DOI: 10.61782/fa.2025.0651

FORUM ACUSTICUM EURONOISE 2025

ON THE NEED FOR HOMOGENEOUS AND STANDARDIZED METHODOLOGIES FOR PERCEPTUAL TESTING IN ACOUSTICS

Daniel de la Prida[1*] **María Larrosa-Navarro**[1]
Luis A. Azpicueta-Ruiz[2] **Antonio Pedrero** [1]
[1] Grupo de Investigación en Acústica Arquitectónica, Universidad Politécnica de Madrid, Madrid, Spain
[2] Dep. Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, Madrid, Spain

ABSTRACT

Although acoustics can be studied from a physical perspective, in many cases its applications are intended to have some effect or interpretation by the human auditory system. In this sense, and until general models of human sound perception are developed, the main way of understanding auditory impressions is through perceptual tests using humans as instruments of evaluation. However, subjective assessments are complex and require controlling a myriad of sources of bias that can affect relevance, validity, and reproducibility. Among these, those related to test design, performance and analysis of results are often the most relevant. A review of the state-of-the-art shows that perceptual testing is often conducted heterogeneously across studies and with great variability in the considered experimental conditions. This communication, primarily intended to stimulate scientific dialogue, will recall factors that can significantly impact perceptual evaluations, commenting on them, and will raise some open questions for further common reflection. In this way, it is hoped that efforts towards establishing homogeneous methodologies for perceptual evaluation in acoustics can be revived, especially in this era where artificial intelligence algorithms increasingly depend on robust and precise data.

Keywords: *listening test, methodology, standardization*

**Corresponding author*: daniel.prida@upm.es.

1. INTRODUCTION

Many acoustical applications and scenarios can be quantified purely in physical terms. However, numerous others often require relating physical acoustic magnitudes to their interpretation by the human brain. Currently, the understanding of hearing mechanisms is deep; however, we still lack systems capable of accurately predicting the complete auditory response, including brain activation and interpretation, for every acoustic scenario.

Consequently, the most common current approach to assess the auditory system's responses and capabilities involves using human listeners themselves as instruments of perceptual evaluation [1] for particular acoustic conditions.

This approach has been commonplace for decades across various domains linking physical sound characteristics to perceptual attributes, including fields such as room and building acoustics, soundscapes, and binaural hearing, among others. However, a review of the state-of-the-art in these areas reveals that similar studies, or even those addressing identical topics, frequently apply markedly diverse evaluation and analytical methodologies as highlighted, for instance, in [2] in the case of soundscape research, in [3, 4] regarding building acoustics, or in [5] in room acoustics.

This heterogeneity has always impeded precise and unbiased comparisons among research outcomes from different research. Nowadays, with the current momentum of artificial intelligence technologies, which are starting to deliver deep neural network models representative of auditory system behavior for specific tasks [6], the need for perceptual evaluation outcomes derived from homogeneous and comparable methodological approaches has

become more critical than ever. Such methodological uniformity is essential to obtain clean and representative input datasets, thereby enabling robust and accurate training of deep learning algorithms.
Encouraging developments can be found in soundscape research, where methodological standardization [7, 8] has facilitated greater homogeneity, as well as in binaural hearing, where some researchers are currently placing significant emphasis on reproducible research [9, 10].
Nevertheless, there remains considerable scope for additional research on methodological issues, essential for achieving full comparability and enabling joint utilization of results in collaborative research endeavors.
The primary aim of this communication is to emphasize the critical need for methodological standardization in perceptual evaluations in acoustics, particularly when conducting listening tests. To illustrate this point, several relevant factors for the designing, performance, and analysis of perceptual experiments are highlighted as examples, to stimulate collective dialogue on this important issue. Given the breadth of the topic and the wide variety of contributing factors, and for reasons of space and focus, many of the examples discussed are drawn from room acoustics, building acoustics, and soundscape research. Nevertheless, the underlying principles and observations are broadly applicable to other areas of acoustics as well.

