



Reduced reproduction levels of outdoor soundscapes are deemed appropriate – even after real-world exposure

Markus von Berg^{1,2,*}, Siegbert Versümer^{1,2}, Joshua Bitta¹, and Jochen Steffens^{1,2}

¹Institute of Sound and Vibration Engineering, Hochschule Düsseldorf – University of Applied Sciences, 40476 Duesseldorf, Germany

²Audio Communication Group, Technische Universität Berlin, 10623 Berlin, Germany

Received 14 March 2025, Accepted 7 November 2025

Abstract – Laboratory experiments in psychoacoustical and soundscape research indicate that participants perceive a reproduction sound level lowered by 8–10 dB as more plausible than the original level. This bias supposedly roots in an adaptation of perceptual loudness scaling to the laboratory environment, that is overall quieter than urban outdoor soundscapes. To gain further insights into the nature of such loudness bias, we conducted a listening experiment in both field and laboratory using a within-subjects design. Thirty-one participants visited a street and listened to the environmental sounds for one minute, while these sounds were also recorded using a dummy head. Thereafter, they listened to the recording in a quiet laboratory nearby and adjusted its level as they remembered it. About half of the sample did this immediately, the other half about 20 min after the recording. Results confirm a bias towards lower levels with a mean of about 8.9 dB, regardless of the time between the recording and the reproduction in the laboratory. Also, participants with higher musical abilities tended to select higher, more accurate levels, whereas noise-sensitive participants deemed lower levels appropriate. Results suggest that the hypothesized adaptation of perceptual scaling to the laboratory happens immediately and is affected by individual factors.

Keywords. Outdoor soundscapes, Laboratory experiments, Reproduction level

1 Introduction

To achieve ecological validity in psychoacoustical experiments that include the presentation of real-world sound in a laboratory, the need to ensure accurate reproduction of sound pressure levels seems rather obvious. Following the widely adopted definition by Keidser et al. [1], ecological validity describes the extent to which findings from a laboratory setting “reflect real-life hearing-related function, activity, or participation”. Thus, an experiment’s ecological validity comprises the selection of participants, the experimental conditions (including the selection and reproduction method of sound stimuli) and the cognitive tasks at hand [2]. This paper focuses on the sound reproduction component with an emphasis on the complex sound compositions of urban soundscapes. Here, previous research has widely adopted three-dimensional spatial reproduction techniques, such as Ambisonics loudspeaker systems [2–7] and binaural headphone auralizations [8–10]. Guastavino et al. [3] found that especially for complex sounds, spatial reproductions evoke mental representations of soundscapes as a single complex back-

ground noise (rather than a composition of single sound sources) that are more similar to results obtained in field studies. Furthermore, research is increasingly incorporating additional visual display of recording sites (see Li and Lau [11] for a review) to convey a sensory impression that is as similar as possible to the real-world experience. As Ambisonics or binaural audio (especially in combination with visual display) seek to ensure a reproduction of the spatial properties of complex sounds that evokes similar impressions as the real-world equivalent, again, it seems reasonable to then carefully calibrate the sound’s playback level to its recording level. However, there is evidence that, in fact, lower reproduction levels are perceived as more appropriate [5–7, 9, 10]. Consequently, environmental sounds presented in a laboratory at physically accurate levels might be perceived louder as they would be in reality, threatening the experiment’s ecological validity.

1.1 Biased evaluation of soundscape levels in laboratory experiments

Regarding the reproduction of urban soundscapes, Oberman et al. [4] and Tarlao et al. [2] observed that participants perceived soundscapes presented with a first

*Corresponding author:
markusmartin.vonberg@hs-duesseldorf.de

order Ambisonics loudspeaker system with physically accurate playback levels as too loud. Furthermore, laboratory reproductions of soundscapes with accurately calibrated playback levels have been rated as more chaotic and unpleasant than in situ [8], and received overall less pronounced pleasantness ratings [2], which could also be at least partially attributable to biased loudness perception.

To correct for such a loudness bias, Fraise et al. [5] lowered the level of soundscapes reproduced via a fourth-order Ambisonics 24-channel loudspeaker system by 4 dB, based on the authors' judgment. By contrast, Davies et al. [6] let the participants adjust the reproduction levels as part of a soundscape synthesis procedure and found that a mean reproduction level of 12.3 dB below the physically accurate level was deemed appropriate. The sounds were presented with an eight-channel first order Ambisonics loudspeaker system, and the selection of lower level was attributed to the quietness of the semi-anechoic chamber where the experiment was carried out.

In a follow-up study by Sudarsono et al. [7], an average level reduction of about 9.5 dB for a soundscape in Manchester, UK, was considered accurate by a sample of Manchester locals, as well as by study participants in Indonesia. Again, an eight-channel first-order Ambisonics system was used for playback, and the experiments in the two countries were conducted in a listening and a recording room, respectively.

Based on the aforementioned study, Yang and Kang [9] lowered the reproduction level in virtual sound walks by 10 dB and confirmed that loudness ratings in virtual sound walks were closer to in-situ ratings, compared to a reproduction with accurate levels.

Recently, Lu and Lau [10] combined dynamic binauralizations of soundscapes via headphones with a visual display of the recorded spaces using a head-mounted display (HMD), and presented fixed reproduction levels of 0 dB, 4 dB, 8 dB, and 12 dB below the original. Here, participants, who had also performed in-situ evaluations of the soundscapes a week earlier, rated a level reduction of 8 dB to be most realistic. The experiment was conducted in an office with about 40 dBA background noise.

In summary, the aforementioned studies suggest that participants perceive lower playback levels as more plausible, regardless of the reproduction systems used (although the only study applying higher-order Ambisonics arrived at a smaller gain reduction [5]), fixed and freely adjustable playback levels, and the degree of familiarity with the soundscapes presented. While previous research acknowledges the existence of the loudness bias and measures its magnitude, the underlying cognitive mechanisms are seldom discussed. Some studies assume an adaptation of loudness perception to laboratory background noise levels [6, 10] and Hermina Cadena et al. [8] suggested that the absence of visual information would change expectations for an urban soundscape's eventfulness. However, Lu and Lau [10] found the loudness bias to persist when participants saw a visual representation of the recording site via an HMD.

1.2 Objectives of the study

According to Braida et al. [12], the perceptual scaling of sound intensity in an experimental setting depends on the range of presented intensities and one or more internal "perceptual anchors". This is in line with the suggested adaptation to laboratory background noise, as this noise level would serve as a perceptual anchor that shifts the scaling baseline from, for example, an urban outdoor area to a sound-proof laboratory. However, Ward [13] observed that the loudness framework participants established based on a set of (pure tone) levels can impact the scaling of a different set of levels on a following day, implying that, even for synthetic tones, perceptual anchors are not limited to the current range of intensities. Therefore, previous exposure to the actual soundscape might reduce the adaptation of perceptual scaling. Lu and Lau [10] included an actual previous exposure to the recording site in their experiment, but the duration of one week between the in-situ evaluation and laboratory reproduction was probably too long to maintain some sort of absolute loudness reference.

Therefore, the present study employed a different design where the selection of an accurate level for reproducing an urban soundscape in the laboratory was performed immediately after an in-situ experience of this soundscape, with either about 20, or only a few minutes in between. The aim was to test if the reported bias towards levels lowered by 8–12 dB would persist when the time span between reproduction and in-situ listening was short enough to possibly maintain an internal reference (or perceptual anchor) of the actual sound level.

