

Guidelines for Jury Evaluations of Automotive Sounds

Norm Otto and Scott Amman, Ford Motor Company, Dearborn, Michigan
Chris Eaton, Ericsson, Inc., Research Triangle Park, North Carolina
Scott Lake, General Motors Corporation, Milford, Michigan

This article provides a set of guidelines intended to be used as a reference for the practicing automotive sound quality (SQ) engineer with the potential for application to the field of general consumer product sound quality. Practicing automotive sound quality engineers are those individuals responsible for understanding and/or conducting the physical and perceptual measurement of automotive sound. This material draws upon the experience of the four authors and thus contains a number of rules-of-thumb which the authors have found worked well in their many automotive related sound quality projects over past years. When necessary, more detailed publications are referenced. The intent here is to provide a reference to assist in automotive sound quality work efforts and to solicit feedback from the general sound quality community as to the completeness of the material presented.

Why is there subjective testing and analysis in automotive sound quality investigations? One might ask why bother with the trouble of conducting subjective testing in the first place? In the authors' experience, conducting subjective jury evaluations of automotive sounds has led to a deeper understanding of those sounds and the way potential customers react to and sometimes appreciate automotive sounds. The following is an attempt to describe subjective testing and analysis as applied to sound quality and its relevance to gaining this deeper understanding. The remainder of this article draws upon the experience of the four authors and as a result, may be biased toward the techniques they have commonly used or have found to work well in their automotive sound quality studies. However, an attempt has been made to address other techniques commonly used by other researchers in the general field of product sound quality. Although not a comprehensive document, it is hoped that this article will provide a set of guidelines which addresses a majority of the issues and techniques used in the field of automotive and general product sound quality. It is hoped that this guide will act as a springboard: a launching point for your own individual investigation into subjective testing and analysis for automotive sound quality.

Definitions

It is appropriate to begin with a few fundamental definitions of terms used throughout this document:

Subjective. In Webster's Dictionary, subjective is defined by the following: . . . peculiar to a particular individual, . . . modified or affected by personal views, experience, or background, . . . arising from conditions within the brain or sense organs and not directly caused by external stimuli, etc. In certain situations, the word subjective conjures up negative connotations, as though subjective results are less valuable pieces of information than objective results. We do not hold that opinion in our treatment of this topic but consider subjective evaluation to be a vital, information-rich portion of automotive sound quality work.

Quality. Again Webster helps to clarify what we are investigat-

Based on paper number 1999-01-1822 © 1999 Society of Automotive Engineers, Inc. presented at the SAE 1999 Noise & Vibration Conference & Exposition, Traverse City, MI, May 1999. The work on this paper was performed while coauthor Chris Eaton was employed by HEAD Acoustics, Inc., Brighton, MI

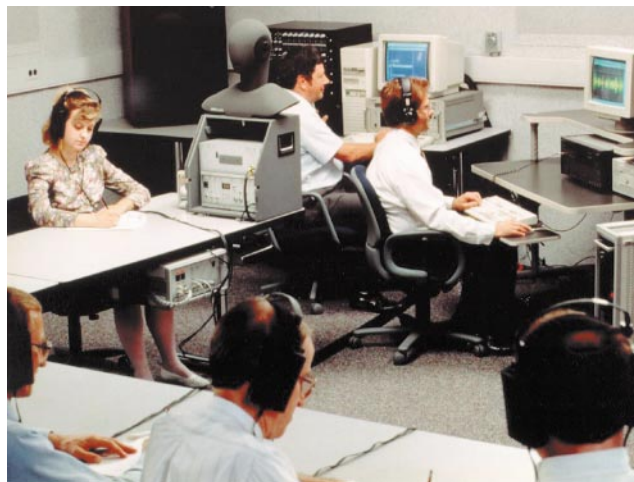


Figure 1. Typical jury testing room setup (Ford Motor Company).

ing. According to Webster quality is: . . . a distinguishing attribute, . . . the attribute of an elementary sensation that makes it fundamentally unlike any other sensation. Notice that 'goodness' or 'badness' does not enter into the definition.

Subjective Testing and Analysis. Subjective testing and analysis involves presentation of sounds to listeners, then requesting judgment of those sounds from the listeners and finally performing statistical analysis on the responses.

Jury Testing. Jury testing is simply subjective testing done with a group of persons, rather than one person at a time. Subjective testing can be done with a single person or many people at a time; both cases have their own set of benefits and caveats (see Figure 1).

The Task of Sound Quality. In automotive sound quality work, one tries to identify what aspects of a sound define its quality. It has been the experience of most persons involved in noise and vibration testing, that analysis of acoustic signals alone does not identify the quality (as defined by Webster) of those signals. Individuals will use words like 'buzzy,' 'cheap,' 'luxurious,' 'weak,' etc., to describe the defining attributes in sounds. Knowing how to design the correct attributes into a vehicle-sound directly impacts the appeal of the vehicle and ultimately impacts the profitability of a vehicle line. No instruments or analysis techniques have to date been able to quantify the descriptive terms mentioned above without the aid of subjective testing of some kind, hence, the need for subjective testing and analysis.

The remainder of this guide will take you through most of the salient issues involved in subjective testing for automotive sound quality work. This article is not intended to cover psychoacoustic testing but rather to provide guidance for the practicing sound quality engineer. Specific topics to be covered include:

- Listening Environment
- Subjects
- Sample (sound) Preparation
- Test Preparation and Delivery
- Jury Evaluation Methods

- Analysis Methods
- Subjective to Objective Correlation.

Before any type of jury evaluation can be conducted an adequate listening space is required. This is the topic of the first section.

Listening Environment

Room Acoustics. If the sounds are presented over loudspeakers in any room other than an anechoic chamber, the frequency characteristics of the room will be superimposed on the frequency characteristics of the loudspeaker and the sounds. Room resonances will then have an effect on the perceived sound. Additionally, if a jury evaluation is being conducted in which the subjects are physically located in different positions within the room, the room acoustics will affect the sound differently for each subject. No longer will all the subjects experience the exact same sound, thus, introducing a bias into the results. If it is necessary to conduct listening evaluations using loudspeakers, adherence to Sections 4.1 (Room size and shape), 4.2.1 (Reflections and reverberation), and 4.2.2 (Room modes) of the AES standard¹ is recommended. It is recommended that loudspeaker arrangement conform to AES20-1996 Section 5.2 (Loudspeaker locations) and location of listeners conform to AES20-1996 Section 5.4 (Listening locations).

Ambient Noise. Control of ambient noise is essential for the proper administration of a subjective listening evaluation. Noise sources within the listening room can be due to: computer fans, fluorescent lighting, HVAC, etc. The influence of these sources can be minimized through the remote location or containment of computers in acoustic enclosures, incandescent lighting and HVAC baffles/sound treatment. High transmission loss into the room is desirable to minimize the influences of outside noise.

Single value dB or dBA levels are generally inadequate in describing the ambient noise levels of indoor environments. ANSI S3.1 defines octave and one-third octave band noise levels for audiometric test rooms.² However, bands below 125 Hz are undefined and everyday sounds with energy below 125 Hz are commonly encountered. As a result, it is recommended that ambient noise levels should conform to NCB (noise criteria) 20 or better³ which specifies allowable levels in the 16 to 8000 Hz octave bands.

During jury evaluations, the station at which the subject is located should be free from influences from the other subjects. Many times partitions are placed between subjects to minimize interaction between subjects. When listening to low level sounds, subjects with colds or respiratory ailments can make it difficult for not only themselves but also adjacent subjects to hear the stimuli.

Decor. The listening room should be a comfortable and inviting environment for the subject. The room should look natural as opposed to “high tech.” The more clinical the room looks, the more apprehension and anxiety the subjects will experience. Neutral colors should be used for the walls and furniture. Comfortable chairs and headphones (if used) are essential to reducing distractions and keeping the subject focused on the task at hand. Moderate lighting should be used. Lighting which is too dim may reduce a subject’s attention to the desired task, especially, during lengthy or monotonous listening evaluations.

Air Circulation, Temperature and Humidity. The listening area should be air conditioned at 72° to 75° F and 45 to 55% relative humidity. Air circulation and filtration should be adequate to prevent distractions due to lingering odors. Construction materials used in the facility should be nonodorous.

Subjects

In this article, the term ‘subject’ is used to refer to any person who takes part in the evaluation of sounds in a listening study. This section discusses the selection and training of these subjects.

Subject Selection. Some of the factors that should be con-

sidered when selecting subjects include subject type, the number of subjects required and how these subjects are obtained.

Subject Type. Subject type is defined based on listening experience, product experience and demographics.

Listening Experience: As a general rule, it is desired that the listening experience level of subjects be appropriate to the task at hand as well as representative of the target customer. An experienced listener may be more capable of judging certain sound attributes than an inexperienced subject. An example is the evaluation of loudspeaker timbre for high end audio systems.⁴ For this task, audiophiles familiar with the concept of timbre are the subjects of choice. An inexperienced listener would no doubt have difficulty in discerning the nuances important to the expert. However, most sound quality evaluations do not require such a high level of expertise. Most automotive sound quality work falls into this category and, generally, subjects are not required to have previous listening experience. In fact, in these cases, using only experts may not be a desirable thing. Experts often pick up little things that are not particularly important to the customer. Generally, screening subjects for hearing loss is not done. To be done properly, hearing tests require skills and equipment usually not at one’s disposal. In addition, such testing may violate subject privacy. Presented in the Test Preparation and Delivery section are methods of detecting poor subject performance and these methods will help to identify any hearing related performance issues.

Product Experience: Listener’s judgments of sounds are always influenced by their expectations. These expectations are, in turn, affected by the experience they have with the product. Thus, it is important that one is conscious of these expectations when selecting subjects. For example, one would not use luxury car owners to evaluate the engine noise of sporty cars. The product experience of the subjects must be matched to the task at hand. Company employees generally have exposure to all product segments and are often immune to this effect. Because of this, company employees may be used as subjects for the majority of listening evaluations. One should use actual customers as subjects when segment specific information is required.

Demographics: In the listening studies the authors have done over the years, a huge dependence of the results on subject demographics has not been observed. Nevertheless, the subject population should contain a demographic mix (age, gender, economic status) that is representative of the customer base for the product. Generally, customers from only one vehicle segment are used to insure proper demographics. Note that a representative mix does not always mean an equal demographic mix. For example, in the luxury vehicle segment, the mean customer age is well over 50 and the owners are predominantly male. When company employees are used as subjects, it is more difficult to control demographics. Usually an attempt is made to use roughly equal numbers of males and females in evaluations.

