DOI: 10.61782/fa.2025.0738

RELATING ROOM ACOUSTIC RATING CONSISTENCY TO BINAURAL MASKING LEVEL DIFFERENCE SCORES

Felix Stärz[1*] **Steven Van De Par**[2,3] **Leon O.H. Kroczek**[4]
Sarah Roßkopf[4] **Andreas Mühlberger**[4] **Matthias Blau**[1,3]
[1] Institut für Hörtechnik und Audiologie, Jade Hochschule Oldenburg, Germany,
[2] Acoustics Group, Carl von Ossietzky University Oldenburg, Germany
[3] Cluster of Excellence 'Hearing4All'
[4] Department of Psychology, Clinical Psychology and Psychotherapy, Regensburg University, Germany

ABSTRACT

Reliable evaluation of room acoustics is a matter of expertise. Evaluators need to know what the room acoustic attributes mean and how they are perceived, while also being able to evaluate small differences consistently. Although these attributes can be explained using audio examples, hearing small differences and interpreting them consistently remains challenging. Listening expertise is not well defined, and a reliable measure does not exist. Selecting participants by questionnaires focusing on musicality or profession may inadvertently exclude individuals who do not meet these criteria but could provide high-quality ratings due to their ability to perceive and consistently classify small differences. As a possible alternative, we explored the use of individual binaural masking level difference (BMLD) scores to predict listening expertise in virtual scenes with binaural audio. BMLD scores were correlated with a consistency score calculated from repeated measurements of room acoustic attributes rated in an interactive virtual environment using head-tracked binaural audio. Results show that the BMLD scores can explain a small to moderate amount of the consistency scores, similar to what has been reported in other approaches. BMLD measurements are a step further in pre-selecting expert listeners; however, further skills represented by expert listeners need to be investigated.

Keywords: *Binaural Masking Level Difference, Expert Listener, Virtual Acoustics, Binaural Audio, Virtual Reality*

1. INTRODUCTION

In listening experiments, one often wants to distinguish between levels of listening expertise and categorise participants as naïve or expert listeners. However, the term "listener expertise" is not clearly defined. Listening expertise can be based on musical training [1] and experience [2] or non-musical related factors such as education [3] or profession [4].

With the aim of reliably assessing "room acoustic listening expertise", von Berg et al. [5] designed a set of listening tests to evaluate participants' performance in identifying rooms with varying reverberation and spectral envelopes. The test results were then correlated with the Goldsmiths Musical Sophistication Index (Gold-MSI) [6], a questionnaire designed to assess the musicality of non-musicians, and other questionnaires to specify musical and acoustic knowledge. For significantly related criteria of musicality and professional expertise explaining the variance of "room acoustic listening expertise," the adjusted R^2 was in the range of 0.11 to 0.28.

In head-tracked binaural applications, it is the binaural aspects of auditory perception in particular that become relevant. While specialist knowledge of room acoustical terms may be acquired through profession or training, it is sometimes even desired to test with untrained participants. In such scenarios, attributes can for instance be explained to naïve listeners with additional audio examples. However, proficiency in binaural percep-

*[*]**Corresponding author**: felix.staerz@jade-hs.de.*

tion and adaptation to binaural audio cannot be assumed, even among potential experts. Therefore, it may be beneficial to pre-select suitable test participants using a task that specifically involves binaural changes.

A paradigm in which the perceived cue is based on binaural differences is the measurement of BMLDs [7]. It is therefore conceivable that BMLD scores could predict listening expertise in experiments involving the binaural rendering of audio.

In this study, we aim to investigate the relationship between BMLD scores and consistency scores. We hypothesise that participants with listening experience, specifically the ability to perceive and reliably rate binaural stimuli, will perform better in BMLD experiments and be more consistent when rating room acoustical attributes in a virtual environment with binaural audio reproduction. Previously, we assessed the consistency of multi-stimulus ratings using the Pearson correlation coefficient as a metric for rating reliability [8]. Here, we examine the relationship between BMLD scores and consistency in a multi-stimulus rating experiment involving head-tracked binaural audio. The first research question we aim to address is:

RQ1: Do higher scores in the BMLD paradigm correlate with higher consistency when rating room acoustic attributes in a binaural virtual reality experiment?

