DOI: 10.61782/fa.2025.0756

FORUM ACUSTICUM
EURONOISE 2025

ROBUSTNESS OF AN EQUAL LOUDNESS PARADIGM TO STIMULI, ENVIRONMENT AND PARTICIPANTS

Michiel Geluykens [1,2,3*] **Herbert Müllner** [2] **Vojtech Chmelík** [3] **Monika Rychtarikova** [1,3]
[1] KU Leuven, Dep. of Architecture, Campus Brussels and Ghent, Belgium
[2] Technologisches Gewerbemuseum TGM, Vienna, Austria
[3] STU Bratislava, Faculty of Civil Engineering, Dep. of Materials Engineering and Physics, Slovakia

ABSTRACT

The equal-loudness matching method is well-suited for assessing the overall loudness of complex, time-varying sounds. In this method, participants adjust the subjective intensity of a stimulus to the same overall loudness as a reference. This study evaluated its sensitivity to methodological factors, focusing on three aspects. First, the impact of the comparisons stimuli' spectra was examined. Typical urban sound recordings were used as references, with adjustable comparisons generated from pink noise filtered to the same average spectrum at varying resolutions (1/1, 1/3 octave bands, or FFT). For the 1/1 and 1/3rd octave bands, the influence of filter shape — flat within each band or interpolated between the centre frequencies — was also analyzed. Second, the test environment's effect was studied in three settings: a living room-like furnished listening room, a semi-anechoic room, and an uncontrolled condition where participants used their own laptops and headphones. Third, responses from acoustics experts and non-experts were compared, given that colleagues are often recruited for listening tests. Results demonstrate the robustness of the equal-loudness matching method across these methodological variables, supporting its reliability for diverse applications.

Keywords: *listening test, equal loudness matching, methodology*

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

1. INTRODUCTION

Excessive noise has a negative impact on people's quality of life [1]. Since we spend most of our time indoors, sound insulation significantly reduces the amount of outdoor noise we experience. The sound-insulating performance of buildings and elements is summarized by single-number ratings, which are described in ISO 717-1 [2]. For these ratings to be effective, their perceptual relevance is crucial, and listening tests play a key role in establishing this link. In our research, we applied such a listening test to evaluate whether the loudness perception of temporally varying sound corresponds to that of a steady-state noise with the same average spectrum and sound pressure level, one of the key assumptions underlying current sound insulation ratings. This application of a listening test is a prime example of how such methods are used not only in fundamental but also applied research. However, to derive meaningful conclusions, the methodological choices of the listening test need to be carefully considered.

The method used in this study was the 'equal loudness matching' paradigm, where in this case, the participants adjust the level of a steady-state comparison noise to equal-loudness of a time-varying outdoor sound as the reference. The steady-state comparison noise is created to have the same average spectrum as the temporally varying reference and is only varied in its overall level. This spectral matching is done for two reasons: Methodologically, matching the spectral contributions in the stimuli pairs is assumed to help the participants make their judgement based on overall loudness and ignore timbral differences. Moreover, by evaluating the relative levels of the equally loud reference and comparison sounds, the responses with this method provide direct insight into the equal loudness of same level and spectrum noise with steady-state and time-varying character.

While the equal loudness matching paradigm is the golden standard in psychoacoustics [3], some methodological factor's impact remain less explored: 1) Participant selection should ideally be random, however, due to resource limitations, studies often resort to colleagues with an acoustic background. These participants may listen differently compared to normal people (e.g. more analytically), or being aware of the research context may make them more subject to experimenters-bias, skewing the findings. 2) The listening tests take place in a range of different acoustic environments: from acoustic laboratories to classrooms and office spaces. Perception is undoubtedly a multisensory phenomenon, and therefore, the visual environment of the listening test should be considered. 4) Moreover, the quality and control over the stimuli reproduction system may affect the participant's responses. 4) Finally, in the equal loudness matching paradigm as described above, filtering the comparison noise to the same average spectrum as the reference could be achieved through filters of different shapes and resolutions. While a filter in full FFT resolution would incorporate more tonal components of the reference in the comparison, the sound insulation rating systems summarize temporally varying sounds in average spectra in 1/1 or 1/3rd octave bands. Moreover, the filters in 1/1 or 1/3rd octave bands could be created with accurate and sharply defined bands, resulting in a spectrum with steep transitions, or a smoothened interpolated filter.

