

Improving Subjective Listening Test Methodology: Investigating Crossmodal Effects of Ambient Lighting and Listener Expertise

SHUYUAN SUN,*   AES Student Member, AND YONG SHEN,   AES Member
(sunshuyuanly@gmail.com) (yshen@nju.edu.cn)

Key Laboratory of Modern Acoustics (Ministry of Education) and Institute of Acoustics, Nanjing University, 22 Hankou Road, Nanjing, 210093, People's Republic of China

In subjective listening tests for evaluating perceptual sound quality, numerous factors can significantly affect outcomes. In addition to well-established variables, recent studies have begun to explore the crossmodal effects of lighting-induced bias. Most current standards for subjective sound quality assessment lack explicit specifications regarding lighting conditions and still rely primarily on video evaluation protocols. This highlights a need to investigate the selection of ambient lighting conditions more thoroughly in listening tests. Furthermore, visual stimuli may modulate loudness perception and auditory thresholds—with these effects appearing to be frequency-dependent. Listener expertise, particularly in detecting linear distortions, emerges as a critical variable that should be rigorously controlled for in the ambient lighting design. In this study, participants with varying levels of experience comparatively evaluated multiple loudspeakers under different ambient lighting conditions, both before and after undergoing the listening training focused on the identification of linear distortions. Comparative analyses of pre-training and post-training results—including overall preference rating, data dispersion, intertrial consistency, and multiple comparison results—are presented. The findings provide evidence-based recommendations regarding appropriate ambient lighting in sound quality evaluations. A comprehensive understanding of lighting-induced bias and listener expertise contributes to the development of more reliable and standardized listening test methodologies.

Keywords: listening test, crossmodal effects, ambient lighting, listener expertise, subjective sound quality, linear distortion listening training

1 INTRODUCTION

In subjective listening tests for the evaluation of perceptual sound quality in sound reproduction systems, numerous factors can significantly affect the measurement outcomes. These include the physical dimensions of the listening room [1], the acoustic characteristics of the listening environment [2–8], the playback sound pressure level or perceived loudness [9, 10], the placement [11, 12] and visual appearance [13–15] of the loudspeakers, and listeners' experience and background [16–19]. Comprehensive summaries and reviews of these well-established variables are available in [20] and [21]. These valuable studies and empirical findings have directly contributed to the development of several

standards for subjective listening tests [22, 23], which now serve as key references for researchers and practitioners in the field.

However, explicit quantitative specifications regarding ambient lighting conditions are rarely included in existing standards. For most standardized acoustic listening tests, ambient lighting requirements are still primarily derived from guidelines originally developed for the subjective evaluation of video quality [24]. According to ITU-R BT.500-15 [24], the environmental illuminance on the screen should be 200 lux for subjective assessments in home environment, while only general guidance of low illumination, without specific numerical recommendations, is provided for laboratory-based assessments.

Moreover, previous studies have demonstrated that, in certain application scenarios, dim environments can be beneficial for enhancing specific sound attributes, including improved sensitivity to auditory details [25, 26], positive effects on perceived pleasantness and eventfulness [27], and enhanced perception of acoustic comfort [28]. Consequently, in certain domains, it is commonly suggested that

*To whom correspondence should be addressed, email: sunshuyuanly@gmail.com

darker ambient environments may be more conducive to enhanced perceived sound quality. Nevertheless, this assumption appears to lack sufficient theoretical grounding and empirical validation. The absence of well-defined ambient lighting specifications not only hampers the development of more scientific and comprehensive standards, but also limits efforts to conduct more reliable and accurate evaluations of perceptual sound quality.

Notably, Bech and Zacharov [20] and Bech [29] offered a preliminary yet highly instructive discussion on the significance of ambient lighting in perceptual audio quality evaluation. In addition, numerous valuable studies have explored multisensory perception within the context of audiovisual interactions [30–43]. A recent study investigating the crossmodal effects of lighting on perceptual sound quality in subjective listening tests [44] yielded several noteworthy initial findings that motivate further research aimed at improving the design of such tests. The published work identified some lighting-induced biases in perceptual sound quality evaluation.

Although ambient lighting environment did not significantly impact the overall subjective sound quality ratings, extremely bright or dim lighting conditions were found to amplify the influence of program material on the evaluation outcomes. Moreover, the interactions between the program materials and the loudspeakers became more pronounced under these extreme lighting conditions. Additionally, the dispersion and distribution of the raw data were also influenced, to some extent, by the lighting conditions in the listening environment. These crossmodal effects induced by ambient lighting may help explain the common misconception that dim environments are inherently more conducive to improved subjective sound quality.

Although the crossmodal effects of ambient lighting on subjective sound quality assessment have been preliminarily explored, several limitations remain in previous research. First, all the listeners involved in the prior study [44] had backgrounds in acoustics and possessed extensive experience in listening experiments. This resulted in a relatively small number of test participants and limited assessment repetitions of each trial. Consequently, it was not possible to further validate the appropriate ambient lighting levels for listening tests by analyzing the crossmodal effects of lighting-induced bias across a more general population. Second, the selection of the ambient lighting conditions was primarily based on existing literature and prior research findings [45–47]. To improve the generalizability of the results, a broader range of conditions—particularly ambient lighting levels beyond the illuminance ranges previously studied—should be considered.

Ultimately, and most importantly, subjective loudness perception [48, 49] and auditory thresholds [50, 51] may be modulated by visual stimuli, with these effects appearing to correlate with the frequency characteristics of the auditory source. Therefore, the selection of ambient lighting conditions in subjective listening tests may be influenced by the listeners' ability to detect linear distortion. This further suggests that listener expertise constitutes a crucial variable that should be rigorously controlled for in the design

of ambient lighting conditions for subjective sound quality evaluations. Furthermore, in real-world engineering applications, listeners often differ in their auditory judgment proficiency. Consequently, examining the differences between trained and untrained listeners holds considerable practical relevance.

To date, generally, the interactions among ambient lighting, auditory stimuli, and listener expertise—and their combined effects on the subjective perception of sound quality—remain insufficiently understood. This renders the selection of appropriate ambient lighting conditions for subjective listening tests a fundamental open question. Since existing standards often lack clarity in this regard, the findings of this study provide valuable empirical evidence to support the refinement of subjective listening test methodologies.

Following this introduction, SEC. 2 describes the design of the expanded and refined listening experiments, including the training procedures. The main results and comparative analyses are detailed in SECS. 3 and 4, respectively. Recommendations regarding suitable ambient lighting levels for conducting perceptual sound quality evaluation are proposed in SEC. 5. Finally, SEC. 6 summarizes the conclusions and outlines directions for future research.

2 LISTENING EXPERIMENTS AND TRAINING

To further investigate the effects of ambient lighting levels on the improvement of subjective listening tests, a series of evaluations were conducted under a variety of ambient lighting conditions. Initially, multiple listeners were asked to comparatively evaluate the overall sound quality preferences of five loudspeakers under varying ambient lighting environments, without any prior training. Following this baseline assessment, the listeners underwent a targeted training focused on the identification of linear distortions. After completing the training, they repeated the same preference-rating task for the same loudspeakers under the identical set of lighting conditions. All experiments were conducted as double-blind listening tests in a standardized listening room. Compared to previous studies [44], the present work employed an expanded and refined experimental design, featuring a more heterogeneous listening panel, a larger sample size of participants, increased repetition of test segments, and a broader range of ambient lighting levels within the listening environment.

2.1 Listening Environment With Controllable Ambient Lighting

All experiments and training described in this study were conducted in a rectangular listening room with a volume of 216 m³ (8 × 7.5 × 3.6 m). The measured reverberation time (RT₆₀) decreased from 0.5 to 0.2 s across the frequency range of 100 Hz–4 kHz. The experimental setup, as illustrated in Fig. 1, featured a ceiling-mounted lighting system with adjustable illuminance, allowing precise control of ambient lighting conditions throughout the tests. To eliminate the influence of loudspeaker appearance on evaluation

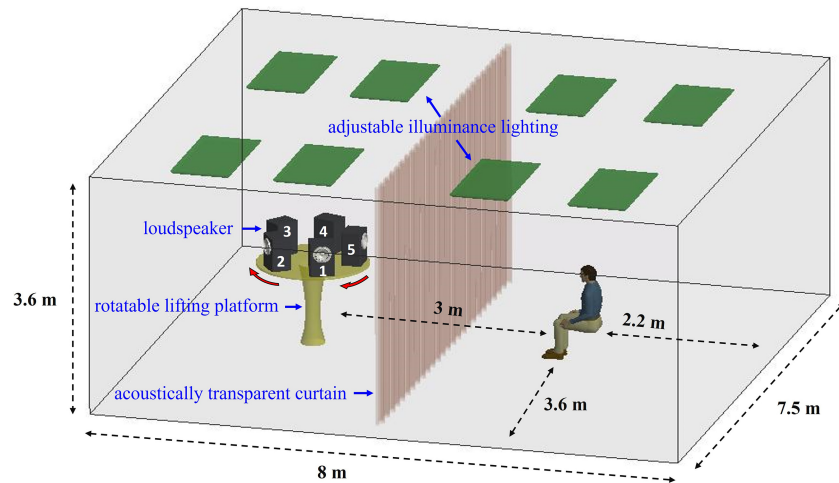


Fig. 1. Diagram of the experimental configuration. The ceiling of the listening room was equipped with a lighting system allowing precise control of illuminance. The listening distance between the loudspeaker and the listener was 3 m. The listening position was located 2.2 m from the rear wall and 3.6 m from the side wall.

outcomes, an acoustically transparent curtain, composed of polyester and microperforated polyvinyl chloride composites, was placed between the listener and the loudspeakers. These curtains are designed to block light while allowing sound waves to pass through with minimal attenuation, and they are widely used in critical listening environments.

The test loudspeakers were positioned on a height-adjustable, clockwise-rotatable (as indicated by the red arrow in Fig. 1) platform to ensure that the acoustic center of each loudspeaker was aligned with the listener's ear level during every listening session. The time interval for switching between any two loudspeakers was kept within 2 s. For instance, the transition from Loudspeaker 5 to Loudspeaker 4 yielded the shortest switching time, while the longest switching time of 2 s was observed for the transition from Loudspeaker 5 to Loudspeaker 1, as shown in Fig. 1. This interval satisfied the assumption that human short-term memory lasts approximately 3–5 s [52].

2.2 Listening Panel and Program Material

In the previous study [44], the listening panel consisted of ten professionals with backgrounds in acoustics. In the present study, the panel was expanded to include 30 listeners with diverse backgrounds, comprising 12 females and 18 males, aged between 18 and 40 years. All participants reported normal hearing. Considering the potential influence of varying listening experience and background on subjective sound quality evaluations [16–19, 53], the 30 listeners were categorized into three groups, as follows:

- Subject Group I (Professionals): individuals with a certain background in acoustics, either as professionals in the audio-related industry or students majoring in acoustics.
- Subject Group II (Audiophiles): individuals who own at least one piece of high-fidelity audio equipment, such as a high-end home audio system or high-

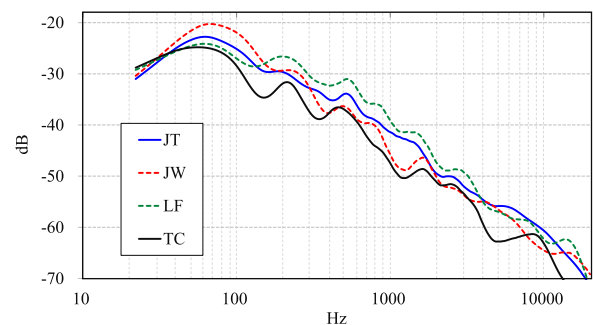


Fig. 2. The long-term spectrum (one-third-octave smoothing) of each program material.

quality headphones, and who regularly update or upgrade their listening devices.

- Subject Group III (General Listeners): Individuals with no formal background in acoustics or not identified as audiophiles.

The distribution of the subject groups within the listening panel was as follows: Subject Group I (Professionals) comprised 40% of the panel (12 out of 30 participants), Subject Group II (Audiophiles) accounted for 20% (six out of 30), and Subject Group III (General Listeners) made up the remaining 40% (12 out of 30).

The listening program materials and training audio samples are outlined below:

- JT - James Taylor, “That’s Why I’m Here”;
- JW - Jennifer Warnes, “Bird on a Wire”;
- LF - Little Feat, “Hangin’ on to the Good Times”;
- TC - Tracy Chapman, “Fast Car.”

All program materials were sampled at 44.1 kHz/16 bit and edited to a duration of 15–35 s. Fig. 2 shows the long-term spectral analysis of each program: Program TC exhibits the largest spectral fluctuations, Programs JW and

Table 1. Center and cutoff frequencies used for equalization during listening training.

Band	Center or Cutoff Frequency	Critical-Band Rate
1	50 Hz	0.5 Bark
2	350 Hz	3.5 Bark
3	700 Hz	6.5 Bark
4	1,170 Hz	9.5 Bark
5	1,850 Hz	12.5 Bark
6	2,900 Hz	15.5 Bark
7	4,800 Hz	18.5 Bark
8	8,500 Hz	21.5 Bark

LF emphasize low and mid frequencies, respectively, and Program JT presents relatively small variation with a balanced frequency distribution. Overall, the selected program materials, characterized by broadband spectral content and sufficient spectral density, have been validated in previous studies as appropriate for use in both subjective listening evaluations [53, 54] and listening training [55], as they facilitate the perceptual discrimination of differences among loudspeakers.

2.3 Listening Training

Substantial and valuable research has been conducted on listening training methodologies, encompassing approaches tailored to specific sound attributes [56–65], as well as approaches designed for various target groups and application scenarios [66–70]. The listening training method adopted in this study primarily follows the methodologies proposed by Olive [55, 71] and Bech [72, 73]. This selection is based on two main considerations: first, their methods have been widely recognized and applied in both academic and industrial contexts; second, both authors have conducted in-depth comparisons of subjective sound quality evaluations between trained and untrained listeners [74, 75]. Therefore, their approaches are particularly well-suited for beginners—namely, individuals without formal listening training or with limited experience in listening experiments.