Furthermore, it is known that training can enhance externalisation [9] and, consequently, the VR experience. It would be interesting to determine whether the exposure to head-tracked binaural audio influences BMLD scores. This leads to our second research question:

RQ2: Does exposure to head-tracked binaural audio in an interactive virtual environment over time influence BMLD scores?

2. METHODS

2.1 Multi-Stimulus Rating

The multi-stimulus rating experiment followed a design comparable to that of Stärz et al. [10], conducted in a small lecture room at Jade Hochschule, Oldenburg, Germany. Audiovisual stimuli were presented using a head-mounted display (HMD) and headphones, creating an interactive virtual environment (IVE).

Participants rated seven auralisation conditions. The attributes to be rated included reverberance, tone colour, loudness, source distance, reproduction quality, plausibility, and externalisation, partly taken from the Room Acoustical Quality Inventory (RAQI) [4]. All attributes were rated three times during one measurement session. Further details can be found in [10]. Based on the repeated ratings, the Pearson correlation coefficient can be used as a measure of consistency in such experiments, see, e.g., Blau et al. [8].

Head-tracked binaural audio was realised using both measured and simulated binaural room impulse responses (BRIRs) for several head rotations. BRIRs were measured in the real room using a Head-and-Torso Simulator (HATS), as well as with one human subject, both equipped with MEMS microphones at the blocked ear canals. Additional simulated BRIRs were generated using the room acoustic simulator RAZR [11], employing either generic or individual Head-Related Impulse Responses (HRIRs).

Furthermore, some BRIR sets were manipulated to introduce perceivable differences in distance, reverberation, and externalisation.

2.2 Binaural Masking Level Difference

If a sinusoidal signal is masked by noise with a distinct bandwidth, a reduction in threshold is observed when the interaural properties of the masker and signal are different instead of the same. The difference between the masker and the signal being interaurally in phase or out of phase is called BMLD. To measure BMLDs we designed our BMLD experiment based on the studies of van de Par and Kohlrausch [7, 12]. We employed the N_0S_π condition, where the masker is interaurally in phase, and the signal is phase-reversed. A masker bandwidth of 25 Hz was used, with two different sinusoidal signal frequencies: 500 Hz (labelled G01) and 4 kHz (labelled G04), with the masker centre frequencies corresponding to the sinusoidal signal frequency. From the literature, it is known large BMLDs are found at 500 Hz, and considerably lower BMLDs are found at 4 kHz [12].

Participants completed three test trials, starting with G01 followed by G04. After the second test trial, participants were asked by the experimenter whether they could perceive the binaural cue. If not, the cue was aurally explained as a wider, more binaural sound impression. Following the trial phase, participants performed three pairs of G01 and G04, one after another. The BMLD test was conducted before the first and after the second multi-stimulus VR experiment, with at least one week between BMLD measurements. Participants were unaware of any connection between the BMLD measurement and the multi-stimulus rating experiment.