This paper explores these methodological aspects of participants, environment and comparison stimuli within the context of the equal loudness matching paradigm. First, in the next section, the listening test paradigm and listening tests are described in more detail. Following, the results and findings regarding the influences of participants, environment and stimuli are presented.

2. EQUAL LOUDNESS MATCHING

In the equal-loudness paradigm, the participants first listened to the reference stimulus (RS), which was always a realistic outdoor recording of a sound event with temporal variation. They then adjusted the level of the comparison noise (CN) to equal subjective loudness. In all listening tests, the participants controlled the CN level through the GUI shown in Figure 1. In previous experiences with this listening test paradigm, it was found that not only the RS but also the direction of adjustment significantly impacted the participants' final response, with differences of over 4 dB in the finally selected CN [4]. Therefore, each matching was made two times for each pair of RS and CN: Starting from the lowest level CN, where the participants needed to increase the level, and vise versa. The participants started the matching at the extreme level of +/-16 dBA of CN relative to RS. They could make adjustments in either smaller steps of +/-2 dB or larger steps of +/-4 dB. The order of the matchings was randomized. The responses were coded as the final level difference between the RS and CN perceived equally loud. In all following analyses, the direction of adjustment and RS were included as factors in the model, in addition to the variable under investigation (participant's background, environment, or CN type).

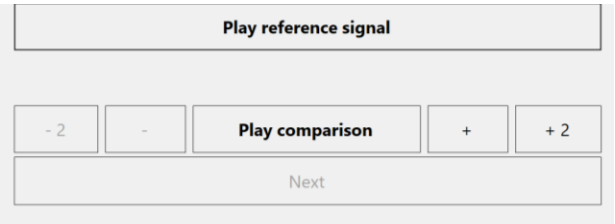


Figure 1. Listening test interface.

In this paper, the results from four listening tests are presented. In total, 65 participants took part. The aim of listening test #1 was to evaluate the impact of participants' background. 16 Participants, 5 non-experts and 11 experts, matched the loudness of a CN to 10 RS in both directions. In listening test #2, those same stimuli were matched by 19 participants in a different laboratory to evaluate the impact of the environment. In listening test #3, 11 participants evaluated a subset of the stimuli of listening test #4 in an uncontrolled fashion. Finally, in listening test #4, 20 participants matched five variations of CN to three RS to investigate the impact of CN creation on the responses.

The listening tests took place in different environments. The test environment for listening tests #2 and #4 at Technologisches Gewerbemuseum (TGM) in Vienna was a dedicated listening room that was acoustically treated for low background noise and reverberation, but furnished to resemble a living room. For listening test #1 at KU Leuven (KUL) the test took place in a semi-anechoic room. Both rooms at TGM and KUL had low background noise levels (<20 dBA) and similar reverberation times. Moreover, at both TGM and KUL, the stimuli were presented over a two-loudspeaker setup (Neumann KH12A) with a reasonably flat frequency response above 50 Hz, which was confirmed through an impulse response measurement. Before each listening test session, the reproduction system's level was calibrated by adjusting the gain on the soundcard so that the level of pink noise matched its intended value. In listening test #3, the participants performed the test in an

uncontrolled fashion: in a conference room using their own laptop and headphones without stimuli calibration. An overview of the listening tests is presented in Table 1.

Table 1. Overview of listening tests

LT	#n	Aim - Impact of:	Environment	RS	CN
1	16	• Participants' background (sect. 3.1) • Environment (sect. 3.2)	KUL	10	1
2	19	• Environment (sect. 3.2)	TGM	10	1
3	10	• Uncontrolled reproduction (sect. 3.3)	Worksh op	3	3
4	20	• CN type (sect. 3.4) • Uncontrolled reproduction (sect. 3.3)	TGM	3	5

3. RESULTS

3.1 Participants

Participant selection should ideally be random and depending on the topic of the study, should be representative of the society. However, as often seen, also in this study colleagues took part in the listening test due to resource limitations. To evaluate the influence of participants' background, the participants in listening test #1 were indicated as expert listeners when they worked in the field of acoustics. The responses grouped by participants' background are presented in Figure 2. The mean response of the expert listeners (mean -1.0 dB, SD 5.0 dB) appears to be higher compared to non-experts (mean -3.8 dB, SD 5.5 dB). The impact of the participants' background was statistically evaluated in a repeated-measures ANOVA with RS and direction of adjustment as within-subject variables, and the participants' background as a between-subject variable. The background was not a significant factor in this model (F(2,13)=[2.324] p=.15), providing no support to the claim that the expert listeners responded differently compared to normal listeners.

