Audio Engineering Society

Conference Paper

Presented at the AES 2026 International Conference on Automotive Audio
2026 July 29-31, Detroit, MI, USA

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (https://aes.org/publications/elibrary), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

A Practical Method for Routine *in situ* Evaluation of Sound Quality in Vehicle Audio Systems

Rafael Kassier[1], Darlene Williamson[2], Kejsiana Kastrioti[2], Basheer Bazoun[2], Jan-Niclas Schneider[1], and Jonathan Pierce[2]

[1] *Harman-Becker Automotive Systems GmbH, Karlsbad, Germany*

[2] *Harman Automotive, Novi, MI*

Correspondence should be addressed to Rafael Kassier: rafael.kassier@harman.com

ABSTRACT

A method is described for measuring and comparing the perceived sound quality of vehicle audio systems *in situ*. The process has been refined through use at company facilities for over two decades and has been applied across a variety of internal use cases within the organization. Panels of experienced evaluators, aligned processes, and dedicated facilities are used at multiple sites around the globe to support consistent data collection. The collected data are benchmarked against a periodically refreshed reference population of recent vehicle evaluations, allowing each vehicle to be positioned relative to the contemporary competitive landscape. Hundreds of current vehicle models are available for comparison at any given time. The resulting information is used within the organization to support sound system target setting and validation, competitor analysis, sales pursuits, and product development and prototyping. More recently, the method has also been explored as a basis for sound-quality predictor modeling. This paper describes the overall method, the associated analysis workflow, and a set of anonymized case studies intended to illustrate practical comparability and discrimination in routine use.

1 Motivation and Overview

The method described in this paper was developed in response to the challenge of characterizing, measuring and comparing the sound quality of vehicle sound systems. The method has evolved over the last two decades, with the latest major method and analysis revision in 2014. Incremental refinements and development have continued and are planned in the future.

Across a typical vehicle program, audio quality judgments arise at multiple stages. Individual sound system designers tune prototype systems, other specialists provide a technical or managerial sign-off, and then additional impressions emerge later through internal vehicle reviews and external media evaluations. Marketing teams and third-party research companies also collect customer self-report feedback, yet the resulting reports do not always make clear which audio system option was experienced or which attributes were being assessed. The result is an evidence chain that is inconsistent in definition and confidence. For this reason, a structured listening evaluation method was developed to provide a more quantitative basis for the process, while preserving expert judgment and producing structured metrics that can support clear engineering trade-offs and sound business decisions.

The purpose of this paper is to document this procedure as a mature industrial workflow for automotive sound-quality evaluation. The contribution of the paper is therefore primarily practical, while also being methodological: it describes how trained-listener evaluations, supporting measurements, metadata collection, and structured analysis are combined in a high-throughput benchmarking process intended for routine product assessment. The paper does not claim complete psychometric validation of the method, nor does it

attempt to establish a universally applicable standard for all forms of automotive audio evaluation. Rather, it presents an approach that has proved operationally useful over an extended period of industrial application and that shows encouraging evidence of consistency and practical sensitivity in use.

Database population statistics and detailed analyses of the performance of specific vehicles are outside the scope of this paper. However, anonymized examples are included to illustrate how the process behaves in practice, how results are interpreted within the workflow, and what sort of comparative insight the method can provide under routine operating conditions.

2 Problem Description and Solutions

Evaluating installed automotive audio systems carries a specific set of challenges. Automotive sound systems do not just consist of the constituent loudspeakers and amplifiers and their configuration, but also include their own unique listening environment: the vehicle cabin interior, including any extraneous noise. Electrical or acoustic noise disturbances can be noticeable even when the vehicle is stationary. When vehicles are moving there will be engine noise, road noise, wind noise and other factors to consider. Furthermore, the interior trim of the vehicle will also contribute acoustically in the form of audible buzzing, squeaking, and rattling, so-called "BSR" [1], which happens in sympathy with audio playback.

2.1 Driving vs. Stationary Testing

For sound quality comparisons we take the most controlled condition: when the vehicle is stationary. For the most part, the static condition has the lowest level of external interference and is therefore used as the primary measurement and comparison point. In other words, each vehicle is assessed under conditions that give it the best opportunity to demonstrate its sound quality performance. If the dynamic tuning of the system is well implemented by the acoustic system designers, it will often only aim to maintain the stationary sound quality performance when driving.

Whilst vehicles are designed to move, the static use-case is far from uncommon [2]. When stopped in traffic or waiting in the vehicle, drivers are able to focus on sound playback with less distraction than when actively driving. In addition, sound systems are commonly approved by auto manufacturer representatives under static conditions, and potential buyers also often audition vehicle-installed sound systems at car dealerships while the vehicle is stationary.

A further consideration is the variability introduced by evaluating sound quality whilst driving a vehicle. Safety concerns include distracted driving and the increased cognitive load associated with evaluating while driving. A parallel-task study has suggested that making sound-quality judgements during a (simulated) driving task may shift ratings, with listeners tending to be less critical at the lower end of the quality scale [3]. In addition, there is limited control over road surface, traffic, and weather conditions, all of which affect the sound playback or its perception. In principle, controlled dynamic evaluation could be carried out on a chassis dynamometer, allowing driving conditions to be simulated under more repeatable conditions. However, the cost and logistical burden of such an approach have generally made it impractical for routine sound-quality evaluation.

2.2 *In situ* testing versus Virtualized Environments

Rather than needing to control the listening environment, as is typically the case for domestic or professional loudspeaker tests [4], the car cabin interior is an integral part of the vehicle sound system and contributes to both the acoustic performance and the subjective experience of the installed system [5]. A key consequence for subjective evaluation of automotive sound systems is therefore the importance of *in situ* testing.

Attempts have been made to allow vehicle sound to be evaluated *outside* the vehicle, either using headphones [6] or a loudspeaker array [7]. Such alternative workflows have a long lineage in automotive audio, from early dummy-head / binaural comparison methods intended to hold listening conditions constant across car systems [8] to later head-tracked binaural car-scanning approaches [9]. For this work we deliberately prioritize *in situ* evaluation in the vehicle over laboratory virtualizations and auralization workflows. The primary reason is what we term *experience fidelity*: preservation of the end-user listening experience in the actual cabin, with the production hardware, packaging, and signal processing operating as intended. By avoiding an additional record/replay chain, we also reduce the risk of introducing artefacts or perceptual side effects that are not part of the product, for example limitations of reproduction or externalization and low-frequency reproduction issues that have been highlighted as shortcomings of headphone-based approaches [7]. Evaluating *in situ*

also includes low-frequency cabin gain effects [10] and cabin modal behavior that loudspeaker-based approaches such as the one outlined in [7] do not reproduce directly. *In situ* testing is also a practical and naturalistic way of obtaining subjective evaluations of in-car sound quality.

A second, equally important reason is throughput. The process is intended for routine product testing, with multiple vehicle evaluations conducted each week throughout the year. Any need to build and validate a virtual model would consume considerable time. Even when virtualization is carefully implemented, it adds acquisition and preparation steps that are inherently additional to the core task of judging the system as experienced in the vehicle [11]. By contrast, *in situ* evaluation allows rapid iteration and decision-making directly on the product, which is essential in an engineering and business context where many variants, updates, and competitive benchmarks must be assessed continuously.

At the same time *in situ* evaluation can be influenced by non-acoustic factors, including sighted bias, brand cues, and other contextual information that may bias subjective perceptions of sound quality. In the present application, however, *experience fidelity* is prioritized because the objective is sound quality assessment in a product rather than an abstracted psychoacoustic experiment. The most salient biases are mitigated through practical countermeasures, such as masking visible audio-system identifiers with tape or applying unobtrusive decoy badges to loudspeaker grilles. This reduces evaluators' ability to infer system grade or option level directly from visual cues, while preserving the practical and decision-speed advantages of *in situ* assessment.

Tactile and haptic transducers are increasingly integrated into automotive seating to enhance perceived low-frequency impact [12,13]. Another advantage of *in situ* evaluation is its ability to assess these systems as part of the combined auditory and tactile experience, including the interaction between body-coupled sensations and acoustic playback. Replicating this combined auditory and tactile experience in a virtualized evaluation is a substantial challenge, because it requires not only accurate acoustic reproduction but also a functionally equivalent tactile reproduction chain, including comparable seat mechanics, transducer placement and coupling, processing, and input stages. In the absence of such equivalence, virtualization risks omitting or altering a perceptually relevant component of the product.

2.3 Blind versus Sighted Tests

Blind and sighted tests are often contrasted because visual and contextual cues can influence sound-quality judgements [14]. The method therefore applies targeted controls that remove the most salient cues with minimal disruption, for example masking or altering visible audio-system identifiers.

In an *in situ* workflow, however, fully blinding an evaluator to the vehicle they are physically sitting in is of limited value and quickly becomes impractical. The evaluator must still interact naturally with the product, for example by skipping tracks and adjusting volume, while also entering subjective ratings through the evaluation interface. If the listener is blindfolded, these actions typically require administrator support, which increases cognitive load, alters the task, increases the possibility of unintended administrator influence on listener responses, and effectively creates a one-to-one staffing requirement throughout the evaluation.