1.3 Loudness recall after short time periods

It has been known for a long time that the ability to accurately recall the level of a tone decays within the first 20 s [14, 15]. Research on the brain activity during pitch recall suggests that this originates from the transition from a purely working-memory-based recall to an activation of long term memory [16]. In the study presented here, at least 2–3 min passed between the initial perception of the soundscape and the recall of its loudness, so that the recall is undoubtedly based on long term memory.

Furthermore, evaluating the plausibility of a soundscape reproduction level does not limit to the recall of a single tone's loudness, but a composition of various complex sounds over a certain period. Several studies suggest that remembered loudness of a soundscape sequence – which forms the basis to select an appropriate reproduction level – does not weight all temporal segments equally. Hellbrück [17] observed that retrospective loudness judgments after being exposed to a soundscape for 2 min correlated more with the momentary loudness ratings during the last 7.5 s of exposure than with the momentary ratings averaged over the entire duration – a phenomenon commonly referred to as recency effect. Dittrich and Oberfeld

[18] also found primacy effects where the beginning of a sound has more impact on the remembered loudness – at least for sounds as short as 900 ms.

Therefore, if a perceptual anchor for a given soundscape is formed during the in-situ experience, distinct temporal segments might be of more importance than others. Thus, in our study, the gain level participants chose in the laboratory are compared to the loudness within different time segments of each individual recording.

1.4 Individual differences

To further explore possible reasons for the bias in loudness judgments, one objective of the present study was to compare the reproduction levels participants considered appropriate to individual differences. Sudarsono et al. [7] observed that participants who had experience in acoustics and had participated in listening tests before were more consistent in the level adjustment of soundscapes, albeit not less biased. If loudness judgments in the laboratory are based on internal perceptual references, it seems plausible that their “robustness against adaptation” can be increased by knowledge on acoustics and experience in critical listening.

For this reason, in our study, we obtained previous experience with acoustics, musical sophistication, which might also be indicative of critical listening ability and auditory sensitivity [19, 20], and individual noise sensitivity, which we found to affect performance in auditory recall tasks in a previous study [21].

1.5 Research questions

The study was guided by two central research questions. As previous research has shown that participants often consider lower soundscape reproduction levels to be realistic, the first question was how accurately participants could reproduce an outdoor soundscape in the laboratory immediately after in-situ exposure to this exact soundscape. In this context, it was also investigated whether a time delay of several minutes between in-situ exposure and recording, where participants were occupied with another listening task, would make a difference.

As explained in the previous section, individual factors such as expertise in acoustics might influence the ability to recall the level of a soundscape recording. The second research question was thus concerned with the importance of individual differences and their possible effects of individual variation in the selected reproduction level.

2 Material and methods

2.1 Participants

To address our research questions, we conducted a combined field and laboratory experiment (within-subjects design) with 32 participants aged between 18 and

68 years (mean: 31.7 years, SD: 14.2 years). The majority, 23 participants, identified as male (71.9%), 8 as female (25%), and 1 as diverse (3.1%). More than half were undergraduates (56.3%), the rest held an university degree or another kind of professional education. No participant reported hearing disorders, except for one report of tinnitus. However, one participant was excluded due to selecting the maximum possible playback level and showing possible symptoms of hearing loss during conversations. Thus, all following analyses refer to a sample of 31 participants.

Participants were recruited via mailing lists targeting students as well as non-university members who registered for listening test invitations at our institute, and by spontaneously approaching people at the Hochschule Düsseldorf campus. Participation was compensated with 5 euros.

2.2 In-situ recordings

During the in-situ listening phase, binaural recordings were made on a vivid street (Münsterstraße) in Düsseldorf, about 80 m away from a street crossing with tram and bus stations (see Fig. 1). The soundscape mainly consisted of passing cars, bicycles, busses, trams, and pedestrians. The recording site was in a broad and rather quiet entrance to the campus of the Hochschule Düsseldorf, so that most sounds stemmed from the street in front of the dummy head and the participant standing next to it. The recordings’ average A-weighted equivalent sound pressure level (L_{Aeq}) was at 69.3 dBA (SD: 2.1 dB).

An HMS II.3 dummy head was used for the recordings, which was connected to a BEQ II preamplifier (HEAD Acoustics), providing a calibrated analog-digital (AD) conversion so that -16 dBFS corresponded to 94 dB SPL. To avoid distortions in the recordings, 3D-printed hemispherical grids covered with nylon fabric were placed as windshields over the dummy head’s ears (see Fig. 2).

The efficiency of these windshields was tested in a wind tunnel that produced a continuous air flow with a speed of 4 m/s (about 14 km/h). Figures 3a and 3b depict the frequency spectrum of recordings of the dummy head’s right ear in the wind tunnel, with the air flow either orthogonal to the ear-axis, or on the ear-axis (the dotted line represent background noise measurements with no air flow). The measurements confirmed a considerable reduction in distortion artifacts, and in a subjective evaluation of the recordings by two of the authors, no audible distortion artifacts were observed.

Furthermore, the dummy head’s transfer function with and without the windshield was measured in an anechoic chamber. A Genelec 1031A loudspeaker was positioned at a distance of about 1.5 m, facing each of the dummy head’s ear, and a logarithmic sine sweep was used as excitation signal. As can be seen in Figures 3c and 3d, the windshields altered the dummy head frequency responses by less than ± 1 dB at frequencies up to 3 kHz and ± 4 dB at higher frequencies relevant for the experiment.



Figure 1. Recording site, showing the tram station and the entrance to campus where participants stood (right).

2.3 Audio playback and level adjustment

For the laboratory reproduction, the binaural recordings were played back via open, circumaural, electrostatic STAX SR-303 headphones, which were connected to a PEQ V interface and a HPS IV preamplifier (HEAD Acoustics), that supplied the polarisation voltage for the headphones. The PEQ V interface provided a custom digital equalization filter of this specific headphone’s transfer function (provided by HEAD Acoustics) and an output level calibration. In line with the dummy head, the output level was set to -16 dBFS corresponding to 94 dB SPL, and it was tested that the output level specified in the PEQ V interface was properly reproduced through the headphones. To do so, recordings made with the dummy head were replayed to the dummy head through the headphones, showing that the replay produced the same input levels as the original recording. The software *ArtemiS* (HEAD Acoustics) was used for the recording and audio playback. To realize the loudness adjustment, the PEQ V interface was not directly connected to the laptop, but through an RME Babyface Pro audio interface that supplied a digital output signal to the PEQ V via ADAT. Participants were given a shuttle wheel, the Contour ShuttleXpress controller, to control the gain level in the Babyface Pro’s control software TotalMix. Here, the Babyface’s gain could be altered in 0.3 dB steps between negative infinity (complete silence) and +16 dB. If this gain level was set to 0 dB in TotalMix, the audio output was forwarded to the PEQ V interface with no additional gain, resulting in the reproduction of the dummy head’s recordings with the

original sound level. Figure 4 illustrates the audio signal chain for the in-situ recording and the reproduction in the laboratory.

The adjustment procedure started with “silent playback”, so that a gain of “negative infinity” was set in TotalMix as the initial value. During the experiment, it was observed that increasing the gain from this initial value by 0.3 dB did not always result in exactly the same next smallest gain value, resulting in minor shifts in the discretizations of gain values between participants. This also meant that a gain of 0.0 dB (exact reproduction of the original level) could not always be exactly selected, resulting in deviations of 0.1 or 0.2 dB. We consider these minor deviations to be of little practical relevance but report them for reasons of transparency and completeness.