The listening training in this study primarily focused on enabling listeners to identify linear distortions across different frequency bands. These linear distortions included peak/dip filters, high-pass/low-pass filters, and high-shelf/low-shelf filters. The selection of center and cutoff frequencies was guided by the critical band principle [76–78] to ensure that the frequency band divisions of the linear distortions within the audible range closely corresponded to the subjective frequency resolution characteristics of the human auditory system [79–83], as summarized in Table 1. The program materials used for the listening training were identical to those described in SEC. 2.2 and were randomly presented throughout the training process.

The listening training was divided into six groups. In each group, participants were tasked with correctly identifying the equalization profile applied by comparing the equalized version of a selected program material to its original version. Each group focused exclusively on one type of linear distortion, as illustrated in Figs. 3(a)–3(f), representing

the six different training groups, respectively. The listening training task involved selecting the correct equalization curve—among eight of the identical type but differing in center or cutoff frequency—that was currently applied to the auditory stimulus. Furthermore, the sequence of training for the six types of equalization was not strictly prescribed, allowing listeners to complete the groups in an order of their own preference.

The playback chain of the listening training setup consisted of a laptop, an audio interface, and a pair of professional over-ear monitor headphones characterized by a smooth frequency response and minimal sound leakage. All equalization processing was conducted on the laptop. By the end of the training, all listeners had acquired the ability to perceptually discriminate spectral changes across eight sub-bands within the audible frequency range. Specifically, for each of the six listening training groups illustrated in Fig. 3, every participant achieved an accuracy rate exceeding 80% in their most recent 40 listening judgments within each group. Notably, participants were not subject to any time constraints or restrictions on the number of judgments during the listening training.

2.4 Experimental Design

The listening tests employed a comparison category rating method to evaluate different loudspeakers using a monophonic playback system, with an emphasis on preference-related sound quality rather than spatial attributes. Listeners were asked to comparatively evaluate five loudspeakers (Loudspeakers 1, 2, 3, 4, and 5) on an 11-point preference scale. Each participant completed four repeated evaluations per program material, resulting in a total of 16 trials (4 program materials $\times$ 4 repetitions) per listening test. The preference scale allowed ratings with a resolution of one decimal place and included semantic descriptors at every second interval, labeled as follows: really dislike (1), dislike (3), neither like nor dislike (5), like (7), and really like (9). Generally, a strong preference between evaluated objects was defined as a rating difference of ≥ 2 points, a moderate preference as ≥ 1 point, and a negligible or no preference as ≤ 0.5 points. All loudspeakers were calibrated according to ITU-R BS.1770 [84], with an average playback level of 80-dB SPL (A-weighted).

Listeners were free to switch between loudspeakers, pause playback at any time, and concurrently complete the evaluations using the software application on a laptop, which was adapted from a MATLAB-based interface originally designed for Multiple Stimuli with Hidden Reference and Anchor listening tests [85], as shown in Fig. 4. To ensure randomization of both loudspeakers and program material presentation across trials, the labels “Speaker A” through “Speaker E” and “Program 1/4” through “Program 4/4” displayed in the graphical user interface (GUI) corresponded to different physical loudspeakers and program materials in each trial. For instance, in one trial, the speakers labeled as A, B, C, D, and E referred to Loudspeakers 1, 2, 3, 4, and 5, respectively; whereas in another trial, the same labels represented Loudspeakers 3, 4, 1, 5, and 2.

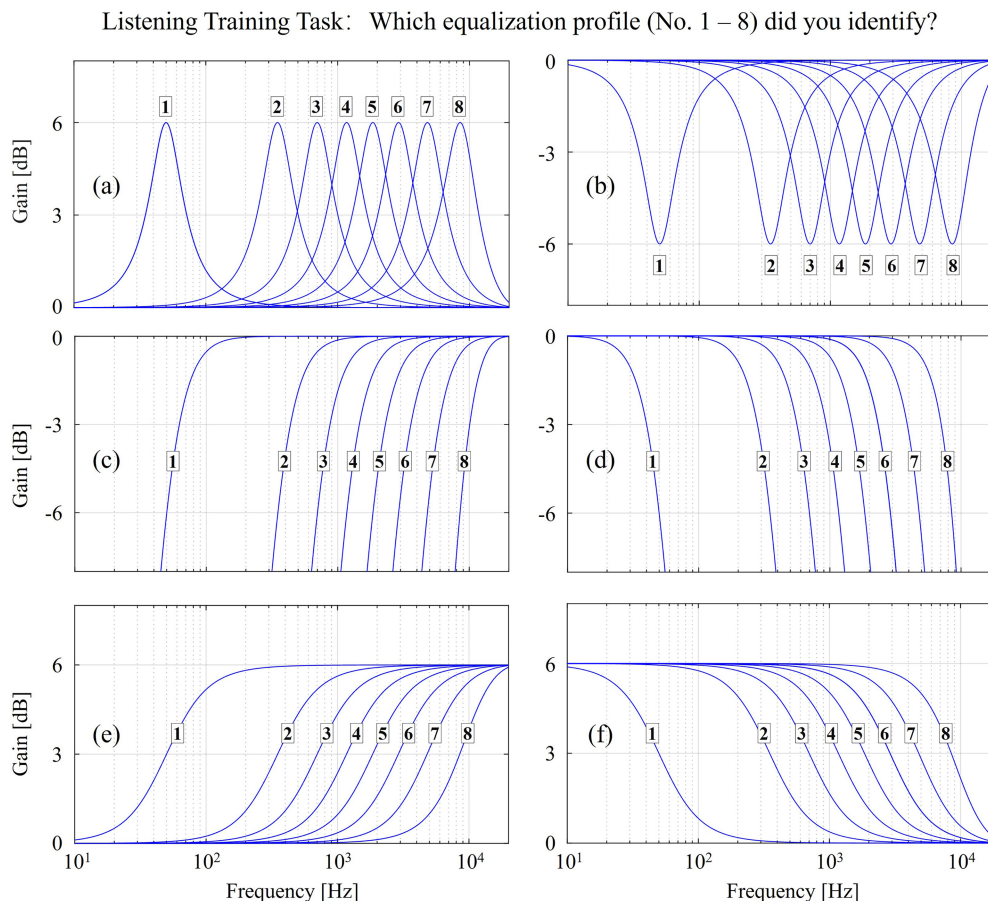


Fig. 3. Equalization curves used during the listening training. The training was divided into six groups: (a) Group 1—peak training; (b) Group 2—dip training; (c) Group 3—high-pass training; (d) Group 4—low-pass training; (e) Group 5—high-shelf training; (f) Group 6—low-shelf training. Each equalization curve was labeled with an index (No. 1–8), denoting that its center or cutoff frequency corresponded to the frequency associated with the same number in Table 1.

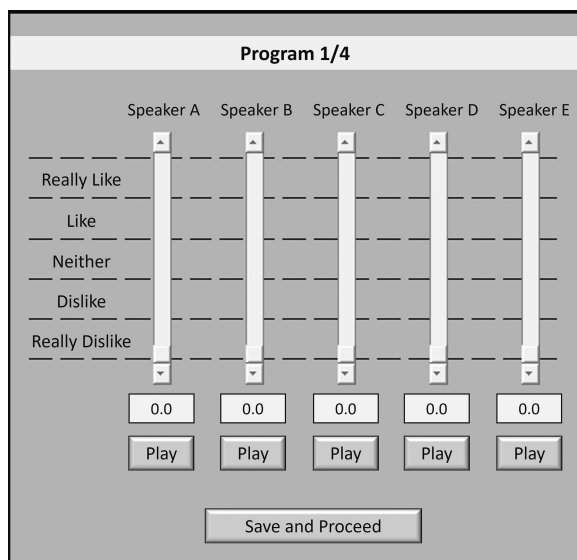


Fig. 4. GUI used in the listening tests.

Prior to the listening training, a total of seven listening tests were conducted. Each test followed the same procedure, differing only in the ambient lighting conditions, as summarized in Table 2. The chromaticity in all experiments was set to D_{65} , as recommended in [24]. The condition with

Table 2. Ambient lighting conditions used in the listening tests.

Test Serial Number	Lighting Level
1	Illuminance = 530 lux
2	Illuminance = 380 lux
3	Illuminance = 250 lux
4	Illuminance = 150 lux
5	Illuminance = 60 lux
6	Illuminance = 6 lux
7	Illuminance = 0 lux

an illuminance level of 0 lux represented a completely dark ambient environment; however, some light still emanated from the GUI. Prior to the experiment, over half of the participants reported that an illuminance level of 530 lux nearly approached the upper limit of what they considered acceptable brightness for a listening room. Although the lighting remained within a tolerable range and did not cause discomfort, the intensity of visual stimulation was sufficiently high that some listeners had to exert conscious effort to fully concentrate on auditory perception.

Regarding subjective perception under different ambient lighting levels, findings from studies on brightness assessment in furnished small rooms or offices environments [45–47] may offer partial reference: most participants perceived

lighting above 300 lux as too bright, whereas lighting below 10 lux was generally rated as too dim. The lighting conditions were randomized for each listener and not presented in any fixed sequence (e.g., from bright to dark or vice versa).

Subsequently, all members of the listening panel underwent a period of training focused on identifying linear distortions across different frequency bands, as described in SEC. 2.3. Upon completion of the training, a second round of seven listening tests was conducted using the same testing methods and ambient lighting conditions as those applied prior to the training. There were no time limits imposed on the evaluations for any listener, and the listening duration was restricted to no more than 30 min per day to minimize the potential effects of listener fatigue [86, 87].

3 RESULTS

3.1 Statistical Model

The primary objective of most listening tests is to investigate the relationship between multiple independent variables and a single dependent variable. Typically, the relationship is analyzed separately for each dependent variable, making univariate analysis of variance (ANOVA) the most commonly employed statistical method. In this study, the normality of the measured data was first assessed using the Kolmogorov–Smirnov test, while homogeneity of variances was evaluated using Levene’s test. In addition, deviations from normality were also examined visually. Although the assumptions required for applying ANOVA were not fully met, minor violations were deemed acceptable due to the robustness of the ANOVA procedure.

Subsequently, a five-way ANOVA was conducted separately for the pre-training and post-training listening tests. For both statistical models, the analyzed data combined evaluation results from all seven listening tests conducted under different ambient lighting conditions. The within-subject fixed factors included **Program** (four levels, corresponding to the four program materials), **Repetition** (four levels, reflecting four repetitions per listening test), **Lighting** (seven levels, representing seven distinct ambient lighting conditions), and **Loudspeaker** (five levels, corresponding to five different test loudspeakers), while between-subject factor was **Subject Group** (three levels, indicating that participants were categorized into three groups). The dependent variable was **Preference Rating**. Additionally, the analysis models included all second-order interaction effects involving the **Loudspeaker** factor, as well as all third-order interactions involving **Loudspeaker** × **Lighting**.

Although each variable has a defined number of levels, not all are ordinal, meaning that their levels cannot be arranged in a meaningful sequence. Only **Lighting** (seven levels corresponding to illuminance values) possesses a natural order. **Program**, **Repetition**, **Loudspeaker**, and **Subject Group** are nominal categorical variables, with levels representing distinct categories without intrinsic ranking. In the ANOVA model, these variables were represented

Table 3. ANOVA results of the pre-training listening test.

Source	df	F	Sig. = <i>p</i> value
Lighting	6	1.167	0.336
Repetition	3	9.353	0.230
Program	3	17.967	0.029
Subject Group	2	9.546	≤ 0.001
Loudspeaker	4	14.348	≤ 0.001
Loudspeaker × Lighting	24	0.605	0.913
Loudspeaker × Repetition	12	5.912	0.558
Loudspeaker × Program	12	15.17	0.003
Loudspeaker × Subject Group	8	50.132	≤ 0.001
Loudspeaker × Lighting × Repetition	90	7.797	≤ 0.001
Loudspeaker × Lighting × Program	90	2.544	≤ 0.001
Loudspeaker × Lighting × Subject Group	60	9.043	≤ 0.001

Note. Dependent variable: Preference Rating.

numerically solely as category codes, without any numerical magnitude. The analysis treated them as categorical factors rather than numeric continuous variables. Dummy (indicator) variables for these categorical factors were automatically generated for the computations. This coding approach does not affect the statistical interpretation, as the model does not utilize the arithmetic relationships among the codes.

Additionally, it is important to note that, given the potential influence of listeners’ experience and background on the outcomes of subjective evaluations, **Subject Group** was treated as a between-subject factor in the analysis model. However, the data collected in this study do not support definitive conclusions regarding the interaction between listeners’ background and ambient lighting conditions in shaping perceived sound quality. This limitation is primarily due to the relatively small sample size of participants. For more scientifically rigorous methods and generalizable findings, reference can be made to the work of Olive et al. [53], which involved over 200 listeners from four different countries with diverse ages, genders, and levels of listening experiences.

3.2 Main Effects in Each Test

As shown in Tables 3 and 4, the ANOVA results of pre-training and post-training listening tests are presented. The analysis indicates that, in all listening tests, **Loudspeaker** was a statistically significant factor influencing subjective sound quality ratings, suggesting that the different loudspeakers were consistently and significantly distinguished by the listeners. In contrast, the effect of **Repetition** was not statistically significant, indicating that the listening test results were reproducible. Additionally, in the pre-training listening tests, significant interaction effects involving the **Program** and **Subject Group** factors were observed. These findings are consistent with those reported in [44] and [53].