Number of Subjects. This section discusses the number of subjects needed for a listening evaluation. This decision is greatly influenced by whether extensive subject training is required as well as by the difficulty of the evaluation task.

Simple Evaluation Tasks: Fairly simple evaluation methods (like those described in the Test Preparation and Delivery section) require little or no subject training. As a result, a number of subjects can take part in the study over a relatively short amount of time. Furthermore, if a facility exists which allows several subjects to listen to the recorded sounds simultaneously, a large number of people can take part in the evaluation. The question is, “How many subjects are required to obtain representative results?” This is an important point because it is often implicitly assumed that results obtained with N subjects would be unchanged if $2N$ or $10N$ subjects were used. The question is, “What is the value of N for which this assumption is approximately true?” If one knew the distribution of the subject responses, then the value of N could be calculated for a given confidence level. There are several limitations to mea-

asuring this distribution. One is simply the time involved in gathering such data. Another is that the response scales used do not always conform to the usual statistical assumptions like those of the central limit theorem. Thus, one must rely on experience for these estimates. In general, the more subjects the better, but time constraints always are a factor. Usually, 25 to 50 is an appropriate number of subjects for listening studies which use company employees as subjects. About 10% of these subjects will have their data removed because of poor performance. If customers are used as subjects, however, 75-100 participants are selected.⁵ Customers have greater variability in their responses than employees and tend to exhibit a higher proportion of poor performers. Finally, always schedule more than the required number of subjects to allow for no-shows (about 20% of the subjects).

Complex Evaluation Tasks: Difficult evaluation tasks generally require some training of the subjects prior to evaluation. This training serves to familiarize the subjects with both the sounds and the evaluation task. While methods requiring extensive training are outside the scope of this article, it is instructive to note that the training time often limits the number of subjects that can participate in the evaluation to a small number, certainly less than 10 and often less than 5. As a result, the inter-subject variability is quite high in these studies.

Recruiting Subjects. Company Employees: Recruiting company employees is generally fairly easy. An announcement asking for volunteers can be posted on the e-mail system and will reach a large number of people. As a rule, employees are generally not paid for their participation in listening clinics. However, company employees have been provided with refreshments in appreciation of their efforts. Most employees are glad to give their time to help improve the company products. When using employees, it is very important to keep each subject's results confidential. These evaluations should not become a contest to see who gets the highest grade.

Customers: Recruiting customers to participate in listening clinics is more involved than recruiting employees. Since the task often focuses on specific product segments, the owners of these products must first be identified. This information is not always readily available particularly if one wants to include owners of competitive products. This is why it may be necessary to work with market research companies when dealing with customers. Using owner registration lists, these companies will identify, recruit and schedule subjects for testing. In addition, they will also gather demographic information. To attract customers, the participants are paid for participation in the clinic. This payment can range from \$50-\$150 dollars depending on the time required and the type of customer. It takes a larger fee to interest luxury car owners than it does for compact car owners.

Subject Training. Training refers to the process of acclimating subjects to both the sounds and the evaluation task. Depending on the difficulty of the task, this training can range from very simple familiarization to extensive regimes.

Simple Evaluation Tasks. A distinction should be made between simple and complex evaluation tasks. Simple tasks often require the subject only to give an opinion about the sounds. These opinions may take the form of choosing which of two sounds the subject prefers or of rating certain sound attributes. Most people are accustomed to giving opinions, so little or no training is needed beyond the simple mechanics of the evaluation. Since it is recommended to use real life product derived sounds that almost everyone has heard before, subjects do not need to be trained to recognize the sounds. For these cases, the subjects are simply familiarized with the sounds and the evaluation process by having a practice block at the start of each session. All the sounds in the study should be included in this practice block so that subjects get an idea of the range covered in the evaluation. This is particularly important when sounds are presented and evaluated sequentially. The evaluation of any given sound can be adversely affected if subjects are not aware of the other sounds in the study.

This is particularly true if sounds are being rated on scales which are bounded. Upon hearing a very good sound, a subject might be tempted to use the top of the scale if they did not know that an even better sound existed in the evaluation samples.

Complex Evaluation Tasks. These tasks are those to which the subject has little or no familiarity and require training to bring their performance to an acceptable level. The more common psychoacoustic methods fall into this category. Methods such as magnitude estimation, pure tone matching techniques and detection tasks are some examples. The principle behind subject training is that performance will improve with increasing exposure to the task. Feedback on performance during training has been shown to increase the acclimation rate. Often this training is very extensive requiring 3 or more hours per day for a number of weeks depending on the difficulty of the task. Training is complete when subject performance reaches an asymptote. Bech^{6,7} gives a very good discussion on subject training.

Sample Preparation

Good Recording/Measurement Practice. Good recording practices are dictated when preparing sound samples to be used in any jury test of "real" product sounds. Adequately prepared samples do not ensure a successful investigation, yet poorly recorded or edited sound samples can ruin an otherwise valid test. Therefore, close attention must be paid to recording practices.

In general, a literal representation of product sounds is desirable so that jurors get the sensation of participating in the original auditory event – whether that event is riding in the passenger seat of an automobile or experiencing a fine orchestra in a renowned concert hall. The most popular way of achieving this authentic near-duplicate of the original event is through the binaural recording technique which employs an artificial head to record the sound onto some form of digital media to be audited later in a controlled environment usually via high quality headphones. This section will be addressed with an emphasis on artificial head recordings, although some of the same considerations can be applied to other recording techniques. Following is a guide for stimuli preparation for presenting real product sounds to jurors for listening tests.

Level Setting and Calibration. In general, digital recordings are considered the standard today and should be used if possible. Currently, digital audio tape recordings are the most popular media for SQ recordings. A high dynamic range (90+ dB) and inexpensive cost per megabyte for data storage are two of the most appealing properties for this media. The following guidelines should be followed for achieving authentic recordings.

Recording Practices. All sounds to be used in a particular listening test should be recorded using the same sensitivity and equalization settings on the measurement system, if at all possible. This reduces the likelihood of operator error being introduced with hardware changes during recording since recording settings are unchanged. Recordings made with the same transducer sensitivity may permit the sounds to be presented via calibrated headphones or loudspeakers at the correct playback volume so they do not need any additional amplitude compensation through software or hardware. Another benefit of this practice is that it ensures the same transducer/recorder self-noise is recorded each time. This is an important issue since different transducer sensitivities will have different background noise levels due to corresponding instrumentation self-noise. This unwanted noise, which varies among recordings made with different sensitivities, may affect the perception of the sound stimuli or be distracting especially when compared back-to-back in A-B fashion.

Recording Range. As when making measurements with other transducers, it is good practice to choose a recording range that yields the highest recorded signal level without measurement system overload so that the maximum number of bits in the

digitized signal are modulated. This ensures that the greatest differential in level exists between the sound under investigation and the noise floor of the measuring system. This is not a trivial task since this should apply to the loudest sound of interest, so you need a bit of intuition to set a recording level that is high enough to capture the loudest signal but not too high that it adds unnecessary noise to the recording. Setting the maximum without overload is easier to accomplish with the advent of improved digital recording technology of 20 bits and beyond. The additional 4 bits provides 24 dB more dynamic range making fine-tuning of recording amplitudes less critical.

Measurement Variation. Measurement (recording) variation not due to instrumentation changes should be controlled as closely as possible.

Recording Variations: Examples of factors that influence the perception and measurement results of a recording are room acoustics (or cabin acoustics as in automotive or airframe), direction and distance to the sound source and also background noise or extraneous noise infiltration (as mentioned in the previous section). Obvious exceptions to these guidelines would be a jury test used for architectural acoustics purposes where the environment is the focus of the investigation or in a test for speech intelligibility where various speech segments are part of the stimulus. In these previous examples, it is advantageous to measure and record different acoustic environments since this is key to the nature of the test. Differences in recording equipment and acoustic environment should be kept to a minimum within a set of sound stimuli, otherwise jurors may be sensitized or cued to unintentionally give a response based on these perceived changes instead of focusing on the important characteristics of the stimuli.

Sample Variations: When making recordings, it is important to “use” or “actuate” products in the same manner, being careful to not bias a test because components were used at different speeds or other test conditions. One problem often encountered is a noticeable pitch discrepancy between sounds recorded at the correct speed and a sound recorded slightly off-speed. Sometimes, however, products must be operated at different speeds to get the same performance such as power output or cooling effect, for example. Whatever the circumstances, it is important to be consistent. For example, if the listening test is for automotive door closing sounds, then the data collection should be consistent across all closing ‘events.’ Each door must be closed the same way with the same physical parameters.

Measurement Position: It is important to have procedures in place that require consistent transducer placement, whether artificial head or microphone, so that consistent measurements are made.

Sample (Sound) Selection. Listening test sample selection should be governed by the purpose of the investigation. For example, if the objective of the test is to find where the sound quality of a product lies in comparison to competitive products, then real product sounds from competitors as found in the marketplace should be used. The sounds, recorded using real products, should be as authentic as possible and reflect what a consumer would hear and experience under normal use conditions. Sometimes products are tested under extreme operating conditions, such as those used in a wide-open-throttle vehicle acceleration test for powertrain sound quality. By testing the extremes using this strategy, it may be easier to pinpoint differences among product sounds and find potential sound quality shortcomings and vulnerable areas that need to be targeted for improvement. If, however, the objective of the investigation is to determine the optimum sound quality for a product, then your collection of sound stimuli should exhibit variation across the relevant attributes (like drone, hum, hiss) for that family of sounds or across sensation continua (loudness, pitch or roughness). By studying the level of the attributes or sensation continua and the impact of their qualities on the whole sound, the optimum overall sound quality may be found. Consumer terms as ‘liking,’ ‘minimum annoyance,’ or ‘comfort,’

etc., can be posed to the jurors to study and pinpoint an optimum attribute mix.

Sample Editing and Preparation. Sample preparation is limited to discussion of audio stimuli and pre-stimuli preparation for listening tests.