The lengths required to remove vehicle cues are illustrated in Shively's double-blind *in situ* testing [15]. The listener is blindfolded in another room and led to the vehicle while wearing headphones playing pink noise, with communication via intercom. The cabin is deliberately de-cued through measures such as scent masking, seat coverings, instructions not to touch the interior, and physical covering of areas that could reveal vehicle identity or configuration. An administrator operates playback outside the vehicle and controls the procedure.

For routine automotive product testing, these measures introduce two important risks. First, obvious concealment can encourage evaluators to spend effort inferring what is being hidden, rather than concentrating on the sound-quality attributes under assessment. In practice, this behavior is occasionally observed when visible brand identifiers in the vehicle under test (VUT) are masked, as discussed in Section 2.2. Second, the procedure itself becomes a substantial departure from *experience fidelity*: the listener is not simply evaluating "the vehicle as experienced", but a constrained, mediated scenario involving blindfolds, masking noise, restricted touch, and remote operation.

There is also a straightforward capacity issue. The preparation, staffing, and repeatable de-cueing steps required for each vehicle are time-consuming, expensive and therefore incompatible with a high-throughput program.

A potential future direction is to control cues without physically blindfolding the evaluator, for example using a "virtual reality" (VR) headset with an integrated rating and control interface and a

genericized cabin visual model. This could reduce vehicle-identifying visual information while preserving a natural interaction loop for track control, level changes, and feedback entry, and may offer a way to strengthen cue control with reduced workflow disruption.

3 Required Resources

Bech [16] outlines the human and physical resources needed for a sustained program of subjective evaluation. He emphasizes that such work depends on both people and infrastructure: a permanent, supported panel of trained evaluators, together with dedicated facilities and stable test setups that reduce workload, variability, and error.

The following section outlines the practical resources required to implement the evaluation method described in this paper. These include the testing facilities, human resources supporting the listening evaluations, audio hardware and software and supporting information technology (IT) infrastructure used for data collection and analysis, and the VUTs. Replication of the method requires appropriate provision of these elements, which together constitute the operational framework for consistent and repeatable subjective evaluation.

3.1 Real Estate

There are two real estate requirements: a reference listening room which provides an auditory anchor for the evaluations, and a vehicle laboratory in which to measure and evaluate the VUT.

3.1.1 Reference Listening Room

A central element of the method is the availability of a stable comparison point against which the sound quality of the VUT can be judged. For this purpose, a dedicated reference listening room is used as an external auditory anchor for the evaluation.

The reference rooms used in this work are based on the Harman Reference Listening Room design described by Olive [17], which was developed to provide a standardized, high quality listening environment representative of a domestic playback space rather than an extremely controlled studio control room. This approach reflects the intended use of many consumer audio products, where the goal is to reproduce music in conditions like those found in typical living rooms. As a result, the reference room provides a perceptually meaningful benchmark for high quality music reproduction.

This characteristic also makes the reference environment particularly relevant for automotive audio systems, especially those marketed under premium audio brands. Such systems are designed with the goal of recreating a high-quality home listening experience within the vehicle cabin. The domestic-style reference listening room therefore serves as a practical target against which the in-vehicle reproduction can be judged.

Although the original intention was to construct physically identical reference rooms at each location, this has not always been possible due to differences in available building spaces and construction methods. Experience has shown that modest differences in room dimensions and construction are less critical than originally assumed, provided that the rooms maintain broadly similar acoustic behavior with calibrated playback systems suitable for critical listening.

To support reliable auditory comparison, the reference listening room must be located close to the vehicle evaluation facility. Minimizing the walking distance between the reference room and the VUT helps evaluators retain as much of the auditory memory of the reference environment during the evaluation process as possible.

Future work may explore the use of an auralized reference listening room reproduced over positionally tracked headphone playback with VR visualization, allowing a reference condition to be experienced directly in the VUT and deployed at sites without dedicated facilities. To date this has not been pursued due to concerns about the *experience fidelity* of auralized environments, the availability of physical reference rooms at existing facilities, and a preference for collecting data under high-quality *in situ* listening conditions.

3.1.2 Vehicle Laboratory

A *relatively* quiet and *relatively* well-soundproofed garage space is required to house the VUT for the duration of the measurement and evaluation phases. Ideally this space will allow the vehicle doors to open on both sides and have enough additional space to house any measurement equipment.

Because the evaluators need to move quickly between referencing in the room and VUT, the vehicle lab must be in close proximity to the reference listening room.

3.2 Human Resources

The personnel required to support the process fall into two groups: *staff* and *listening panelists*.

3.2.1 Staff

A permanent member of staff, referred to as the *site administrator*, is required at each evaluation location. This individual is responsible for recruitment, training and maintenance of the listener panel. The role also includes planning and executing evaluations, performing objective measurements, collecting metadata, and producing reports.

In addition, a management team oversees the program. This group is responsible for data analysis, methodological development and the presentation and dissemination of results to stakeholders.

3.2.2 Listening Panelists

Bech and Zacharov describe listening tests as a principal means of studying perceptual audio characteristics and of assessing the quality of audio systems [18]. Perceived sound quality is not directly measurable by physical means alone and therefore requires listening-based evaluation. In practice, this requires a panel of trained listeners who act collectively as the measurement instrument.

Our listening evaluations are conducted by employees who have been trained for the task and who develop increasing experience through continued participation. These evaluators must be available on site for both scheduled and *ad hoc* listening sessions. This requirement depends on a supportive management culture that allows employees to participate in the program. The most active benchmarking sites typically maintain a pool of approximately thirty to fifty listeners. The actual number per site varies by the size of the employee population within each benchmarking location. This is generally sufficient to support two listening sessions per week with an initial sample of twelve to fifteen scheduled listeners for each VUT. It also helps keep participation to no more than a single one-hour session per listener, while retaining margin for absences and fluctuations in employee workload.

3.3 Audio Hardware and Software

The implementation of the evaluation method also depends on supporting audio hardware and software. These resources fall into three main areas: equipment used in the listening room to establish a reference listening experience, software and facilities used for listener training, and objective measurement tools used to document the acoustic performance of the VUTs.

3.3.1 Listening Room Equipment

The main requirement for the reference room at each benchmarking location is a two-channel stereo playback system designed to support playback quality at the “amazing” anchor of the scale described in Section 6.2. As part of achieving that, we follow the recommendations in [17], using appropriate sound absorption and acoustic diffusion hardware. The playback system combines prosumer-grade home theater control hardware with loudspeakers, amplifiers, and DSP components from consumer and professional audio applications, enabling selection between computer-based playback, streaming services, and traditional media players.

Over time, the reference listening rooms have evolved alongside industry playback formats, progressing from two-channel stereo through 5.1 and 7.1 surround sound to the current immersive playback systems with height channels. Bass management is implemented using four subwoofers, one in each corner of the room, to ameliorate room resonances and improve bass uniformity across seats. Time-of-arrival alignment is applied to all loudspeaker channels relative to the center-front seat, and the system response is tuned to a target curve with bass lift and treble roll-off similar to that published in [19]. The current target curve specifies a 5 dB boost below 100 Hz with a slope of 12 dB per octave and a 1.5 dB cut above 1 kHz with a slope of 6 dB per octave. Future work should evaluate whether the room target reported in [19] remains applicable when assessed across all five global sites, using a broader and more geographically diverse response base.

At present, vehicle audio systems are evaluated with stereo program material, whether reproduced directly or processed by “upmix” algorithms within the VUT. Accordingly, the default reference condition in the room is stereo playback through the front left and right speakers with bass management to the four subwoofers.

A consistent method for benchmarking immersive-format playback in vehicles has not yet been finalized, largely due to difficulties obtaining and replaying consistent program material across the various immersive audio streaming platforms that are available in VUTs. The multichannel playback capabilities of the reference rooms are still used during listener training, where listeners are encouraged to experience and describe upmixed stereo, surround sound, and immersive-format playback alongside the traditional stereo reference condition.

3.3.2 Listener Training Software and Facilities

Formal listener training has been shown to improve the effectiveness of listening panels. Earlier implementations of the method used the software training applications described in [20] and [21], while more recent work has focused on internal multichannel training software for the reference listening rooms, allowing more complex spatial scenarios to be demonstrated. Similar "training vehicles" with variable acoustics have also begun to be introduced to accelerate listener experience in describing automotive sound reproduction scenarios. Accurate description of a range of playback scenarios in such vehicles is expected to form part of a future certification program for validating listener performance. Further details of automotive listener training may be presented in a future publication.

3.3.3 Vehicle Lab Equipment and Software

Vehicle benchmarking also requires a set of measurement and documentation tools to capture the acoustic performance of the sound system and record relevant metadata describing the VUT.

An appropriate microphone array [22], together with acoustic measurement hardware and software, is required to document and capture the acoustic performance of the sound system (see Section 7). A sound pressure level (SPL) meter and microphone calibrator are also essential.

A dedicated photographic camera is used to capture metadata relating to the VUT and its sound system, including exterior and interior views, infotainment settings pages, media interfaces and controls, and loudspeaker grille appearance and locations (see Section 8). A measuring ruler may be placed in-frame to aid documentation of loudspeaker dimensions.