2.4 Questionnaires

To test for individual differences in noise sensitivity, participants were administered the German version of the reduced noise sensitivity questionnaire (NoiSeQ-R) by Schütte et al. [22] (see Tab. II.9 in [23] for the questions). The NoiSeQ-R comprises different sub-scales of noise sensitivity that are specific to the contexts of *sleep* (e.g., “I need an absolutely quiet environment to get a good night’s sleep.”), *work* (e.g., “I need peace and quiet to do difficult work”), and *habitation* (e.g., “I am very sensitive to neighbourhood noise.”) that are answered on a 4-point Likert agreement scale.

Also, questions on musicality and previous experience with acoustics were included to test if training in focused



Figure 2. (a) Dummy head with mounted windshields on both ears. (b) The windshield's plastic grid (left) covered with nylon fabric (right).

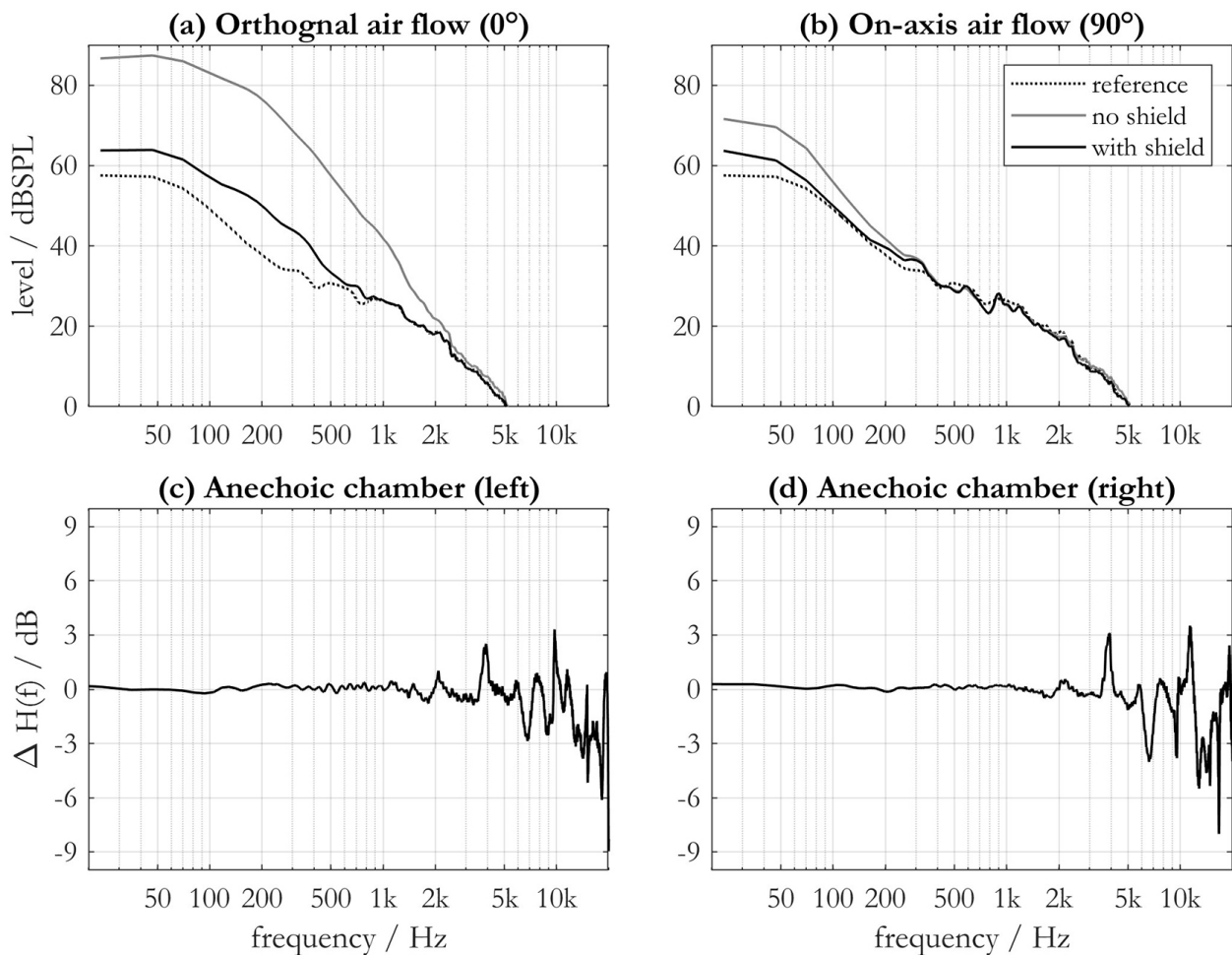


Figure 3. (a) Recording of the dummy head's right ear in a wind tunnel with an air flow orthogonal to the on-ear axis, with (black line) and without (grey line) windshields. The dotted line is a reference measurement with the wind tunnel turned off. (b) Same as (a), but with wind flow on the ear-axis. (c) Difference between transfer functions of the left ear in an anechoic chamber with and without the windshields. (d) Same as (c), but for the right ear.

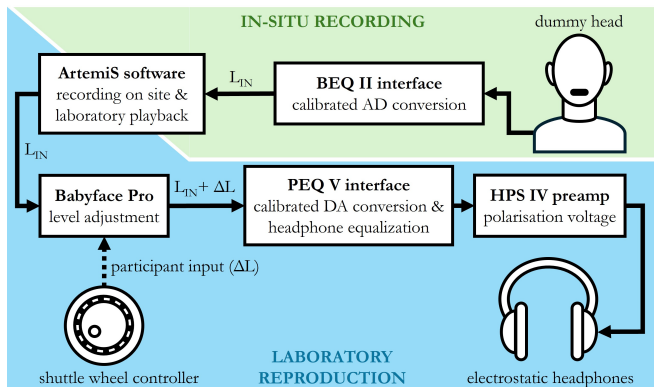


Figure 4. Schematic overview of the audio signal chain. The analog-digital (AD) and digital-analog (DA) conversions were calibrated to the same level of -16 dBFS corresponding to 94 dB SPL.

listening would have any effect on loudness reproduction accuracy. Musical sophistication was assessed by means of two sub-scales of the Goldsmiths Musical Sophistication Index (Gold-MSI) by Müllensiefen et al. [24]. These sub-scales measured the participants' *perceptual abilities*, comprising nine self-report items on abilities to spot mistakes in own and other people's musical performances (e.g., "I can tell when people sing or play out of tune."), and their *musical training* consisting of seven items on formal musical training (e.g., "I have had formal training in music theory for . . . years"). Following the original publication [24], the Gold-MSI questions were answered on 7-point Likert agreement scales, except for those that inquire numerical information such as years of training on an instrument, where 7 response options (e.g., "0" to "10 or more") were available. Finally, previous experience with acoustics was assessed by means of two questions on dealing with acoustics as part of an academic career, and recording activities, respectively, that have been previously used to measure acoustical expertise by von Berg et al. [20]. Here, the amount of engagement for each question was inquired using an ordinal 5-point frequency scale from "never" to "very often".