It is noteworthy that, in the pre-training listening tests, a significant third-order interaction effect was observed among **Loudspeaker**, **Lighting**, and **Repetition**, indicat-

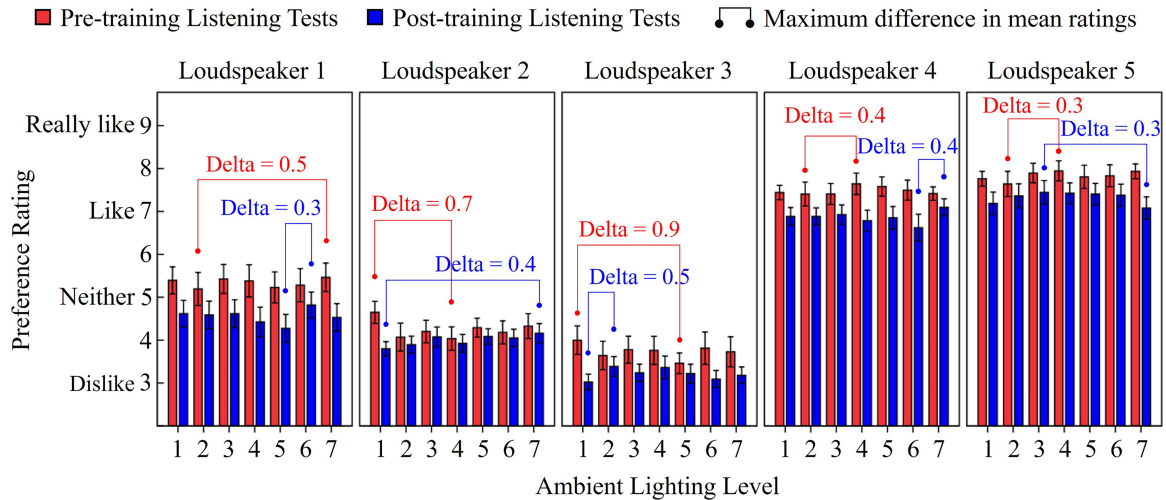


Fig. 5. Mean preference ratings of overall subjective sound quality in both the pre-training and post-training listening tests, along with the corresponding 95% confidence intervals. Each column of subplots represents a tested loudspeaker. In each subplot, the x-axis denotes the ambient lighting condition, with values from 1 to 7 corresponding to illuminance levels of 530, 380, 250, 150, 60, 6, and 0 lux, respectively (i.e., Tests 1–7). The y-axis indicates the subjective sound quality ratings. For each loudspeaker, the maximum differences in overall subjective sound quality ratings observed under different ambient lighting conditions are visualized by differently colored solid lines with dot markers at both endpoints, separately for the pre-training and post-training conditions.

Table 4. ANOVA results of the post-training listening test.

Source	df	F	Sig. = <i>p</i> value
Lighting	6	0.332	0.918
Repetition	3	4.625	0.429
Program	3	5.324	0.326
Subject Group	2	1.216	0.346
Loudspeaker	4	18.788	≤ 0.001
Loudspeaker × Lighting	24	0.783	0.743
Loudspeaker × Repetition	12	2.794	0.668
Loudspeaker × Program	12	7.206	0.006
Loudspeaker × Subject Group	8	50.971	≤ 0.001
Loudspeaker × Lighting × Repetition	90	1.050	0.355
Loudspeaker × Lighting × Program	90	0.527	1.000
Loudspeaker × Lighting × Subject Group	60	14.765	≤ 0.001

Note. Dependent variable: Preference Rating.

ing that the reproducibility of the perceptual sound quality ratings for the loudspeakers was influenced by the ambient lighting conditions. This finding has not been reported in previous studies. In fact, during the pre-training phase of this study, listeners required a certain number of repeated evaluations to provide consistent subjective ratings for the same loudspeaker under varying ambient lighting conditions. Further analysis and discussion of this effect are presented in SEC. 4.

Overall, a comparison of the ANOVA results from the pre-training and post-training test data revealed several key findings. First, the statistical significance of the **Subject Group** factor was notably reduced after the listening training, indicating the training's effectiveness in aligning perceptual evaluations across listeners with varying backgrounds. Second, the interaction effects of **Loudspeaker**

× **Lighting** × **Repetition** and **Loudspeaker** × **Lighting** × **Program** were also substantially reduced. This suggests that the listening training had a beneficial effect in reducing the perceptual variability caused by repeated evaluations under different ambient lighting conditions, as well as in minimizing the influence of program material under varying ambient lighting environments. These results underscore the importance that ambient lighting conditions should be tailored differently for trained and untrained listeners. Further analysis and discussion of this phenomenon are provided in SEC. 4.

Both before and after the listening training, the interaction effects involving the **Subject Group** factor remained statistically significant. This suggests that, even after training, the crossmodal effects of ambient lighting on subjective sound quality perception are still influenced by individual differences—likely shaped by the interaction between a listener's unique background and the characteristics of the auditory stimuli. Unfortunately, due to the limited number of participants in this study, no definitive conclusions can be drawn regarding whether listening training is more effective for specific subject groups. Further in-depth research with larger and more diverse participant populations is required to investigate this issue more comprehensively. Nevertheless, the results related to the **Subject Group** factor in this study may still provide useful analytical insights and serve as a reference for future research. To this end, raw data are provided in APPENDIX A.1 without additional discussion or interpretation, allowing interested readers to examine and analyze them independently.

3.3 Mean Opinion Scores of Each Test

As shown in Fig. 5, the mean opinion scores of overall preference ratings, along with the corresponding 95% confidence intervals, are presented. Each subplot illustrates

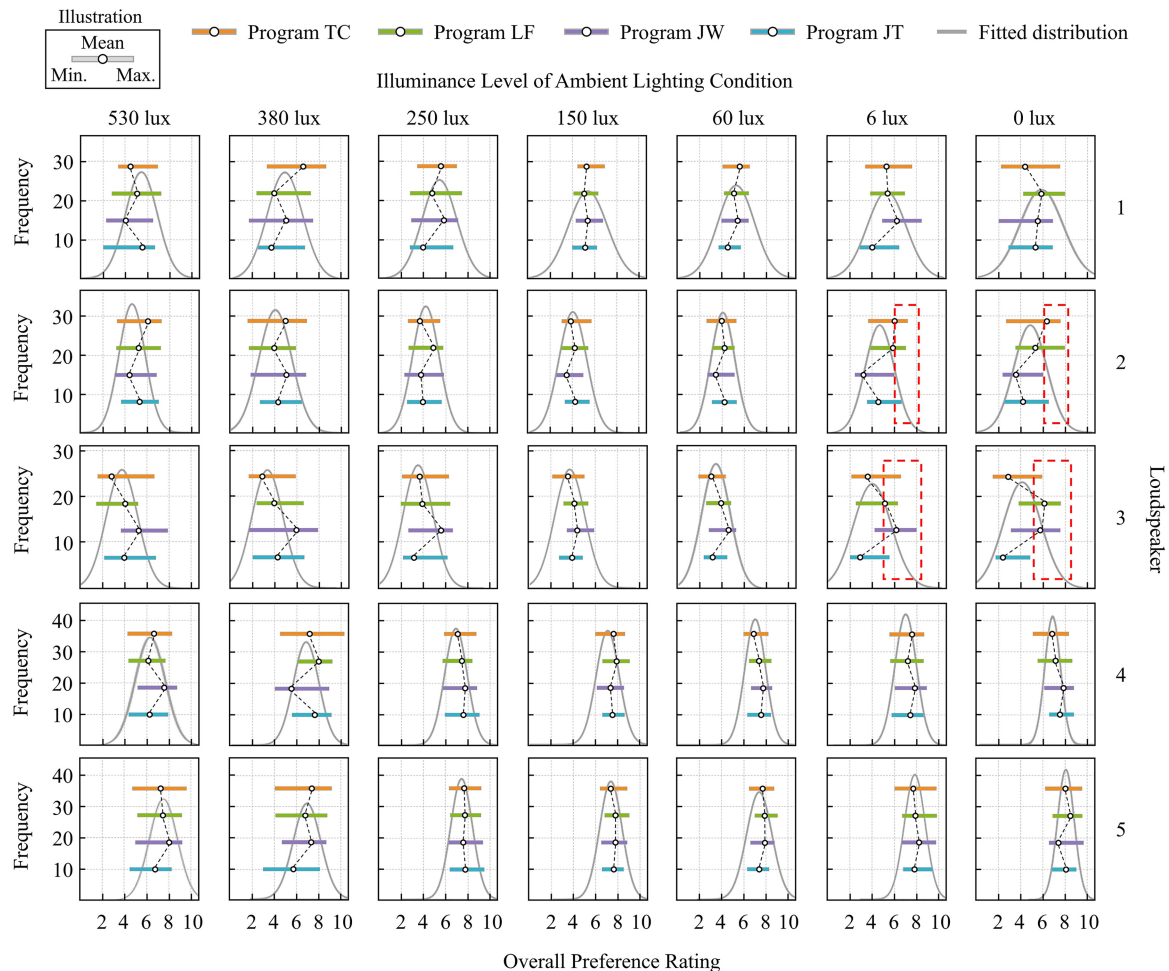


Fig. 6. Preference ratings of overall subjective sound quality for different loudspeakers under various program materials in the pre-training listening tests. Subplots in different columns correspond to the tests conducted under different ambient lighting conditions, while subplots in different rows represent evaluation results for different loudspeakers. The x -axis shows the preference rating, and the y -axis depicts the distribution of raw rating data. The red dashed box highlights the region of overly high ratings. To facilitate a clear comparison, the colored program material bars are set at different heights on each subplot, and the y -axis has no correlation with the colored bars.

the subjective evaluations of the same loudspeaker under different ambient lighting conditions, both before and after listening training. Overall, Loudspeakers 2 and 3 were rated as low in subjective sound quality, with most scores falling below 4 points, reflecting a generally negative perception. Loudspeaker 1 was perceived as medium sound quality, with most ratings around 5 points, indicating a relatively neutral perception. Loudspeakers 4 and 5 were perceived as high sound quality, with ratings around 7 points or higher, corresponding to a relatively positive perception.

No statistically significant differences were observed in the subjective sound quality ratings of the loudspeakers across various ambient lighting conditions when comparing the results of pre-training and post-training listening tests. Specifically, the differences in mean preference scores across lighting environments for each loudspeaker did not reach the threshold for strong preference (≥ 2 points); most of the differences can be categorized as either a moderate preference difference (≥ 1 point) or negligible preference (≤ 0.5 points). These findings further support the conclusion that variations in ambient lighting environments do not have

a statistically significant effect on perceptual sound quality in subjective listening tests, which is consistent with the results reported in [44]. Moreover, the presented results also challenge the commonly held assumption that darker environments are substantially more conducive to improved perceived sound quality. Based on a systematic and comprehensive listening test procedure, different ambient lighting conditions appeared to exert only a limited influence on overall subjective sound quality ratings.

However, in the subjective listening tests conducted prior to the listening training, greater variability and inconsistency in evaluation results across different ambient lighting conditions were observed for sound reproduction systems with lower perceived sound quality (Loudspeakers 2 and 3), compared to those with higher perceived quality (Loudspeakers 4 and 5). As illustrated in Fig. 5, the maximum differences in overall subjective sound quality ratings observed between listening tests conducted under different ambient lighting conditions were more pronounced for Loudspeakers 2 and 3. Their mean preference ratings exhibited greater fluctuations, and the corresponding 95%

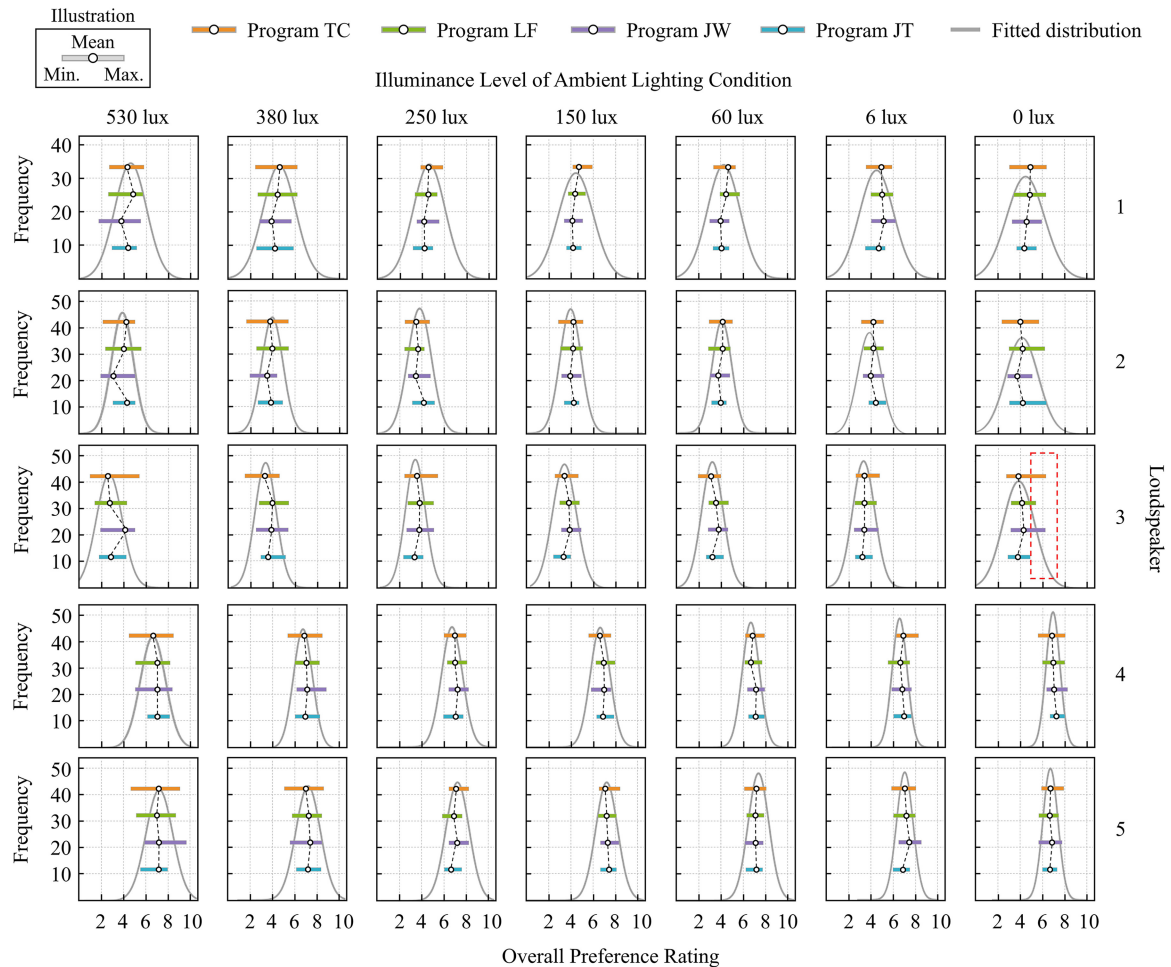


Fig. 7. Preference ratings of overall subjective sound quality for different loudspeakers under various program materials in the post-training listening tests. Subplots in different columns correspond to tests conducted under different ambient lighting conditions, while subplots in different rows represent evaluation results for different loudspeakers. The x-axis shows the preference ratings, and the y-axis depicts the distribution of raw rating data. The red dashed box highlights the region of overly high ratings. To facilitate a clear comparison, the colored program material bars are set at different heights on each subplot, and the y-axis has no correlation with the colored bars.

confidence intervals were noticeably wider than those of Loudspeakers 4 and 5. These observations suggest that, prior to the listening training, the subjective evaluation results of sound reproduction systems with low perceived quality were more susceptible to the influence of ambient lighting environments. This finding is also consistent with the results reported in [44].