A 360° camera and panorama "stitching" software are also used to generate a "driver's-eye panorama" of the cabin interior using a Mercator projection. This image is incorporated into the evaluation interface, allowing listeners to indicate the perceived locations of sound stage elements (see Section 6.4). Small workshop lights are useful for providing indirect illumination of darker areas of the cabin during image capture.

3.4 IT Infrastructure

Subjective data, objective data, and test metadata must be collected, stored, analyzed, visualized, and reported. These functions can be implemented using a variety of software tools and system architectures. For the purposes of this discussion, it is not necessary to describe any specific implementation, but rather to outline the role of each application and the capabilities required to support the evaluation process.

3.4.1 Database

All collected data are stored in an appropriate database containing fields for all relevant metadata, along with storage for both objective measurements and subjective responses. The database is reliably backed up and accessible from each site where data is collected, processed, and reviewed. It also makes the stored information readily available for analysis, reporting, and export. Where data, analyses, or reports are captured or generated outside the main database environment, they are imported into, or linked with, the corresponding database records.

3.4.2 Subjective Data Collection Software

Subjective data (see Section 6) are collected during listening tests in which trained evaluators are guided through the assessment of a vehicle sound system using a laptop running a custom-built test interface. The evaluators' responses are recorded, stored in the database, and linked with the corresponding metadata and objective measurements.

3.4.3 Data Analysis and Reporting Software

Depending on the capabilities of the database and associated software, data relating to each evaluation may be analyzed and cross-referenced, and reports may be generated, either directly within the database or using external software. Reports may describe metadata, objective and subjective measurements, and their relationships for individual vehicle systems or groups of systems.

3.5 Vehicles Under Test

Ready access to vehicles is an essential requirement for any vehicle sound evaluation program, especially one that aims to sample and build a population based on the entire market. This requires practical local arrangements for sourcing vehicles, typically through relationships with dealerships, rental companies, fleet providers, or other vehicle-finding services available at each test location. Budget must also be allocated for the rental, transport, and general handling of these vehicles as part of normal benchmarking activity.

Vehicles may be requested for several purposes, including concept studies, demonstration activities, and prototype development. However, the majority of vehicles tested are current production vehicles obtained for competitive analysis and product validation. In these cases, the preferred VUT is one that is representative of what an end customer would

receive, rather than a prototype, engineering mule, or manufacturer press vehicle.

A healthy set of current production vehicle evaluations should be maintained on an ongoing basis. This serves several purposes: it provides a relevant contemporary reference against which newly tested vehicles can be compared, supports ongoing monitoring of market trends, and ensures that the test data remain grounded in the products that customers are buying and using. As a general guideline, vehicles are typically requested with fewer than one hundred miles on the odometer, so that the VUT is new, or as close to new as practical.

The vehicle set used in benchmarking is intentionally broad so that it reflects the range of audio system offerings present in the market and provides an appropriate basis for competitive analysis, product validation, and market performance benchmarking. This includes vehicles equipped with competitor systems as well as vehicles containing our own systems, across base-grade, non-branded premium, and premium branded audio options. Maintaining this breadth is important because it allows the program to assess how our systems compare with the market at multiple trim and feature levels, rather than only against a narrow set of directly branded premium competitors.

In addition to production vehicles under test, access to a dedicated *training vehicle* is beneficial. This vehicle should be equipped to support listener training and method development, for example through selectable loudspeakers and DSP-based control of acoustic conditions. Such a platform allows evaluators to be exposed to controlled changes in system behavior and provides a practical means of developing and maintaining panel sensitivity.

4 Listener Selection and Training

The use of trained evaluator panels in listening tests, similar to those used in sensory-evaluation practice, has long been recognized as a practical means of improving consistency and efficiency. Bech reports that listeners with substantial previous experience in listening tests are more consistent in their ratings than inexperienced listeners, and further states that “a trained subject can replace up to seven untrained subjects if only statistical aspects are considered” [23]. The benchmarking method described here has therefore always relied on panels of trained evaluators.

At each location, participants are drawn exclusively from employees. With the support of their line managers, this employee-based approach allows a stable listener pool to be maintained over time. The precise methods used to recruit, select, and train listeners vary across the global sites according to local labor laws and regional practice. Nevertheless, several common principles apply across all locations. Initial training is intended primarily to give new evaluators enough information to begin participating in the process, including familiarity with the method and a framework for evaluating the listening experience using the defined attributes and rating scales. In that sense, the individuals used in the benchmarking program may reasonably be described as *trained listeners*, because they have received sufficient instruction to take part in formal evaluations. In practice, however, the greater value of a listener develops through repeated participation in evaluations of many different vehicles and through a progressively deeper understanding of how the range of listening experiences should be interpreted and scaled. For this reason, *experienced listener* is often the more useful description in the context of the present benchmarking program.

Sound system designers are not excluded from the listener pool, unless they have personally worked on the VUT. In routine benchmarking, however, they typically account for no more than one or two listeners within an evaluation panel of around twelve to fifteen subjects. The emphasis is instead placed on maintaining as diverse a listener pool as practical. Employees in both technical and non-technical positions are encouraged to participate, and reasonable efforts should be made to improve representation of demographic groups that may otherwise be under-represented. At the same time, the exact composition of an individual evaluation panel is more difficult to control in practice. Participant availability is constrained by normal patterns of employee workload and absence, and it is not always possible to select listeners from the full pool on an ideal basis. Some individuals therefore occasionally participate in more than the usual limit of one evaluation per week in order to ensure a full panel for a session. These are practical realities of using human listeners as measurement instruments in an industrialized subjective evaluation program.

Historically, audiometric screening formed part of listener selection in some benchmarking locations, but this was later discontinued. One reason for this was practical: audiometric records constitute employee health data and therefore impose a substantial burden in relation to collection, storage, access control, and legal compliance. A second reason was empirical within the program itself: high-performing listeners, as judged by training-software acuity and by the quality of their listening-test

responses, did not always align with the audiometric criteria that had been applied. In the absence of evidence that those criteria materially improved panel quality for the present application, they were removed. This adjustment reduced barriers to participation.

Listener development does not depend on initial training alone. Trends in individual rating behavior are also documented over time in order to track outlier incidence (see Section 9.3), identify possible needs for additional training, confirm that listeners are judging each VUT independently of those previously evaluated, and verify that their ratings follow the same general trend as the panel average over time. Confidential periodic feedback reports summarizing participation history, outlier incidence, and rating trends are shared with listeners as part of this process. We therefore employ a practical approach in which initial training is kept deliberately limited, broad employee participation is encouraged, and the greater part of listener development occurs through repeated use of the method in routine benchmarking, supported by ongoing review of listener performance.

5 Evaluation Criteria

The benchmarking program relies on evaluation criteria that provide a practical description of the main perceptual dimensions of automotive sound quality. Because the program spans various vehicle classes and includes base-grade, non-branded premium, and premium branded audio offerings, the evaluations repeatedly encounter substantial differences in system architecture, tuning, vehicle acoustics, and system integration. The fact that these recurring differences can be meaningfully described within a single evaluation framework supports the continued practical utility of the attribute structure.

5.1 Spectral

The spectral attribute is a broad quality term for perceived tonal balance and the quality of bass, midrange, and treble playback. In automotive systems, spectral differences commonly arise from loudspeaker number, type, bandwidth, and placement, as well as from crossover design, equalization, and the acoustic interaction between the loudspeakers and the vehicle cabin.

Descriptors such as *balanced*, *full*, *thin*, *bright*, and *harsh* relate to aspects of the spectral attribute.

In training, listeners are instructed to focus on spectral balance and the quality of bass/mid/treble when judging spectral quality.

5.2 Spatial

The spatial attribute is a broad quality term for perceived sound stage position, width, extent, stability, spaciousness, and envelopment. In automotive systems, spatial differences commonly arise from loudspeaker layout, channel count, center-channel implementation, and spatial or immersive processing. Particular attention is often paid to the center image because perceptually important elements, such as lead vocals, are typically intended to appear there. The preferred location of that center image has long been a matter of debate in automotive audio [24].

Descriptors such as *centered*, *focused*, or *stable image*, *narrow* or *wide stage*, and *closed*, *spacious*, or *enveloping* relate to aspects of the spatial attribute.

In training, listeners are instructed to focus on how the loudspeakers work together to create the illusion of a sound stage, a stable center image, and a sense of spaciousness when judging spatial quality.

5.3 Dynamics

The dynamics attribute is a broad quality term for the system's ability to reproduce program material with convincing impact, contrast, and level-dependent playback quality. In automotive systems, dynamic differences commonly arise from amplifier power, loudspeaker displacement capability, compression or limiting strategy, subwoofer provision, loudness correction, and general system headroom.

Descriptors such as *powerful*, *compressed*, or *distorted* relate to aspects of the dynamics attribute, as do descriptions such as *bass gets distorted when loud* or *exaggerated treble when quiet*.

In training, listeners are instructed to focus on how the playback changes as volume is raised or lowered from normal listening levels, with particular attention to behavior at higher volume steps.