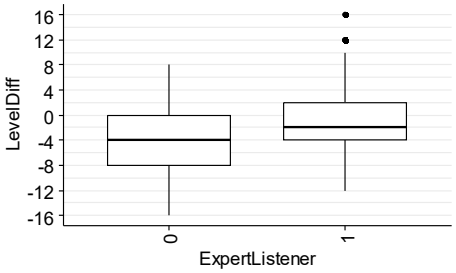


Figure 2. Impact of participants' background.

3.2 Environment

To evaluate the impact of the listening test environment, the test at KUL was also done at TGM. While the acoustic conditions (background noise, reverberation and reproduction) were similar in the rooms at TGM and KUL, they presented a completely different visual environment: The listening room at TGM appeared like a comfortable living room, in contrast, the semi-anechoic was a more unusual visual experience for the participants. In total, 35 participants took part in the two listening tests, of which 16 at KUL and 19 at TGM. Although no statistical evidence for an impact participants' background was found in the previous section, it must be noted that in this analysis, the variables of background and environment were largely confounded as at TGM all but one participants were non-experts, while at KUL the majority were experts. The mean of responses at TGM (mean -0.9 dB, SD 5.5 dB) was slightly higher compared to KUL (mean -1.9 dB, SD 5.3 dB), see Figure 3. The impact of the environment was analyzed in a RM-ANOVA with RS and direction of adjustment as within-subject variables, and the environment as a between-subject variable. The environment was not a significant factor in this model (F(1,32)=[0.543] p=.46), presenting no indication that the listening test environment affected the responses.

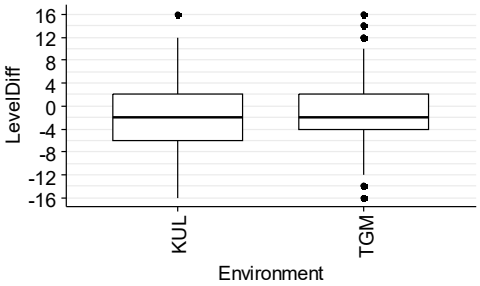


Figure 3. Impact of environment.

3.3 Uncontrolled conditions

At a workshop, 11 participants performed the listening test in an uncontrolled fashion. The participants matched the loudness of CN with different filter resolutions (1/1 octave bands, 1/3rd octave bands and FFT resolution, stepped filters only) to three RS. The same RS and CN stimuli were also evaluated by 20 participants in the listening room at TGM as part of a controlled experiment in listening test #4. The mean of responses of the uncontrolled condition (mean 0.7 dB, SD 4.9 dB) and the controlled (mean 0.3 dB, SD 3.9 dB) are presented in Figure 4 and are similar. Again, a RM-ANOVA with sound source, CN type, and direction as

within-subject variables, and listening test conditions as the between-subject variable was constructed. All assumptions were met. The data did not indicate the uncontrolled conditions to be an influencing factor regarding the responses (F(1,28)=[0.804] p=.37).

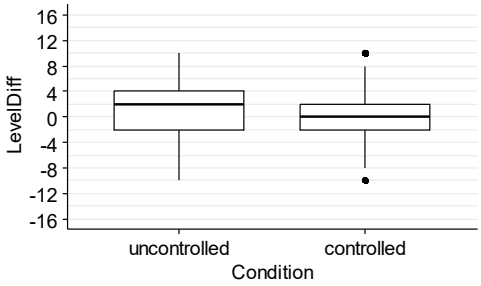


Figure 4. Impact of listening test conditions.

3.4 Stimuli

While the RS were recordings of real sound events, the CN were derived from random pink noise filtered to have the same average spectrum as the RS. To evaluate the impact of the filter resolution for the CN creation, noises were created with a filter based on the average spectra of RS in full octave bands, 1/3rd octave bands, or the FFT resolution. The resulting A-weighted FFT spectra of the CN created with these three resolutions are compared in Figure 5 for a single RS.