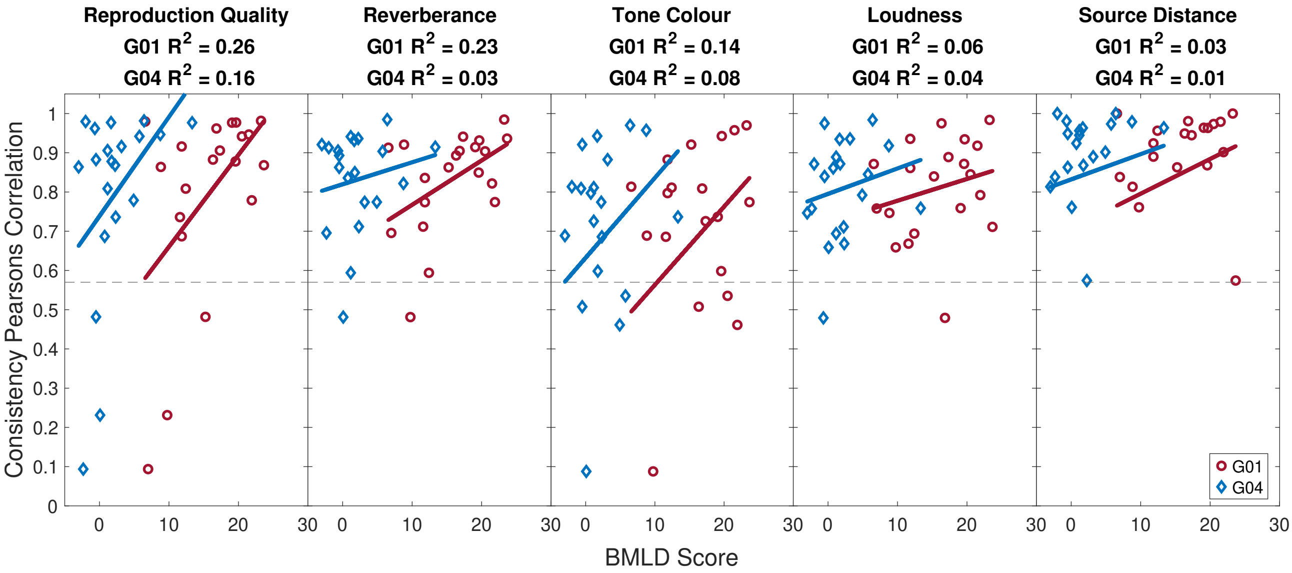

Figure 1. Consistency (as measured by Pearson correlation coefficient) as a function of the BMLD score for two different masker frequencies. G01: Masker frequency of 500 Hz and a masker bandwidth of 25 Hz. G04: Masker frequency of 4 kHz and a masker bandwidth of 25 Hz. Each individual scatter point represents a single participant. The dashed line indicates the consistency threshold; below this, participants would be considered as inconsistent.

2.3 Listening Test

A total of 20 participants (4 female, 16 male, median age 27.5 years, tested normal hearing) took part in the listening test. The participants were not considered naïve, given their background in hearing research, familiarity with listening tests, and experience with head-tracked binaural audio.

Initially, a BMLD measurement was conducted. Following this, with at least one day in between, the multi-stimulus rating as described in Sec. 2.1 was performed twice on two different days. After several days had passed since the first BMLD measurement, the same BMLD measurement procedure was repeated.

3. RESULTS

3.1 BMLD and Consistency

Fig. 1 illustrates the relationship between the Pearson correlation coefficient and the BMLD score for all subjects and masker frequencies. There is a wide spread of BMLD scores for G01, ranging from 5 dB to 25 dB. BMLD scores for G04 were comparably smaller, as was the overall variance. While the lowest BMLD score for G01 is close to 5 dB, most participants' scores were near or even above 0 dB for G04. Only a few participants achieved a BMLD score above 5 dB for G04.

The fitted regression lines tend to follow the hypothesised direction for each room acoustic attribute, indicating that participants with higher BMLD scores can be expected to be more consistent. The R^2 value is up to 0.26 for reproduction quality and G01. Depending on the sinusoidal signal frequency and the room acoustic characteristics, a small to moderate proportion of the data can be explained. The best fit is found for reproduction quality for G01 (β_1 = 0.023, SE = 0.009, t(18) = 2.517, p = 0.021), as well as reverberance (β_1 = 0.011, SE = 0.005, t(18) = 2.318, p = 0.032).

It is also worth mentioning that there are a few inconsistent scores from different participants, mostly from those with comparably low BMLD scores.