5.4 Integrity

The integrity attribute is a broad quality term for the extent to which reproduced sound remains free from audible artefacts introduced by the sound system integration in the vehicle, or by the vehicle itself. Integrity is generally judged in terms of the presence or absence of negative artefacts rather than positive qualities. In automotive systems, integrity problems may arise from trim excitation, buzzing, rattling, resonances, or other issues related to component design, packaging, or vehicle integration. It is also the appropriate place to note any electrical or acoustic noises.

Descriptors such as *maintains integrity but does not get loud*, *loses integrity at higher volumes*, *instrument panel rattle*, or *buzzing in the driver door* relate to aspects of the integrity attribute.

In training, listeners are instructed to focus on the extent to which the car adds unwanted sound to the intended sound of the audio system. The best sound systems are able to reproduce high playback levels without causing BSR issues.

5.5 Calculated Overall Rating and Listener Summary Rating

Two summary metrics are used to characterize overall system performance for each listener: Calculated (Cal.) Overall and the listener's Overall rating.

Cal. Overall is not an additional listener rating, but a summary score used to represent performance across Spectral, Spatial, and Dynamics for each evaluated track and each individual listener. It is calculated as the unweighted average of those three ratings and is hidden from listeners during the evaluation.

This structure reflects the different roles of the attributes. Spectral, Spatial, and Dynamics describe attributes of the sound system and its reproduced sound, so their average provides a compact summary of overall sound system performance. Integrity, by contrast, reflects the extent to which the vehicle or its integration introduces unwanted artefacts into that reproduced sound. For this reason, Integrity is treated separately from Cal. Overall rather than incorporated into it.

Listeners provide an Overall summary rating of system performance across all four attributes, including Integrity. This single rating allows listeners to capture their overall impression of the combined listening experience, supports debrief discussions with benchmarking administrators, and provides a useful cross-check against Cal. Overall. As a check, the relationship between Cal. Overall and the Overall summary rating is periodically monitored. Across all current vehicles (N=490), a strong correlation was observed (Pearson r = 0.89), indicating that Cal. Overall ratings align closely with listener summary ratings. Note that the Statistical Analyses described in Section 9 do not incorporate the listeners' summary ratings.

6 Subjective Measurements

"Subjective measurements" [25] are the central outcome of the evaluation process. In the present method, they provide the primary means by which the otherwise diffuse psychological construct of "sound quality" is represented as a structured and repeatable set of numerical and descriptive observations, supported by controlled procedures, trained listeners, and consistent test conditions.

6.1 Sources and Nominal Volume

Short, edited music loops of twenty to thirty seconds have been used in published automotive subjective work [9,24]. The current procedure uses seven program items, each an excerpt of approximately thirty to forty seconds from a commercially available music track. The selection is updated occasionally to maintain listener engagement, typically by replacing approximately one evaluation track per year.

To create a stable evaluation stimulus, each excerpt is edited and repeated to form a playback block of approximately five minutes. This gives listeners enough time to evaluate the content without excessive distraction from short, repeated loops, obvious loop-transition discontinuities, or the need to manually restart playback.

The tracks are arranged with a noise file placed immediately before each music item. Each noise file is adjusted so that its loudness, measured in LKFS/ LUFS [26], is aligned to that of the evaluation track it precedes. In practice, this noise signal acts as a stable proxy for the loudness of the associated music excerpt. By measuring sound pressure level at the listening position in both the reference room and the VUT, playback level can be matched by adjusting system volume.

Following internal development work, a nominal playback range of 70 to 76 dBA has been adopted. In the reference room, the volume settings required to achieve this level for each track pair are documented for the listeners. In the VUT, an SPL meter is provided so that listeners can, if necessary, check and re-establish the intended playback level during the evaluation.

6.2 Scaling

The scales used in the process contain 101 points and run from 0.0 to 10.0. As in MUSHRA-style listening tests [27], the scales are divided into labeled quality regions. The present implementation extends this approach with two additional categories at the extremes and greater resolution towards the center of the scale, where most production vehicles tend to be rated:

- 7.5 to 10.0: *Amazing*
- 6.5 to 7.5: *Excellent*
- 5.5 to 6.5: *Good*
- 4.5 to 5.5: *Average*
- 3.5 to 4.5: *Poor*
- 2.5 to 3.5: *Bad*
- 0.0 to 2.5: *Terrible*

During training, listeners are told that it will generally be reasonable to regard playback in the reference room as lying within the "*Amazing*" region of the scale. They are also encouraged to leave space at the top of the scale in case a future listening experience is judged to be superior. To reinforce this calibration, listeners are asked once or twice per year to provide a formal rating of the reference room itself. This is distinct from the routine reference-room listening that takes place at the beginning of each sound quality evaluation – see Section 6.3.

Any VUT judged to be on par with, approaching, or exceeding the reference room for a given attribute should likewise be rated within the "*Amazing*" region. In practice, this is rare for production vehicles and is more commonly approached by one-off aftermarket, prototype, or demonstration systems, where tuning can be optimized for the exact vehicle and component configuration under evaluation.

At the lower end of the scale, one or more low anchors are presented during training. Examples have included smartphone loudspeakers or dictation devices, which can demonstrate reproduction quality that, at least for some attributes, falls within the *Terrible* region. By contrast, production automotive systems in mature markets are not typically rated *Terrible* unless something is malfunctioning or seriously misconfigured.

The intermediate regions are interpreted progressively in terms of perceived quality. During training, examples are used to illustrate these distinctions. An *Average* system may be unremarkable rather than engaging. A *Good* system is enjoyable and may be entirely satisfactory for everyday ownership. An *Excellent* system is notably impressive. A *Poor* system usually has some aspect of reproduction that is irritating, distracting, or clearly below expectation, while a *Bad* system is sufficiently objectionable that a listener might prefer not to use it for music playback at all.

6.3 Referencing

Before entering the VUT, each listener individually spends approximately ten to fifteen minutes in the reference listening room listening to the evaluation tracks. This allows them to reacquaint themselves with the program material and refresh their auditory memory with high-quality sound reproduction immediately prior to assessing the vehicle.

During training, listeners are instructed to listen mainly at the nominal playback level when forming judgments of spectral and spatial performance. They are also encouraged to vary playback level in order to understand how the system behaves in terms of dynamics and integrity. With experience, this referencing period also serves a secondary practical function by allowing listeners to disengage from day-to-day work activity and reorient their attention toward careful listening.

6.4 VUT Evaluation

Following the reference-room session and transfer to the VUT, approximately 45 minutes of the one-hour evaluation session is devoted to evaluating the tracks in a randomized sequence. Evaluations are conducted primarily from the driver seat using the factory-default playback configuration, although additional assessments from alternative seating positions or playback modes may occasionally be performed when required.

For each track, listeners assign a score for each attribute using the defined scales and record comments explaining the basis for their judgments. These comments may be attached to a specific attribute or sub-attribute, or recorded in a general text field available throughout the session for broader observations.

As in the reference-room procedure, listeners are instructed to judge spectral and spatial performance primarily at the nominal playback level, while using higher and lower playback levels as needed to assess dynamics and integrity.

In addition to attribute ratings and written comments, spatial performance is documented using the panorama interface (see Section 3.3.3). Here, listeners indicate perceived stage presentation by placing colored circular or oval markers representing the left, center, and right regions of the sound stage, with marker size reflecting the perceived focus or diffuseness of the image.

Once all tracks have been evaluated, listeners are given the opportunity to complete any remaining overall comments and to provide an overall rating using the same attribute-style scale. The resulting dataset therefore combines structured numerical judgments with written observations and spatial annotations, providing both quantitative and qualitative accounts of system performance.

6.5 Comment Summarization and Spatial Visualization

While the statistical analysis of the attribute ratings is presented in Section 9, the text comments and panorama data are also examined and reported after each evaluation to support interpretation of the numerical results.

Historically, comment analysis required substantial manual effort, and certain forms of cross-linking between track-specific and overall observations were difficult or impractical to perform consistently. With the availability of large language models (LLMs), it is now possible to carry out a much more thorough analysis of the comment data, linking observations across listeners, specific tracks, and overall impressions in order to produce a more joined-up summary of each evaluation than was previously feasible. These methods can also be applied consistently across regions and response languages, providing a level of comparative capability that was not previously available.

The panorama data are also reviewed after the evaluation. At present, analysis of these spatial placements remains largely descriptive. Automated methods for interpreting the spatial pictures are under investigation, but the current approach is similar to a traditional heat map, allowing the stage and center-image placement across all listeners to be visualized together. This provides a practical way of describing the perceived spatial presentation of the system as experienced by the average listener. Examples of heat maps for different spatial conditions can be found in Section 10.2.

7 Objective Measurements

Objective measurements are collected for every vehicle evaluation to document the acoustic conditions under which the subjective ratings were obtained. They provide a repeatable acoustic description of system performance in the cabin, support comparison across vehicles, sites, and time, help identify physical causes of perceived differences, and assist verification of field quality issues. In this method, objective data accompany subjective evaluation and help characterize measurable aspects of system behavior, while the perceptual significance of any measured difference must still be established through listening.