2.5 Individual recording loudness

To test for effects of intensity differences in the individual recordings on the remembered loudness, potential acoustical predictors were analyzed. Besides the mean L_{Aeq} , the mean loudness level according to ISO 532-1 [25] LLz, which we found to be a more accurate predictor of perceived loudness [26], was calculated for the each recording. Furthermore, the loudness level of the first and last 7.5 s [LLz7.5(f) and LLz7.5(l)] and 15 s [LLz15(f) and LLz15(l)], respectively, were calculated to account for recency or primacy effects. The time interval of LLz15(l) was chosen as the last quarter of the recording while the duration of 7.5 s agrees with the final period of a recording for which Hellbrück [17] observed a recency effect in loudness judgments.

2.6 Experimental procedure and design

The experiment was conducted with one participant at a time. At the beginning, the experimenter led the participant to the recording site, where participants were instructed to carefully listen to the street noise for one minute and remember its loudness. At the same time, the environmental noise was recorded with the dummy head placed next to the participants.

After the outside recording, participants were brought to a sound-proof laboratory with a background noise level of about 22 dBA. The walk from the recording site to the laboratory took about one minute, with about 2/3 of the way leading along the same road where the recordings were made and 1/3 through the hallway of the institute, which had a noise level of about 47 dBA.

In the laboratory, there was a short explanation of the experimental procedure, then each participant was presented the binaural recording of the acoustic environment that they just listened to outside. They were given the shuttle wheel controller to adjust the playback level, and instructed to use the controller to make the replayed sound as loud as they remembered it from the in-situ listening. The controller was unlabeled and there was no visual feedback on a computer screen or on the audio interfaces' displays.

As mentioned above, the playback started at a gain of negative infinity (complete silence), and the recording was looped so that there was no time limit for the adjustment procedure. Most participants needed 2–5 loops (of one minute each) to perform the level adjustment, and everyone listened to the entire recording more than once. After finding a level they deemed appropriate, the experimenter noted this level.

Each participant completed only one level adjustment for the recording that they had previously attended in-situ.

To test for effects of the time passed since the recording on the ability to correctly remember the soundscape's loudness in the laboratory environment, half of the participants (15) performed the loudness reproduction immediately after having returned to the laboratory, which was about two minutes after the recording, and were afterwards administered questionnaires on demographic information, noise sensitivity, and musical sophistication (see Sect. 2.4). The other 16 participants first completed the questionnaires, as well as another listening experiment of about 20 min, where they had to rate different concert hall auralizations that were also presented via headphones, with a mean playback level of 63 dBA, similar to the average level of the street where recordings were made.

The reason for presenting a listening test with different audio stimuli in this 20 min time-span was to ensure an identical and controlled sound exposure for all participants without presenting other soundscapes (that would most-likely bias loudness scaling of the previously recorded one) or a definitely more pleasant auditory stimulus like music. Moreover, when instructed to just wait

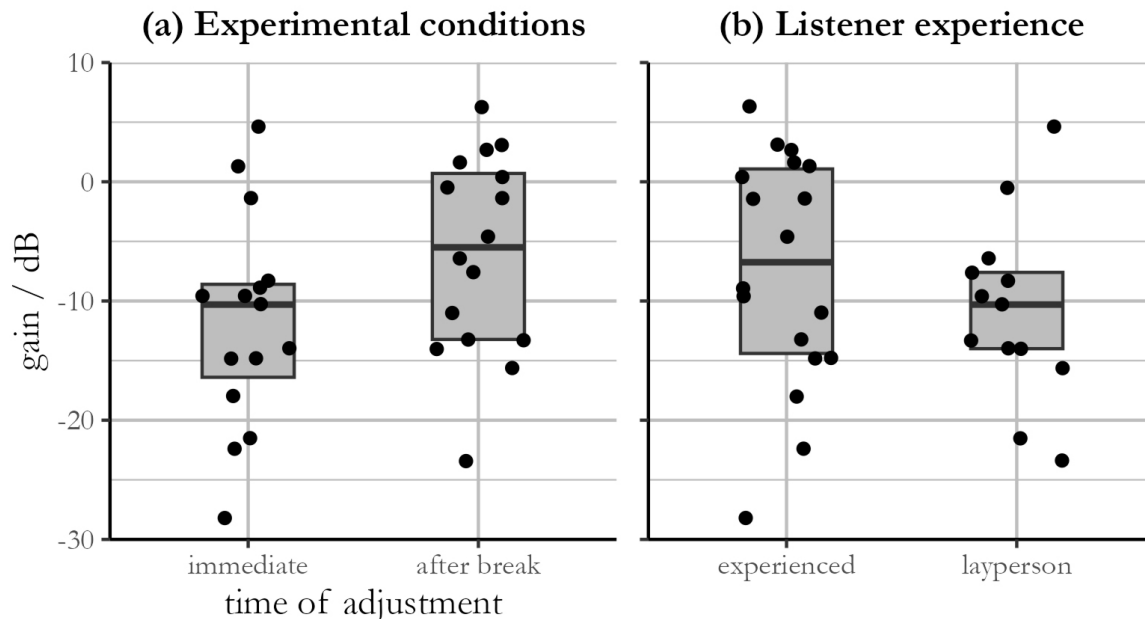


Figure 5. (a) Adjusted gain in dB relative to original level and experimental conditions regarding the time of adjustment. (b) Adjusted gain for both conditions, depending on participants’ experience with acoustics. The box plots represent the inter-quartile ranges and the black horizontal line marks the median value.

in silence, participants might use their phones or engage in conversation with the experimenter, leading to an uncontrollable sound exposure.

3 Results

3.1 Playback gain adjustments

Except for the one person who was excluded due to signs of hearing loss, no participant reached the maximum gain adjustment of 16 dB above the original recording level during the adjustment procedure, meaning that they practically encountered no upper limit. The lower limit, on the other hand, was inaudibility, which is certainly below the level range participants would want to explore while searching for an appropriate reproduction level. We therefore assume that the gain adjustments can be treated as unbound continuous data in the statistical analysis.

Twenty-four of 32 participants (75%) selected a playback level that was below the original level. In detail, playback level adjustments ranged from -28.2 dB to $+6.3$ dB with a mean value of -8.9 dB (SD: 9.0 dB) across both experimental conditions. Since most participants selected lower reproduction levels, higher levels also represent more accurate reproductions in most cases. The most “accurate” adjustment was 0.4 dB more than the original level. Figure 5 depicts the distribution of gain level adjustments relative to the original level of 0 dB, separately for the two experimental conditions (loudness reproduction either immediately after the recording or

with a 20 min break in between). All box plots show inter-quartile ranges as grey boxes with median values as solid horizontal lines as well as single observations as black dots.

Results thus suggest that the experimental group which performed the loudness reproduction task immediately after the recording tended to select lower playback levels, resulting in larger deviations from the original levels (i.e., smaller reproduction accuracies).

3.2 Gain adjustment and recording level

Table 1 reports bivariate Pearson correlations between the gain and all acoustical measures. Interestingly, none of the loudness measures was statistically significantly associated with the reproduction gain selected by participants.