Overall, a comparison of the results from the pre-training and post-training listening tests indicates that no significant differences were observed between the test outcomes before and after the listening training. Under most conditions, the differences in mean preference ratings were within 1 point, and the ranking order of sound quality preference among the tested loudspeakers remained largely unchanged, except for Loudspeakers 4 and 5 under the ambient lighting condition with an illuminance level of 0 lux (Test 7), detailed discussions are provided in SEC. 5.3.2. Trained listeners tended to assign lower ratings, and the confidence intervals for the loudspeakers with lower perceived quality became narrower, which is consistent with the results reported in [74].

The key difference is that lighting-induced bias was mitigated in the subjective listening tests conducted after the listening training. As shown in Fig. 5, compared to the pre-training results, the variability and fluctuations in the mean preference ratings for Loudspeakers 1, 2, and 3 under different ambient lighting conditions were reduced across the tests. Additionally, the confidence intervals of subjective evaluation results across most testing conditions became narrower for reproduction systems with lower perceived sound quality (Loudspeakers 2 and 3). These findings also demonstrate the beneficial effect of the listening training methodologies proposed in [55, 71–73] in minimizing lighting-induced bias, thereby enabling listeners to provide more consistent evaluation results across varying ambient lighting environments.

4 COMPARATIVE ANALYSES

The selection of ambient lighting conditions for subjective listening tests on perceptual sound quality should fully consider the potential influence of varying ambient light-

ing environments on evaluation outcomes, especially in the development of standardized testing protocols. The present findings suggest that, across varying illuminance levels, both program material–related interaction effects and those pertaining to measurement repeatability may significantly affect the results of overall preference ratings. This section presents a comparative analysis of the subjective listening test results obtained before and after the listening training, with a focus on the aforementioned aspects. The aim is to offer empirical evidence and theoretical rationale for selecting appropriate ambient lighting conditions in subjective listening tests, thereby contributing to the refinement and standardization of perceptual sound quality evaluation procedures.

4.1 Program-Related Interactions

The analysis results revealed that the interaction effects of **Loudspeaker** $\times$ **Lighting** $\times$ **Program** on overall preference ratings differed between the pre-training and post-training listening tests. The reduced significance of these interactions following the training suggests that trained listeners are more capable than untrained listeners of mitigating the combined influence of ambient lighting and program material on perceptual sound quality evaluations. Ideally, subjective assessments of sound reproduction systems should be independent of the specific program material used. Therefore, this finding underscores the importance that ambient lighting conditions should be tailored differently for trained and untrained listeners to account for their varying susceptibility to contextual factors.

A comparative analysis of the test results under different ambient lighting conditions and program materials was conducted before and after the listening training, as illustrated in Figs. 6 and 7. In each subgraph, the mean opinion scores of overall preference ratings for each program material are shown as open circles, accompanied by the corresponding range between the maximum and minimum values, which represents the distribution of program ratings. The top and bottom points represent the average of the five highest and five lowest ratings in each experiment, respectively, thereby minimizing the influence of outliers. A detailed illustration is provided in the figures. For comparison purposes, different colors are used to indicate the evaluation results under different program materials. In each subgraph, the fitted distribution curves of preference ratings obtained in each experiment are also displayed.

As shown in Fig. 6, it was initially observed that under conditions of extremely high or extremely low ambient lighting intensity, listeners tended to give overly positive or negative subjective sound quality ratings for certain test loudspeakers when paired with specific program materials in the pre-training experiments. For example, under the ambient lighting condition with an illuminance level of 530 lux, the rating distributions for Loudspeakers 3, 4, and 5 under Program TC all exceeded a range of 4 points and sometimes approached 5 points. Similarly, when the ambient lighting level was set to 0 lux, the ratings for Loud-

speakers 1 and 2 under Program LF also range around 4 points.

These results suggest that, under either very bright or very dim environments, the interaction between ambient lighting and program material exerts a more pronounced influence on the outcomes of subjective evaluations. This influence manifests as either a positive effect—enhancing perceived sound quality—or a negative one—degrading it. Regardless of the direction, such interactions increase the uncertainty in listeners' subjective judgments, thereby complicating the process of conducting reliable listening tests.

Secondly, it was observed that Loudspeakers 2 and 3 were more likely to receive higher ratings in Tests 6 and 7, which were conducted under ambient lighting conditions with illuminance levels of 6 and 0 lux, respectively. As indicated by the red dashed boxes in Fig. 6, under these two dimmest ambient lighting conditions, listener ratings for Loudspeaker 2 were more likely to exceed 6 points compared to other lighting conditions. Similarly, Loudspeaker 3 showed a higher likelihood of receiving ratings above 5.5 points under the same low-light conditions. These findings suggest that, in dim lighting environments, listeners tend to assign higher subjective ratings to sound reproduction systems with lower perceived sound quality, whereas systems with higher perceived sound quality appear to be less affected by this bias.

One possible explanation is that different visual stimuli can influence the auditory perception capabilities of the human ear [50, 51]. Moreover, this influence is strongly correlated with the acoustic characteristics of the auditory stimuli, particularly their frequency response, resulting in significant interaction effects among the lighting environment, the tested loudspeaker, and the program material. These interactions may alter the outcomes of subjective listening tests. This is one reason why people often mistakenly believe that dim lighting environments improve the perceived sound quality of loudspeakers. A more scientific and rigorous interpretation would be that, under specific combinations of optical and acoustic environmental stimuli, subjects may experience enhanced subjective perception. However, such scenarios should be avoided in perceptual sound quality evaluations of loudspeakers, as they may introduce bias. Conversely, in soundscape applications involving auditory-visual interactions, these effects could be harnessed to produce positive and beneficial outcomes.

By contrast, in the post-training listening tests, the experimental data demonstrate a reduced difference between the maximum and minimum evaluation results for each program material under most lighting conditions, as shown in Fig. 7. This suggests that trained listeners avoided giving excessively high or low subjective sound quality ratings to the tested loudspeakers under specific combinations of program materials and ambient lighting environments, consequently, improving the precision of subjective sound quality evaluation in single or limited trial measurements across different program materials and ambient lighting conditions.

The data distribution of evaluation results for Loudspeakers 2 and 3 under the ambient lighting condition with an

illuminance level of 6 lux was significantly improved. Unlike the pre-training test results, there was no apparent tendency for ratings to shift toward higher subjective scores, indicating that trained listeners exhibited a greater ability to mitigate the inflation of ratings for loudspeakers with low perceived quality under dim lighting conditions. However, the distribution of subjective ratings for these loudspeakers under the ambient lighting level of 0 lux showed little improvement after training. This implies that such lighting conditions are particularly unsuitable for conducting reliable subjective listening tests.

Moreover, the consistency of evaluation results was enhanced for each program material across different ambient lighting conditions. As shown in Fig. 7, in most experiments, the variability of the mean evaluation scores for each program material was significantly reduced. For example, under the ambient lighting condition with an illuminance level of 250 lux, the maximum difference in mean preference ratings for Loudspeaker 3 across program materials was 3 points prior to training in Fig. 6 (mean rating of Program JT = 2.8; mean rating of Program JW = 5.8). In contrast, this difference was reduced to less than or equal to 1 point in Fig. 7.

However, it is worth noting that under the ambient lighting condition with an illuminance level of 530 lux, only the difference between the maximum and minimum evaluation results for Program JT showed a significant reduction, while the evaluation results for the other program materials exhibited no notable improvement. Furthermore, for sound reproduction systems with low perceived quality (Loudspeakers 2 and 3), the consistency of evaluation results for across program materials did not improve significantly, indicating that this lighting condition is particularly unsuitable for conducting subjective listening tests. A more detailed discussion on the recommended ambient lighting conditions for subjective listening tests is provided in SEC. 5.

Based on the present results, both high and low ambient light levels can affect subjective evaluations, especially under certain stimulus. This implies that, in terms of program material selection, the interaction effects of ambient lighting should be carefully considered. Fig. 8 shows the maximum differences in mean preference ratings under different programs across various tests conducted under different ambient lighting conditions. Overall, Program JT exhibited the smallest variation in mean preference ratings across the experiments, whereas Program TC showed the largest fluctuations. The consistency of subjective evaluations for Programs JW and LF fell between those of Programs JT and TC across the different ambient lighting environments.

This trend remained consistent before and after the listening training, except for the ratings of Loudspeaker 3 under Program TC in the pre-training test. Further examination of the data in Fig. 8(c) may provide additional insights. Although the ratings of Loudspeaker 3 under Program TC varied little across different ambient lighting conditions, they were generally lower than the final ratings, which ranged from approximately 3.8 to 4.2 points, as shown in Fig. 5. This suggests that Loudspeaker 3 may be highly unsuitable for reproducing Program TC, and the high consistency in

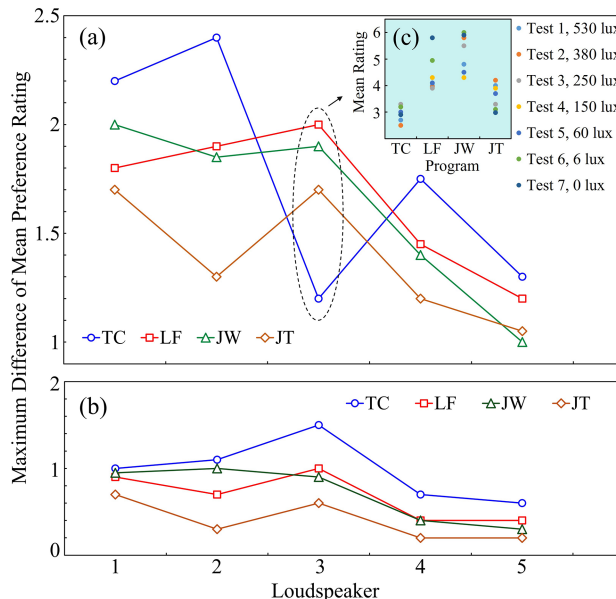


Fig. 8. Maximum differences in mean preference ratings under different programs across various tests conducted under different ambient lighting conditions: (a) pre-training tests, (b) post-training tests, and (c) mean preference ratings of Loudspeaker 3 under different program materials across all pre-training tests. The curves represent the range between the maximum and minimum mean ratings for each program material across listening tests. For example, in (a), the result for Loudspeaker 3 under Program JT was calculated from the difference between Program JT ratings in Tests 2 and 7, as shown in (c). Loudspeaker 3 exhibits a trend in (a) that differs markedly from the other curves; therefore, its mean preference ratings under different program materials across all pre-training tests are displayed in (c) for detailed analysis.

listeners' ratings of overly low sound quality likely limits the interaction effect of ambient lighting on the test results.

Overall, several patterns could be observed from the long-term spectral analysis of the program materials, as shown in Fig. 2. Program TC exhibits the largest fluctuations in its spectral curve, whereas Programs JW and LF emphasize the reproduction of low and mid frequencies, respectively. In contrast, the curve fluctuation of Program JT is relatively small, with a more balanced distribution of high, mid, and low frequencies. Therefore, for the selection criteria of program materials, it is recommended to prioritize the smoothness of the power spectrum, followed by the spectrum balance. Such considerations are particularly important for minimizing the interaction effects between ambient lighting and program material on evaluation outcomes.

This outcome is not unexpected. Previous studies have demonstrated that frequency-dependent auditory thresholds, or the just noticeable difference (JND) in auditory perception, may be modulated by visual stimuli [50, 51]. As a result, listeners' subjective sensitivity to sounds at different frequencies may vary across distinct ambient lighting conditions. In this context, if the excitation source contains excessive energy within a particular frequency band or exhibits pronounced fluctuations in its power spectrum, it is

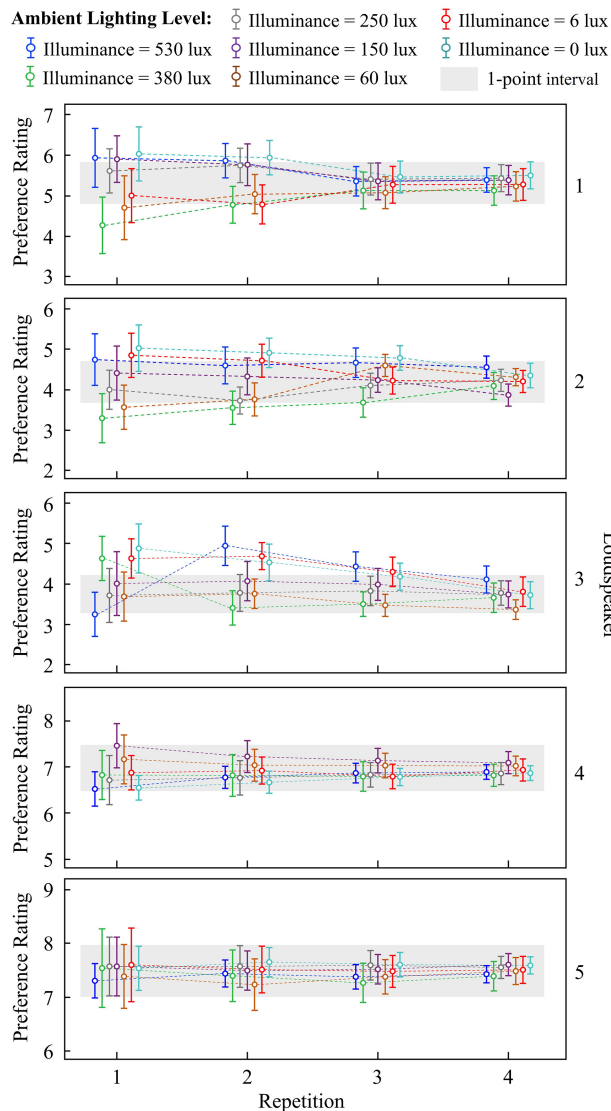


Fig. 9. Mean preference ratings under different repetition conditions in the pre-training listening tests, along with the corresponding 95% confidence intervals.

likely to induce divergent perceptual responses under different ambient lighting environments. To mitigate these confounding effects and to ensure that the evaluation primarily reflects the intended auditory factors rather than artifacts introduced by ambient lighting conditions, program material with a relatively balanced and smooth power spectrum is recommended for use in subjective listening evaluations.