The current procedure includes a core set of routine measurements captured at the driver seat, principally "full-system balance" responses, that is, in-cabin frequency-response measurements of the complete production audio system at different volume steps, as well as a maximum SPL measurement. These are in general agreement with the current AES Technical Committee on Automotive Audio (TC-AA) white paper *In-car Acoustic Measurements*, particularly in relation to basic system-level characterization and use of the recommended microphone array [28].

Additional methods are being explored for their value in explaining subjective scores, supporting prediction, and assisting diagnosis of field quality issues. Current areas of investigation include alternative approaches for evaluating total harmonic distortion (THD) and offline sine-sweep measurements more generally. Future work is also expected to include evaluation of both the impulsive distortion method described in the TC-AA white paper [28] and the enhanced perceptual rub-and-buzz method described by Temme in [29].

8 Metadata Collection

Alongside the subjective results and objective measurements, vehicle and audio-system metadata are collected for each VUT. These metadata provide a structured description of the evaluated vehicle, its installed audio system, and the audio-performance-relevant physical and interface features associated with it. This documentation is necessary for accurate reporting of benchmark outcomes and for meaningful comparison between vehicles. It also supports later database filtering, investigation of field sound-quality issues, and analysis of possible relationships between vehicle or system characteristics and measured or perceived performance. Examples of collected metadata include vehicle identity and classification, odometer reading, trim and body style, vehicle configuration, audio-system branding and trim level, loudspeaker configuration, and photographic documentation of the vehicle, interior, and relevant control interfaces.

9 Statistical Analysis and Reporting

Statistical analysis is a key part of the benchmarking workflow because it converts trained-listener responses into structured, robust, and decision-useful summaries for engineering and stakeholder reporting. In the present application, the analysis is intended primarily to support routine industrial benchmarking rather than to provide full psychometric validation of the method. The emphasis is therefore on structured aggregation, practical robustness, and clear reporting of vehicle audio-system performance within the maintained benchmark framework.

9.1 Data Capture and Integrity

Subjective data are collected using a dedicated test interface that constrains response format, value range, and session completion at the point of capture. As a result, missing, non-numeric, out-of-range, and incomplete-response records are largely prevented by design. Exported data are then inspected for any remaining duplicate or anomalous records before analysis.

9.2 Sample Size

Each in-vehicle audio system is evaluated by a minimum of ten trained listeners. Under the standard evaluation design, each listener rates seven tracks on four attributes, yielding at least 280 raw attribute ratings per evaluation. The resulting dataset has a repeated-measures structure, with multiple ratings contributed by the same listener within a vehicle evaluation. These ratings are then aggregated to produce the reported attribute summaries for the VUT. Because the observations are not statistically independent, the figure of 280 is reported here as the number of raw ratings collected rather than as a direct statement of effective sample size.

9.3 Outlier Identification

Because routine benchmarking is based on modest trained-listener panels and is intended to support timely engineering and stakeholder reporting, the analysis includes an evaluation-specific listener-screening step. The purpose of this step is to reduce the influence of highly atypical listener responses on the final reported panel summary for a given vehicle evaluation. It is used as a practical robustness measure within the reporting workflow, rather than as a claim that atypical responses are necessarily invalid. Our listener training program is structured to include several practice evaluations, specifically learning how to compare track performance in the room to the car. The ability to do this, along with becoming acclimated to the attribute scales and anchors takes time to develop. During this time, their ratings will be statistically examined to determine if they have had enough practice, including exposure to high and low performing systems. Until their listening skills are honed, new listeners are likely to be statistical outliers.

Even without outliers and new listeners gaining practice, we strive to retain a final sample of at least ten listeners. Outlier rates are routinely tracked. It is worth noting that no single listener has been consistently or routinely found to be an outlier. In fact, a listener may be an outlier on one evaluation but not others, regardless of the time lapses between evaluations. Importantly, there have been benchmarked evaluations without any statistical outliers.

Descriptive statistics, specifically the mean and standard deviation of Cal. Overall, are used to establish the range of acceptability for identifying statistical outliers in the VUT. This range includes the full panel of listeners and has a lower and upper end. The lower end is established as the average Cal. Overall rating, aggregated across all tracks and listeners, *minus* the standard deviation. The upper end of the range is the average Cal. Overall rating *plus* the standard deviation:

1. **Flagging criterion:** Listeners are flagged if their aggregated Cal. Overall mean rating falls outside the established acceptability range for the evaluation.
2. **Statistical confirmation:** Each flagged listener's Cal. Overall ratings are then tested individually using Student's two-sample t-test with unequal variances as a practical confirmation step within the evaluation-specific screening workflow. A significance threshold of $\alpha \leq 0.05$ is used to determine whether the flagged listener's ratings differ sufficiently from the remainder of the panel to justify exclusion from the final reported summary.
3. **Panel consensus:** A case study of 100 vehicle audio system evaluations across all benchmarking sites is shown below to illustrate the effect of outlier removal on panel consensus.

The corresponding histograms below (Figure 1 and Figure 2) show that the distribution of deviations is centered near zero and relatively narrow, with the majority of observations falling within approximately ±1 rating unit. The reduction in spread after outlier removal is consistent with the operational goal of improving panel consensus.

This procedure is specific to the evaluation under consideration and is intended to stabilize the reported panel result in a high-throughput industrial setting. It should not be interpreted as a general judgment of listener quality, nor as proof that agreement with the majority is inherently correct. Rather, it is an operational step used to limit undue influence of a small number of unusually divergent responses on the final benchmark summary.

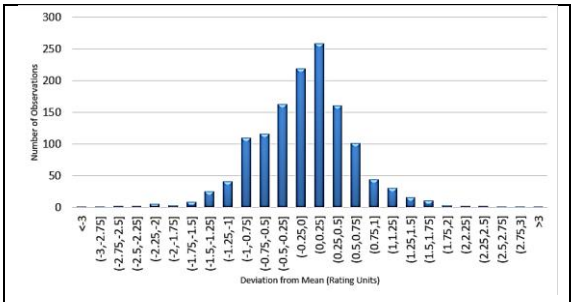


Figure 1: Distribution of Listener Deviation from the Panel Mean (Outliers Included)

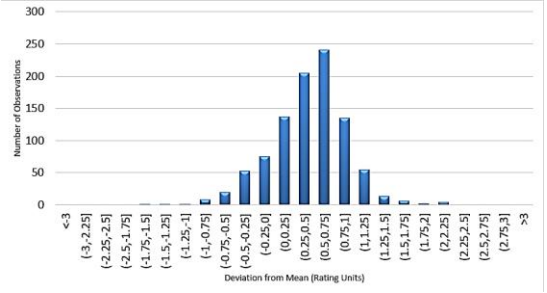


Figure 2: Distribution of Listener Deviation from the Panel Mean (Outliers Excluded)

9.4 Standardized Scoring

For operational reporting, the aggregated attribute result for each VUT is mapped to a proprietary, population-referenced benchmark score on a 0-100 scale. After removal of outliers, listener ratings for each attribute, namely Cal. Overall, Spectral, Spatial, Dynamic, and Integrity, are aggregated to obtain a vehicle-level score, X_{attr}. Each vehicle-level score is then standardized relative to a maintained reference population using a z-score transformation shown in Equation 1:

$$Z_{attr} = \frac{X_{attr} - \mu_{attr}}{\sigma_{attr}}$$

Equation 1: Z-Score Standardization Equation

where $\mu_{attr}$and σ_{attr}denote the mean and standard deviation, respectively, of the reference population for that attribute. This standardization step is important because it expresses each vehicle's performance relative to the population distribution for that attribute. For reporting, the standardized value is converted to a percentile using the cumulative distribution function (CDF) of the standard normal distribution, $\Phi(\cdot)$. The resulting score is a population-referenced index of comparative standing within the benchmark dataset rather than an absolute measure of perceptual magnitude. The benchmark population is refreshed periodically, typically every three to five years, to reflect current market offerings, technologies, and audio brands while retiring outdated models. Benchmarked scores are therefore intended principally for comparison within the current reference population.

9.5 Population-Referenced Reporting

Results are reported in two complementary ways:

1. Actual standardized scores (0-100)
2. Scores within each vehicle segment are divided into quartiles within the VUT's relevant market segment (currently using "IHS Markit" global vehicle segmentation classification [30]).

Taken together, these two views provide stakeholders with a clear and compelling picture of vehicle performance and segment-specific competitiveness.

10 Some Case Studies

The following anonymized examples are included to illustrate how the method behaves in practice and how the resulting data are used within the benchmarking workflow. They are not intended as a complete validation of repeatability, generalizability, or other psychometric properties of the method. Rather, they are presented as practical examples showing the kinds of agreement, discrimination, and diagnostic insight that can arise under routine operating conditions.

Different examples serve different purposes within the paper. Some are included principally to illustrate the closeness of repeated or comparable evaluations, while others show how the method can support more detailed comparative interpretation. Together, they are intended to illustrate practical utility and the kinds of agreement that may be observed in use, rather than to provide exhaustive proof of methodological performance. In particular, the repeated or closely comparable examples are not presented as a formal repeatability study, since such evaluations are limited in routine operation and complete control of all relevant conditions is not always possible.