Extraneous Noise: Since test accuracy is directly influenced by the accuracy of the sounds themselves, careful screening of the sound stimuli is important. The purpose of screening is to ensure that the sound is true to the original situation. Screening is accomplished by simply auditing the sounds as they are to be presented. The sound samples should be prepared so that they include no extraneous noise. If the sounds are to be presented in a quiet room via headphones, then screening should be performed under the same conditions. For example, where product sounds are concerned, extraneous sounds, like talking or environmental noise, that are not intended to heard by jurors should be eliminated from the recordings by editing or choosing a different data segment that is not contaminated. Recordings should be made in quiet environments, absent of noise to distract from, or interfere with, the product sound. Only the intended sounds should be audible. The stimuli should be free of unwanted noise that might interact with or distract from the test. Sounds should be scrutinized for unwanted noise before they are administered to a jury.

Equalizing Loudness: Sounds may be amplitude adjusted so that each sound in a group gives the same calculated loudness value. This is a useful technique when it is necessary to have listeners focus on other important quality aspects of the sounds other than loudness level. This can be very effective when using pair-comparison test strategies.

Sample Length: In general, for steady sounds, the sound stimuli should be between 3 and 5 sec long. For transient events, such as door closure sounds, the single event may need to be repeated per comparison if greater differentiation is sought between short duration sounds.

Binaural Recording Equalization: Equalization is achieved during recording and playback. Its purpose is twofold – to make recordings sound authentic and to allow measurement compatibility. It is important to be consistent with equalization. Improper or unmatched record-playback equalization pairs can affect the tone color of sound presented using the system. Use the same record equalization and playback equalization since they are a matched set. There is a paradox in the use of artificial head equalization in some measurement systems that use software to equalize headphone playback. Most modern artificial heads used for sound quality work have only an ear-canal entrance (or a cavum conchae only) and no ear canal that would otherwise normally extend to a microphone placed at the eardrum location. The ear canal is not needed since it does not add any directional cues to the sound signal. Since there is no ear canal, it is not appropriate to use these artificial heads to measure headphone characteristics to generate an equalization curve, since an erroneous acoustic impedance of the coupling of headphone to ear canal will result. This makes measured headphone equalization and any sound heard through this system inaccurate or different from the original.

Free Field Equalization (FF). Free field equalization is suitable only when the artificial head is directly in front of the test object (0 deg azimuth, 0 deg elevation) and in a reflection free environment. FF provides a 2-stage equalization that nulls the cavum and ear canal entrance resonance and normalizes the measurement to give a flat amplitude response for frontal sound incidence. FF playback equalization is necessary to flatten the headphone response and reverse the effect of the normalizing equalization for frontal sound incidence.

Diffuse Field Equalization (DF). Diffuse field equalization assumes that all incoming sound frequencies and directions are weighted equally. The inverse of the average artificial head response is applied. While not truly measurement microphone compatible because of the averaging effect, it is included as a feature across manufacturers of artificial heads.

Independent of Direction Equalization (ID). ID is a general-

purpose equalization that nulls the cavum and ear canal entrance resonance so that instrumental measurements such as frequency spectra or loudness measurements do not show these resonances. This gives measurement results consistent with what an omnidirectional microphone might yield while also providing realistic playback capability. In general, ID equalization should be used whenever the sound field is spatially distributed whether due to source location, number of sources or room/cabin acoustics issues. In most cases, sound fields in recording environments for product sound quality are neither diffuse nor free so ID equalization is the correct choice.

Other Recording Issues. Sampling Frequency: Typical sample frequencies are: 44.1 kHz, the compact disc standard, and 48 kHz the digital audio tape standard. There are other sample rates, such as 32 kHz, available on some DAT machines and computer sound cards. 44.1 and 48 kHz are the most popular sample rates since their frequency response spans the audible frequency range. A 44.1 kHz sample should be employed if CD audio files are desired in the future. Many companies have a standardized method for recording using only one sample rate (44.1 kHz, for example). This ensures that any computational results based on the audio files do not differ because of sample rate differences. Also, depending on the hardware, clicks may result on the audio output when sounds with different sample rates are played back-to-back. Sounds with differing sample rates should be thought of as incompatible since they have different time resolution, which will usually affect any data analysis performed on them. Using a re-sampling algorithm, sample rates can be changed to make data compatible for playback and analysis.

Other Inputs: Embedded tachometer signals are used to record a speed or rotational signal related to a product's sound. This signal is particularly useful for modifying (removing, reducing, amplifying or extracting) parts of the sound that are related to the speed. This feature helps sound quality personnel solve problems and also play what-if games by modifying existing product sounds to create new ones. The embedded tachometer is registered by modulating the least significant bit of the digital word and storing it in the sound signal. The tachometer signal should not be audible in the digital signal if the playback equipment is capable of ignoring the least significant bit.

Quantization: Quantization refers to the number of levels in A/D conversion. Usually in sound quality work 16 bit recordings (or 15 bit plus pulse signal) or better (20 bit) are used for jury work. This maintains an important level of accuracy necessary for transmitting these signals.

Emphasis: In a digital recorder, emphasis is provided by a pair of analog shelving filters, one before A/D conversion and one after D/A conversion to provide a 1:1 input versus output signal. The emphasis circuit provides a signal with a high frequency boost, as much as 9 dB at 20 kHz. The purpose of 'emphasis' is to improve the signal to noise ratio of a recording by yielding a 'hotter' signal with greater high frequency content. An 'emphasized' signal must be 'de-emphasized' for either analysis or listening tests via hardware or software. In general emphasis should be avoided in light of compatibility issues and improvements in digital recording technology that affords quieter measurement systems and higher dynamic range

Test Preparation and Delivery

Sound Presentation (Play) Order. Incorporating a method of controlling sound presentation order is important for reducing experimental error due to biases as much as possible. Guidelines follow for testing strategies mentioned.

Paired Comparison Tests. Forced Choice Task: For a paired comparison or preference test, usually $t(t-1)$ pairs are presented, where t is the number of sounds in the study. This is known as a two-sided test. One kind of "presentation order" effect is removed since each sample appears with each other sample twice, but in opposite order. For example, the stimuli pair blue-red would be repeated in reverse order as red-blue.

The two-sided test introduces other niceties such as the ability to check subject consistency or see if subjects' overall agreement improves during the test. The two-sided test is, of course, usually roughly twice the length of a one-sided test, which is its only disadvantage. To optimize the play order to further reduce presentation effects, pairs should be evenly spread out so that no two identical pairs are near each other and so that no two adjacent sounds in the play order are the same. Strategies for determining optimum presentation order can be found in Reference 8. Ideally, each juror would hear a different play order, but this is usually prohibited by available hardware and test setup time.

Scaling Task: For a paired comparison of similarity or difference test, usually t^2 pairs are prepared for presentation. Unlike the paired comparison of preference, the same-sound pairs are presented such as red-red, blue-blue. The same general guidelines can be followed as in the description for paired comparison of other test methods as well (response scales) outlined above, although the data scaling and interpretation are much different.

Presentation of Samples. *Pacing or Timing.* Self-paced Tests: Self-paced means that the user has control of the test and can play the sounds as many times as necessary. Using this methodology it is possible to deliver the same sounds to each juror but with a different play order. This benefit can be used to minimize the effects of play order on the test results. This is usually executed using a computer based jury system.

Paced Juries: Paced jury tests present one set of stimuli to several jurors at once, typically through headphones.

Sample Size. The number of samples included in a test are usually chosen based on the test length constraints and the number needed to reach some desired level of product variation. Another consideration is that as the number of samples increases (and they better represent the variation to be expected among that product sounds), the likelihood of building a more robust or accurate model relating to consumer preference goes up. Clearly, the most important first consideration is the test methodology to be used (Jury Evaluation Methods section). This can govern the number and range of stimuli to be presented.

Test length. The length of the test is important due to the potential for juror fatigue which, in turn, depends on the level, duration and annoyance of the stimuli. Also, the health of the juror must be kept in consideration since exposure to high noise levels can cause long term hearing damage. In addition, a test that is too long produces results that are less discriminating. In general, try to limit the maximum test length to 30-45 min.

Sound Reproduction Method. Loudspeakers: Sounds may be audited through loudspeakers. Presentation of the same stimuli to all jurors is difficult, however, because the speaker type, position and listening room will influence the sound. Loudspeaker playback is more appropriate for products recorded in a reflection free environment (free-field) whose source is not spatially distributed.

Headphones: Headphone auditioning can be an effective way to ensure that each juror hears the same stimuli under the same conditions. Headphones can be level calibrated and equalized so that their responses are equivalent. Headphone auditing also allows flexibility in jury facility setup. Since the sound each auditor hears is presented over headphones, it is not influenced by the room's acoustic properties or by listener positioning.

Low Frequency Phenomena with Headphones: Jurors may report that the sound that they hear via headphones is louder or has 'more bass' than the 'real' product sound they are accustomed to, although the sound is delivered at the correct calibrated playback level. This discrepancy is introduced when our mind tries to assimilate and process what is seen in conjunction with what is heard. Under normal hearing conditions there is no problem since our visual and auditory scenes agree. However, when listening to a recorded acoustic image while experiencing the visual cues of a listening room, a mismatch

exists. For best results and the most authentic experience, sounds should be played back over headphones in the original sound environment or a simulation of the environment or some sort of mockup. Most often this is impractical though and listening tests are usually limited to a jury room. Anything to improve the visual context in which the sound is to be experienced is a plus.

Headphones with Subwoofer: Headphone listening can be enhanced through the use of a subwoofer system. A subwoofer system can replace missing low frequency sound energy that normally impinges on the human body during exposure to sound events. The subwoofer system can improve the realism of sound playback by a giving a sense of the importance of the vibration behavior of the device under scrutiny and a better sense of the low frequency sound characteristics of the product.

Visual Stimuli. Context improvement possibilities exist to help the jurors get a sense of “being there” at the site of the original recording. By demonstrating the product in use in a video or even a still picture the expectation of the jurors is better controlled and focused on the product under test.

Data Collection Environment. *Forms processing.* Using software and a scanner, a test administrator can design a form to be used to collect juror responses. The “bubble-in” type survey is a very familiar interface and should require little or no training for the jurors to use it. The test administrator may elect for the forms data to be sent to a master database once form images are scanned into the computer. From the database, statistics can be applied and conclusions made about the data. Form entry, while being the most flexible data collection device, requires a form to be designed each time a new type of test is taken. Fortunately, templates can be created so that the design time can be greatly reduced.