3.2 Effect of Exposure

To evaluate the effect of listening to binaural auralisations, Figure 2 illustrates the results from the first and second BMLD measurement sessions for both masker frequen-

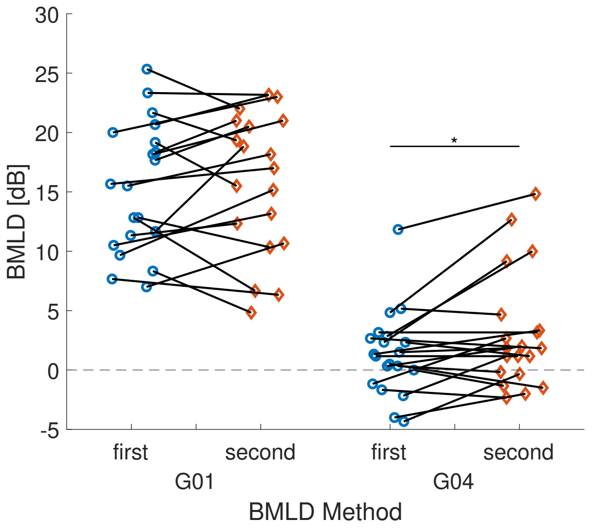

Figure 2. BMLD scores for repeated measurements for both G01 and G04 condition. Straight black lines indicate the intra-individual shift.

cies. Intra-individual changes from the first to the second BMLD score can be traced by following the straight lines. It is apparent that only a few individuals exhibit an increase of more than 5 dB for repeated BMLD scores. Some participants even demonstrated worse performance in the second session. Overall, there is no significant improvement for G01 after the multi-stimulus VR experiment.

In contrast, a significant increase in BMLD score can be observed for G04 after the multi-stimulus experiment (t[19]=2.627, p=0.017). The highest BMLD score improvements of approximately 8 dB are predominantly found for participants who initially had a BMLD score greater than 0 dB in the first session.

4. DISCUSSION

4.1 BMLD as a Predictor for Consistency

BMLD scores explain a statistically significant small to moderate amount of consistency in ratings of reproduction quality and reverberance in a VR experiment with head-tracked binaural audio. The R^2 values are comparable to those reported by von Berg [5]. While von Berg designed a test battery aimed at “room acoustic listening expertise,” our focus was on the consistent adaptation, perception, and interpretation of binaural cues.

Van de Par [12] reported mean BMLD scores of 24 dB for the G01 condition and 10 dB for G04. In the present study, mean scores for G01 increased from 14 dB to 16 dB between the first and second measurements, and from 1 dB to 3 dB for G04. The higher scores reported by van de Par were achieved by only a small number of our participants. Notably, for G04, only four participants attained comparable BMLD scores after the second measurement. One possible explanation is that van de Par’s study included only three participants, who were highly trained in this specific BMLD paradigm, as the authors themselves participated.

We asked the participants if they had ever performed a BMLD measurement. Only one participant had extensive experience with psychoacoustic listening tests and BMLD measurements. Unsurprisingly, this participant also obtained the highest BMLD scores for the G04 condition. This raises the question: Is proficiency in BMLD measurements solely a matter of training, or is it possible that some participants are inherently better at perceiving and interpreting binaural cues?

A tendency was observed where only participants who could perceive the binaural cue in the first measurement were able to achieve even higher scores in the second measurement, approaching those reported in the literature [12]. However, many participants who initially scored below the threshold of 0 dB did not improve to the extent of reaching 10 dB. It is possible that not only the initial perception of the cue but also the time taken to adapt to and learn this unfamiliar signal plays a crucial role in becoming an expert listener concerning BMLDs.

4.2 Improvement in BMLD Scores over Time

An effect of improved BMLD scores can be observed over time for the G04 condition, with a significant increase in BMLD scores for the second test. Two potential explanations for this phenomenon are considered.

The first is a training effect within the BMLD measurement itself. By repeatedly performing this measurement, participants may improve over time. By the start of the second BMLD test, participants had already completed six repetitions of this measurement and could potentially have learned to recognise the cues. However, given that several days elapsed between the two BMLD measurements, participants would have needed to retain this knowledge.

The second explanation is that exposure to head-tracked binaural audio may have aided participants in perceiving and interpreting binaural cues, leading to improved performance (at least for the more challenging G04 condition) if they had listened to head-tracked binaural audio between the first and second BMLD sessions.

The precise reason for the BMLD improvement cannot be definitively explained. A comparison group is needed, performing this repeated BMLD experiment while listening to a placebo condition between the two BMLD measurements, i.e., any stimulus other than binaural audio.