2. A SAMPLE OF SIGNIFICANT FACTORS

Perceptual experimentation in acoustics involves a sequence of interconnected steps. Among these, some of the most crucial are the experimental design, the performance of the tests, and the subsequent analysis of results. Throughout each of these stages, researchers must make numerous decisions, each potentially introducing specific biases. Consequently, varying decisions made during these stages can lead to considerable divergence in the outcomes, even among experiments sharing identical research objectives.

2.1 Experiment design

The experimental design stage is one of the most demanding in terms of both time and decision-making. During this phase, researchers select the auditory samples to be used in the tests, the acoustic scenarios and their variations to be assessed through these samples, and the experimental protocol itself, including the sequences followed for stimuli presentation, among other things. In addition, decisions are made regarding any corrections needed to compensate for the effects introduced by the sound playback systems.

2.1.1 Auditory samples

The influence of auditory samples and their temporal and spectral features, on participants' discrimination abilities and preferences in perceptual tests is well-documented and has been demonstrated extensively in the literature. Mentions to this influence can be found in publications in many areas, such as, for instance, [11] in building acoustics and in room acoustics [12, 13]. Therefore, it is critically important either that study conclusions be explicitly constrained to the auditory samples used as base for the test stimuli or that auditory samples considered representative for each practical application be selected accordingly. However, in many areas of acoustics, such as those of room or building acoustics, to the best of our knowledge, there is no established catalogue of reference sounds for use in perceptual testing that would ensure homogeneity across evaluations. Such a catalogue might not be strictly necessary if the sole objective were the creation of a large dataset for training deep learning models. However, given the current state of the field, where heterogeneity extends far beyond just this aspect, a standardized catalogue of reference sounds could be of considerable value.

2.1.2 Test protocol

The test protocol is one of the most influential aspects of a perceptual study, as it determines how stimuli are presented to participants and how their responses are collected [14]. The chosen protocol has a substantial impact on discrimination performance, with some protocols being operationally more powerful than others. Furthermore, different protocols vary in their susceptibility to learning, fatigue, and sequence effects, among other sources of bias. While these issues have been thoroughly studied in other areas of sensory evaluation [15], acoustics has seen the widespread use of a diverse range of protocols [5], often without consistency. Some studies have shown that these protocols differ in their operational power and vulnerability to bias [14,16]. Nevertheless, despite emerging trends favoring certain approaches, much research is still needed and there is still no widely accepted, robust, and standardized protocol in acoustics, unlike in the field of audio engineering, where a clearer methodological consensus has been reached [17, 18].

Looking more closely at commonly used protocols, particularly within the two major families, *difference testing* and *scaling*, several methodological trends arise that may introduce additional biases beyond those already mentioned, related to operational power and usual effects. For instance, it is not uncommon to find the use of attribute-related difference tests (i.e., tests in which participants are asked to evaluate the difference between stimuli regarding a specific attribute) even when the stimulus variation is inherently multifactorial. A clear example in the field of room acoustics is the use of attribute-related difference tests focused on particular attributes, such as "reverberation". For instance, changes in reverberation typically involve simultaneous changes in other auditory attributes (e.g., clarity, etc.), making it uncertain whether the perceived differences reported by the participants are exclusive to reverberation. While in other fields of sensory evaluation it is clearly recommended to use attribute-related tests for single-factor variations [19] and overall when they are multi-factorial and complex, it seems that there is currently no clear consensus in acoustics.

Regarding the use of scaling protocols, it is not uncommon to find studies employing even-numbered scales and those with a high number of response points in scales. However, findings from other domains of sensory perception suggest that such scales, and particularly those with more than seven points, may be inappropriate [20], as they can be cognitively demanding for participants and often fail to yield additional meaningful information. Furthermore, it is common to find scaling tests in which participants are not informed in advance about the full range of stimuli variability to assess during the test. This omission has been shown, in other fields, to result in a lack of use of the extreme ends of the scale [21], thereby biasing the sensitivity of the measurements.