Computer Display. A computer can be utilized to setup jury tests. This system uses a desktop computer and a professional sound card to deliver listening tests. The test may be conducted on an individual workstation where the test may be run separately for each panelist. An evaluation may also be conducted with a group where all panelists are presented the samples at the same time for evaluation.

Hand Held Devices. Hand held devices, controllers and PDAs, could be set up to collect data into a computer.

Subject Instructions. The instructions given to subjects are very important when trying to obtain good subjective data without inadvertently biasing the jury. Samples of instructions given for paired comparison, semantic differential, attribute intensity scaling (response scaling) and magnitude estimation tasks are provided in the appendix. Every evaluation is unique and may require significant alterations of the examples given. These examples are intended to provide a starting point for the evaluation organizer. The jury evaluation methods will now be discussed in the next section.

Jury Evaluation Methods

This section discusses the methods that are used to elicit opinions of sounds. The term jury evaluation is meant to be synonymous with other like descriptors such as listening test, and more generally subjective evaluation. These methods define both the presentation and evaluation format. In addition, they may also imply particular analysis methods (Analysis Methods section).

Definition of Scope. No attempt is made to discuss every possible psychometric method in this section. For testing consumer products, it is important that the subjective results be representative of customer opinion. Since most consumers are neither sound quality experts nor audiophiles, the scope will be limited to those methods appropriate to inexperienced, relatively untrained subjects (see **Subjects** section). This excludes many traditional psychoacoustic methods (like matching techniques⁹ and Levitt procedures¹⁰) from this discussion.

Methods. Several jury evaluation methods appropriate for inexperienced, untrained subjects are discussed in the sections

that follow. While, each method has its strengths and weaknesses, it is important to note that no one method works best for every application. It is very important to choose the method which best fits the application.

Rank Order. Rank ordering of sounds is one of the simplest subjective methods. Subjects are asked to order sounds from 1 to N (where N is the number of sounds) based on some evaluation criteria (preference, annoyance, magnitude, etc.). The sounds are presented sequentially. The subjects often have the option in this method of listening to a sound as many times as they want. However, since the complexity of the ordering task grows combinatorially with the number of sounds, the sample size is usually kept low (six or less). The major disadvantage of this method is that it gives no scaling information. While ranking will tell one that sound A is, for example, preferred to sound B, it does not reveal how much more preferred A is over B. Because of this lack of scaling information, rank order results are not useful for correlation with the objective properties of the sounds. Rank ordering is used only when one simply wants a quick idea of how the sounds compare, for example, when evaluating customer preference for alternative component designs.

Response (Rating) Scales. In general, the term response scale refers to any evaluation method in which subject's responses are recorded on a scale. However, for this discussion, the subject will be limited to numbered response scales, like the familiar 1-10 ratings. The discussion on descriptive scales is deferred until later. Numbered response scales (e.g., 1-10) are very familiar to most people. Subjects rate sounds by assigning a number on a scale. The sounds are presented sequentially with, generally, no option to replay. This method is quick and, on the surface, easy. Scaling information is directly provided. However, rating scales can be difficult for inexperienced, untrained subjects to use successfully. Some of the reasons for this are given below.

1. Numbered response scales do not allow the subjects to express their impressions in an easy and natural way. Inexperienced subjects have no idea what a '3' or a '5' or a '8' rating means in terms of their impressions. When, for example, people normally listen to an engine sound, they do not say that the engine sounds like a '3' or a '5' or a '8'. Instead, they would describe their engine as loud, rough, powerful, etc.

2. Different subjects use the scales differently. Some use only a small rating range, while others may use most of the scale. An example of this effect is given by Kousgaard,¹¹ in which four different loudspeakers were rated by five subjects on a 0-9 scale. The rating ranges for each subject are given below.

Subject 1.....	3.0-7.0
Subject 2.....	6.0-7.2
Subject 3.....	6.5-8.5
Subject 4.....	0.0-8.0
Subject 5.....	6.0-8.4

While subjects 2, 3 and 5 have reasonably similar ranges, subjects 1 and 4 use the scale very differently. Another source of inter-subject rating variability is that different subjects use different sound attributes as the basis for their judgments. Placing the attribute to be rated on the scale can eliminate this problem. In any case, with intrinsic rating differences like those shown above, statistics like the average may be misleading.

3. The extremes of the scales (e.g., '1' and '10') are generally not used. Because sounds are usually rated sequentially, subjects avoid extreme ratings for the current sound just in case an upcoming sound is better (or worse). An example of this effect is taken from a study of powertrain sound quality.¹² A paired comparison of similarity was conducted in which subjects rated similarity on a 1-10 scale, with 1 being the most dissimilar and 10 the most similar. The results showed that the rating extremities were never used even when a sound was compared to itself! The numbered scale was then replaced with an unnumbered line labeled “Very Dissimilar” and “Very Similar” at the ends of the line. With this method, subjects readily used the extremes.

4. There is absolutely no reason to believe that ratings on an arbitrary interval scale should correlate with the objective characteristics of the sounds. While trendwise agreement between subjective ratings and objective values can be achieved, correlation requires ratings that are proportional to the objective characteristics of the sounds. This is rarely achieved with rating scales.

In summary, rating scales are fraught with difficulties for untrained subjects and should be used with caution.

Paired Comparison Methods. Paired comparison (PC) methods are those in which sounds are presented in pairs and subjects asked to make relative judgments on the sounds in the pair. Of course, this same basic paradigm can be extended to more than two sounds but we will limit our discussion to the paired presentation. For paired sound presentations, a number of evaluation tasks have been developed. Three of these will be discussed in some detail, the first two are forced choice procedures where the subject must choose one sound in the pair while the last is a scaling task.

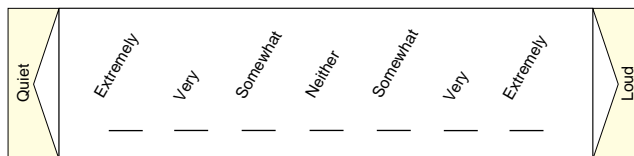
Detection Task: In this implementation of the paired comparison method, the subject must choose which of the sounds in the pair contains the signal to be detected. This method is often used for determining detection thresholds. For example, detection of a tone masked by broadband noise. One of the sounds in the pair is the masker alone and the other is the tone plus masker. The level of the tone is varied from pair to pair. Since the experimenter knows the ‘answer’ (which sound contains the tone), subject performance (psychometric function) is measured by the percentage of correct answers. The tone level at which the subject’s percentage correct equals 75% is defined as the threshold.¹⁰ Difficult detection tasks require extensive subject training.

Evaluative Tasks: In this method, subjects make relative judgments (pick A or B) on sounds presented to them in pairs based on some evaluative criterion. The criterion for these paired judgments can be virtually anything. Preference is often used as the basis for judgment. If all the samples are unpleasant, preference seems inappropriate and annoyance is often used. Attributes like loudness and roughness can be used but care must be taken to insure that subjects can readily distinguish the attribute among the different sounds being evaluated. This pair judgment process is repeated until all possible pairs have been evaluated (complete block design). Very often, a replicate experiment is conducted. A variation on this procedure is to include scaling information as part of the judgment.¹¹ An example might be both to pick which sound in the pair you prefer and to rate how much more you prefer that sound on a 1-10 scale. The scaling is generally unnecessary because PC models like Bradley-Terry (see the Analysis section) will give you the same information with much less work. Since judgments are relative, not absolute, subjects never have to worry about previous or future judgments. The PC method is very natural and easy for untrained subjects to use because it reflects something they do in everyday life, make comparisons. A number of successful studies using the paired comparison method have been conducted.¹³⁻¹⁵ One disadvantage of the paired comparison method is that the number of pairs can be quite large since they grow as the square of the number of sounds. More specifically, the number of pairs in a full paired comparison is $t(t-1)/2$, where t is the number of sounds. This means that the evaluation can be quite tedious if there is a large number of sounds. An incomplete block design can sometimes alleviate this problem. In this design, only some of the pairs are evaluated and these results used to infer how the other pairs would have been judged. This approach will work well only if one chooses the presented pairs appropriately. This is best done adaptively, where the next pair presented is based on the current results.

Similarity Tasks: Unlike detection and evaluation, similarity is not forced choice but a scaling task. Sounds are again presented in pairs, but instead of choosing one of the sounds, an estimate of their similarity is made. Similarity judgment is

rated on an unnumbered line, labeled only at the extremities as “very dissimilar” and “very similar.” By not numbering the scale, some of the problems associated with response scales are avoided. All possible pairs are evaluated in this manner. In addition, each sound is paired with itself to help judge how well subjects are performing. After the evaluation, a numbered grid is placed over the line and the markings converted to numbers, usually 1-10. Similarity scaling is useful for determining how well subjects discriminate among the sounds in the study. Combined with proper analysis techniques such as multidimensional scaling and cluster analysis, this method can determine the number of perceptual dimensions that underlie the judgments as well as giving clues to the important objective properties of the sounds.^{5,12}

Semantic Differential. While the paired comparison method focuses on one attribute of the sounds (preference, annoyance, similarity, etc.) the semantic differential (SD) technique allows evaluation of multiple sound attributes. Subjects evaluate sounds on a number of descriptive response scales, using bipolar adjective pairs. A bipolar pair is simply an adjective and its antonym. The adjectives generally consist of attributes (quiet/loud, smooth/rough) or impressions (cheap/expensive, powerful/weak) of the sound. These lie at opposite ends of a scale with several gradations. The gradations are labeled with appropriate adverbs that allow the subject to rate the magnitude of their impressions. Five, seven and nine point scales are common. A typical seven-point scale is shown below for the quiet/loud category.



Subjects choose whichever gradation best fits their impression of the sound. In choosing the semantic pairs, it is very important that they are appropriate to the application. This can be done in several ways. Newspapers and magazines are often good sources of semantic descriptors for consumer product sounds. Focus groups are another source. In general, technical engineering lingo is not good because customers are usually not familiar with these terms. Finally, pairs chosen for Americans are probably not ideal for Europeans or Japanese. Of course, the converse is also true. It is also important to avoid pairs that are closely associated with each other (associative norms). For example, quiet/loud and peaceful/noisy will almost always be correlated. Thus, there is no value in using the second pair since quiet/loud will serve the same purpose. By using both, you are ‘wasting’ a pair, which could be better used for other, unique descriptors. This is especially important when you consider that the practical limit of semantic pairs is 8-12.