4.2 Measurement Repeatability

As indicated by the ANOVA results in Table 3, the interaction of **Loudspeaker** $\times$ **Lighting** $\times$ **Repetition** had a significant effect on the outcomes of the pre-training listening tests. This suggests that the measurement repeatability of subjective sound quality ratings for the tested loudspeakers was influenced by the ambient lighting conditions. Fig. 9 illustrates the evaluation results of different loudspeakers under various ambient lighting environments and repetition conditions prior to the listening training. Each row of subplots corresponds to a different tested loudspeaker. The

mean preference ratings for each test are marked by open circles and connected by dashed lines.

In each subplot, the x-axis indicates the repetition condition, and the y-axis represents the evaluation results. Repetition = 1 refers to the results obtained from listeners' first complete round of evaluations for all program materials under each ambient lighting condition. Repetition = 2 corresponds to the results after listeners completed two full rounds of evaluations under the same conditions. Repetition = 3 and Repetition = 4 follow the same logic. To facilitate clearer comparisons of evaluation differences under various ambient lighting conditions, the figure also highlights the 1-point interval region based on the final evaluation results, representing a moderate preference rating difference of 1 point.

As shown in Fig. 9, in the pre-training subjective tests, sound reproduction systems with lower perceived sound quality—specifically Loudspeakers 2 and 3—exhibited significant variability in evaluation results across different ambient lighting conditions prior to the full four-round repeated assessments. In some cases, the corresponding 95% confidence intervals did not substantially overlap, and the maximum difference in mean preference ratings under across ambient lighting environments exceeded 1 point (moderate preference difference), even approaching 2 points (strong preference difference). Such results indicate that the consistency was insufficient, as acceptable consistency in sound quality evaluations typically requires not only that the mean ratings show no statistically significant differences across tests (i.e., within a 1-point interval in the present experiment) but also that the 95% confidence intervals of all test results substantially overlap. The present results suggest that, for these types of sound reproduction systems, at least four repetitions are necessary to achieve acceptable consistency.

For the system with medium perceived sound quality, Loudspeaker 1, three repetitions were sufficient to obtain consistent evaluation results, while for the systems with high perceived sound quality, Loudspeakers 4 and 5, only two repetitions were needed. These findings suggest that the lower the perceived sound quality of a sound reproduction system, the more repetitions required to ensure reliable and consistent subjective sound quality measurements under varying ambient lighting conditions. Conversely, high-quality systems demonstrate greater robustness and require fewer repetitions.

Moreover, as shown in Fig. 9, for the system with the highest perceived sound quality, Loudspeaker 5, a single test appears sufficient to yield consistent evaluation results across different ambient lighting conditions. Nevertheless, it is still recommended to conduct at least one additional repetition to verify the reproducibility of the subjective results and to reduce the width of the confidence interval. These experimental findings further support the conclusion that the lower the quality of a loudspeaker, the more susceptible its subjective sound quality ratings are to variations in ambient lighting conditions. For loudspeakers with low perceived sound quality, a single evaluation conducted under certain ambient lighting conditions may result in either

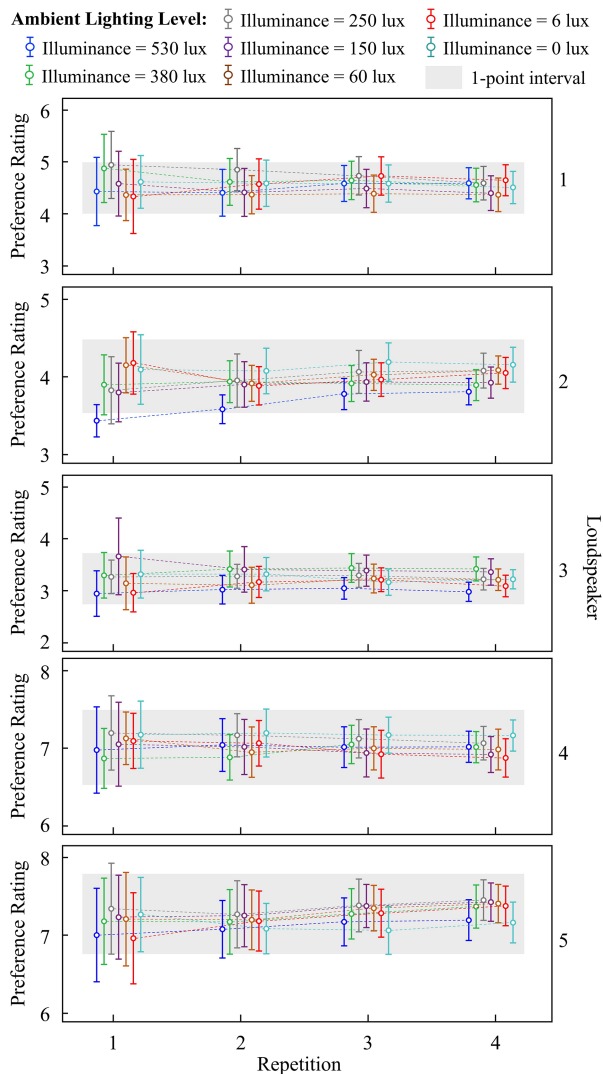


Fig. 10. Mean preference ratings under different repetition conditions in the post-training listening tests, along with the corresponding 95% confidence intervals.

excessively high or low ratings. This may also contribute to the common misconception that dim listening environments enhance the perceived sound quality of loudspeakers.

The post-training listening test results are presented in Fig. 10. It can be observed that trained listeners were able to provide a relatively consistent set of subjective evaluations for the tested loudspeakers during the first round of listening tests across all program materials and ambient lighting conditions. With the exception of Loudspeaker 2 under the ambient lighting level of 530 lux, the confidence intervals did not show substantial overlap with those obtained under 60 and 6 lux. This suggests that an ambient lighting level of 530 lux may also be unsuitable for conducting subjective listening tests, even for trained listeners.

As the number of repetitions increased, the corresponding 95% confidence intervals for the evaluation results became progressively narrower. On one hand, the present results demonstrate that the repeatability requirements for subjective sound quality measurements under varying ambient lighting conditions are reduced when evaluations are conducted by trained listeners. On the other hand, they in-

directly support the validity of previous studies in which the listening panels primarily consisted of trained listeners, thereby suggesting that conducting two repetitions per condition is both reasonable and sufficient.

In subjective listening tests, it is generally desirable to achieve consistent results with a minimal number of repeated assessments. Based on the results of this study, different ambient lighting conditions significantly affect the repeatability requirements of the measurements, potentially due to human adaptation mechanisms [88, chap. 1; 89]. Humans manage to compensate for many variations contributed by listening environments. Although such adaptation is not instantaneous and requires an exposure time, the duration of which may be modulated by different ambient lighting environments, individuals tend to ultimately attend to and appreciate the essential acoustic qualities of a sound source, even under conditions involving complex multisensory interactions. Whether the observed effects of ambient lighting on measurement repeatability are indeed attributable to these adaptation mechanisms remains to be further investigated. Nevertheless, selecting appropriate lighting conditions clearly contributes to reducing the repeatability demands of subjective listening tests. Further discussion is provided in SEC. 5.

5 PROPOSED LIGHTING LEVELS

Regarding the selection of ambient lighting conditions in subjective listening tests, the following principles may serve as useful guidelines: (1) The interaction effects between ambient lighting and program material should have minimal impact on evaluation outcomes. (2) Limited repeated testing—typically two repetitions—should be sufficient to obtain subjective results minimally affected by lighting-induced bias. (3) Ambient lighting should not reduce evaluation consistency, and the dispersion of rating data should be low. (4) In comparative evaluations involving multiple test objects, ambient lighting conditions should not excessively interfere with listeners' judgments of subjective sound quality, so that they can clearly distinguish the subjective differences among the tested loudspeakers under appropriate ambient lighting conditions.

Among these, the findings related to principles (1) and (2) have already been presented in SEC. 4. Accordingly, this section focuses on a comparative analysis of the subjective listening test results under different ambient lighting conditions with respect to principles (3) and (4). Based on these analyses, recommendations for selecting appropriate ambient lighting conditions in subjective listening tests are proposed, with a distinction made between trained and untrained listeners.

5.1 Data Dispersion

The dispersion of individual loudspeaker preference ratings in each listening test was quantified using the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean. The CV value is commonly used as a normalized measure of dispersion in probability or

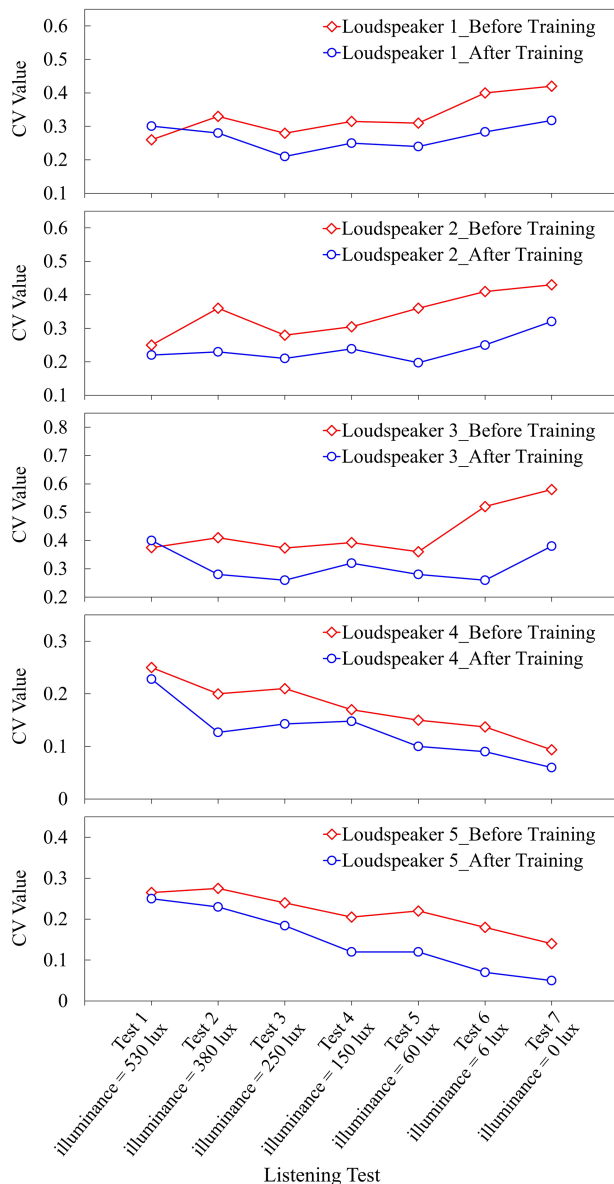


Fig. 11. CV values for the listening tests.

frequency distributions [90–92]. Minimizing experimental data dispersion is essential for achieving accurate and consistent results in subjective listening tests. The results are presented in Fig. 11. It was observed that, in the pre-training experiments, dimmer lighting conditions tended to reduce the data dispersion for sound reproduction systems with high perceived quality (i.e., Loudspeakers 4 and 5). In contrast, for systems with low perceived quality (i.e., Loudspeakers 2 and 3), excessively dim environments led to increased data dispersion. For example, in the experiments conducted under ambient lighting conditions with illuminance levels of 6 and 0 lux, the CV values were the highest. This phenomenon is consistent with the experimental results reported in [44].

A comparison of the CV values before and after the listening training shows that the data dispersion was reduced in most post-training experiments, indicating that trained listeners demonstrated a greater ability to concentrate sub-

jective evaluation results for the tested loudspeakers under varying ambient lighting conditions. However, in Test 1, conducted under the ambient lighting condition with an illuminance level of 530 lux, the CV values for Loudspeakers 1 and 3 increased rather than decreased. This occurred because, in Test 1, although the absolute dispersion of the test data remained unchanged, the mean evaluation scores for the tested loudspeakers were lower in the post-training experiments, resulting in higher CV values. This finding is also consistent with the results shown in Figs. 6 and 7.

5.2 Significant Difference Between Tested Loudspeakers

In comparative evaluations involving multiple sound reproduction systems, the loudspeaker test statistic F_L is commonly used to determine whether the tested loudspeakers are perceived as significantly different by listeners, as recommended in [72]. A higher F_L value indicates that the perceptual differences in sound quality among the loudspeakers are more readily distinguishable to listeners. As illustrated in Fig. 12, the figure presents the multiple comparison results from the listening tests conducted before and after the training, with different colors used to distinguish between the pre-training and post-training results.

First, as shown in the figure, during the pre-training listening tests, listeners' ability to discriminate perceptual differences in sound quality between loudspeakers varied depending on the ambient lighting conditions. The values of the loudspeaker test statistic F_L differed across tests conducted under various ambient lighting environments. The highest F_L values were observed under ambient lighting conditions with illuminance levels ranging from 60 to 150 lux, suggesting that this lighting range is more conducive to enabling listeners to significantly differentiate the subjective sound quality among the tested loudspeakers. When the ambient lighting level exceeded or fell beyond this range, the listeners' ability to discriminate perceptual sound quality differences between loudspeakers declined.