Another key factor to consider when designing the listening test is the vocabulary used to describe the items to be evaluated [22, 23]. It is of the utmost importance to clearly define the term to ensure all participants understand it in a consistent and accurate way. Failing to do so could have a negative impact on the results of the listening test. There are a number of ways to ensure that the terminology is understood. Those with experience in professional fields such as acoustic research or music may have acquired the relevant knowledge through practical experience. However, it is important to note that the same term might carry different meanings in different areas of acoustics or between acoustics and other fields of knowledge. Individuals with no prior background in acoustics might find the terminology unfamiliar, in which case, it is essential that they receive specific training prior to the commencement of the test [24]. A glossary of acoustic terminology [25] already employed in numerous perceptual tests represents a fundamental step in the standardization of terminology and definitions employed in perceptual evaluation in acoustics.

2.2 Experiment performance

As previously mentioned in subsection 2.1, the experimental design phase is one of the most critical of a perceptual study. Nevertheless, during the actual performance of the experiment, several practical considerations may arise that are highly relevant in terms of bias control. For example, if the selected protocol is prone to learning effects, it becomes essential to provide participants with adequate training (for the protocol) beforehand, so they can familiarize themselves with the nature of the evaluation task. Similarly, if the protocol is likely to induce fatigue, or if the duration of the test is long, it is advisable to include rest periods between blocks to maintain concentration and response reliability. While such factors are generally acknowledged in acoustic perceptual testing, they are often insufficiently documented in the papers, which can hinder reproducibility and methodological transparency.

Regarding the test environment and its characteristics, two main factors, among many others, frequently emerge as potential sources of bias in laboratory-based acoustic perceptual experiments. On the one hand, the presence of acoustically inadequate test environments, such as those with excessive background noise or reverberation, is typically well controlled in contemporary research and does not pose significant concerns for reproducibility. On the other, many laboratory perceptual tests aim to evaluate real-world acoustic situations within controlled environments that often lack the full context of real-life perception. In such cases, additional sensory modalities, temporal dynamics, and cultural context may significantly influence perception. This discrepancy can lead to a mismatch between the real-world scenario being studied and the experience recreated in the laboratory. Although this issue has been recognized and addressed, in different degrees of depth, in fields such as soundscape [26, 27], room acoustics [28], and sound insulation [11, 29], we believe that further investigation is essential to better understand and mitigate this gap.

Also, the experimental setup invariably includes electroacoustic transducers, either loudspeakers or headphones, to

deliver the test stimuli to participants. These playback devices have their own acoustic responses, which must be properly compensated to ensure accurate stimulus reproduction. Although this factor is increasingly considered in the literature, it is important to emphasize its significance and to acknowledge the valuable contributions of the research community in providing publicly available datasets, such as [30], of the acoustic responses of commonly used transducers in perceptual acoustics testing.

The relevance of the interaction between listening test participants and experimenters is an aspect that has been acknowledged in some studies, although its handling remains inconsistent. In certain cases, experimenters provide verbal instructions, which may unintentionally introduce bias into the test results [31]. Other studies use written instructions to minimize this risk. However, detailed descriptions of this interaction are generally lacking in the literature. Also related to the interaction between the experimenter and the participant are the strategies used for participant recruitment, which are often simplified by relying on volunteers from the academic community. However, this approach may not always yield a sample that adequately represents the target population under investigation. We believe these aspects deserve greater attention and more systematic documentation.

Participant training and experience are critical considerations in the design of listening tests. In tonal discrimination research, for example, particular emphasis has been placed on comparing the analytical listening abilities of musically trained individuals with those of untrained participants, in an effort to identify the most suitable profiles. In the field of room acoustics, both trained and untrained listeners are frequently included in perceptual evaluations. Despite this, there remains a lack of targeted research examining the specific impact of musical or acoustical education on auditory perception. Notable contributions addressing this gap include some recent work [12, 24].