Magnitude Estimation. Magnitude estimation is a method where subjects assign a number to some attribute of the sound (how loud or how pleasant it is). There is generally no limit to the range of numbers a subject may use. Magnitude estimation is basically a scaling task without a bounded scale. This method may offer some advantages over bounded response scale methods (numbered or semantic) in that the subject need never run out of scale. A major disadvantage of this method is that different subjects may give wildly different magnitude estimates. Thus, a key element of this technique is subject training. Magnitude estimation is more difficult, initially, for subjects to accomplish. They must be given a period of experimentation and practice in the technique before the actual data gathering exercise. This is accomplished by providing the subjects with a series of practice sounds and asking them to perform the magnitude estimation upon those sounds. There is no general rule to the amount of practice that is appropriate, although subject progress can be monitored. Subjects with prior experience in magnitude estimation often require less training and, thus, this technique is probably more appropriate for ‘ex-

pert' evaluators. Subject-to-subject variability can be addressed in a number of other ways as well. One is to present a reference sound with a specified magnitude (like 100) and have all other sounds rated relative to that reference (ratio estimation). A variant of this technique again uses a reference, but no value is given to that reference. In either case, presentation of a reference with each test stimuli effectively doubles the length of the evaluation.

Analysis Methods

Analysis methods for sensory evaluation data are numerous. It is the intent of this section to outline methods which have been used successfully in sound quality jury evaluation techniques described in the Jury Evaluation Methods section. Details of each analysis technique will not be given, but references will be cited in which the technique is used or demonstrated. The techniques discussed here, along with many others used for sensory data analysis, can be found in Meilgaard, et. al.¹⁶ Additionally, Malhotra¹⁷ provides a comprehensive examination of statistical analysis used in the evaluation of market research data. Both of these texts are easy to read and provide a good overview of techniques used for the interpretation and analysis of customer data.

Magnitude Estimation, Rating and Semantic Differential Scales. Magnitude estimation, rating and semantic differential scales fall into the category called interval scaling. An interval scale contains all the information of an ordinal scale but also allows the computation of differences between objects. The data generated by magnitude estimation, rating and semantic differential scales can be analyzed by a number of different methods described in the following sections. Before outlining the common analysis techniques used for these types of data, a few words must be said about the normalization of magnitude estimation responses.

Since magnitude estimation involves the subjects creating their own scales, a method of normalization of responses is needed before any statistical analysis can be performed. There are two basic methods for normalizing the results of a magnitude estimation exercise. The first involves creation of a geometric average for each stimulus across the range of subjective magnitude estimation scores for that stimulus. This is done by multiplying all individual scores for a particular stimulus together, then raising that result to the $1/n$ power, where n is the number of subjective magnitude estimation scores. The geometric averaging process ensures that each person's scale is given equal importance in the overall average for the particular stimuli.

A second technique commonly used involves a transformation of each subject's range of scores for the stimuli to a percentage scale, then the percentages from each subject are averaged together for a particular stimulus. An individual's maximum score given in the set of stimuli is set to 100%, their minimum score given in the set of stimuli is set to 0% and all values in between are scaled accordingly. This is done for each subject in the evaluation and then the percentages are averaged together for a particular stimulus. Bisping¹⁸ gives an accounting of this technique with a comparison to absolute scaling.

Distribution Analysis. Measures of Location: Measures of location are used to quantify the central tendency of a distribution of responses. The common measures of location are the mean, median and mode.

Mean. The mean is simply the value obtained by summing all responses and dividing by the number of responses. The mean can be somewhat deceptive in that it can be heavily influenced by outliers in a data set. A given data set should always be screened for outliers to determine if the mean is being unduly influenced by just one or two data values and may not be representative of the majority of the population.

Median. The median is the value above which half of the responses fall and below which half of the responses fall. The median is sometimes used over the mean because it is less sensitive to the influence of outliers in the response data.

Mode. Another measure of central tendency is the mode. The mode is the value which occurs the most in a sample distribution and is commonly applied to data which are categorical or which have been grouped into categories.

Measures of Variability: Measures of variability are used to describe the spread in a data distribution. Measures of central tendency mean very little without knowing something about the spread of a given set of data.

Range. The difference between the largest and smallest values in a set of responses. The range can be greatly impacted by outliers and should be used with caution.

Interquartile Range. The range in distribution covering the middle 50% of the responses. This measure is much more robust to outliers in a data set.

Variance and Standard Deviation. The variance is the mean squared deviation of all the values from the mean. The standard deviation is the square root of the variance. The variance and standard deviation are the most commonly used measures of distribution spread.

Measures of Shape: Measures of shape can be used to quantify the nature of distributions resulting from response data. Common shape measures include the skewness and kurtosis.

Skewness. Skewness is the characteristic of a distribution which describes the distribution's symmetry about the mean. It is the third central moment of the distribution data. For a normal distribution the skewness is zero. If the skewness is positive, the distribution is skewed to the right. If it is negative, the distribution is skewed left. Skewing of scaling data at the scale extremes is common and can be quantified using skewness.

Kurtosis. The kurtosis is a measure which quantifies the peakedness or flatness of the distribution. It is the fourth central moment of the distribution data. The kurtosis of a normal distribution is zero. If the distribution data contain a large number of outliers, the distribution tails will be thick and the kurtosis will be positive. However, if the data are closely centered about the mean, the distribution will be peaked and the kurtosis will be negative.

Test for Normality: Since many test procedures assume that the data distribution is normal, it is desirable to have tests which will indicate whether the response data actually belong to the family of normal distributions. The Kolmogorov-Smirnov (K-S) test statistic is commonly used to test how good a sample distribution fits a particular theoretical distribution. This statistic is defined as the maximum vertical absolute deviation of the sample cumulative distribution function (CDF) from the theoretical CDF. For a normal distribution, the actual population mean and variance must be known in order to generate the theoretical CDF. Of course, this information is not known. Lillefor's variation of the K-S test statistic substitutes the sample mean and variance for the actual mean and variance. The sample CDF can then be compared to the normal CDF generated from the sample mean and variance.

Graphical Techniques. Enormous amounts of quantitative information can be conveyed by using graphical techniques. Too often experimenters are quick to jump into quantitative statistics before exploring their data qualitatively. Graphical techniques can provide insight into the structure of the data that may not be evident in non-graphical methods. Numerous graphical techniques have been developed for areas such as distribution and data relationship analysis. Only a few of these techniques will be outlined here. For a comprehensive treatment, the text by Chambers¹⁹ is excellent.

Scatter Plots: Scatter plots can be used to explore relationships among data sets. Two subjective as well as one subjective and one objective sets of data may be plotted against one another to investigate relationships. Scatter plots can reveal information on data relationships that may not be apparent when using numerical approaches. Whether a relationship is linear, logarithmic, etc., can often be guessed at by an initial investigation of a scatter plot. Additionally, data outliers are also readily apparent.

Quantile-Quantile and Normal Probability Plots: Quantile-Quantile or Q-Q plots are commonly used to compare two data distributions to see how similar the two data distributions are. When quantiles for two empirical data sets are plotted against one another the points will form a straight line if the data distributions come from the same family. Sometimes the quantiles from an empirical data set are plotted against the quantiles from a theoretical distribution to investigate how closely the empirical data set matches that of the theoretical distribution. Normal probability plots are a special case of this situation in which the theoretical distribution used is that of a normal distribution. As a result, the normal probability plot can be used as a graphical means of checking for data normality.

Histograms: Another way to summarize aspects of a data distribution is to use a histogram. The histogram divides the data into equal intervals and plots the number of points in each interval as a bar where the bar height is representative of the number of data occurrences. Histograms are very easy for non-technical people to understand; however, the selection of interval size can be very important in determining the conclusions drawn from the histogram. Intervals that are too wide may hide some of the detail of the distribution (multi-modality, outliers, etc.) while intervals which are too narrow remove the simplicity of the display.

Others: There are many other graphical techniques commonly used in exploratory data analysis. Stem-and-leaf, box plots, bar charts, etc., are examples of graphical techniques commonly used to initially investigate response data.

Confidence Intervals. The confidence interval is the range of values of a population parameter that is assumed to contain the true parameter value. The population parameter of interest here is the true mean value of the response data (as opposed to the sample mean). Normally distributed data are assumed. Confidence levels are chosen and specify the probability that the true mean value will be covered by the confidence interval. Confidence levels of 0.90, 0.95 and 0.99 are commonly chosen. For example, if a confidence level of 0.95 is chosen, it can be said that we are 95% confident that the true response mean is contained within the confidence intervals calculated from the sample responses.

Large confidence intervals indicate large variability in the response data. If responses obtained for two different sounds have confidence intervals with significant overlap the true mean values of the responses may not be different. Observing the overlap of confidence intervals obtained from subject responses is a visually appealing and intuitive method of determining whether significant differences exist in the ratings. However, this process is very qualitative. Tests such as the *t*-test for pairwise comparisons or analysis of variance for group testing must be performed if rigorous significance testing is desired.

Testing and Comparing Sample Means (t-Test). The *t*-test is a parametric test used for making statements about the means of populations. It is based on the Student's *t* statistic. The *t* statistic assumes that the data are normally distributed and the mean is known or assumed known and the population variance is estimated from the sample. The *t* distribution is similar to the normal distribution and approaches the normal distribution as the number of data points or responses goes up. For sample sizes > 120 the two distributions are nearly identical.

One Sample t-Test: A one sample *t*-test is commonly used to test the validity of a statement made about the mean of a single sample (in our case the responses from a single sound). The one sample *t*-test can be used to determine if the mean value is significantly different from some given standard or threshold. For example, is the mean value of the responses significantly different from zero? Additionally, if the threshold of acceptance of a sound has previously been determined to be 7 on a 10 point scale and the sample mean is 7.9, what is the probability that the true mean is greater than 7?

Two Sample t-Test: Very often it is desirable to test whether the means of the responses given for two different sounds are

significantly different. The two-sample *t*-test tests the equality of means of the two sets of independent responses where independence implies that the responses generated for one sound have no effect on those generated for the second sound.

Comparing Equality of Means for k Samples (ANOVA). An extensive treatment of ANalysis Of VAriance (ANOVA) is beyond the scope of this article. Only a brief overview will be given here. Application of ANOVA applied to subjective listening evaluations can be found in Reference 20.