In contrast, during the post-training listening tests, the variability of F_L values across various listening tests conducted under different ambient lighting conditions was significantly reduced, and all F_L values obtained after training were higher than those observed in the pre-training tests. This indicates that trained listeners, compared to untrained listeners, were better able to clearly distinguish the subjective sound quality differences among the tested sound reproduction systems across a wider range of ambient lighting conditions. This observation further suggests that the selection of appropriate ambient lighting conditions for conducting subjective listening tests should take into account whether the listeners have received prior listening training.

Secondly, the results of the least significant difference (LSD) analysis further reveal that, although no consistent upward or downward trend was observed in the pairwise comparison outcomes before and after the listening training, in Tests 3, 4, 5, and 6—conducted under ambient lighting conditions with illuminance levels of 250, 150, 60, and 6 lux, respectively—the confidence intervals of the LSD re-

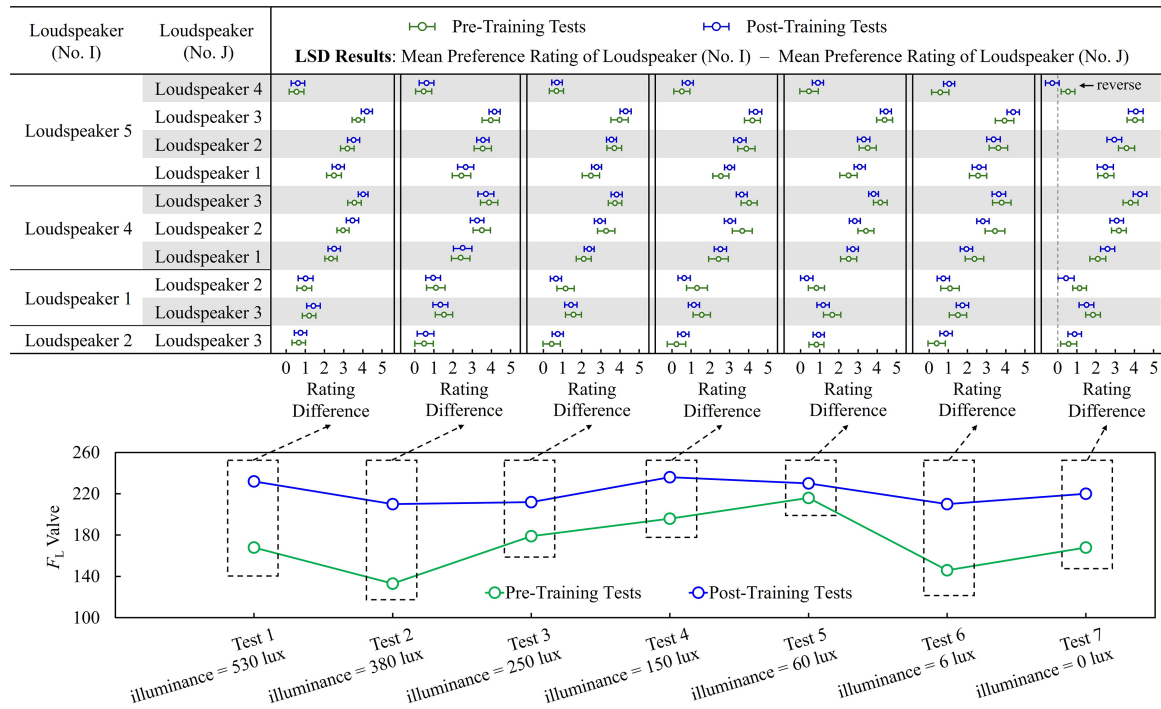


Fig. 12. Multiple comparison results from the listening tests conducted before and after training, with different colors used to distinguish between the pre-training and post-training results. The values of loudspeaker test statistic F_L in different tests conducted under various ambient lighting conditions are shown at the bottom of the figure. The detailed results of the LSD analysis for each listening test, along with the corresponding 95% confidence intervals, are presented at the top. In the LSD result plots, the subplots in different columns represent the outcomes of the listening tests conducted under different ambient lighting conditions. The x-axis indicates the differences in mean preference ratings between each pair of tested loudspeakers, while the y-axis denotes the specific loudspeaker pair being compared. Arrows are used to link each LSD result to the corresponding F_L values under the same test condition.

sults became narrower. This suggests that linear distortion discrimination training enhances the precision of listeners' ability to distinguish sound quality differences under the aforementioned ambient lighting conditions. However, this improvement does not extend to all ambient lighting environments. For instance, the LSD results for the pairwise comparison between Loudspeakers 1 and 2 in Test 1, and between Loudspeakers 5 and 1 in Test 7, showed almost no change following the training.

5.3 Discussion on Proposed Lighting Levels

Based on the analyses presented in previous sections, it is strongly recommended that the selection of ambient lighting conditions for subjective sound quality evaluation should be differentiated according to listener expertise. It should be noted that the listener expertise discussed here refers specifically to the listeners' ability to detect linear distortion. In the present study, it pertains solely to the distinction between untrained (naïve) and trained (expert) listeners and does not necessarily reflect the level of test-specific training, which is typically implemented during the tests.

5.3.1 For Untrained Listeners

For most participants of the listening panel who have not undergone relevant listening training, ambient lighting conditions with illuminance levels between 60 and 150 lux are recommended for conducting perceptual sound quality

evaluations. This recommendation is consistent with the findings reported in [44]. On one hand, under these ambient lighting conditions, the influence of different program materials on the evaluation of loudspeakers with varying sound quality was relatively small, as shown in Fig. 6. The mean differences in the evaluation results across program materials were within 1 point, making it relatively easier to select appropriate program materials for the listening tests. Moreover, listeners did not exhibit a tendency to assign higher ratings to loudspeakers with lower perceived sound quality.

On the other hand, as shown in Fig. 9, untrained listeners require at least four repeated trials to mitigate the influence of ambient lighting on evaluation results. However, in most subjective listening tests, repeated trials are difficult to implement due to time constraints, particularly when the composition of the listening panel is not fixed. Therefore, minimizing the number of repeated trials while still reducing the impact of varying ambient lighting conditions is a critical criterion in selecting appropriate ambient lighting conditions for subjective listening tests. Notably, under ambient lighting conditions with illuminance levels of 60 and 150 lux, reliable results were obtained for all tested loudspeakers after only two repeated trials, with subjective ratings falling within a 1-point interval of the final evaluation outcomes, as illustrated in Fig. 9.

Finally, under these ambient lighting conditions, overall, the dispersion of subjective measurement data was relatively low across all tested loudspeakers, as shown in

Fig. 11. Additionally, such ambient lighting environments also facilitated listeners' ability to perceptually distinguish differences in sound quality among the tested sound reproduction systems. As shown in Fig. 12, the highest F_L values were observed in the listening tests conducted under ambient lighting conditions with illuminance levels of 60 and 150 lux.

5.3.2 For Trained Listeners

For most participants of the listening panel who have undergone relevant listening training, ambient lighting conditions with illuminance levels between 6 and 250 lux are recommended for conducting perceptual sound quality evaluations. First, as shown in Fig. 7, the interaction effect involving the **Program** factor was not substantially improved in Test 1, which was conducted under the ambient lighting condition with an illuminance level of 530 lux, even for trained listeners. Second, although the listening training partially mitigated the negative influence of program materials in Test 2, conducted under the ambient lighting condition with an illuminance level of 380 lux, the difference between the maximum and minimum evaluation results across program materials still exceeded 2 points, indicating pronounced preference separation. Therefore, ambient lighting conditions with illuminance levels of 530 and 380 lux are not recommended for conducting subjective listening tests, even after listeners have received relevant listening training.

Penultimately, under the ambient lighting condition with an illuminance level of 0 lux, listeners continued to assign higher subjective ratings to systems with lower perceived sound quality, even after undergoing listening training. This effect is illustrated by the red dashed box in Fig. 7, which highlights the region of overly high ratings. Listener ratings for Loudspeaker 3 were still more likely to exceed 5.5 points for 0 lux compared to other ambient lighting conditions. Moreover, for reproduction systems with high perceived sound quality, the relative preference ranking between Loudspeakers 4 and 5 exhibited a slight reversal before and after the listening training. As indicated by the arrow in Fig. 12, the preference rating difference between Loudspeakers 4 and 5 changed sign between the pre-training and post-training listening tests.

This suggests that, for high-quality systems with perceptual sound quality that are close to each other, preference ratings may be subject to bias depending on listeners' training status under extremely dim lighting conditions. In addition, under this ambient lighting condition, the data dispersion of subjective ratings for reproduction systems with lower perceived sound quality also increased significantly, as shown in Fig. 11. Therefore, the ambient lighting condition with an illuminance level of 0 lux is strongly discouraged for conducting subjective listening tests.

The same pair of tested loudspeakers could produce different LSD results under different lighting conditions. For instance, in pre-training tests, the mean difference in preference rating between Loudspeakers 1 and 3 was around 1 point (moderate preference difference) under ambient light-

ing condition of 530 lux, whereas it was nearly 2 points (strong preference difference) under 0 lux, as shown in Fig. 12. One possible explanation is that optical stimuli may influence listeners' JND in auditory perception [50, 51]. This influence is known to be frequency-dependent. Since the subjective sound quality of a loudspeaker is primarily related to its frequency response, changes in JND under varying illuminance levels of ambient lighting environments may affect listeners' ability to discriminate perceptual differences between loudspeakers.

Furthermore, JND variations induced by optical stimuli may also shift listeners' subjective preference for spectral balance across lighting conditions. As shown in Fig. 6, the same loudspeakers and program materials may yield significantly different results under different lighting environments, which indirectly supports this hypothesis. For instance, there was a strong preference separation (≥ 2 points) between the mean preference ratings of Loudspeaker 3 under Program LF in the pre-training tests of Test 4 (4.2 points) and Test 7 (6.3 points). Unfortunately, due to the limited data currently available, further investigation is required to substantiate this explanation.

Finally, overall, compared to the pre-training tests, the improvement in the LSD results of pairwise comparisons between the tested loudspeakers was more pronounced in Tests 3, 4, 5, and 6 than in Tests 1, 2, and 7. As shown in Fig. 12, the confidence intervals of the LSD results for the pairwise comparisons between loudspeakers became narrower, indicating the enhanced precision in listeners' ability to distinguish sound quality differences under these specific ambient lighting conditions. Therefore, when the majority of listeners have undergone relevant listening training, ambient lighting conditions with illuminance levels between 6 lux and 250 lux are recommended for conducting subjective sound quality evaluations.

However, it should be noted that when the ambient lighting environment is set to an illuminance level of 250 lux, greater caution is required when selecting program materials for evaluating systems with low perceived sound quality. This is because, under such ambient lighting condition, the difference between the maximum and minimum evaluation scores for certain program materials may still be large and statistically significant (≥ 2 points), as exemplified by Loudspeaker 3 under Program TC in the post-training test of Test 3, as shown in Fig. 7. The long-term spectral analysis of Program TC showed the largest fluctuations in spectral content among the tested materials, as shown in Fig. 2. This further supports the recommendation that it is advisable to prioritize the smoothness of the power spectrum for the selection criteria of program materials. This guideline is also consistent with the suggestions made in [44, 54].

6 CONCLUSION

To improve subjective listening test methodologies by accounting for the crossmodal effects of ambient lighting and listener expertise on perceptual sound quality evaluation, a series of listening tests were conducted under various ambient lighting conditions both before and after the lis-

tening training. Loudspeakers with systematically varied performance—ranging from low to high perceived sound quality—were selected to facilitate a comprehensive investigation into the selection of appropriate ambient lighting conditions for subjective listening tests. The results were compared to examine the impact on several factors, including program-related interaction effects, measurement repeatability, data dispersion, and listeners' ability to identify significant perceptual differences between tested loudspeakers.

The results of this study indicate that based on a systematic and comprehensive listening test procedure, different ambient lighting conditions had only a limited impact on overall perceived sound quality. However, extremely bright or dim lighting conditions significantly amplified the program material–related interactions on the evaluation outcomes. Untrained listeners were more likely to assign overly positive or negative ratings to certain loudspeakers when exposed to such extreme ambient lighting, particularly under specific program materials.

In comparison, trained listeners demonstrated a greater ability to mitigate these interaction effects. Nonetheless, except under completely dark ambient lighting environments, a tendency remained for their ratings to shift toward higher subjective scores. Moreover, trained and untrained listeners showed notable differences in both the dispersion of subjective ratings across ambient lighting conditions and their ability to significantly differentiate the perceived sound quality among the tested sound reproduction systems. These findings collectively underscore the importance of considering listener training status when selecting appropriate ambient lighting conditions for subjective listening tests.

Based on the experimental results presented in this study, for listening panels composed primarily of untrained listeners, the recommended ambient lighting conditions for conducting perceptual sound quality measurement are within the illuminance range of 60–150 lux. For panels consisting predominantly of trained listeners, this illuminance range can be extended to between 6 and 250 lux. However, under ambient lighting condition with an illuminance level of 250 lux, program materials should be selected with greater caution. In particular, when evaluating loudspeakers with low sound quality, program materials with relatively smooth power spectra are preferable.

Overall, under the recommended ambient lighting conditions, the results of the subjective listening tests were less influenced by program-related interactions, exhibited lower data dispersion, and enabled clearer differentiation in the perceived sound quality among the tested systems. In addition, the requirement for test repeatability was relatively low, as conducting two repeated trials was generally sufficient to obtain reliable evaluation outcomes.

Although the selection of ambient lighting conditions for subjective listening tests has been initially explored in this study, further in-depth investigation remains necessary. On one hand, the interactions among program materials, tested loudspeaker and ambient lighting suggest the potential existence of crossmodal effects, whereby perceptual sound

quality is not solely determined by the acoustic properties of the reproduction system but is instead shaped by multi-sensory integration. Such effect remains subject to individual differences of listener's unique background. Whether such a mechanism exists—and if so, how it may be instantiated physiologically or computationally—remains a fundamental open question.