3.3 Individual differences

Regarding individual differences, the sample was first split into “laypersons” and “experienced” listeners, following the work by Sudarsono et al. [7]. To do so, 18 participants (58%) who reported to have dealt with acoustics “from time to time” as either part of an academic career or professional recordings activities were categorized as experienced with acoustics. Figure 5b depicts the similarity of gain adjustments within both groups and shows that the ratings of the experienced participants were less consistent than those of the laypersons. By contrast, the majority of the experts’ judgments were closer to the

Table 1. Bivariate Pearson-correlations between the adjusted gain and the loudness measures. Statistically significant correlations are marked (* $p < 0.05$, *** $p < 0.001$).

	Gain	L_{Aeq}	LLz	LLz ₁₅ (f)	LLz _{7.5} (f)	LLz ₁₅ (l)
L_{Aeq}	0.061					
LLz	0.136	0.833***				
LLz ₁₅ (f)	-0.140	0.440*	0.156			
LLz _{7.5} (f)	-0.240	0.335	0.069	0.895***		
LLz ₁₅ (l)	-0.114	0.348	0.326	0.002	0.085	
LLz _{7.5} (l)	-0.112	0.414*	0.335	0.158	0.271	0.866***

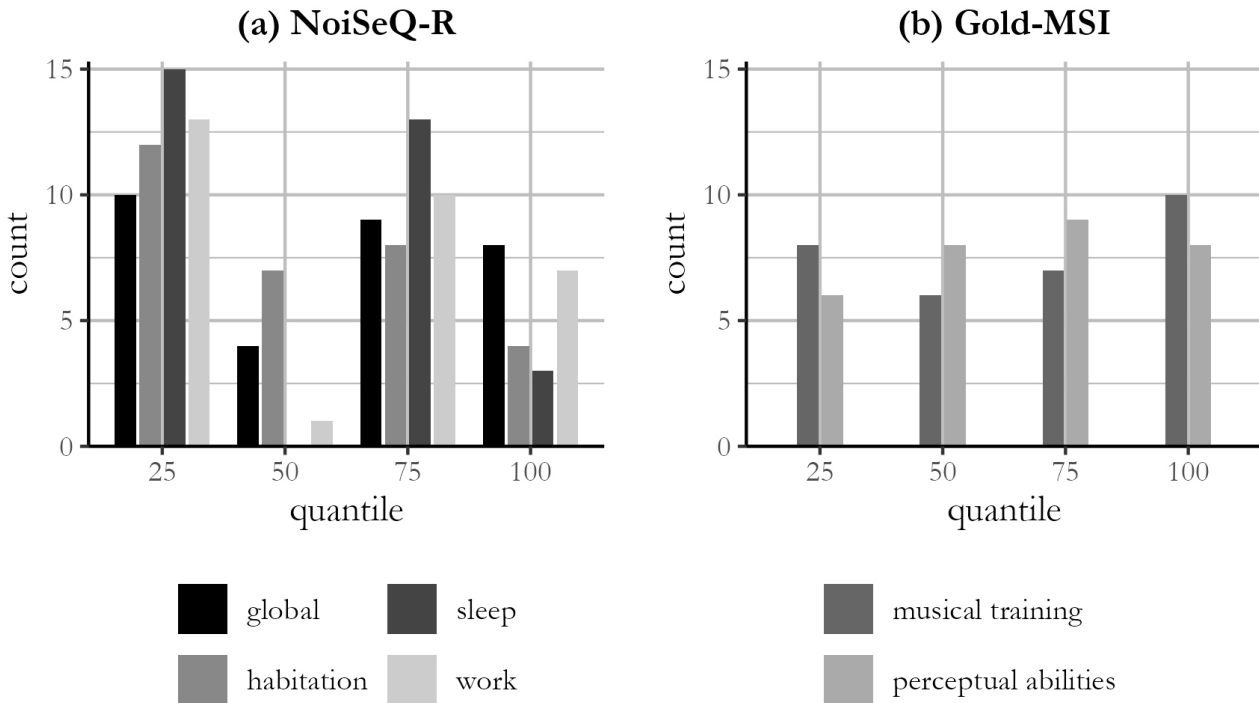


Figure 6. Quantiles for each sub-scale of the NoiSeQ-R (a) and Gold-MSI (b) inventories based on data norms from the respective original publications.

physically accurate reproduction gain compared to those of the laypersons.

The aggregate scores of the sub-scales of the NoiSeQ-R and the Gold-MSI were calculated by averaging the items belonging to each sub-scale according to the factor structures proposed in Schütte et al. [22] and Müllensiefen et al. [24], respectively. To compare the individual scores of our sample with those of the original studies of the two inventories, all scores were assigned to population quantiles according to data norms derived from larger samples of 288 participants in the case of the NoiSeQ-R [22], and 147 633 participants in the case of the Gold-MSI [24]. Note that the NoiSeQ-R quantile limits are based on the complete inventory, that includes more items and the two additional sub-scales leisure and communication. Figure 6 illustrates how many participants of our sample fell within which population quantile of the reference studies.

The Gold-MSI scores are evenly distributed among the four quantiles with the largest portions in the 100%-quantile of *musical training*, and the 75%-quantile of *perceptual abilities*, respectively. This is not surprising as the sample included several audio engineering students whose curriculum includes studying a musical instrument. To test for statistically significant interrelations between individual differences and the selected reproduction gain, we conducted bivariate correlations between NoiSeQ-R and Gold-MSI sub-scales, the adjusted gain value, and participants' age, as shown in Table 2. Because several variables, especially the NoiSeQ-R sub-scales, did not meet the assumption of normal distribution, Spearman correlations are reported.

Within the NoiSeQ-R and Gold-MSI, the respective sub-scales are expectably highly correlated (note that the general NoiSeQ-R score comprises the three sub-scales). However, there seems to be no linear association between

Table 2. Spearman correlations between the adjusted gain, NoiSeQ-R (N.S.) general score (g) and the sub-scales *habitation* (h), *sleep* (s), *work* (w), the Gold-MSI sub-scales *perceptual abilities* (perc. abil.), *musical training* (mus. train.), and participants' age. Statistically significant correlations are marked (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

	Gain	N.S. (g)	N.S. (h)	N.S. (s)	N.S. (w)	Perc. abil	Mus. train
NoiSeQ-R (g)	-0.407*						
NoiSeQ-R (h)	-0.556**	0.790***					
NoiSeQ-R (s)	-0.282	0.862***	0.565*				
NoiSeQ-R (w)	-0.237	0.849***	0.567*	0.560*			
Perc. abil	0.213	0.019	0.016	-0.032	-0.016		
Mus. train	0.292	0.011	0.047	0.020	-0.113	0.520*	
Age	-0.288	0.244	0.168	0.231	0.271	-0.311	-0.548*

noise sensitivity and musical sophistication as captured by these inventories.

There was a notable negative correlation between *musical training* and participants' age, which is most likely attributable to the above-mentioned portion of musically well-trained relatively young audio engineering students in our sample. General noise sensitivity according to the NoiSeQ-R, as well as its sub-scale *habitation*, show strong to moderate, statistically significant, negative correlations with adjusted gain. The other person-related factors exhibit small, statistically non-significant correlations below 0.3.

3.4 Effects of time delay and individual differences on adjustment gain

The present study's two research questions were whether the time delay between recording and reproduction on the one hand, and individual differences on the other hand, would affect the adjusted reproduction level. Therefore, linear regression models were fit to test to what extent the observed differences in adjustment gain could be explained by the time delay between reproduction and recording and the observed individual differences. Since previous research only found gain adjustments to be more concise among participants with experience in experimental procedures [7], the selection of individual differences predictors followed an exploratory approach based on best model fit. To avoid issues with multicollinearity, the models included only one sub-scale of each the NoiSeQ-R and the Gold-MSI. More precisely, the NoiSeQ-R sub-scale *habitation*, which showed the largest correlation with the adjusted gain, the two Gold-MSI sub-scales, and participants' age were tested as predictors. The time delay was always included as a predictor.