Data relationships from interval scaled data can be explored using ANOVA. ANOVA can be used to determine if the mean values given to various sounds are indeed significantly different. ANOVA analysis makes the assumption that the distributions of the scale values for each sound are normal. It also assumes that standard deviations are equal. These assumptions can be tested using some of the previously described techniques. Variations from these two assumptions can result in erroneous conclusions. The test statistic used to determine significance is the 'F' statistic. Typically the test significance used is $\alpha = 0.05$.

In so-called "designed experiments" ANOVA may also be used to determine if certain factors inherent in the design are influential in the mean values given to the sounds. For instance, the loudness of the sound could have an influence on the annoyance ratings given to sounds in a sample. The loudness could be determined either subjectively or objectively. In this case, only one factor is of interest and thus the analysis is referred to as a one-way ANOVA. When *m* factors are of interest the analysis is referred to as a *m*-way ANOVA. The factors influencing the mean values do not necessarily need to be intrinsic properties of the sound. Typically, the subjects themselves will impart their own variation in the form of scale use differences. In this case, the evaluators themselves are a factor which causes variation in the scale rating. Since subjects may use the scale differently, factors in the experiment may not prove to be significant due to large variances in the responses for individual sounds. If each evaluator's response is normalized by the mean value of all the sounds rated by that evaluator, the evaluator variance can be removed. This is commonly referred to as a 'within subject' ANOVA. The disadvantage of the 'within subject' ANOVA is that only relative values are obtained and the absolute scale values are lost. For example, one could say that in presenting a sound with a loudness of 20 sones and one with a loudness of 25 sones the influence on the annoyance ratings of all the evaluators (using a within subject ANOVA) was an increase of 3 (10 point scale); however, it could not be said that the absolute rating would go from a 5 to an 8 since for some evaluator's ratings it could go from a 2 to a 5, or 3 to 6, etc.

Fisher's LSD: Once significance has been determined among average values given for a sample of sounds, *post hoc* pairwise testing may be done to determine if significant differences occur for individual pairings of sounds. The most common of the pairwise comparison techniques is Fisher's Least Significant Difference (LSD) method. This is essentially the application of the two-sample *t* test to each pair of means. Other *post hoc* pairwise tests include Duncan's Multiple Range, Newman-Keuls and Tukey's HSD.

Linear Regression Analysis. Regression analysis is a technique used to assess the relationship between one dependent variable (DV) and one or several independent variables (IV). Assumptions made in any regression analysis include: normal distribution of the variables, linear relationships between the DV and IVs and equal variance among the IVs. Regression analysis, as it applies to the prediction of subjective responses using objective measures of the sound data, is discussed in the next section. A thorough treatment of regression techniques may be found in Reference 21. Regression analysis may also be used to explore relationships among subjective responses. For instance, the annoyance ratings from different aspects of a set of sounds may be used to predict an overall preference rating. The regression equation coefficients can then be used

to quantify the relative importance of each of the sound aspects to the overall preference of the sound. An example as applied to automotive windshield wiper systems is given in Reference 15.

It should be mentioned that although regression analysis can reveal relationships between variables, this does not imply that the relationships are causal. A measure used to quantify the regression model fit is called the *coefficient of determination* or R^2 . R^2 takes on values from 0 to 1 with 0 indicating no relationship and 1 showing perfect correlation. The tendency is to add IVs to the regression model in order to improve the fit. Caution should be used since the addition of more variables may not actually have a significant contribution to the prediction of the dependent variable. P -values for the individual independent variables are commonly used to assess how meaningful the contribution is – 1 minus the p -value is the probability that a variable is significant. Generally, variables with p -values greater than 0.20 may be considered as questionable additions to the regression model. In regression analysis a large ratio of cases (number of subjective responses) to independent variables is desired. A bare minimum of 5 times more cases than IVs should be used.

Residual analysis can be useful when using regression techniques. Residuals are the difference between an observed value of the response variable and the value predicted by the model. Residual plots can be used to validate the assumptions of normality, linearity and equal variance. Residual analysis can also identify outliers which can have a strong impact on the regression solution.

Factor Analysis. Factor analysis is a general term for a class of procedures used in data reduction. Only an overview of the basic principles of factor analysis will be given here. References 22 and 23 are excellent introductory texts on factor analysis. References 24-26 are case studies in which factor analysis is used with automotive powertrain sounds.

Factor analysis is a statistical technique applied to a single set of variables to discover which sets of variables form coherent subsets that are relatively independent of one another. In other words, factor analysis can be used to reduce a large number of variables into a smaller set of variables or factors. It can also be used to detect structure in the relationship between variables. Factor analysis is similar to multiple regression analysis in that each variable is expressed as a linear combination of underlying factors. However, multiple regression attempts to predict some dependent variable using multiple independent variables, factor analysis explores the association among variables with no distinction made between independent and dependent variables.

The general factor model is given as:

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + \dots + A_{im}F_m + V_iU_i \quad (1)$$

where,

- X_i = i th standardized variable
- A_{ij} = standardized multiple regression coefficient of variable i on common factor j
- F = common factor
- V_i = standardized regression coefficient of variable i on unique factor i
- U_i = the unique factor for variable i
- m = number of common factors

The unique factors are uncorrelated with each other and with the common factors. The common factors themselves can be expressed as linear combinations of the observed variables.

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \dots + W_{ik}X_k \quad (2)$$

where,

- F_i = estimate of i th factor
- W_{ij} = weight or factor score coefficient for factor i on standardized variable j

Figure 2 shows the general process used in factor analysis. For the case of applying factor analysis to scaling data derived from a jury evaluation, the first step is the acquisition of the

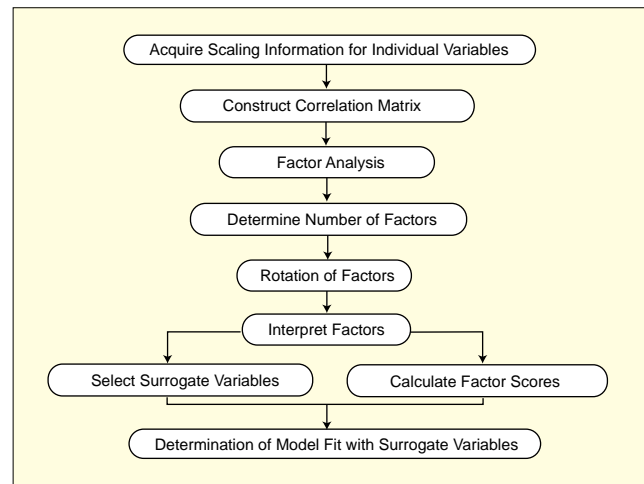


Figure 2. Process flowchart for factor analysis.

data. From that data a correlation matrix is derived. This matrix shows the simple correlations between all possible pairs of variables. Next, a method of factor analysis is chosen. The two main types of factor analysis are common factor analysis (CFA) and principal component analysis (PCA). In PCA you assume that all variability in a variable should be used in the analysis, while in CFA only the variability in a variable that is common to the other variables is considered. PCA is generally used as a method for variable reduction while CFA is usually preferred when the goal of the analysis is to detect structure. In most cases, these two methods yield similar results.

One of the goals of factor analysis is variable reduction. The question of how many factors should be extracted may be based on prior knowledge of the structure in question. Other guidelines used in practice include the Kaiser Criterion and the Scree Test. Both methods use the eigenvalues of the correlation matrix to determine the appropriate number of factors. The references contain detailed information on these and other techniques for determination of the number of factors to be extracted.

The factor analysis generates a factor pattern matrix which contains the coefficients used to express the standardized variables in terms of the factors. These coefficients are called factor loadings and represent the correlations between the factors and the variables. Initially, this matrix contains factors which are correlated to many variables which makes interpretation difficult. Factor pattern matrix rotation reduces the number of variables with which a factor is correlated and makes interpretation easier. Orthogonal rotations such as the *varimax*, *equimax* and *quartimax* procedures result in factors which are uncorrelated to one another. Oblique rotations may provide a clearer reduction in the variables but lead to factors which are correlated.

A factor is generally made up of variables which have high factor loadings for that particular factor. If the goal of factor analysis is to reduce the original set of variables to a smaller set of composite variables or factors to be used in subsequent analysis, the factor scores for each respondent can be calculated using Equation (1). Sometimes the researcher is interested in selecting a variable which represents the factor. This variable can then be used in subsequent listening evaluations to represent the factor of interest.

The final step is to determine the factor model fit to the actual data. The correlation between the variables can be reproduced from the estimated correlation between the variables and factors. The difference between the observed correlation (correlation matrix) and the reproduced correlations (estimated from the factor pattern matrix) can be used to measure model fit.

In general, there should be at least four or five times as many observations as variables. For example, if factor analysis is used to reduce the variables that result from a semantic differential

evaluation in which there were twelve semantic pairs, approximately 48-60 subject responses would be required.

Paired Comparisons. In one approach to paired comparison evaluations, the subjects are presented with two sounds and asked to select one based on some criterion. This is referred to as a forced choice method. As a result, the data obtained are ordinal. Paired comparison of similarity has also been used in sound quality evaluations. Similarity is a special application of the PC paradigm. Sounds are again presented in pairs, but instead of choosing one of the sounds, an estimate of their similarity is made, thus, providing interval scaled data.

Forced Choice Paired Comparisons. A thorough treatment of the methods discussed in this section can be found in David.⁸ A very readable synopsis of the paired comparison analysis including application to automotive wind noise listening evaluations is given by Otto.⁵ Selected applications of the methods discussed here can be found in References 15, 20, 27-29.

Tests of Subject Performance: Performance measures are used to judge subject suitability. These measures reveal how each subject, as well as the population as a whole, perform. Because paired comparison is a forced choice method, an answer will always be obtained. However, it is not known whether that answer is valid or just a guess. The performance measures help us find out this information. Repeatability and consistency are the two measures used most for PC evaluations.

Subject Repeatability. Repeatability is defined as simply the percentage of all comparisons which are judged the same the first and second time through the PC. Subjects who are doing poorly, for whatever reason, will usually not have high repeatability. Subject average repeatability should be 70% or greater. Subjects with individual repeatabilities of 60% or below usually have their data removed from further consideration. If the subject average repeatability falls much below 70%, then there is usually a problem with the evaluation itself. For example, having sounds which are not easily discriminable can often give rise to poor repeatability.

