figure 5, in some comparisons, there was only a marginal difference in votes, while for others the one stimulus clearly dominated. This raises concerns about the applicability of Thurstonian models in the context of our experiment.

Experiment 2: Listening test environment

To explore the influence of the listening test environment, the stimuli were evaluated with the scaling method in a listening room as an 'ideal acoustic environment', and a seminar room as a 'non-ideal acoustic environment'. As the stimuli and method were the same for both environments, the responses can be directly compared. The mean ratings of the stimuli in each environment are presented in Figure 6. Furthermore, a two-way ANOVA with the environment and stimulus as factors was constructed. The assumption of normality was met, however, Levene's test indicated unequal variances. The variance in the responses per stimulus was systematically lower in the listening room, although rarely with significance according to Levene's test. The unequal variance was accounted for in the ANOVA model. Evidently, the model indicated a significant effect of the stimulus ($p < 0,0001$), but more interestingly, also of the environment ($p < 0,0001$), and the interaction between the stimulus and environment ($p < 0,0001$). In the listening room, the stimuli were

consistently rated as louder, this is most likely due to the higher background noise and presence of multiple participants in the seminar room. However, a post-hoc pairwise independent sample t-test, with p-values Bonferroni corrected showed that the difference was only significant for 5 out of 16 stimuli. The responses in the listening and seminar room were only moderately correlated with a Pearson R^2 of 0,51.

**Figure 6:** Mean of responses in the different environments

Conclusions

Although the number of participants was limited, the three listening tests in this study revealed interesting insights on the listening test methodology.

Regarding the response collection method, the comparison method revealed perceptual differences between stimuli more effectively. However, the analysis of this method was restricted to non-parametric methods, while the scaling method is compatible with parametric analysis. The latter allows to identify influencing factors in more detail. Moreover, the scaling method allowed to quantify the extent of the perceptual difference between stimuli, which can be exploited to model the influence of sound insulation characteristics on the perception.

Regarding the listening test environment, the mean results of the same listening test collected in a seminar room and dedicated listening room were only moderately correlated. The variance of the responses was lower in the listening room, although without statistical significance.

These results support the scaling method as the preferred method for the proposed context. Moreover, the results suggest that the particular acoustic conditions of dedicated listening test environments matter if the accuracy of listening tests for building acoustics issues should be increased.

References

- [1] Geluykens, M.; Müllner, H.; Chmelík, V.; Rychtarikova, M.: Airborne sound insulation and noise annoyance: Implications of listening test methodology. Forum Acusticum, Torino, Italy. 2023.
- [2] Bech, S.; Zacharov, N.: Perceptual Audio Evaluation – Theory, Method and Application. ISBN: 978-0-470-86923-9, 2006
- [3] de la Prida, D.; Pedrero, A.; Navacerrada, M. A.; Díaz-Chyla, A.: Methodology for the subjective evaluation of airborne sound insulation through 2-AC and Thurstonian models. Applied Acoustics 157 (2020), 107011.