Subjective case-study results are presented using the population-referenced reporting scale described earlier, as this is the form used operationally within the benchmarking workflow. The reported values indicate comparative standing within the maintained reporting framework rather than direct raw-score distances on the original listener scale.

10.1 Comparability Case Studies

Within our high-throughput industrialized benchmarking process, repeated evaluations of the same vehicle system have generally been avoided because it consumes time and effort that would otherwise be spent expanding market coverage. Over the years, however, a small number of evaluations have occurred, either by design or by chance, where, to the best of our knowledge, comparable versions of the same system were evaluated more than once at the same location or across different locations. A selection of these is presented here as anonymized case studies to examine the degree of comparability observed in selected pairs of evaluations within and between sites.

Across the case studies, performance designations, derived from segment-based population benchmarks, remained stable within a defined repeatability tolerance of ±5% (Integrity attribute is excluded from calculations). Minor objective differences were observed in some cases. Student's t-tests assuming unequal variances were conducted to compare listener ratings between the systems. No statistically significant differences were found in the overall performance of the VUTs, confirming comparable performance between them.

10.1.1 Case Study 1

A premium compact SUV with 12-speaker premium branded system was evaluated as a right-hand drive VUT at our site in India (1A) and a left-hand drive VUT at our site in Germany (1B).

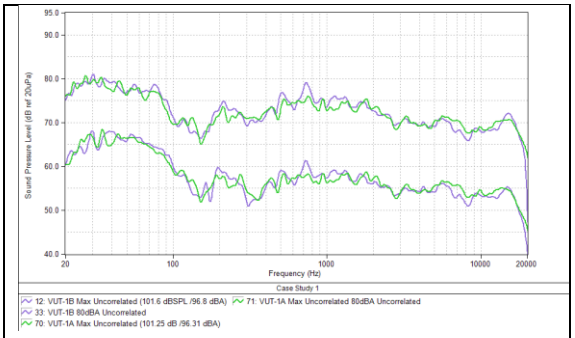


Figure 3: Case Study 1 Objective Measurements (showing a comparison of 80dBA and maximum level uncorrelated pink noise curves)

Figure 3 shows a very tight agreement between the measurements at nominal and maximum levels between the two different VUTs (the driver seat was used each time for the objective and subjective measurements).

VUT	Region	Overall	Spectral	Spatial	Dynamics	Integrity
1A		49	49	50	48	54
1B		48	48	53	44	59

Table 1: Case Study 1 Subjective Measurements

The reported attribute scores are also in tight agreement (see Table 1). The small differences in the attribute scores for left vs. right hand drive performance were tested and not found to be statistically significant.

10.1.2 Case Study 2

A non-premium small SUV with a 7-speaker base-grade system was evaluated in our USA (2A) and Germany (2B) benchmarking sites.

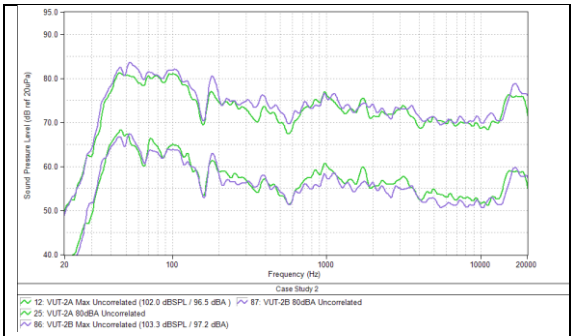


Figure 4: Case Study 2 Objective Measurements

Figure 4 shows agreement between the measurements at nominal levels between the two different VUTs. 2B having a slightly higher maximum level.

VUT	Region	Overall	Spectral	Spatial	Dynamics	Integrity
2A		30	26	39	29	46
2B		32	33	33	35	42

Table 2: Case Study 2 Subjective Measurements

The overall reported scores are also in close agreement (see Table 2). The performance of each attribute was tested for US vs. Germany. The 7-point difference in Spectral performance was the only attribute found to be statistically significant at the 95% confidence interval (CI).

10.1.3 Case Study 3

A non-premium compact SUV with a 9-speaker premium branded system was evaluated in the USA (3A) and in Germany (3B).

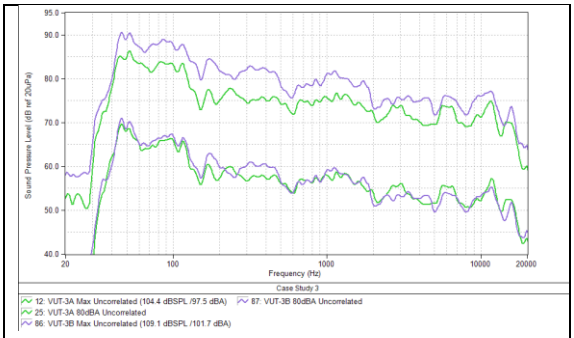


Figure 5: Case Study 3 Objective Measurements

Figure 5 shows broad agreement between the measurements at nominal level. At maximum level 3B is allowed to go louder than 3A indicating that the two systems are not fully comparable.

VUT	Region	Overall	Spectral	Spatial	Dynamics	Integrity
3A		52	42	60	54	46
3B		47	44	58	40	62

Table 3: Case Study 3 Subjective Measurements

The Overall, Spectral, and Spatial reported scores show relatively close agreement (see Table 3). Differences in the high-level performance are reflected mainly in the Dynamics and Integrity ratings, and in this case the difference in Dynamics also contributes to the difference in Overall performance.

Statistical testing of each attribute revealed significant differences in Dynamics and Integrity performance for the USA vs. Germany (95% confidence interval). This is expected because listeners provide these ratings at high listening levels. The 5-point difference in Overall performance was not found to be statistically significant due to very similar Spectral and Spatial performances in the two benchmarking sites.

10.1.4 Case Study 4

A non-premium midsize car with a 9-speaker premium branded system was evaluated twice in the USA (4A and 4B) within a five-month period. Note that this is the only case study reported in this paper that featured the exact same car brought back to test the reliability of our evaluation process.

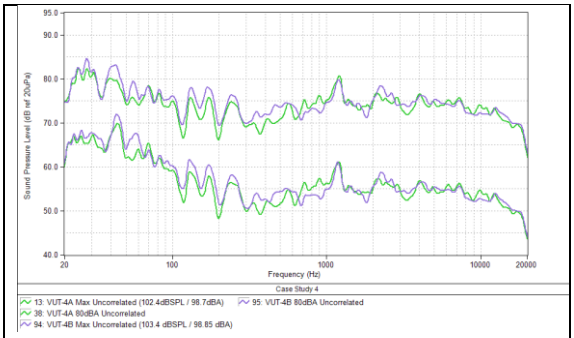


Figure 6: Case Study 4 Objective Measurements

Figure 6 shows very close agreement between the measurements of the two VUTs at nominal and maximum levels. However, the slight differences in the objective measurements are likely to reflect practical factors such as slight differences in microphone array positioning, seat position, or ambient conditions.

VUT	Region	Overall	Spectral	Spatial	Dynamics	Integrity
4A		60	57	60	60	53
4B		60	52	62	62	47

Table 4: Case Study 4 Subjective Measurements

The Overall reported scores are the same (see Table 4). The performance of each attribute was tested for statistically significant differences, and none were found (95% CI). Given the time lapse between the two evaluations of the same car (five months) and the

fact that the majority of listeners had evaluated only one of the two VUTs, the similarity of the benchmarked performance provides encouraging evidence of the repeatability of the subjective evaluation process.

10.1.5 Comparability Case Studies Summary

These case studies provide informative examples of the degree and character of comparability observed in selected pairs of evaluations within a single site and across different benchmarking locations. Two of the four pairs are very closely matched, while the other two cases still show close agreement given the practical realities of repeated in-vehicle testing.

10.2 Evaluation Sensitivity Case Study

The following investigation is included to illustrate the ability of the method to resolve meaningful differences within a group of low-performing systems. Six semi-trucks, evaluated without their semi-trailers and representing multiple manufacturers, models, and audio-system options, were included as part of the initial capture of a new market segment. The vehicles were assessed in no deliberate order over a four-month period, with panel composition varying between evaluations. Owing to their size, the vehicles could not be fully accommodated within the vehicle lab and were therefore positioned immediately outside the garage door for evaluation. For ease of reference, the vehicles are identified here as VUTs 7A to 7F in descending order of Overall score (see Table 5). In this case, the vehicles occupy a similar general region of reported performance, making the comparison a useful test of how sensitively the workflow can distinguish relative strengths and weaknesses within a constrained range. The investigation is presented as an example of practical discrimination in use, rather than as a formal validation of fine-grained rank ordering.