It is also important to acknowledge that the concept of ”expertise” in listening tests is multifaceted and often subject to debate. Listener performance may be shaped by various factors, such as perceptual sensitivity, knowledge of acoustics, musical background, or experience with audio recording. The relevance of each of these factors can vary depending on the specific task being assessed. As suggested in [24] for room acoustics, conventional criteria for identifying expert listeners may be insufficient. Therefore, we believe further research in this area would be highly beneficial.

2.3 Analysis of the results

The analysis phase of perceptual evaluation has the relative advantage that participant-related bias no longer plays a direct role. In this stage, the most critical issues concern how researchers process the experimental data to extract meaningful insights from participants' judgments. Often, when the tests yield quantitative outcomes, and independently of the level of measurement, analysis in acoustic research often relies on traditional parametric methods, such as the calculation of means and standard deviations, ANOVA, or linear regression.

However, consulting classical literature on statistics suggests that parametric statistic metrics such as mean and standard deviations, upon which widely used techniques like ANOVA and linear regression are based, can be considered to be valid, in a broad sense, under certain assumptions. For example, ANOVA, among others, requires the data to meet conditions of normality, homoscedasticity, and independence [32]. These assumptions are frequently not verified, or not reported, in many published studies in acoustics. Additionally, it is common practice to compute means from ordinal data, such as Likert-type responses, despite such data not being inherently continuous [33]. Sometimes it is argued that, with sufficiently large sample sizes, the distribution of aggregated scores may approximate a normal distribution, but this assumption might not always hold and some research suggests that it would be advisable that this be checked or non-parametric analysis to be conducted instead. Furthermore, perceptual research in acoustics has shown that the perceptual spacing between adjacent points on an ordinal scale may not be uniform [34], potentially further undermining the validity of using always parametric statistics. This concerns have been well-documented historically in the literature, although other well-documented studies suggest that parametric statistics may still be appropriate for the analysis of ordinal data. These studies argue that the violation of assumptions such as normality may have minimal, if any, impact on the validity of the conclusions drawn [35]. In view of this, we believe that it would be valuable, at least, to establish a common framework of analysis in perceptual acoustics so that the results of similar research can be properly compared and used together as data sets.

A parallel concern arises in difference testing protocols, where results are often interpreted in terms of the proportion of correct responses (p_c). However, as it has been demonstrated in other fields of sensory evaluation, this metric is inherently biased [36, 37]. It depends not only on

the actual perceptual discriminability between stimuli but also on the specific protocol used [37] and the cognitive decision strategies [36] employed by the participants, factors which are rarely examined in perceptual research in acoustics. Signal Detection Theory (SDT) provides more appropriate alternatives, such as d-prime (*d'*), which offer a bias-independent measure of sensitivity [36].

For these reasons, we argue that a common framework for the statistical methods, parametric and non-parametric, used in perceptual research in acoustics is essential. Such a framework would facilitate inter-study comparability, promote methodological homogeneity, and reduce potential biases introduced during the data analysis phase. We believe this is one of the factors that could be most readily addressed, for example through the development of statistical analysis packages created collaboratively and based on standardized tools for processing perceptual data obtained from specific types of listening tests.

3. CONCLUSIONS

This work has highlighted the widespread methodological variability that still characterizes perceptual testing in acoustics, even in fields with decades of accumulated experience. From the design of stimuli and protocols to the treatment of data and the selection of participants, decisions made at each stage of the process can significantly influence outcomes and often in ways that hinder comparability and reproducibility across studies.

While some domains, such as soundscape research, are moving toward methodological convergence, many others continue to rely on ad hoc decisions. This limits the broader utility of perceptual data, especially in a context where artificial intelligence is emerging as a powerful tool to model auditory perception. Without homogeneous, high-quality input data, the potential of these models may remain challenging.