On the other hand, it is still uncertain whether changes in JND caused by different optical stimuli may result in shifts in subjective preference for spectral balance under varying ambient lighting conditions. Addressing these issues would provide a more objective basis for understanding how visual stimuli modulate auditory spectral weighting and provide valuable empirical evidence to support the advancement of theoretical frameworks of audiovisual integration in human perception. It is hoped that this study can offer some initial insights into the refinement of subjective listening test methodologies, and stimulate further and deeper investigation into conducting accurate and valid perceptual sound quality evaluation.

7 ACKNOWLEDGMENT

The authors would like to thank all the listeners who participated in this study. This work was supported by National Natural Science Foundation of China (Grant No. 12374446).

8 REFERENCES

- [1] R. Walker, "Optimum Dimension Ratios for Studios, Control Rooms and Listening Rooms," Tech. Rep. BBC RD 1993/8 (1993 Jan.).
- [2] F. E. Toole, "Loudspeakers and Listening Rooms, Part II: The Perception of Reflections," *J. Acoust. Soc. Am.*, vol. 119, no. 5 (Suppl), p. 3251 (2006 May). <https://doi.org/10.1121/1.4786061>.
- [3] S. E. Olive and F. E. Toole, "The Detection of Reflections in Typical Rooms," *J. Audio Eng. Soc.*, vol. 37, no. 7/8, pp. 539–553 (1989 Jul.).
- [4] S. Bech, "Timbral Aspects of Reproduced Sound in Small Rooms. I," *J. Acoust. Soc. Am.*, vol. 97, no. 3, pp. 1717–1726 (1995 Mar.). <https://doi.org/10.1121/1.413047>.
- [5] S. Bech, "Timbral Aspects of Reproduced Sound in Small Rooms. II," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3539–3549 (1996 Jun.). <https://doi.org/10.1121/1.414952>.
- [6] Y. Ando, "Subjective Preference in Relation to Objective Parameters of Music Sound Fields With a Single Echo," *J. Acoust. Soc. Am.*, vol. 62, no. 6, pp. 1436–1441 (1977 Dec.). <https://doi.org/10.1121/1.381661>.
- [7] G. A. Soulodre, N. Popplewell, and J. S. Bradley, "Combined Effects of Early Reflections and Background Noise on Speech Intelligibility," *J. Sound Vibr.*, vol. 135, no. 1, pp. 123–133 (1989 Nov.). [https://doi.org/10.1016/0022-460X\(89\)90759-1](https://doi.org/10.1016/0022-460X(89)90759-1).
- [8] A. J. Watkins, "Perceptual Compensation for Effects of Reverberation in Speech Identification," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 249–262 (2005 Jul.). <https://doi.org/10.1121/1.1923369>.

- [9] G. A. Soulodre, "Evaluation of Objective Loudness Meters," presented at the *116th Conv. of the Audio Engineering Society* (2004 May), paper 6161.
- [10] A. Gabrielsson, B. Hagerman, T. Bech-Kristensen, and G. Lundberg, "Perceived Sound Quality of Reproductions With Different Frequency Responses and Sound Levels," *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1359–1366 (1990 Sep.). <https://doi.org/10.1121/1.399713>.
- [11] S. Bech, "Perception of Timbre of Reproduced Sound in Small Rooms: Influence of Room and Loudspeaker Position," *J. Audio Eng. Soc.*, vol. 42, no. 12, pp. 999–1007 (1994 Dec.).
- [12] S. E. Olive P. L. Schuck, S. L. Sally, and M. E. Bonneville, "The Effects of Loudspeaker Placement on Listener Preference Ratings," *J. Audio Eng. Soc.*, vol. 42, no. 9, pp. 651–669 (1994 Sep.).
- [13] T. -A. Karandreas and F. Christensen, "Influence of Visual Appearance on Loudspeaker Sound Quality Evaluation," presented at the *124th Conv. of the Audio Engineering Society* (2008 May), paper 7491.
- [14] T. -A. Karandreas and F. Christensen, "Influence of Visual Appearance on Loudspeaker Sound Quality Evaluation," *J. Audio Eng. Soc.*, vol. 63, no. 9, pp. 684–697 (2015 Sep.). <https://doi.org/10.17743/jaes.2015.0067>.
- [15] T.-A. Karandreas, *Audiovisual Interaction: Influence of Visual Appearance on Loudspeaker Sound Quality Evaluation*, Ph.D. thesis, Aalborg University, Aalborg, Denmark (2010 Nov.).
- [16] M. Schoeffler, and J. Herre, "The Relationship Between Basic Audio Quality and Overall Listening Experience," *J. Acoust. Soc. Am.*, vol. 140, no. 3, pp. 2101–2112 (2016 Sep.). <https://doi.org/10.1121/1.4963078>.
- [17] A. Smimite, A. Beghdadi, and K. Chen, "Investigating 'The Experience Effect' in Audio Quality Assessment," in Proc. of the *21st Telecommunications Forum (TELFOR)*, pp. 769–772 (Belgrade, Serbia) (2013 Nov.). <https://doi.org/10.1109/TELFOR.2013.6716343>.
- [18] T. Walton and M. Evans. "The Role of Human Influence Factors on Overall Listening Experience," *Qual. User Exp.*, vol. 3, no. 1, paper 1 (2018 Jan.). <https://doi.org/10.1007/s41233-017-0015-4>.
- [19] J. Kreiman, B. R. Gerratt, and K. Precoda. "Listener Experience and Perception of Voice Quality," *J. Speech Lang. Hear. Res.*, vol. 33, no. 1, pp. 103–115 (1990 Mar.). <https://doi.org/10.1044/jshr.3301.103>.
- [20] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application* (Wiley, Chichester, UK, 2006).
- [21] F. E. Toole, Loudspeakers and Rooms for Sound Reproduction –A Scientific Review," *J. Audio Eng. Soc.* vol. 54, no. 6, pp. 451–476 (2006 Jun.).
- [22] IEC, "Sound System Equipment - Part 13: Listening Tests on Loudspeakers," Tech. Rep. 60268-13 (1998 Mar.).
- [23] ITU-R, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," *Recommendation ITU-R BS.1116-1* (1997 Oct.).
- [24] ITU-R, "Methodologies for the Subjective Assessment of the Quality of Television Images," *Recommendation ITU-R BT.500-15* (2023 May).
- [25] J. Kang and M. Zhang. "Semantic Differential Analysis of the Soundscape in Urban Open Public Spaces," *Build. Environ.*, vol. 45, no. 1, pp. 150–157 (2010 Jan.). <https://doi.org/10.1016/j.buildenv.2009.05.014>.
- [26] M. E. Nilsson and B. Berglund. "Soundscape Quality in Suburban Green Areas and City Parks: Recent Advances in Soundscape Research," *Acta Acust. un. Acust.*, vol.92, no. 6, pp. 898–902 (2006 Nov.).
- [27] O. Axelsson, M. E. Nilsson, B. Berglund. "A Principal Components Model of Soundscape Perception," *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 2836–2846 (2010 Nov.). <https://doi.org/10.1121/1.3493436>.
- [28] J. Y. Jeon, P. J. Lee, J. You and J. Kang. "Perceptual Assessment of Quality of Urban Soundscapes With Combined Noise Sources and Water Sounds," *J. Acoust. Soc. Am.*, vol. 127, no. 3, pp. 1357–1366 (2010 Mar.). <https://doi.org/10.1121/1.3298437>.
- [29] S. Bech, "Listening Tests on Loudspeakers: A Discussion of Experimental Procedures and Evaluation of the Response Data," in Proc. of the *8th Intl. AES Conf.: The Sound of Audio* (1990 May), paper 5427.
- [30] L. Galbrun and F. Calarco, "Audio-Visual Interaction and Perceptual Assessment of Water Features Used Over Road Traffic Noise," *J. Acoust. Soc. Am.*, vol. 136, no. 5, pp. 2609–2620 (2014 Nov.). <https://doi.org/10.1121/1.4897313>.
- [31] S. Agrewal, S. Bech, K. B. Bærentsen, K. De Moor, and S. Forchhammer, "Method for Subjective Assessment of Immersion in Audiovisual Experiences," *J. Audio Eng. Soc.*, vol. 69, no. 9, pp. 656–671 (2021 Sep.). <https://doi.org/10.17743/jaes.2021.0013>.
- [32] Ľ. Hládek, A. R. Seitz, and N. Kopčo, "Auditory-Visual Interactions in Egocentric Distance Perception: Ventriloquism Effect and Aftereffect," *J. Acoust. Soc. Am.*, vol. 150, no. 5, pp. 3593–3607 (2021 Nov.). <https://doi.org/10.1121/10.0007066>.
- [33] S. Agrawal, S. Bech, K. De Moor, and S. Forchhammer, "Influence of Changes in Audio Spatialization on Immersion in Audiovisual Experiences," *J. Audio Eng. Soc.*, vol. 70, no. 10, pp. 810–823 (2022 Oct.). <http://dx.doi.org/10.17743/jaes.2022.0034>.
- [34] J. Wang, A. Zheng, Y. Yan, R. He and J. Tang, "Attribute-Guided Cross-Modal Interaction and Enhancement for Audio-Visual Matching," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 4986–4998, (2024 Apr.). <http://dx.doi.org/10.1109/TIFS.2024.3388949>.
- [35] M. Y. Finkelstein, K. Banai, and O. S. Weiss, "Accuracy of Auditory and Auditory-Visual Voice Quality Perception by Speech Language Pathologists and Ear, Nose, and Throat Specialists," *J. Voice*, ahead of print (2025 Feb.). <https://doi.org/10.1016/j.jvoice.2025.01.030>.
- [36] J. Zhou, X. Shen, J. Wang, et al., "Audio-Visual Segmentation With Semantics," *Int. J. Comput. Vis.*, vol. 133, no. 4, pp. 1644–1664 (2025 Apr.). <https://doi.org/10.1007/s11263-024-02261-x>.

- [37] L. A. Passos, J. P. Papa, J. Del Ser, A. Hus-sain, and A. Adeel, "Multimodal Audio-Visual Information Fusion Using Canonical-Correlated Graph Neural Network for Energy-Efficient Speech Enhancement," *Inf. Fusion*, vol. 90, pp. 1–11 (2023 Feb.). <https://doi.org/10.1016/j.inffus.2022.09.006>.
- [38] S. Zhang, Y. Yang, C. Chen, et al., "Deep Learning-Based Multimodal Emotion Recognition from Audio, Visual, and Text Modalities: A Systematic Review of Recent Advancements and Future Prospects," *Expert Syst. Appl.*, vol. 237, part C, paper 121692 (2024 Mar.). <https://doi.org/10.1016/j.eswa.2023.121692>.
- [39] H. Fan, X. Zhang, Y. Xu, et al., "Transformer-Based Multimodal Feature Enhancement Networks for Multimodal Depression Detection Integrating Video, Audio and Remote Photoplethysmograph Signals," *Inf. Fusion*, vol. 104, paper 102161 (2024 Apr.). <https://doi.org/10.1016/j.inffus.2023.102161>.
- [40] Z. Li, Z. Li, J. Zhang, Y. Feng, and J. Zhou, "Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2476–2483 (2021 Mar.). <https://doi.org/10.1109/TASLP.2021.3065823>.
- [41] S. Mai, S. Xing, and H. Hu, "Analyzing Multimodal Sentiment via Acoustic- and Visual-LSTM With Channel-Aware Temporal Convolution Network," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1424–1437 (2021 Mar.). <https://doi.org/10.1109/TASLP.2021.3068598>.
- [42] A. S. Hatem and A. M. Al-Bakry, "Designing a Multimodal Emotion Recognition System Using Visual and Audio Modalities," *AIP Conf. Proc.*, vol. 3229, no. 1, paper 040022 (2024 Nov.). <https://doi.org/10.1063/5.0236119>.
- [43] J. G. Beerends and F. E. De Caluwe, "The Influence of Video Quality on Perceived Audio Quality and Vice Versa," *J. Audio Eng. Soc.*, vol. 47, no. 5, pp. 355–362 (1999 May).
- [44] S. Sun, S. Chen, and Y. Shen, "The Effects of Lighting on Perceptual Sound Quality in Subjective Listening Tests," *J. Audio Eng. Soc.*, vol. 73, no. 5, pp. 293–302 (2025 May). <https://doi.org/10.17743/jaes.2022.0201>.
- [45] W. van Bommel, *Interior Lighting: Fundamentals, Technology, and Application* (Springer, Cham, Switzerland, 2019). <http://dx.doi.org/10.1007/978-3-030-17195-7>.
- [46] L. Loe, K. P. Mansfield, and E. Rowlands, "Appearance of Lit Environment and Its Relevance in Lighting Design: Experimental Study," *Light. Res. Technol.*, vol. 26, no. 3, pp. 119–133 (1994 Sep.). <http://dx.doi.org/10.1177/096032719402600301>.
- [47] R. M. Kirsch, *Lighting Quality and Energy Efficiency in Office Spaces*, Ph.D. thesis, Technische Universität Berlin, Berlin, Germany (2015 May).
- [48] H. Fastl, "Audio-Visual Interactions in Loudness Evaluation," in Proc. of the 18th Intl. Congress on Acoustics (ICA), pp. 1161–1166 (Kyoto, Japan) (2004 Apr.).
- [49] C. Patsouras, T. Filippou, and H. Fastl, "Influences of Color on the Loudness Judgement," presented at the 3rd EEA European Congress on Acoustics Forum Acusticum (Sevilla, Spain) (2002 Jan.).
- [50] Z. Zhao and L. Xie, "Measurement on Just Noticeable Difference of Auditory Perception With Visual Stimuli," *Chin. J. Acoust.*, vol. 38, no. 5, pp. 624–631 (2013 Sep.). <https://doi.org/10.15949/j.cnki.0371-0025.2013.05.017>.
- [51] J. J. O'hare, "Intersensory Effects of Visual Stimuli on the Minimum Audible Threshold," *J. Gen. Psychol.*, vol. 54, no. 2, pp. 167–170 (1956 Apr.). <https://doi.org/10.1080/00221309.1956.9920271>.
- [52] K. Jensen and J. Hjortkjær, "An Improved Dissonance Measure Based on Auditory Memory," *J. Audio Eng. Soc.*, vol. 60, no. 5, pp. 350–354 (2012 May).
- [53] S. E. Olive, T. Welti, and E. McMullin, "The Influence of Listeners' Experience, Age, and Culture on Headphone Sound Quality Preferences," presented at the 137th Conv. of the Audio Engineering Society (2014 Oct.), paper 9177.
- [54] S. Sun, Y. Shen, Z. Liu, and X. Feng, "The Effects of Recording and Playback Methods in Virtual Listening Tests," *J. Audio Eng. Soc.*, vol. 63, no. 7/8, pp. 570–582 (2015 Jul.). <https://doi.org/10.17743/jaes.2015.0058>.
- [55] S. E. Olive, "A Method for Training Listeners and Selecting Program Material for Listening Tests," presented at the 97th Conv. of the Audio Engineering Society (1994 Nov.), paper 3893.
- [56] R. Quesnel, "Timbral Ear Trainer: Adaptive, Interactive Training of Listening Skills for Evaluation of Timbre," presented at the 100th Conv. of the Audio Engineering Society (1996 May), paper 217.
- [57] T. Letowski, and A. Miskiewicz, "Development of Technical Listening Skills for Sound Quality Assessment," in Proc. of the 24th INTER-NOISE, pp. 917–920 (Newport Beach, CA) (1995 Jul.).
- [58] T. Letowski, "Development of Technical Listening Skills: Timbre Solfeggio," *J. Audio Eng. Soc.*, vol. 33, no. 4, pp. 240–244 (1985 Apr.).
- [59] F. L. Martin, "Sound Localization Training and Auditory Adaptation: A Review," presented at the 150th Conv. of the Audio Engineering Society (2021 May), paper 10464.
- [60] Z. Liu, F. Wu, and Q. Yang, "The Practice and Study of Ear Training on Discrimination of Sound Attributes," presented at the 122th Conv. of the Audio Engineering Society (2007 May), paper 7114.
- [61] R. Kassier, T. Brookes, and F. Rumsey, "Designing a Spatial Audio Attribute Listener Training System for Optimal Transfer," presented at the 120th Conv. of the Audio Engineering Society (2006 May), paper 6819.
- [62] M. McKinnon-Bassett, W. Martens, and D. Cabrera, "Experimental Comparison of Two Versions of a Technical Ear Training Program: Transfer of Training on Tone Color Identification to a Dissimilarity-Rating Task," in Proc. of the AES 50th Intl. Conf.: Audio Education (2013 Jul.), paper 2-4.
- [63] T. Rosciszewska and A. Miskiewicz, "Timbre Solfege: Development of Auditory Cues for the Identification of Spectral Characteristics of Sound," presented at the 138th Conv. of the Audio Engineering Society (2015 May), paper 9351.