VUT	Overall	Spectral	Spatial	Dynamics	Integrity	# Speakers	Max SPL (dBSPL)	THD% @Max Level
7A	20	29	16	24	35	7	**101**	**6.5**
7B	4	4	8	6	25	10	**89**	**3.3**
7C	4	4	6	6	23	8	**99**	**4.4**
7D	2	4	8	2	21	6	**100**	**23.5**
7E	1	1	3	1	14	5	**94**	**5.8**
7F	0	0	0	0	11	4	**94**	**9.1**

Table 5: Evaluation Sensitivity Case Study

10.2.1 VUT 7A

VUT 7A clearly stood apart from the remainder of the group in the subjective results, and also achieved the highest maximum SPL, reaching 101 dB SPL while maintaining a relatively low 6.5% THD at maximum level. The system featured seven comparatively high-quality loudspeakers, including a capable subwoofer. As shown in Figure 7, the tuning more closely resembled that of a passenger-car audio system, with smooth treble and a *relatively balanced* bass boost.

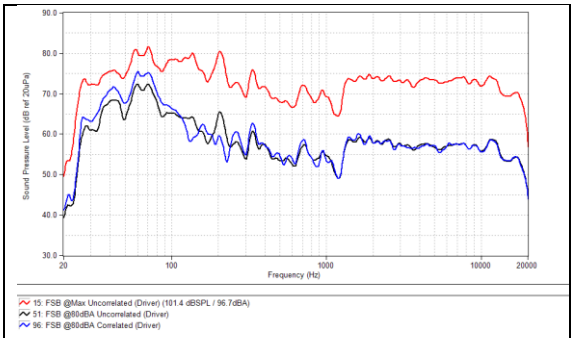


Figure 7: VUT 7A Objective Measurements showing correlated and uncorrelated noise measurements at 80dBA playback level (blue and black curves respectively), and uncorrelated noise at maximum playback level (red curve). 12th octave smoothing.

As shown in Figure 8, the stage is perceived relatively high, at or near eye level, and the center image is reasonably well focused. However, the center image is also biased to the left relative to the steering wheel. Overall, the spatial performance is not especially strong, although it retains some positive qualities.

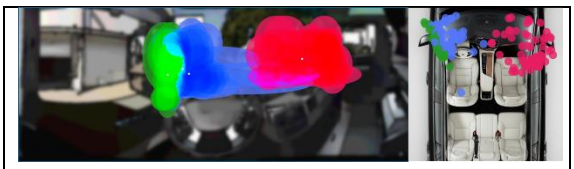


Figure 8: VUT 7A Spatial Heatmap

10.2.2 VUT 7B

Although VUT 7B featured the largest number of loudspeakers in the group, it produced the lowest maximum output, reaching only 89 dB SPL, while remaining relatively clean at maximum level with 3.3% THD. As shown in Figure 9, the frequency response shows an almost complete absence of low bass, but is otherwise less uneven across the spectrum than the other systems in this comparison.

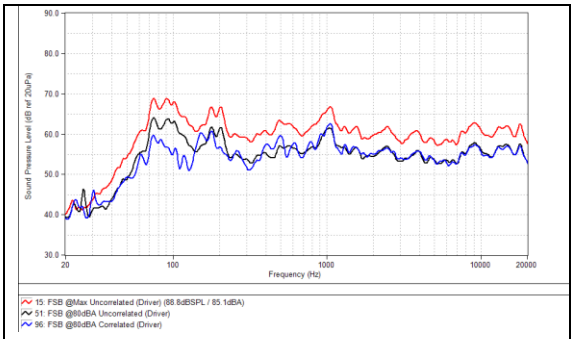


Figure 9: VUT 7B Objective Measurements

As shown in Figure 10, the staging is perceived as quite diffuse and wide, with a mostly left-biased center image.

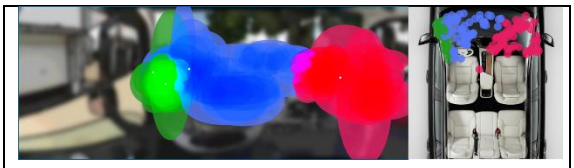


Figure 10: VUT 7B Spatial Heatmap

10.2.3 VUT 7C

VUT 7C was a relatively powerful system achieving 99 dB SPL and a relatively low 4.4% THD at maximum level. It featured eight loudspeakers, including a subwoofer. As shown in Figure 11, the frequency response is dominated by a pronounced bass peak, likely corresponding to a cabin resonance, and is otherwise relatively uneven.

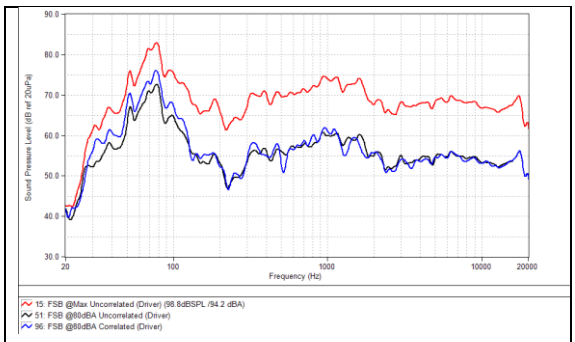


Figure 11: VUT 7C Objective Measurements

As shown in Figure 12, the staging is also perceived as quite diffuse and wide, with a mostly left-biased center image.

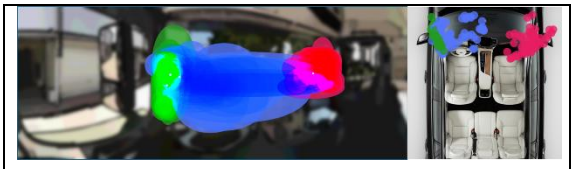


Figure 12: VUT 7C Spatial Heatmap

10.2.4 VUT 7D

VUT 7D is capable of high maximum output, reaching a level close to that of VUT 7A. However, this is accompanied by 23.5% THD at maximum level, which is extremely high and would be considered unacceptable in normal playback. As shown in Figure 13, the system also shows little evidence of controlled tuning, with substantial unevenness across the frequency response.

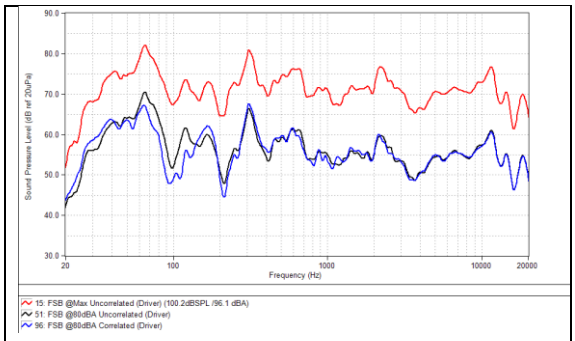


Figure 13: VUT 7D Objective Measurements

As shown in Figure 14, the staging is also perceived as quite diffuse and wide, with a mostly left-biased center image.

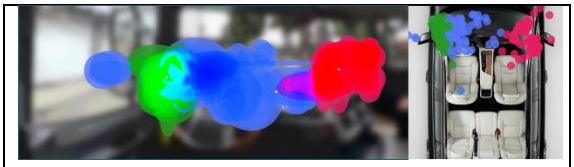


Figure 14: VUT 7D Spatial Heatmap

10.2.5 VUT 7E

As shown in Figure 15, VUT 7E is a weak system with effectively no bass reproduction beyond the boomy peak likely associated with a cabin resonance. Although it is not as quiet as VUT 7C, playback remains relatively distorted, with 5.8% THD at maximum level.

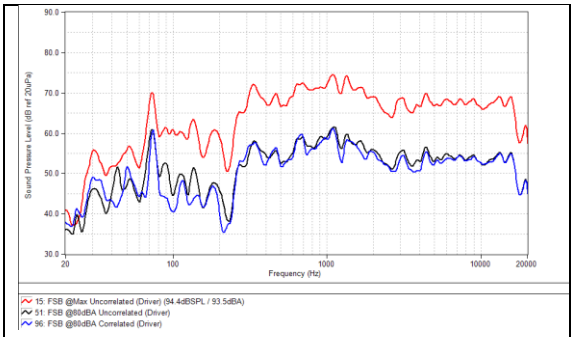


Figure 15: VUT 7E Objective Measurements

As shown in Figure 16, the staging is also perceived as quite diffuse and wide, with a mostly left-biased center.

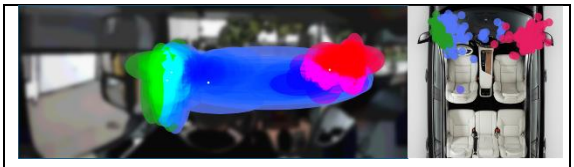


Figure 16: VUT 7E Spatial Heatmap

10.2.6 VUT 7F

As shown in Figure 17, VUT 7F is a weak system with little bass reproduction. Although its maximum SPL is comparable to that of VUT 7E, playback is substantially more distorted, with 9.1% THD at maximum level.

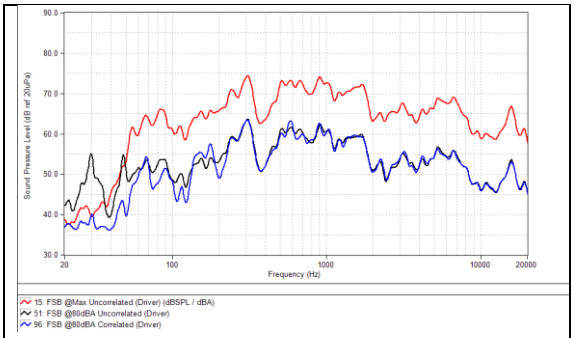


Figure 17: VUT 7F Objective Measurements

As shown in Figure 18, the staging is perceived as very low and not well defined.