Rather than proposing a one-size-fits-all solution, which we believe is unfeasible without much discussion across many researchers, this paper advocates for conscious, evidence-based methodological decisions, better documentation, and above all, collective efforts to develop shared protocols, reference datasets, and analytical tools. The goal is not only to improve the quality and transparency of individual studies, but to empower the field to build interoperable datasets that can serve both traditional research needs and the training of perceptually informed AI systems.

4. REFERENCES

[1] N. Zacharov, ed., *Sensory evaluation of sound*. Boca Raton, Florida: CRC Press, 1st ed., 2018.

[2] F. Aletta, J. Kang, and Ö. Axelsson, "Soundscape descriptors and a conceptual framework for developing predictive soundscape models," *Landscape and Urban Planning*, vol. 149, pp. 65–74, 2016.

[3] N.-G. Vardaxis and D. Bard, "Review of acoustic comfort evaluation in dwellings: Part iii—airborne sound data associated with subjective responses in laboratory tests," *Building Acoustics*, vol. 25, no. 4, pp. 289–305, 2018.

[4] M. Geluykens, H. Muellner, V. Chmelík, and M. Rychtarikova, "Airborne sound insulation and noise annoyance: Implications of listening test methodology," in *Proceedings of Forum Acusticum 2023*, (Torino, Italy), pp. 155–162, European Acoustics Association, 2023.

[5] D. de la Prida, A. Pedrero, L. A. Azpicueta-Ruiz, and M. Ángeles Navacerrada, "Does the method matter? a review of the main testing methods for the subjective evaluation of room acoustics through listening tests," in *Proc. of the 23rd International Congress on Acoustics*, (Aachen, Germany), pp. 7871–7878, 2019.

[6] G. Tuckute, J. Feather, D. Boebinger, and J. H. McDermott, "Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions," *Plos Biology*, vol. 21, no. 12, p. e3002366, 2023.

[7] *ISO/TS 12913-2:2018 - Acoustics — Soundscape — Part 2: data collection and reporting requirements*. Geneva, Switzerland: International Organization for Standardization, 2018.

[8] *ISO/TS 12913-3:2019 - Acoustics — Soundscape — Part 3: data analysis*. Geneva, Switzerland: International Organization for Standardization, 2019.

[9] P. Majdak, M. J. Goupell, and B. Laback, "3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention, perception, & psychophysics*, vol. 72, no. 2, pp. 454–469, 2010.

[10] R. Barumerli and P. Majdak, "Frambi: A software framework for auditory modeling based on bayesian inference," *Neuroinformatics*, vol. 23, no. 2, p. 20, 2025.

[11] V. Hongisto, D. Oliva, and J. Keränen, "Subjective and objective rating of airborne sound insulation–living sounds," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 848–863, 2014.

[12] A. Kuusinen and T. Lokki, "Recognizing individual concert halls is difficult when listening to the acoustics with different musical passages," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1380–1390, 2020.

[13] M. Larrosa-Navarro, D. de la Prida, and A. Pedrero, "Influence of musical stimulus on the perception of clarity in rooms and its relation to c80," *Applied Acoustics*, vol. 208, p. 109370, 2023.

[14] D. de la Prida, A. Pedrero, L. A. Azpicueta-Ruiz, and M. Á. Navacerrada, "Listening tests in room acoustics: Comparison of overall difference protocols regarding operational power," *Applied Acoustics*, vol. 182, p. 108186, 2021.

[15] M. A. Stocks, D. van Hout, and M. J. Hautus, "Cognitive decision strategies adopted by trained judges in reminder difference tests when tasting yoghurt, mayonnaise, and ice tea," *Food Quality and Preference*, vol. 34, pp. 14–23, 2014.

[16] E. Parizet, N. Hamzaoui, and G. Sabatie, "Comparison of some listening test methods: a case study," *Acta Acustica united with Acustica*, vol. 91, no. 2, pp. 356–364, 2005.