Subject Consistency. Consistency is a measure of how well the pair judgments map into higher order constructs. Of general use is the Kendall Consistency³⁰ which is based on triads. For example, a triad of sounds A, B, C is said to be consistent when the paired comparison shows that if $A > B$ and $B > C$ then $A > C$ and inconsistent if $C > A$ where $>$ can mean preferred, more annoying, louder, rougher, etc. The Kendall Consistency is defined as the percentage of consistent triads relative to the total number of triads. A subject average consistency value of 75% or higher is considered acceptable. Individuals who have poor repeatability often show poor consistency as well.

Scores: Once the PC data have been adjusted for subject performance, the next step is to analyze the data. Under the assumption of transitivity it is possible to convert paired comparison data to rank order data. One very straightforward way to obtain the rank order result is to use the scores. The score for a given sound is simply the total number of times that the sound is chosen, summed over all paired comparisons. Scores are a quick and easy way to look at the results, but are not appropriate for correlation with objective metrics. This is because scores are a measure of how each sound compares against the rest of the sound population. Scores tell you nothing about how one sound was judged against another. In order to determine if significant differences exist among the scores, a test statistic such as Friedman's T can be calculated. This statistic is analogous to the F statistic for equality of treatment means in an ANOVA. Pairwise testing of the scores can also be done using the nonparametric least significant difference method. Again, this is analogous to the paired *t*-test. It should also be mentioned that since scores provide rank order data, they can be subjected to the same methods discussed in the section on rank order data analysis.

Bradley-Terry and Thurstone-Mosteller Models: It is possible to derive scale data from paired comparison data using statistical models proposed by Thurstone-Mosteller^{31,32} and Brad-

ley-Terry.³³ To perform each of these models the scores must be converted into a random variable having either a normal (Thurstone-Mosteller) or logistic (Bradley-Terry) distribution. The values of these random variables are then further analyzed to arrive at a scale value for each sound in the study. These scale values are appropriate for correlation with objective properties of the sounds. The Bradley-Terry model has met with a great deal of success in the automotive sound quality industry. As a result, the mechanics for derivation of the Bradley-Terry scale values will be outlined here. This material is taken from David.⁸

Data on how one sound compares to another are available in the form of pair probabilities. For example, in a paired comparison of preference study with 100 subjects, if 80 subjects prefer sound *i* to sound *j*, the probability that sound *i* is preferred to sound *j*, p_{ij} is 0.8 while the probability that *j* is preferred to *i*, p_{ji} is 0.2. In general, the relationship is $p_{ij} = 1 - p_{ji}$. These pair probabilities are the basis for the models of the paired comparison evaluation. These are statistical models of the paired comparison process that take pair probabilities as input and produce a single valued measure for each sound.

All the linear PC models follow the same basic premise, that for each sound in a PC study, there exists a value, called the merit value, that underlies all the pair judgments. These merit values lie along a linear scale and the relative position of each value is indicative of how the sounds will be judged in the PC. Sounds with merit values that are close together should have pair probabilities near 0.5 while sounds with vastly different merit values should have pair probabilities approaching 1 (or 0). Like all models, this is an abstraction, but it's useful in analyzing the results. If merit values determine pair probabilities, then one can reverse this process to use pair probabilities (which are known) to find merit values (which are unknown). Otto²⁷ gives a straightforward summary of the mathematical formulation for the Bradley-Terry model.

In most sound quality studies, the maximal range in merit values is +3 to -3. Note that once the merit values have been determined, these values can be used to 're-predict' the pair probabilities. These predicted pair probabilities can be compared to the probabilities measured in the paired comparison evaluation. The correlation between these two sets of probabilities gives an estimate of how well the Bradley-Terry model fits the experiment. Usually a correlation (R^2) of 0.9 or better is found. No linear PC model will do well in an experiment where a number of inconsistencies is found. From the above discussion, it should be apparent that merit values fit very well with the consistent judgments (if $A > B > C$ then $A > C$) but cannot describe inconsistent data (if $A > B > C$ but $C > A$).

Paired Comparison of Similarity. Performance measures for this type of evaluation include histograms of subjects' numerical ratings to insure the entire scale is being used and the rating differences between replicate judgments. Generally, rating differences of 1-2 (10 point scale) are considered acceptable.

Analysis of paired comparison of similarity evaluations is done using non-metric Multi-Dimensional Scaling (MDS).³⁴ An application to engine valve noise is given by Namura.³⁵ This well-known technique is available in most statistical software packages. MDS is closely related to methods which reduce dimensionality such as factor and principal component analysis. A common use of MDS in similarity experiments is to display the sound samples on a two perceptual dimension plot. Sounds which have similar characteristics will cluster together. However, it is up to the researcher to determine the meaning of the perceptual dimensions. Some of the other previously described evaluation techniques can be used for this purpose.

Input to this technique consists of a matrix of similarity ratings for each pair. The analysis takes these data and produces an *n* dimensional map in which the samples are placed based on their similarity. Samples which are close together in this map were judged similar, while samples which are far apart in the map were judged dissimilar. The dimensions of the map are up to the user. Generally, we restrict ourselves to two or

three dimensions for ease of interpretation. The axes of the map have no physical meaning. There is a measure called Kruskal Stress, values range from zero to one, which is an indicator of how well the map matches the similarity data.

Rank Order. Rank order data fall into the category of *ordinal* scaling and as a result are subject to nonparametric statistical analysis. This is a comparative scaling technique in which the subjects are presented with several sounds and asked to rank them according to some criterion. Values generated from rank order data indicate the relative positions of the objects (sounds) but not the magnitude of the differences between them. One way to present rank order results is to calculate the average rank for each sample. However, it must be remembered that those averages have no meaning outside the given experiment.

Significance Tests. Friedman's test is commonly used for significance testing of rank order data.¹⁶ The Friedman statistic is analogous to the F-statistic for the analysis of rating data. If the Friedman test determines that the overall distribution of the rank totals is significantly different then a multiple comparison procedure can be applied to determine which samples differ significantly. A nonparametric version of Fisher's LSD can be used for this purpose.¹⁷

Correlation Tests. Kendall's Tau: Kendall's Tau is similar to the common correlation coefficient in that it assumes a linear relationship between dependent and independent variables. However, this measure operates on rank order data. This measure can also be extended to multivariate situations.

Spearman Rank Correlation: Also measures the relationship between rank order data sets, but cannot be applied to the multivariate case.

The Contingency Coefficient: The contingency coefficient C is one of the most broadly used correlation measures for rank order data. It has a number of advantages over other nonparametric methods. The contingency coefficient makes no assumption on the distribution of the data. It does not require any specific relationship in the data (i.e., linearity). Also, it can be extended to multivariate situations.

Subjective to Objective Correlation

Purpose. The logic behind performing subjective to objective correlation centers on the concept that one can possibly replace subjective testing with mere objective characterizations of the stimuli. By doing this one can reduce subjective testing that is costly from a time, equipment, facilities and general logistics standpoint. If one can reliably replace subjective testing and the costs involved with objective characterizations of the stimuli which are usually less costly, then that gives a significant impetus to finding strong correlation between the two views of the stimuli. The eventual goal, of course, is to guide the establishment of design and testing specifications for your product that will guarantee the product is the one that the marketplace desires. The following sections will deal with increasingly more complicated and powerful correlation methods.

Scatter Plots. If the sound event at hand is strongly correlated to a single, physical dimension of the stimuli, a simple scatter plot often will yield significant insight into relationships between the subjective responses and the physical stimuli. The process merely involves plotting the subjective response for a stimuli against some scalar on the vertical axis versus some objective measure of the stimuli on the horizontal axis. Most popular spreadsheet and statistics packages provide for some means of generating scatter plots.

Linear Regression. Linear regression takes scatter plots one level of information higher: provision of a mathematical relationship between the subjective and objective characterizations of the stimuli. The mathematical relationship in this case is a straight line, fitting the response data to some class of objective characterization. The process involved is technically known as least squares estimation and works on the premise of the model regression equation:

$$y = B_0 + B_1x + \varepsilon \quad (3)$$

where x is the objective characterization of the stimuli or regressor, y is the response, B_0 the intercept and B_1 is the slope of the straight line relationship between y and x . The error term ε is considered random with mean of zero and an unknown variance. To find the values of the coefficients on the regressors, one goes through the process of least squares estimation. One looks for the objective characterization, whether physical or psychophysical, that gives the strongest correlation and hence, the higher coefficient of determination R^2 . There are many good texts on linear regression and it is not the purpose of this guide to give all of the details of these methods. For more detailed information, refer to an introductory text on linear regression.²¹

Multiple Linear Regression. Going one step further than single variable linear regression, one can perform multiple linear regression where the straight line relationship is now between the subjective response data and some linear combination of scalar, objective characterizations of the stimuli. In this case, the model regression equation is:

$$y = B_0 + B_1x_1 + B_2x_2 + \dots + B_nx_n + \varepsilon \quad (4)$$

Again, the goal is to find the relationship that gives the strongest relationship between the regressors and the response, yielding the highest R^2 . We again refer those interested to texts on Linear Regression, and the text recommended from the previous section provides an excellent discussion of this method. Issues which need to be addressed when performing Multiple Linear Regression are: 1. feasibility of using multiple regressions in setting test and design specifications and 2. colinearity of regressors.

Non-Linear Regression. Of course, relationships between subjective responses and objective characterizations are not limited to the linear type. They can be nonlinear relationships; nonlinear combinations of the objective characterizations that correlate to the subjective response data. The mathematics involved here can be more complicated than for linear regressions. No attempt will be made to discuss this topic as it cannot be handled efficiently in a guide such as this. The authors merely want to call attention to the method as a possible way to draw information from your subjective and objective characterizations of the stimuli. Given an intelligent survey of available and self-designed objective characterizations and a well-posed subjective test and analysis, one can hope for some linear relationships to surface between the subjective response and objective characterizations of a set of stimuli.

Neural Networks. Yet another method to draw out relationships between subjective response and objective characterizations of a set of stimuli is the use of neural networks. This method is relatively untapped in the field of automotive sound quality but does offer some exciting potential in the areas of pattern recognition. Neural networks are advertised as an attempt to model the neural structures in the human brain, to perform artificial learning in a system. Much like the mysteries involved with how the brain organizes information, neural networks operate as a black box with known input and output – but little usable information about the sequence and logic of events that happen inside. Typically, these networks require a great deal of input/output data (training) to produce reliable results.