Consequently, four models were compared. Model 1 had only the time delay and the *habitation* sub-scale of the NoiSeQ-R as predictor. The second, Model 2A, additionally included the Gold-MSI score for *perceptual abilities* and Model 2B the score for *musical training*. Model 3 comprised the predictors time delay, *habitation*, *perceptual abilities*, and participants' age. This four-predictor model was only tested with the Gold-MSI sub-scale

perceptual abilities, as combining *musical training* and age in the same model would be susceptible to multicollinearity. Table 3 displays standardized regression coefficients for the models' individual predictors, as well as each models' R^2_{adj} , Akaike information criterion (AIC), and Bayesian information criterion (BIC). The binary variable "time delay" was dummy-coded so that the coefficient represents the change in the dependent variable upon including the time delay. For model comparison, analyses of variance (ANOVA) were calculated to test for significant model improvement by including additional predictors. To do so, Models 2A and 2B were each compared to Model 1, whereas Model 3 was compared to Model 2A. Table 3 displays each ANOVA's null model, the increment in explained variance (ΔSS), the increment in degrees of freedom (Δdf), and the F test statistic with its p value.

As mentioned before, the longer time delay between recording and reproduction seems to result in participants selecting a higher reproduction gain at first glance (see Fig. 5a). However, if individual differences are taken into account, the effect of the delay on the reproduction gains turns statistically non-significant in all tested models.

In terms of explained variance described by the models' R^2_{adj} values, Model 2A and Model 2B achieve nearly identical values, and outperform Model 1. The addition of age as a predictor in Model 3 slightly increased R^2_{adj} by about 0.01, but neither benefitted model fit according to the AIC and BIC, nor statistically significantly improved the model according to the ANOVA ($p = 0.244$). Therefore, both Model 2A and Model 2B seem to be equally good compromises between parsimonious models and a high variance explanation. Our data thus implies that the adjusted reproduction gain is negatively associated with noise sensitivity specific to personal habitation and positively associated with self-reported perceptual abilities in critical music listening as well as musical training. It should be kept in mind that, in the present study, higher reproduction gain tended to also mean a closer approximation to the physically accurate level. In both models, these individual differences account for 34.8% and 34.5%, respectively, of the variance in selected reproduction gain. The corresponding effect size measurements of Cohen's $f^2 = 0.534$ and Cohen's $f^2 = 0.527$, respectively, are commonly considered indicative of a strong effect [27].

Table 3. Regression model overview showing standardized regression coefficients (β) and probability values (p) for each predictor, and each model’s R_{adj}^2 , AIC and BIC. For “time delay”, the β represents the change in the dependent variable upon including the time delay. The last section displays ANOVAs testing model improvement due to inclusion of additional predictors, reporting the increment in sum of squares (ΔSS), the increment in degrees of freedom (Δdf), the F statistic and its p value.

Predictors	Model 1		Model 2A		Model 2B		Model 3	
	β	p	β	p	β	p	β	p
Time delay	0.340	0.305	0.261	0.404	0.527	0.109	0.273	0.379
NoiSeQ-R (h)	-0.477	0.002	-0.528	0.002	-0.468	0.006	-0.487	0.005
Perc. abil	-	-	0.329	0.036	-	-	0.281	0.080
Mus. train	-	-	-	-	0.333	0.039	-	-
Age	-	-	-	-	-	-	-0.184	0.245
<i>Model fit</i>								
R_{adj}^2	0.259		0.348		0.345		0.358	
AIC	83.533		80.416		80.566		80.772	
BIC	89.269		87.586		87.736		89.376	
<i>ANOVA</i>								
Null model	-		Model 1		Model 1		Model 2A	
ΔSS	-		3.158		3.072		.909	
Δdf	-		1		1		1	
F	-		4.920		4.787		1.416	
p	-		0.035		0.038		0.245	

4 Discussion

The present study investigated whether a bias in soundscape reproduction levels that participants deem appropriate found in previous literature also exists in an experiment where participants listened to the soundscape in-situ a few minutes before. The mean reproduction gain was 8.9 dB below the original level, and did not seem to depend on whether the reproduction was presented after 2 min, or after 20 min of completing another listening task. Instead, the reproduction gain was statistically significantly associated with individual noise sensitivity specific to habitation as well as with musical expertise in terms of critical listening or musical training.

The observed average level reduction of 8.9 dB is in line with results from previous research that did not include immediate previous in-situ exposure where the participants stated that gain reductions of 8 dB [10] and 9.5 dB [7] were the most appropriate. Previous research has suggested that this loudness bias roots in the quieter background levels in the lab, compared to the recording sites [6, 10]. Indeed, in both aforementioned studies where similar gain biases were observed, the experiment took place in listening rooms or quiet offices, arguably comparable to our sound-proven, quiet laboratory [7, 10].

In the semi-anechoic chamber utilized in the study by Davies et al. [6], participants selected an even larger gain reduction of 12.3 dB. Also, Fraisse et al. [5] employed a continuous noise exposure between stimulus presentations by presenting an uneventful baseline soundscape. Based on the subjective judgment of the experimenters, they found a comparatively small level reduction of 4 dB to be sufficient. These findings support the hypothesis that biased loudness judgments root in the adaptation of per-

ceptual scaling to the momentary acoustic environment. As a consequence, the loudness of a soundscape reproduction (especially when participants quickly rose the level from silence during the adjustment) would not just be evaluated based on the memory of the site, but also as a difference to a baseline established by the momentary environment. Our results suggest that the adaptation happens within seconds after entering the room.

Besides this hypothesized influence of lower background noise, the presented audio material itself, (i.e., the presentation of outdoor soundscapes in a laboratory) might have affected the loudness judgments. Here, previous research has observed that also pleasantness ratings were different and less pronounced in laboratory evaluations as compared to in-situ results [2, 8]. Hermida Cadena et al. [8] attributed this to multi-sensory integration, where visual information on noise coming from a distance would positively affect ratings of auditory pleasantness. Tarlao et al. [2] additionally suggested that contextual information in a broader sense, such as appreciating the historical significance of an urban space, might also mitigate perception of auditory pleasantness on site. In the present study, recordings were made at an ordinary busy street with a lot of traffic and a few shops and restaurants. However, participants could still have been more accepting of noise as a necessary consequence of traffic (including the streetcars and buses that they might use themselves every day) during the in-situ listening phase, and thus may not have remembered the traffic being as loud in the laboratory.

Moreover, while the 8–10 dB gain offset applies to the reproduction of outdoor soundscapes, Meyer-Kahlen et al. [28], by contrast, found that participants chose a fairly lower deviation mean of 3.6 dB below the original

for perceptually accurate reproduction levels of speech. One explanation for this discrepancy could be that a single speaker has presumably less variability in sound level than a soundscape, making both encoding loudness and actively determining an accurate reproduction level over the entire duration of a stimulus easier. In the present study, differences in the original sound level of the recordings did not systematically affect the selected reproduction gain – neither the average of the entire recordings, nor the first or last seconds. However, the L_{Aeq} standard deviation of about 6 dB indicates that there was overall little level variability in the soundscape recordings.