[17] *ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems.* Geneva, Switzerland: International Telecommunication Union, 2015.

[18] *ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems.* Geneva, Switzerland: International Telecommunication Union, 2015.

[19] S. McClure and H. T. Lawless, "Comparison of the triangle and a self-defined two alternative forced choice test," *Food Quality and Preference*, vol. 21, no. 5, pp. 547–552, 2010.

[20] L. J. Simms, K. Zelazny, T. F. Williams, and L. Bernstein, "Does the number of response options matter? psychometric perspectives using personality questionnaire data.," *Psychological assessment*, vol. 31, no. 4, p. 557, 2019.

[21] P. Chevret and E. Parizet, "An efficient alternative to the paired comparison method for the subjective evaluation of a large set of sounds," in *Proceedings of the 19th International Congress on Acoustics (ICA 2007), Madrid*, (Madrid, Spain), pp. 1–5, 2007.

[22] R. Hawkes and H. Douglas, "Subjective acoustic experience in concert auditoria," *Acta Acustica united with Acustica*, vol. 24, no. 5, pp. 235–250, 1971.

[23] A. Farina, "Acoustic quality of theatres: correlations between experimental measures and subjective evaluations," *Applied acoustics*, vol. 62, no. 8, pp. 889–916, 2001.

[24] M. von Berg, J. Steffens, S. Weinzierl, and D. Müllensiefen, "Assessing room acoustic listening expertise," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2539–2548, 2021.

[25] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (saqi)," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 984–994, 2014.

[26] C. Guastavino, B. F. Katz, J.-D. Polack, D. J. Levitin, and D. Dubois, "Ecological validity of soundscape reproduction," *Acta Acustica united with Acustica*, vol. 91, no. 2, pp. 333–341, 2005.

[27] C. Tarlao, D. Steele, and C. Guastavino, "Assessing the ecological validity of soundscape reproduction in different laboratory settings," *Plos one*, vol. 17, no. 6, p. e0270401, 2022.

[28] B. N. Postma and B. F. Katz, "The influence of visual distance on the room-acoustic experience of auralizations," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. 3035–3046, 2017.

[29] V. Chmelík, M. Rychtáriková, H. Müllner, K. Jambrošić, L. Zelem, J. Benklewski, and C. Glorieux, "Methodology for development of airborne sound insulation descriptor valid for light-weight and masonry walls," *Applied acoustics*, vol. 160, p. 107144, 2020.

[30] B. Boren, M. Geronazzo, P. Majdak, and E. Choueiri, "Phona: a public dataset of measured headphone transfer functions," in *Proceedings of the 137th Audio Engineering Society Convention 2014*, (Los Angeles, USA), 2014.

[31] M. C. Meilgaard, B. T. Carr, and G. V. Civille, *Sensory evaluation techniques*. CRC press, 1999.

[32] A. Agresti and B. Finlay, *Statistical methods for the social sciences*. Upper Saddle River, NJ: Pearson Education, 4 ed., 2009.

[33] A. Agresti, *Categorical Data Analysis*. Hoboken, NJ: Wiley, 3 ed., 2013.

[34] M. Lionello, F. Aletta, A. Mitchell, and J. Kang, "Introducing a method for intervals correction on multiple likert scales: A case study on an urban soundscape data collection instrument," *Frontiers in psychology*, vol. 11, p. 602831, 2021.

[35] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in health sciences education*, vol. 15, pp. 625–632, 2010.

[36] J. Frijters, "The paradox of discriminatory nondiscriminators resolved," *Chemical Senses*, vol. 4, no. 4, pp. 355–358, 1979.

[37] H.-S. Lee, M. O'Mahony, *et al.*, "Sensory difference testing: Thurstonian models," *Food Science and Biotechnology*, vol. 13, no. 6, pp. 841–847, 2004.