Putting One's Instincts to Work. From the perspective of a person that is at least partly responsible for automotive sound quality, this person will most likely have some knowledge and suspicions about the relationships that might exist between your particular sound events and the mathematical transformations of those same signals. Performing a proper subjective to objective correlation will give one the chance to test and challenge those suspicions. However, one must guard against, "Worshipping at the altar of R^2 ," a quote from Norm Otto of Ford. This means that just because a correlation seems strong, don't put your full faith in that R^2 ; the relationship between

the input and output data must make sense. A further anecdote to underline this point. It has been told that in Germany, the birth rate of human babies and the Stork population are highly correlated. Do you believe the correlation has any merit?

One will need to defend any derived relationship when using it to establish design and test specifications. There are always chance relationships which will, from time to time, appear in correlation. This arises usually from too small of a number of stimuli and/or subjective evaluations being used. Remember, it is very easy to fit straight lines to a small number of data points. One also needs to guard against using multiple regressions that use dimensions that are highly colinear. For example, if a multiple regression shows that a 3 regressor equation gives a high R^2 value, but those regressors are all colinear (i.e., strongly correlated), which is to say they are all telling one nearly the same thing about the stimuli, the regression is probably no more useful for setting test and design specifications than a single dimension regression using one with the 3 regressors. The only information proven by the 3 regressor correlation was that each regressor gives a slightly different view of the stimuli. No correlation that is performed should be thought of as proof of causality between the stimuli and the subjective response.

Conclusion

The authors have presented this material in the belief that it will aid the reader to become familiar with successful techniques currently applied to jury evaluations of automotive sounds. They encourage the reader to apply the techniques presented here and also to add to this methodology with further development and critique. Readers are encouraged to contact the authors using the contact information provided.

Acknowledgments

The authors would like to acknowledge the ANSI Sound Quality Work Group (ANSI S12/WG 36) from which this work originally initiated.

References

1. AES20-1996, AES recommended practice for professional audio – Subjective evaluation of loudspeakers,” Audio Engineering Society Standard, New York, 1996.
2. ANSI S3.1-1991, Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms,” American National Standard.
3. ANSI S12.2-1995, Criteria for Evaluating Room Noise,” American National Standard.
4. S. Bech, “Planning of a Listening Test – Choice of Rating Scale and Test Procedure,” Symp. on Perception of Reproduced Sound, Denmark, 1987.
5. N. Otto, “Listening Test Methods for Automotive Sound Quality,” Proceedings of the Audio Engineering Soc., New York, 1997.
6. S. Bech, “Selection and Training of Subjects for Listening Tests on Sound Reproducing Equipment,” *J. Audio Eng. Soc.*, 40, 590, 1992.
7. S. Bech, “Training of Subjects for Auditory Experiments,” *acta acustica*, 1, 89, 1993.
8. H. David, *The Method of Paired Comparisons*, Oxford University Press, 1988.
9. S. Stevens, *Psychophysics*, John Wiley and Sons, New York, 1975.
10. H. Levitt, “Transformed Up-Down Methods in Psychoacoustics,” *J. Acoust. Soc. Am.*, 49, 467, 1971.
11. N. Kousgaard, “The Application of Binary Paired Comparisons to Listening Tests,” Symp. on Perception of Reproduced Sound, Denmark, 1987.
12. N. Otto and G. Wakefield, “The Design of Automotive Acoustic Environments: Using Subjective Methods to Improve Engine Sound Quality,” Proceedings of the Human Factors Society, Atlanta, 1992.
13. N. Otto and B. Feng, “Wind Noise Sound Quality,” Proceedings of SAE Noise and Vibration Conference, Traverse City, MI, 1995.
14. N. Otto and B. Feng, “Automotive Sound Quality in the 1990s,” Third Int. Congress on Air- and Structure-Borne Sound and Vibration, Montreal, 1994.
15. S. Amman and N. Otto, “Sound Quality Analysis of Vehicle Windshield Wiper Systems,” Proceedings 1993 SAE NVH Conference, Traverse City, MI, Paper 931345, 1993.
16. M. Meilgaard, G. Civile and B. Carr, *Sensory Evaluation Techniques*, CRC Press, Boca Raton, FL, 1991.
17. N. Malhotra, *Marketing Research – An Applied Orientation*, Prentice Hall, Englewood Cliffs, NJ, 1993.
18. R. Bisping, S. Giehl and M. Vogt, “A Standardized Scale for the Assessment of Car Interior Sound Quality,” Proceedings 1997 SAE NVH Conference, Traverse City, MI, Paper 971976, pp. 843-847.
19. J. Chambers, W. Cleveland, B. Kleiner and P. Tukey, *Graphical Methods for Data Analysis*, Chapman and Hall, NY, 1983.
20. M. Blommer, S. Amman and N. Otto, “The Effect of Powertrain Sound on Perceived Vehicle Performance,” Proceedings 1997 SAE NVH Conference, Traverse City, MI, Paper 971983, pp. 891-896.
21. D. Montgomery and E. Peck, *Introduction to Linear Regression Analysis*, John Wiley.
22. J. Kim and C. Mueller, *Introduction to Factor Analysis-What It Is and How To Do It*, Sage Publications, Beverly Hills, CA, 1978.
23. J. Kim and C. Mueller, *Factor Analysis-Statistical Methods and Practical Issues*, Sage Publications, Beverly Hills, CA, 1978.
24. R. Bisping, “Emotional Effect of Car Interior Sounds: Pleasantness and Power and Their Relation to Acoustic Key Features,” Proceedings 1995 SAE NVH Conference, Traverse City, MI, Paper 951284, pp. 1203-1209.
25. H. Murata, et al., “Sound Quality Evaluation of Passenger Vehicle Interior Noise,” Proceedings 1993 SAE NVH Conference, Traverse City, MI, Paper 931347, pp. 675-681.
26. H. Takao, et al., “Quantification of Subjective Unpleasantness Using Roughness Level,” Proceedings 1993 SAE NVH Conference, Traverse City, MI, Paper 931332, pp. 561-570.
27. N. Otto and G. Wakefield, “A Subjective Evaluation and Analysis of Automotive Starter Sounds,” *Noise Control Engineering Journal*, Vol. 94, No. 3, pp. 377-382, 1993.
28. A. Champagne and S. Amman, “Vehicle Closure Sound Quality,” Proceedings 1995 SAE NVH Conference, Traverse City, MI, Paper 951370, pp. 1109-1114.
29. H. Staffeldt, “Correlation Between Subjective and Objective Data for Quality Loudspeakers,” *Audio Engineering Society Journal*, Vol. 22, No. 6, 1974.
30. M. Kendall, “Further Contributions to the Theory of Paired Comparisons,” *Biometrics*, 11, 1955.
31. L. Thurstone, “The Prediction of Choice,” *Psychometrika*, 10, 237, 1945.
32. F. Mosteller, “Remarks on the Method of Paired Comparisons. I. The Least Squares Solution Assuming Equal Standard Deviation and Equal Correlations,” *Psychometrika*, 16, 3, 1951.
33. R. Bradley and M. Terry, “Rank Analysis of Incomplete Block Designs I. The Method of Paired Comparisons,” *Biometrika*, 13, 51, 1957.
34. J. Kruskal and M. Wish, *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA, 1978.
35. T. Nemura, N. Adachi and K. Suzuki, “Research in Regard to Sensory Characteristics Measuring for the Impulse Noise of the Engine Valve System,” Proceedings 1991 SAE Int. Congress and Exp., Paper 910620.



Authors may be contacted at:
 Norm Otto – notto@ford.com
 Scott Amman – samman@ford.com
 Chris Eaton – chris.eaton@ericsson.com
 Scott Lake – scott.a.lake@gm.com

Appendix – Script for Jury Instruction Examples

Paired Comparison.

Instructor:

In this evaluation you will be presented pairs of [fill in type of sound] sounds. These sounds were recorded [describe the recording location]. You are asked to indicate which of the pair you think is most/least [fill in preferred, annoying, loud, sharp, . . .]. Select sound A if you think the first sound is more/least [fill in preferred, annoying, loud, sharp, . . .] or select sound B if you think the second sound is more/least [fill in preferred, annoying, loud, sharp, . . .].

Since the objective of the experiment is to understand the individual’s reaction towards the sounds, there are no right or wrong answers. Please feel free to select the sound as you hear it. We do request that you select one sound out of each pair, no ties are allowed. You will hear [fill in number of blocks] blocks of data, each consisting of [fill in number of pairs/block] pairs of sounds. There will be a pause following each block.

There will be a practice block of [fill in number of pairs] pairs to familiarize you with the selection process. We will now begin the practice block.

Semantic Differential.

Instructor:

In this evaluation you will be presented with [fill in type of sound] sounds. You are asked to evaluate the sound based on [fill in number of adjective pairs] adjective pairs. The sounds

are to be evaluated using a scale which is divided into [*fill in number of scale divisions*] parts. For example, on the quiet/loud [*use an example; use any pair out of your set*] scale a sound can be evaluated as extremely, very or somewhat quiet or extremely, very or somewhat loud based on how you perceive that sound. If you do not think that the sound is either quiet or loud, the part of the scale labeled neither should be marked. As each sound is announced, evaluate the sound based on the seven categories between the adjective pair. To assist in your judgment of the sounds, each sound will be repeated [*fill in number of repeats*] times before advancing to the next sound. After the last sound is played, the next adjective pair will be announced [*scale example provided in the Jury Evaluation Methods/Semantic Differential section*].

Attribute Intensity (Response) Scaling.

Instructor:

In this evaluation you will be presented with [*fill in type of*

sound] sounds. For each sound that is presented, your task is to judge the placement of the sound on the scale provided for the various sound categories. To assist in your judgment of the sounds, each sound will be repeated [*fill in number of repeats*] times before advancing to the next sound.

Magnitude Estimation.⁹

Instructor:

In this evaluation you will be presented with [*fill in type of sound*] sounds. They will be presented in irregular order. Your task is to tell how intense they seem by assigning numbers to them. Call the first sound any number that seems appropriate to you. Then assign successive numbers in such a way that they reflect your subjective impression. There is no limit to the range of numbers that you may use. You may use whole numbers, decimals or fractions. Try to make each number match the intensity as you perceive it.