Furthermore, Meyer-Kahlen et al. [28] observed the error in adjusted reproduction level to be larger for louder speech (i.e., yelling) and smaller for recordings of familiar speakers. According to Susini et al. [29], loudness is a crucial factor of the recognition of sound sources, pointing towards some sort of reference encoding using a source prototype. This could explain that speech and in particular speech by known speakers is encoded with a more precise loudness reference than diffuse soundscapes. This internal reference might also remain more stable when the relative scaling range shifts due to the adaptation of perceptual anchors, following the concept of intensity perception being a compromise of relative and absolute scaling processes [13].

A dependency of the stability of loudness encoding on the sound(scape)’s familiarity and complexity also explains the effects of musicality and noise sensitivity on the individual ability to correctly reproduce a soundscape. Both Gold-MSI sub-scales were found to be positively associated with reproduction level. The two scales comprise questions specific to critical listening abilities, such as spotting mistakes in the rhythmical synchrony or tonal accuracy of a musical performance on the one hand, and formal musical training on the other. Training in critical listening and in playing an instrument could, on the one hand, improve the stability of loudness encoding in general, or, on the other hand, help participants to disentangle the soundscape’s acoustical complexity. This in turn could help identifying single sources within the scene that can be closely monitored and internally validated for plausible loudness—which would be an advantage specific to the reproduction paradigm employed in the present study.

The second personality trait that seemed to affect the individual reproduction gain was noise sensitivity, measured using a scale targeting the aspect of noise sensitive behavior in the context of habitation. For example, participants were asked whether they would be willing to live at a loud street. The recorded soundscapes were dominated by street noise, so it seems plausible that the *habitation* sub-scale shows the strongest correlations with adjusted reproduction gain. The correlation of this sub-scale as compared to general noise sensitivity demonstrated the advantage of measuring noise sensitivity as a multidimensional construct.

There are several explanations, why more noise-sensitive participants tended to select lower, less accurate

reproduction levels. First, the adjustment procedure that started by gradually increasing the level from inaudibility towards the remembered loudness might have triggered some sort of “protective” behavior. Tarlao et al. assumed that immersive reproductions of urban soundscapes in quiet labs are an “uncomfortable reminder of how pervasive traffic is in the city” [2]. Noise-sensitive participants might be particularly susceptible to this and thus not have increased the levels above a certain point to avoid discomfort caused by loud traffic noise.

Second, if perceptual scaling is affected by momentary background noise levels, more noise-sensitive participants might be more susceptible to this adaptation, contrary to the effect of musicality. There is also evidence that noise sensitivity might be linked to less efficient auditory processing [30], possibly making loudness encoding less stable and promoting adaptation effects.

Sudarsono et al. [7] found a sample of acoustics experts to be more consistent in the reproduction gain adjustment, albeit not more accurate, than non-experts. In the present study, a group with higher expertise (i.e., experience with acoustics) in fact tended to be less consistent in their responses. A study by Kreiman et al. [31] on the recognition of symptoms of dysphonia in voice recordings showed that experts used a greater variety of strategies to evaluate the recordings. The same might apply to recalling and reproducing soundscape loudness, where experienced listeners might choose different strategies such as either focusing on the overall loudness or specific sources, creating a greater variance in judgments.

4.1 Limitations

The design of the present study poses a few limitations that, at times, only allow rather speculative interpretation of the results. First, the experimental group that performed the reproduction level adjustment 20 min after the in-situ recording were administered another listening test in between. We explained our motivation to do so in Section 2.6, but different effects of this time-span might have been observed if participants would have been presented with another, non-auditory task performed in silence.

Second, the responses (i.e., gain adjustments) we obtained from participants are a result of both their ability to recall the loudness and maintain this recalled loudness during the adjustment procedure itself, where participants were fed new, changing loudness information that could, additionally to the laboratory background noise, continuously shift the references or anchors for loudness scaling. In addition, the recordings always started in silence, so that this continuous recalibration might have altered the recalled loudness in favor of lower levels during the adjustment procedure itself. We chose not to start at a randomly lower or higher gain level instead, as this fixed level might bias participants even more in direction of this reference. Also, the selected reproduction gains observed here did not seem to be critically affected by the adjustment procedure

as other studies with fixed level adjustments showed similar results [9, 10].

Finally, technical properties of the reproduction might have also contributed to loudness bias. The recordings contained a lot of traffic noise, including buses and street-cars that produce a substantial amount of low frequency rumble, that might be underrepresented when recordings are reproduced via headphones. This lack of low frequency content could have resulted as the remaining sound energy at higher frequencies appearing disproportionately loud to participants. Previous studies employing Ambisonics loudspeaker reproduction reported similar loudness biases [6, 7], but it might be worthwhile investigating whether different reproduction formats of the same recordings affect the reproduction level that participants deem appropriate.

4.2 Outlook for future research

The results of the present study as well as its limitations suggest several improvements in the experimental design. As the loudness or pleasantness of the soundscape were not assessed, a follow-up study where participants would be asked to rate unpleasantness and loudness during the in-situ recording could further investigate the importance of the soundscape's affective quality. Blind-folding participants appears also promising. Even though the findings by Lu and Lau [10] indicate that the lack of a visual stimulus in laboratory experiments is not the cause for biased loudness reproduction, it could help participants focus on the soundscape during the recording and closer align the conditions for sound exposure in-situ and in the laboratory.

Since, in the present study, each participant had to reproduce the loudness of one only recording, it would be interesting to either extend the paradigm to several recording sites, and to test if, for example, the observed association of reproduction level and noise sensitivity generalize to different types of soundscapes. Also, one could present recordings from the same site at different times during the day to the same person (e.g., with more or less traffic). As the street crossing where we made the recordings is an access path to the university campus well-known by many participants, the formation perceptual anchors for loudness scaling could also be influenced by long-term experience with this particular site.

Furthermore, it might be rewarding to let participants adjust a gain value starting from silence and then validate this adjustment, for example, by pausing the playback for some time – possibly with some background noise in between, as in previous studies [2, 5] – and to test after restarting the playback if participants would still agree with this level. Finally, one might argue that, in our results (and those of previous studies), the reproduction level and its accuracy are confounded, as louder levels usually correspond to a physically more accurate reproduction. To investigate whether the adjustments in the lab are indeed due to the adaptation to the environmental noise, it would be interesting to test if this bias also

works the other way around, that is, if recordings from a quiet site would be reproduced at higher gain level if the reproduction environment was louder.

5 Conclusion

The present study reproduced a previously observed bias in the sound level of laboratory-based reproduction of soundscapes, that participants deem plausible, in an experiment where participants experienced exactly the same soundscape in-situ minutes before. The persistence of this bias supports the hypothesis from previous research that perceptual loudness scaling adapts to the background noise in the lab. Moreover, the observed effects of individual musicality and noise sensitivity implies that the robustness of loudness encoding might be crucial to what extent the internal reference that participants use to judge loudness would ‘shift’ towards the background noise of the laboratory. These considerations on the cause of this bias are rather of theoretical interest for perceptual loudness encoding and scaling. From a practitioner’s perspective, our results further corroborate findings that soundscape playbacks in a quiet laboratory are perceived as more plausible when they are presented at a level of 8–10 dB below the original recording level. This threatens the ecological validity of experiments that rely on calibrated original playback levels, as these levels may seem too loud and implausible to participants, despite being physically accurate.