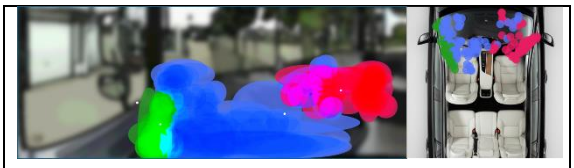


Figure 18: VUT 7F Spatial Heatmap

10.2.7 Sensitivity Case Study Summary

Taken together, these six semi-truck evaluations show that the method can resolve meaningful differences even within a uniformly low-performing group. The observed pattern in the reported results was broadly consistent with the accompanying objective measurements, system architectures, and subjective character of the individual systems. This investigation is therefore presented as a practical illustration of evaluation sensitivity in use, rather than as a formal validation of complete rank ordering across the full set.

11 Discussion

The method described here is intended as a practical industrial workflow for *in situ* vehicle-audio benchmarking, and the claims made in this paper should be interpreted within that scope. Routine application prioritizes evaluation under realistic use conditions, throughput, and timely stakeholder reporting over the tighter experimental control that would be expected in a dedicated psychoacoustic study. Within that context, the material presented here documents the workflow and illustrates how it operates in practice.

The anonymized case studies are intended as examples of practical utility and comparability across selected pairs of evaluations, rather than as exhaustive validation of all methodological

properties. Across the cases presented, the workflow showed useful sensitivity and interpretable relationships among the reported subjective results, relevant system characteristics, and supporting objective measurements. Taken together, these results suggest that the method provides a credible and operational basis for benchmarking, engineering interpretation, and product-development support in routine use.

Future work will investigate the use of mixed model analysis as a complementary framework for further examining listener, track, site, and vehicle-related sources of variation in the collected data. The method is also expected to be refined through additional repeated evaluations, continued development of objective-subjective interpretation, and further work on predictive modeling of subjective outcomes (see [31]).

Acknowledgments

The authors acknowledge Kristina Busenitz, Mae Nutley, and Lori Duprey for their contributions to the early development and continuity of the automotive benchmarking activity. We also thank past and present leadership, and the managers who enabled participation in vehicle evaluations, for their sustained support and for recognizing the program's value in competitive analysis, product validation, and market understanding. We are grateful to the other members of the current benchmarking team, Papa Adu-Boaheng, Shriniwas Chivate, Ken Huang, Wing Liu, and Louis Yu, for their expertise, care, and sustained effort in supporting the continued operation and development of the program. Finally, we gratefully acknowledge the many hundreds of dedicated listeners, past and present, whose commitment to providing careful and candid evaluations has been fundamental to the benchmarking effort.

References

[1] G. Cerrato-Jay, J. Gabiniewicz and J. Gatt, "Automatic Detection of Buzz, Squeak," Presented at the 2001 Noise and Vibration Conference, Traverse City, MI, 2001.

[2] H. Grythe, A. Nicińska, A. Drabicki and G. Sousa Santos, "How idling and maneuvering affect air quality: Case study of school commutes," Transportation Research Part D: Transport and Environment, vol. 146, 2025.

[3] M. Schäfer, J. Holub, J. Reimes and T. Drábek, "Subjective Testing of Car Audio Systems With and Without Parallel Task," Presented at DAGA 2018, Munich, Germany, 2018.

[4] F. E. Toole, "Subjective Evaluation: Identifying and Controlling the Variables," Presented at the AES 8th International Conference, Washington, D.C., 1990.

[5] M. Ziemba, "Test Signals for the Objective and Subjective Evaluation of Automotive Audio Systems," Presented at the 110th AES Convention, Amsterdam, The Netherlands, 2001.

[6] M. Schäfer, "Sound Quality Assessment of Car Audio Systems," Presented at the AES 2022 International Conference on Automotive Audio, Dearborn, MI, 2022.

[7] N. Kaplanis, S. Bech, S. Tervo, J. Pätynen, T. Lokki, T. van Waterschoot and S. H. Jensen, "A Method for Perceptual Assessment of Automotive Audio Systems and Cabin Acoustics," Presented at the AES 60th International Conference, Leuven, Belgium, 2016.

[8] M. Hibino, F. Hayakawa and K. Shima, "Subjective Evaluation and Acoustic Measurement of Sound Quality for Car Radio Systems," SAE Technical Paper 850161, presented at the SAE International Congress and Exposition, Detroit, MI, 1985.

[9] S. E. Olive and T. Welti, "Validation of a Binaural Car Scanning System for Subjective Evaluation of Automotive Audio Systems," Presented at the AES 36th International Conference, Dearborn, MI, 2009.

[10] P. Hegarty, M. Møller, M. Olsen, M. Lydolf and J. A. Pedersen, "Room Gain in a Car," Presented at the AES 48th International Conference on Automotive Audio, Munich, Germany, 2012.

[11] F. Postel, P. Hegarty and S. Bech, "A Listening Test System for Automotive Audio: PART 5 – The Influence of Listening Environment on the Realism of Binaural Reproduction," Presented at the 130th AES Convention, London, UK, 2011.

[12] E. Stanhope, S. G. Oxnard and L. J. Hobden, "Tactile Signal Derivation Part 1: Evaluation of Bandwidth Extension Methods," Presented at the 154th AES Convention, Espoo, Finland, 2023.

[13] S. G. Oxnard, E. Stanhope and L. J. Hobden, "Tactile Signal Derivation Part 2: Extracting Haptic Information from Traditional Audio," Presented at the 154th AES Convention, Espoo, Finland, 2023.

[14] F. E. Toole and S. E. Olive, "Hearing Is Believing vs. Believing Is Hearing: Blind vs. Sighted Listening Tests, and Other Interesting Things," Presented at the 97th AES Convention, San Francisco, CA, 1994.

[15] R. Shively, "Subjective Evaluation of Reproduced Sound in Automotive Spaces," Presented at the AES 15th International Conference, Copenhagen, Denmark, 1998.

[16] S. Bech, "Methods for Subjective Evaluation of Spatial Characteristics of Sound," Presented at the AES 16th International Conference on Spatial Sound Reproduction, Rovaniemi, Finland, 1999.

[17] S. E. Olive, "A New Reference Listening Room for Consumer, Professional and Automotive Audio Research," Presented at the 126th AES Convention, Munich, Germany, 2009.

[18] S. Bech and N. Zacharov, Perceptual Audio Evaluation: Theory, Method and Application. Chichester, UK: Wiley, 2006.

[19] S. E. Olive, T. Welti and E. McMullin, "Listener Preferences for In-Room Loudspeaker and Headphone Target Responses," Presented at the 135th AES Convention, New York, NY, 2013.

[20] S. E. Olive, "A New Listener Training Software Application," Presented at the 110th AES Convention, Amsterdam, The Netherlands, 2001.

[21] S. E. Olive, "Harman's 'How to Listen' - A New Computer-based Listener Training Program," May 30, 2009. [Online]. Available: https://seanolive.blogspot.com/2009/05/harmans-how-to-listen-new-listener.html [Accessed: March 8, 2026].

[22] E. Geddes and H. Blind, "The Localized Sound Power Method," Journal of the Audio Engineering Society, vol. 34, no. 3, pp. 167-173, 1986.

[23] S. Bech, "Listening Tests on Loudspeakers: A Discussion of Experimental Procedures and Evaluation of the Response Data," Presented at the AES 8th International Conference, Washington, D.C., 1990.

[24] P. Dennis, "In-Vehicle Audio System Sound Quality Preference Study," Presented at the 139th AES Convention, New York, NY, 2015.

[25] F. E. Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," Journal of the Audio Engineering Society, vol. 33, no. 1/2, pp. 2-32, 1985.

[26] Radiocommunication Sector, "Recommendation ITU-R BS.1770-5: Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level," International Telecommunication Union, 2023.

[27] Radiocommunication Sector of ITU, "Recommendation ITU-R BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," ITU-R, 2015.

[28] Audio Engineering Society Technical Committee on Automotive Audio, "In-Vehicle Acoustical Measurements White Paper (Beta Version 1.0)," 2023. [Online]. Available: https://aes.org/community/technical-council/aes-technical-committee-automotive-audio [Accessed: March 8, 2026].

[29] S. Temme, "Enhanced Perceptual Rub & Buzz Measurement for Testing Automotive Loudspeakers," Presented at the AES 2022 International Conference on Automotive Audio, Dearborn, MI, 2022.

[30] S&P Global Mobility, IHS Markit Global Vehicle Segmentation Classification. Southfield, MI, USA: S&P Global Mobility.

[31] R. Vlasov, I. Shishalov, A. Filimonov and M. Johansson, "Towards a Virtual Listener Panel for Car Audio System Evaluation," Presented at the AES 2026 International Conference on Automotive Audio, Detroit, MI, 2026.