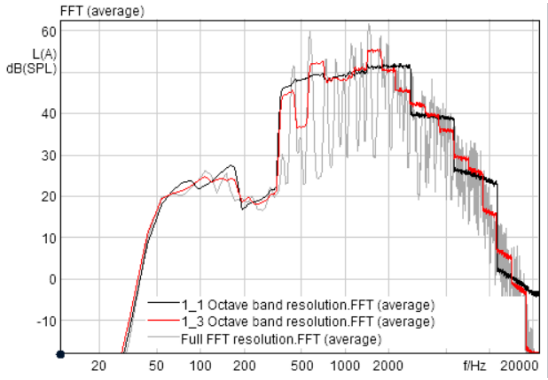


Figure 5. Comparison of filter resolution for CN

Moreover, for the 1/1 and 1/3rd octave bands, the influence of the shape of the filter is explored through a flat frequency spectrum within each band, resulting in a stepped shape, or interpolated between the center frequencies, presented for the 1/1 octave band resolution in Figure 6 (but also applied to the 1/3rd octave band filter).

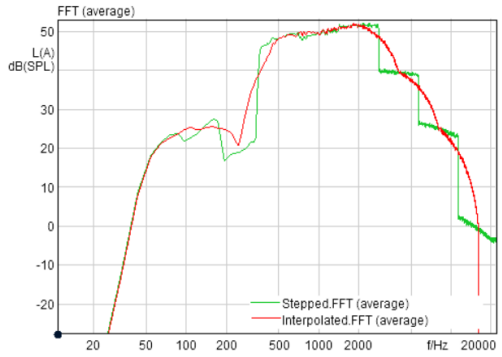


Figure 6. Comparison of filter shape for CN

In listening test #4, these 5 CN types (1/1 and 1/3rd octave bands with stepped and interpolated filter, and the FFT resolution) were used in equal-loudness matchings of three RS stimuli by 20 participants. Differences in responses due to CN type were analyzed in a RM-ANOVA with RS, direction of adjustment and filter type as within-subject variables. As the full FFT resolution did not have two types of filters, all filter resolutions and shapes were first included as separate levels of one combined within-subject factor. The model indicated the filter type as a significant factor (F(4,72)=[5.579] p<.0001). Assumption of normality was met and there were no significant interactions with other factors. To investigate the influence of filter resolution and shape separately, another RM-ANOVA was constructed with resolution and shape as separate factors. The data of the full-spectrum FFT comparisons was left out to achieve a crossed design. In this second model, resolution had a significant influence on the responses (F(1,18)=[9.124] p<.01), while the shape did not (F(1,18)=[2.449] p=.14).
The responses for the different filter types are presented in Figure 7. Following the ANOVA with all filter resolutions and shapes as separate levels of one combined within-subject factor, post-hoc pairwise t-tests between all filter types initially revealed significant differences between the interpolated filter in 1/1 octave bands and both stepped (p=.01) and interpolated (p=.016) 1/3rd octave band filters, as well as the FFT filter (p=.006). However, the p-values were above the .05 threshold after Holm correction (p=.094, p=.124, and p=.056, respectively). To investigate the effect of filter resolution in the second ANOVA, the difference responses for 1/1 and 1/3rd octave band filters, grouping across filter shapes, was on average 0.7 dB, with a higher-level noise signal selected for the 1/3rd octave band noises. In other words, there was only a small impact of filter resolution. The matchings made with noises generated with any of the filter types can be assumed to largely correspond.