Acknowledgments

The authors would like to thank Alexander Mann for modeling and 3D-printing the windshields for the dummy head.

Conflicts of interest

The authors declare that they have no conflict of interest to disclose.

Data availability statement

Participants response data as well as level envelopes of the recordings are available in Zenodo, under the reference <https://doi.org/10.5281/zenodo.17663268>.

Author contribution statement

M. v. B., S. V., and J. S. equally contributed to the experimental design and wrote the final paper. S. V. designed the dummy head windshields. J. B. supervised the experiment and did preliminary data analysis. M. v. B. performed the final data analysis.

Informed consent

Participants provided written informed consent prior to participating in the study.

References

1. G. Keidser, N. Graham, D. Brungart, A. Caduff, J. Campos, S. Carlile, M.G. Carpenter, G. Grimm, V. Hohmann, I. Holube, S. Launer, T. Lunner, R. Mehra, F. Rapport,

- M. Slaney, K. Smeds: The quest for ecological validity in hearing science: What it is, Why it matters, and How to advance it. *Ear and Hearing* 41 (2020) 5S–19S.
2. C. Tarlao, D. Steele, C. Guastavino: Assessing the ecological validity of soundscape reproduction in different laboratory settings. *PLoS One* 17 (2022) e0270401.
 3. C. Guastavino, B.F.G. Katz, J.-D. Polack, D.J. Levitin, D. Dubois: Ecological validity of soundscape reproduction. *Acta Acustica united with Acustica* 91 (2005) 333–341.
 4. T. Oberman, K. Jambrošić, M. Horvat, B. Bojanić Obad Šćitaroci: Using virtual soundwalk approach for assessing sound art soundscape interventions in public spaces. *Applied Sciences* 10 (2020) 1–27.
 5. V. Fraisse, N. Schütz, M.M. Wanderley, C. Guastavino, N. Misdariis: Using soundscape simulation to evaluate compositions for a public space sound installation. *The Journal of the Acoustical Society of America* 156 (2024) 1183–1201.
 6. W.J. Davies, N.S. Bruce, J.E. Murphy: Soundscape reproduction and synthesis. *Acta Acustica* 100 (2014) 285–292.
 7. A.S. Sudarsono, Y.W. Lam, W.J. Davies: The effect of sound level on perception of reproduced soundscapes. *Applied Acoustics* 110 (2016) 53–60.
 8. L.F. Hermida Cadena, A.C. Lobo Soares, I. Pavón, J.L. Bento Coelho: Assessing soundscape: comparison between in situ and laboratory methodologies. *Noise Mapping* 4 (2017) 57–66.
 9. T. Yang, J. Kang: Perception difference for approaching and receding sound sources of a listener in motion in architectural sequential spaces. *The Journal of the Acoustical Society of America* 151 (2022) 685–699.
 10. Y. Lu, S.-K. Lau: Examining the ecological validity of VR experiments in soundscape and landscape research. *Computers in Human Behavior* 162 (2025) 108462.
 11. H. Li, S.-K. Lau: A review of audio-visual interaction on soundscape assessment in urban built environments. *Applied Acoustics* 166 (2020) 107372.
 12. L.D. Braida, J.S. Lim, J.E. Berliner, N.I. Durlach, W.M. Rabinowitz, S.R. Purks: Intensity perception. XIII. Perceptual anchor model of context-coding. *The Journal of the Acoustical Society of America* 76 (1984) 722–731.
 13. L.M. Ward: Remembrance of sounds past: memory and psychophysical scaling. *Journal of Experimental Psychology: Human Perception and Performance* 13 (1987) 216–227.
 14. M.C. Botte, C. Baruch, S. Mönikheim: Memory for loudness: the role of loudness contour, in: *Advances in the Auditory Physiology and Perception: Proceedings of the 9th International Symposium on Hearing held in Carcens, France, 9–14 June, 1991*, pp. 305–311.
 15. S. Clément, L. Demany, C. Semal: Memory for pitch versus memory for loudness. *The Journal of the Acoustical Society of America* 106 (1999) 2805–2811.
 16. S. Kumar, S. Joseph, P.E. Gander, N. Barascud, A.R. Halpern, T.D. Griffiths: A brain system for auditory working memory. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 36 (2016) 4492–4505.
 17. J. Hellbrück: Memory effects in loudness scaling of traffic noise – How overall loudness of short-term and long-term sounds depends on memory. *Acoustical Science and Technology* 21 (2000) 329–332.
 18. K. Ditttrich, D. Oberfeld: A comparison of the temporal weighting of annoyance and loudness. *The Journal of the Acoustical Society of America* 126 (2009) 3168–3178.
 19. H. Lee, D. Müllensiefen: The Timbre Perception Test (TPT): a new interactive musical assessment tool to measure timbre perception ability. *Attention, Perception & Psychophysics* 82 (2020) 3658–3675.
 20. M. von Berg, J. Steffens, S. Weinzierl, D. Müllensiefen: Assessing room acoustic listening expertise. *The Journal of the Acoustical Society of America* 150 (2021) 2539–2548.
 21. M. von Berg, H. Himmelein, J. Steffens: Effects of noise sensitivity and listening effort on perceptual ratings of background noise. *JASA Express Letters* 4 (2024) 084401.
 22. M. Schütte, A. Marks, E. Wenning, B. Greifahn: The development of the noise sensitivity questionnaire. *Noise & Health* 9 (2007) 15–24.
 23. T. Eikmann, A.Z. Nieden, D. Ziedorn, K. Römer, A. Lengler, S. Harpel, J. Pons-Kühnemann, H. Hudel, J. Spilski: Blood pressure monitoring, in: Vol. 5, Final report of NORAH: Research programm on Noise-Related Annoyance, Cognition and Health: A Transportation Noise Effects Monitoring Program in Germany, 2015.
 24. D. Müllensiefen, B. Gingras, J. Musil, L. Stewart: The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS One* 9 (2014) e89642.
 25. ISO 532-1: Acoustics – Methods for calculating loudness–Part 1: Zwicker method, International Organization for Standardization, Geneva, Switzerland, 2017.
 26. S. Versümer, J. Steffens, S. Weinzierl: Day-to-day loudness assessments of indoor soundscapes: exploring the impact of loudness indicators, person, and situation. *The Journal of the Acoustical Society of America* 153 (2023) 2956–2972.
 27. J. Cohen: *Statistical Power Analysis in the Behavioral Sciences*. Routledge, 1988.
 28. N. Meyer-Kahlen, S. de las Heras Pérez, T. Lokki: Expected levels of reproduced speech, in: *AES 5th International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2024. <https://aes2.org/publications/elibrary-page/?id=22661>.
 29. P. Susini, O. Houix, L. Seropian, G. Lemaitre: Is loudness part of a sound recognition process? *The Journal of the Acoustical Society of America* 146 (2019) EL172.
 30. M. Kliuchko, M. Heinonen-Guzejev, P. Vuust, M. Tervaniemi, E. Brattico: A window into the brain mechanisms associated with noise sensitivity. *Scientific Reports* 6 (2016) 39236.
 31. J. Kreiman, B.R. Gerratt, K. Precoda: Listener experience and perception of voice quality. *Journal of Speech and Hearing Research* 33 (1990) 103–115.