4.3 Impact of Hearing and Cognition

As the test is designed to perceive and rate binaural cues, hearing loss could potentially impact the ratings. However, hearing thresholds were measured, and according to the World Health Organisation (WHO), all participants had normal hearing. Furthermore, both tests are well above the threshold in quiet, such that the basic audibility of the stimuli was not impaired.

The BMLD measurement is perceptually demanding because auditory cues tend to be very subtle. The primary challenge is maintaining concentration and focus on the binaural cue, which could be affected by cognitive abilities. Since both tests are approximately the same length, we assume an effect of concentration on the test would be evident in both measurements.

5. CONCLUSION AND OUTLOOK

BMLD scores account for a small to moderate proportion of the consistency in a multi-stimulus rating experiment within a VR environment, explaining up to 26 % of the variance. However, they are not reliable enough to be used to pre-select expert listeners in terms of consistency.

It would be interesting to correlate BMLD scores with the Gold-MSI [6] or other questionnaires, such as the one used by von Berg et al. [5] that assesses listener or acoustic expertise to better understand the relationship between binaural cue perception and broader listener expertise.

A between-subjects design with a placebo group would be necessary to gain insight into the observed time effect leading to improved BMLD scores for the G04 condition. This would help to disentangle the impact of training, adaptation, and prior experience. In addition, a larger and more diverse sample would be required to investigate the possibility of a bimodal distribution of BMLD scores, which could potentially distinguish expert listeners from non-experts. The current study is limited by the inclusion of participants with some prior knowledge, and future work should aim to recruit naïve listeners as well.

6. ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under the project ID 422686707, SPP2236 - AUDICTIVE, subproject 444832396. We want to thank our subjects for their participation in the study.

7. REFERENCES

[1] G. A. Soulodre and J. S. Bradley, “Subjective evaluation of new room acoustic measures,” *The Journal of the Acoustical Society of America*, vol. 98, pp. 294–301, July 1995.

[2] M. Barron, “The subjective effects of first reflections in concert halls—The need for lateral reflections,” *Journal of Sound and Vibration*, vol. 15, pp. 475–494, Apr. 1971.

[3] H. Wilkens, “A Multidimensional Description of Subjective Judgement of Concert-Hall Acoustics,” *Acta Acustica united with Acustica*, vol. 38, no. 1, pp. 10–23, 1977. Publisher: European Acoustics Association.

[4] S. Weinzierl, S. Lepa, and D. Ackermann, “A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI),” *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1245–1257, 2018.

[5] M. Von Berg, J. Steffens, S. Weinzierl, and D. Müllensiefen, “Assessing room acoustic listening expertise,” *The Journal of the Acoustical Society of America*, vol. 150, pp. 2539–2548, Oct. 2021.

[6] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, “The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population,” *PLoS ONE*, vol. 9, p. e89642, Feb. 2014.

[7] S. van de Par and A. Kohlrausch, “A new approach to comparing binaural masking level differences at low and high frequencies,” *The Journal of the Acoustical Society of America*, vol. 101, no. 3, pp. 1671–1680, 1997.

[8] M. Blau, A. Budnik, M. Fallahi, H. Steffens, S. D. Ewert, and S. van de Par, “Toward realistic binaural

auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario," *Acta Acustica*, vol. 5, 2021.

[9] F. Klein, S. Werner, and T. Mayenfels, "Influences of Training on Externalization of Binaural Synthesis in Situations of Room Divergence," *Journal of the Audio Engineering Society*, vol. 65, pp. 178–187, 2017.

[10] F. Stärz, S. Van De Par, L. O. H. Kroczek, S. Roßkopf, A. Mühlberger, and M. Blau, "Comparison of binaural auralisations to a real loudspeaker in an audiovisual virtual classroom scenario: Effect of room acoustic simulation, HRTF dataset, and head-mounted display on room acoustic perception." In press, 2025.

[11] T. Wendt, S. Van De Par, and S. D. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *Journal of the Audio Engineering Society*, vol. 62, no. 11, pp. 748–766, 2014.

[12] S. Van De Par and A. Kohlrausch, "Dependence of binaural masking level differences on center frequency, masker bandwidth, and interaural parameters," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1940–1947, 1999.