- [64] J. Merimaa and W. Hess, "Training of Listeners for Evaluation of Spatial Attributes of Sound," presented at the *117th Conv. of the Audio Engineering Society* (2004 Oct.), paper 6237.
- [65] T. Neher and T. Brookes, "Training of Listeners for the Evaluation of Spatial Sound Reproduction," presented at the *112th Conv. of the Audio Engineering Society* (2002 Apr.), paper 5584.
- [66] S. I. Iwamiya, Y. Nakajima, K. Ueda, K. Kawahara, and M. Takada, "Technical Listening Training: Improvement of Sound Sensitivity for Acoustic Engineers and Sound Designers," *Acoust. Sci. Technol.*, vol. 24, no. 1, pp. 27–31 (2003 Jan.). <https://doi.org/10.1250/ast.24.27>.
- [67] A. Miskiewicz, "Timbre Solfege: A Course in Technical Listening for Sound Engineers," *J. Audio Eng. Soc.*, vol. 40, no. 7/8, pp. 621–625 (1992 Jul.).
- [68] N. C. Otto, "Listening Test Methods for Automotive Sound Quality," presented at the *103th Conv. of the Audio Engineering Society* (1997 Sep.), paper 4586.
- [69] D. Martin and G. Massenburg, "Advanced Technical Ear Training: Development of an Innovative Set of Exercises for Audio Engineers," presented at the *139th Conv. of the Audio Engineering Society* (2015 Oct.), paper 217.
- [70] R. E. Shively and W. N. House, "Listener Training and Repeatability for Automobiles," presented at the *104th Conv. of the Audio Engineering Society* (1998 May), paper 4660.
- [71] S. E. Olive, "A New Listener Training Software Application," presented at the *110th Conv. of the Audio Engineering Society* (2001 May), paper 5384.
- [72] S. Bech, "Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment," *J. Audio Eng. Soc.*, vol. 40, no. 7/8, pp. 590–610 (1992 Jul.).
- [73] S. Bech, "The Influence of Room Acoustics on Reproduced Sound, Part 1: Selection and Training of Subjects for Listening Tests," presented at the *87th Conv. of the Audio Engineering Society* (1989 Oct.), paper 2850.
- [74] S. E. Olive, "Differences in Performance and Preference of Trained Versus Untrained Listeners in Loudspeaker Tests: A Case Study," *J. Audio Eng. Soc.*, vol. 51, no. 9, pp. 806–825 (2003 Sep.).
- [75] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "Relationships Between Experienced Listener Ratings of Multichannel Audio Quality and Naïve Listener Preferences," *J. Acoust. Soc. Am.*, vol. 117, no. 6, pp. 3832–3840 (2005 Jun.). <https://doi.org/10.1121/1.1904305>.
- [76] J. A. Swets, D. M. Green, and W. P. Tanner, Jr., "On the Width of Critical Bands," *J. Acoust. Soc. Am.*, vol. 34, no. 1, pp. 108–113 (1962 Jan.). <https://doi.org/10.1121/1.1908978>.
- [77] E. Zwicker and H. Fastl, "On the Development of the Critical Band," *J. Acoust. Soc. Am.*, vol. 52, no. 2B, pp. 699–702 (1972 Aug.). <https://doi.org/10.1121/1.1913161>.
- [78] E. De Boer, "Note on the Critical Bandwidth," *J. Acoust. Soc. Am.*, vol. 34, no. 7, pp. 985–986 (1962 Jul.). <https://doi.org/10.1121/1.1918235>.
- [79] D. D. Greenwood, "Auditory Masking and the Critical Band," *J. Acoust. Soc. Am.*, vol. 33, no. 4, pp. 484–502 (1961 Apr.). <https://doi.org/10.1121/1.1908699>.
- [80] D. D. Greenwood, "Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane," *J. Acoust. Soc. Am.*, vol. 33, no. 10, pp. 1344–1356 (1961 Oct.). <https://doi.org/10.1121/1.1908437>.
- [81] B. C. J. Moore, "Relation Between the Critical Bandwidth and the Frequency-Difference Limen," *J. Acoust. Soc. Am.*, vol. 55, no. 2, p. 359 (1974 Feb.). <https://doi.org/10.1121/1.1914512>.
- [82] E. Zwicker, "Subdivision of the Audible Frequency Range Into Critical Bands (Frequenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, no. 2, p. 248 (1961 Feb.). <https://doi.org/10.1121/1.1908630>.
- [83] E. Zwicker and E. Terhardt, "Analytical Expressions for Critical-Band Rate and Critical Bandwidth as a Function of Frequency," *J. Acoust. Soc. Am.*, vol. 68, no. 5, pp. 1523–1525 (1980 Nov.). <https://doi.org/10.1121/1.385079>.
- [84] ITU-R, "Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level," *Recommendation ITU-R 1770.3* (2012 Aug.).
- [85] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary Guidelines for Subjective Evaluation of Audio Source Separation Algorithms," in *Proc. of the UK ICA Research Network Workshop*, pp. 93–96 (Liverpool, UK) (2006 Sep.).
- [86] R. Schatz, S. Egger, and K. Masuch, "The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings," *J. Audio Eng. Soc.*, vol. 60, no. 1/2, pp. 63–73 (2012 Jan.).
- [87] D. Schwarz, G. Lemaitre, M. Aramaki, and R. Kronland-Martinet, "Effects of Test Duration in Subjective Listening Tests," in *Proc. of the Intl. Computer Music Conf. (ICMC)*, pp. 515–519 (Utrecht, Netherlands) (2016 Sep.).
- [88] F. E. Toole, *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms* (Routledge, New York, NY, 2017), 3rd ed. <https://doi.org/10.4324/9781315686424>.
- [89] F. E. Toole, "Beneficial Loudspeaker/Room/Listener Interactions," *J. Acoust. Soc. Am.*, vol. 150, no. 4 (Suppl), p. A75 (2021 Oct.). <https://doi.org/10.1121/10.0007678>.
- [90] S. Barbieri, K. Brkić Bubola, A. Bendini, et al., "Alignment and Proficiency of Virgin Olive Oil Sensory Panels: The OLEUM Approach," *Foods*, vol. 9, no. 3, paper 355 (2020 Mar.). <https://doi.org/10.3390/foods9030355>.
- [91] L. Sipos, Á. Nyitrai, G. Hitka, L. F. Friedrich, and Z. Kókai, "Sensory Panel Performance Evaluation—Comprehensive Review of Practical Approaches," *Appl. Sci.*, vol. 11, no. 24, paper 11977 (2021 Dec.). <https://doi.org/10.3390/app112411977>.
- [92] Z. Jalilbal, A. Amiri, P. Castagliola, et al., "Monitoring the Coefficient of Variation: A Literature Review," *Comput. Ind. Eng.*, vol. 161, paper 107600 (2021 Nov.). <https://doi.org/10.1016/j.cie.2021.107600>.

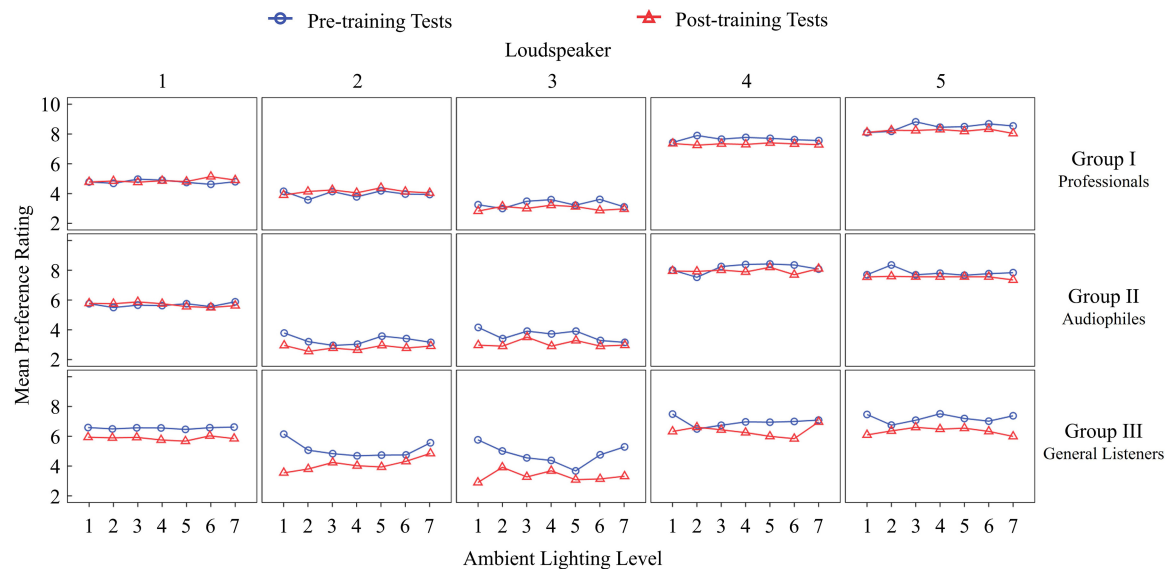


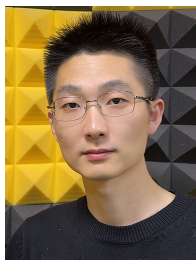
Fig. 13 Mean preference ratings of overall subjective sound quality in both pre-training and post-training listening tests.

A.1 RESULTS RELATED TO THE SUBJECT GROUP FACTOR

Fig. 13 shows the mean preference ratings of overall subjective sound quality in both pre-training and post-training listening tests. Each row of subplots corresponds to a different subjective group, while subplots in different columns represent evaluation results for different loudspeakers. In

each subplot, the x -axis denotes the ambient lighting condition, with values from 1 to 7 corresponding to illuminance levels of 530, 380, 250, 150, 60, 6, and 0 lux, respectively (i.e., Tests 1–7). The y -axis indicates the subjective sound quality ratings. In view of the limitations discussed in the main text, the data are presented to facilitate independent examination and interpretation by readers.

THE AUTHORS



Shuyuan Sun

Shuyuan Sun obtained his B.Sc. and M.Sc. degrees in acoustics from Nanjing University in 2011 and 2015, respectively. He has worked at Apple Inc. and AAC Technologies Inc., primarily focusing on the development of audio-related solutions. He led a team involved in various research and development activities, including acoustic user experience research, auditory perception measurement, architecture design of audio system, acoustic transducer development, audio power amplifier design, audio algorithm development, and audio system testing and tuning. He holds over 30 authorized patents in China, the United States, and Japan, covering various fields such as consumer electronics, automotive, television, and wear-



Yong Shen

able devices. He is currently pursuing a Ph.D. degree at the Institute of Acoustics at Nanjing University under the supervision of Professor Shen.

• Professor Yong Shen has been an AES member since 2004. He received B.Sc., M.Sc., and Ph.D. degrees in acoustics from Nanjing University where he continued his career in research and education. Yong works on research and development of electroacoustics and room acoustics. Yong has organized the International Symposium on Electro Acoustics Technology, a biannual symposium held in Shenzhen, since 2007.