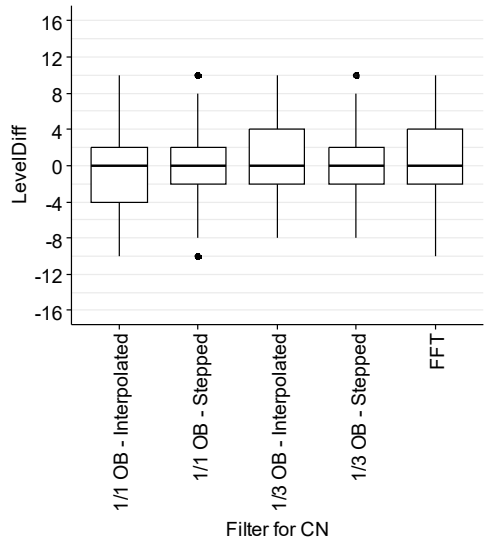


Figure 7. Comparison of filter resolution for CN

4. CONCLUSION

In four listening tests, the impact of participant's background, listening test environment, listening test conditions and comparison noise creation were investigated. While the difference in mean responses of experts and non-experts was notable, there was no significant difference between these groups. The small and unbalanced sample size should be considered here, and therefore it cannot be said with certainty that the responses of experts and non-experts are consistent. In the comparison of listening test environment and conditions, the mean responses were consistent and statistical tests presented no support for an influence. Finally, while the filter shape for creation of the comparison noises did not affect the responses, there was a significant impact of the filter resolution. Nevertheless, the size of this effect was small (0.7 dB), and negligible in most contexts. These findings demonstrate the robustness of the equal-loudness matching method across these methodological variables, supporting its reliability for diverse applications.

5. ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon Europe research & innovation programme under the HORIZON-MSCA-2021-DN-01 grant agreement No. 101072598 – "ActaReBuild".

6. REFERENCES

[1] World Health Organization, *Environmental noise guidelines for the European region*. Regional Office for Europe, 2019.

[2] ISO 717-1:2020. Acoustics - Rating of sound insulation in buildings and of building elements: Part 1: Airborne sound insulation.

[3] M. Florentine, A. N. Popper and R. R. Fay: *Loudness*. New York: Springer, 2011.

[4] M. Geluykens, H. Müllner, V. Chmelík, M. Rychtáriková, "Towards a reference spectrum for façade sound insulation: on average spectra representing temporally varying sounds", in *Proc. of International Conference ACOUSTICS 2024 High Tatras*, (Štrbské Pleso, Vysoké Tatry, Slovakia), pp 33-36, 2024

Additional notes:

I'm not happy with this paper as it is mostly reporting non-significant results. As in the words of Daniel Lakens:

"When you perform a statistical test, and the outcome is a p-value larger than the alpha level (α), the only formally correct conclusion is that the data are not surprising, assuming the null hypothesis is true. It is not possible to conclude there is no effect – our test might simply have lacked the statistical power to detect it."

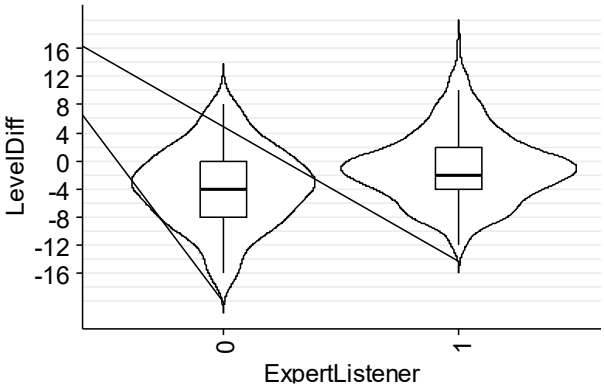
Or in another quote:

"absence of evidence is not evidence of absence"

Therefore, I wanted to revisit the analysis in hopes of coming up with a more meaningful conclusion.

6.1 Participants

A notable difference of 2.8 dB in mean responses was found between expert and non-expert listeners, however, it was not a significant factor in the RM ANOVA.

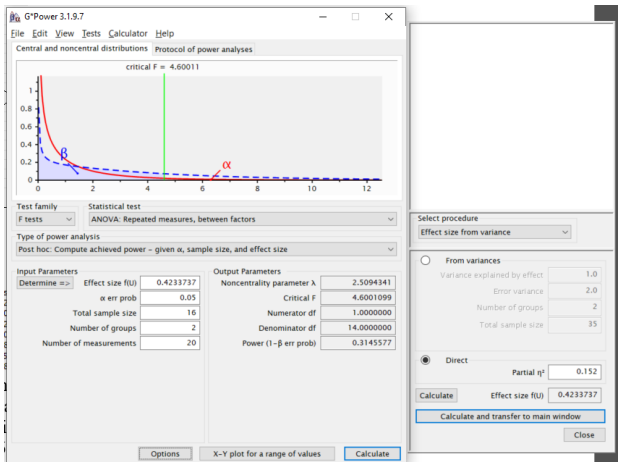


RM-ANOVA:

ANOVA Table (type III tests)

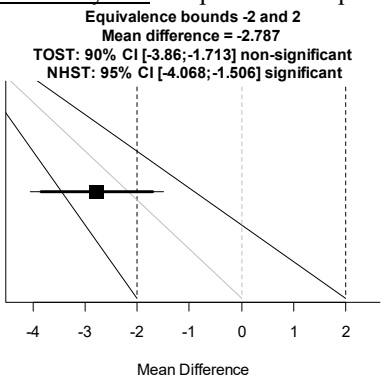
	Effect	DFn	DFd	F	p	p<.05	pes
1	ExpertListener	1	13	2.324	1.51e-01		0.152
2	SoundSource	9	117	5.323	4.39e-06	*	0.290
3	Direction	1	13	30.599	9.68e-05	*	0.702
4	ExpertListener:SoundSource	9	117	1.944	5.20e-02		0.130
5	ExpertListener:Direction	1	13	0.103	7.54e-01		0.008
6	SoundSource:Direction	9	117	3.554	6.24e-04	*	0.215
7	ExpertListener:SoundSource:Direction	9	117	0.517	8.60e-01		0.038

Entering the partial Eta-squared (0.152), alpha level (0.05), sample size (16), number of groups (of between-factor levels = 2), and number of measurements (within-factor levels = 20) into G*power reads an achieved power of 31 (31% probablility to find a significant result if there is a true effect)



Note that the total sample size was not equally distributed over the between-participant groups (experts and non-experts), therefore the power is likely still overestimated. In conclusion, the non-significance is not very surprising.

Alternative analysis 1: simple t-test & equivalence tests



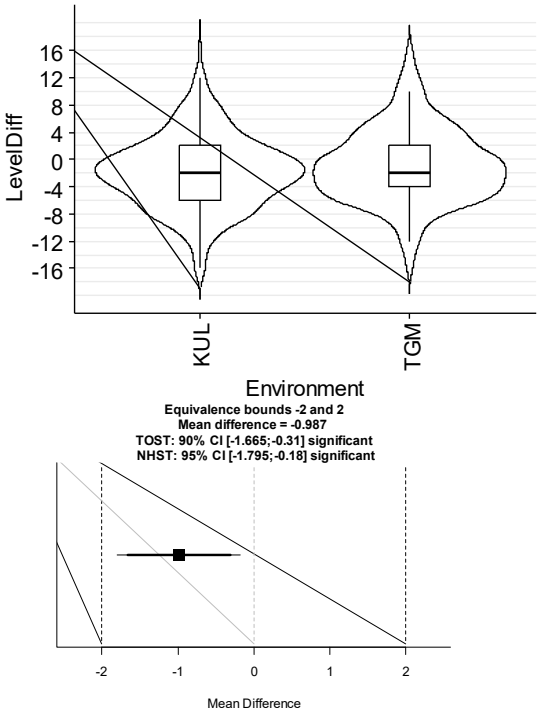
(Based on simple t-tests, assumption of independence of observations violated)

Alternative analysis 2: equivalence test for ANOVA

The equivalence test has not yet been defined for within-subject or mixed design ANOVAs [https://bpspsychub.onlinelibrary.wiley.com/doi/epdf/10.1111/bmsp.12201].

6.2 Environment

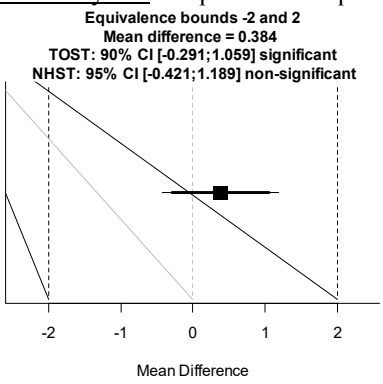
Alternative analysis 1: simple t-test & equivalence tests



According to the equivalence test, the effect of environment is within 2 dB (Based on simple t-tests, assumption of independence of observations violated, moreover, environment in confounded with participants background)

6.3 Conditions

Alternative analysis 1: simple t-test & equivalence tests



According to the equivalence test, the effect of uncontrolled conditions is within 2 dB (Based on simple t-tests, assumption of independence of observations violated)