

Conducting Listening Tests

Introduction

Considering that listening tests are time- and cost-intensive and often difficult to evaluate, one might ask: Why conduct listening tests at all?

The buying decision and the satisfaction of customers are to a great degree depending on subjective impressions – including acoustic ones. Therefore, it is the responsible acoustics engineer's task to create an appropriate sound matching the image of a product. Especially in the NVH market, the increasing demands of customers often cannot be satisfied by a simple reduction of noise levels. If a sound does not meet the user's expectations, in the worst case the entire product can be rejected or the negative sound impression can be interpreted as a malfunction. But which sound is appropriate? Listening tests are an essential basis for answering this question to the customers' satisfaction. And an appropriate preparation and selection of the listening tests to be conducted can save a lot of time and effort during the test as well as in the evaluation phase.

This application note will provide you with an overview of possible test types, suitable test environments and test signals, the selection of test participants (subjects) and meaningful test evaluations. This knowledge can facilitate the responsible test operator's work and provide him with useful suggestions and help. The following examples and screenshots were created with the Sound Presentation and Evaluation Studio SQuare. This software was developed by HEAD acoustics specifically for conducting listening tests.

Test Types

For listening tests, a wide variety of test types is available. Depending on the requirements and objectives, a suitable test method must be chosen. The following test types described are especially suited for the areas of sound quality and benchmarking, and represent only a limited selection of all possible test types.

Ranking

In a ranking test, the subject is asked to sort N sound samples in a ranking order from 1 to N according to a certain criterion (e.g. annoyance). This task becomes more difficult as more sound samples have to be sorted. Therefore, the maximum number of sounds offered for such a ranking test should be six.

This test method offers an easy and uncomplicated way to check the first impression of the sounds; for example, using customer preference as an attribute. The disadvantage of this evaluation method is that the subject only specifies an order for the individual sounds, but no information is provided about the "distance" between the sounds on this quality scale. Therefore the results of a ranking test are not necessarily suitable for calculated correlations with the results of technical measurements or analyses.

Figure 1 shows how such a ranking test could look. By clicking on the sound buttons (Door01-Door06), the subject can play the respective sound and change its position in the ranking list with the arrow buttons.

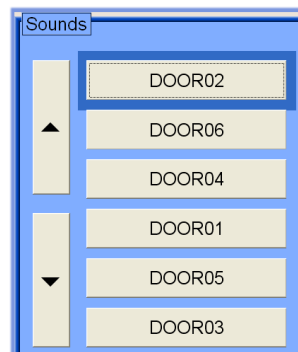


Figure 1: Sound ranking test

Paired Comparison

In a paired comparison test, two sound samples are presented to the subject in succession. The subject is asked to judge these two sounds according to a certain criterion (e.g. loudness). Either two or three answers are possible. If a decision is to be enforced (“forced choice”), only two possible answers are offered: $A > B$ (e.g. A louder than B) and $B > A$ (e.g. B louder than A). Otherwise, the selection $A = B$ (A and B equally loud) can be offered as well.

Some subjects tend to avoid a decision and will frequently choose the answer $A = B$, which makes the evaluation of the test more difficult. This can be avoided with the “forced choice” variant of the test. On the other hand, this variant may put the subject under pressure, because a decision is demanded even in cases where the subject does not hear a difference, so the subject is forced to make a judgment not matching his or her actual perception. However, these two effects can be easily avoided by a suitable introduction and test instructions.

The paired comparison test is suitable for discovering differences in very similar sounds. However, this can easily lead to an overvaluation of these differences. In real life, for example when judging the interior noise of a vehicle, there is no possibility for a direct comparison. The sounds can only be judged one by one with a considerable time in between. The human hearing system is capable of remembering a sound level in the short-term memory, so in a paired comparison with sound samples following quickly after each other, even small level differences can be detected. The acoustic long-term memory, on the other hand, mainly stores sound patterns. Therefore, if the sounds are not presented in direct succession, the characteristics of the sound based on the contained patterns are much more significant for the judgment than the absolute sound level.

Therefore, before a listening test is conducted, the objective must be defined: Is it important to find small differences between sounds, or should the test resemble the real-life experience? Based on this question it is possible to decide whether the paired comparison method is suitable. Figure 2 shows an example of the SQuare user interface for a paired comparison test. With the playback control buttons, the two sounds A and B can be played, and the judgment is entered with the three buttons below.

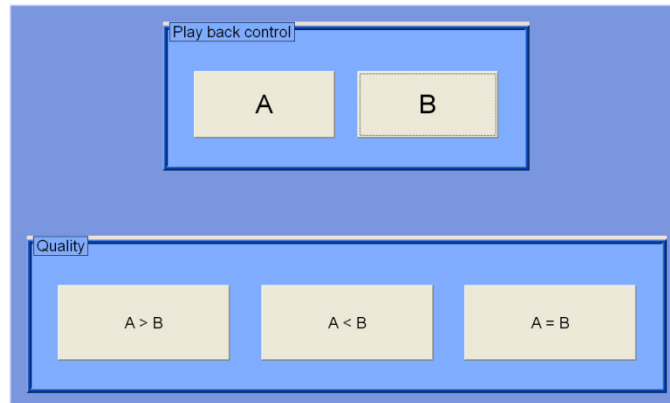


Figure 2: Paired comparison test using SQuare

A disadvantage of this test method is the duration of the test, which increases considerably with the number of sounds to be evaluated due to the many possible pair combinations.

Category Judgment

In a category judgment test, the sound samples are presented one by one to the subject, who rates the respective sound on a scale with several categories according to a certain criterion. The criterion could be, for example, the sharpness of the sound. For the judgment, the five-point Rohrman scale with the categories “not at all”, “slightly”, “moderately”, “very” and “extremely” is often used. Another frequently used scale is the ten-point scale for the judgment of disturbing noise in vehicles as specified in the VDI Guideline 2563. Table 1 shows an overview of the ten categories.

Category	Description
1	not acceptable
2	hardly acceptable
3	considered a severe fault by all persons
4	considered a fault by all persons
5	considered disturbing by all persons
6	considered disturbing by some persons
7	noticeable by all persons
8	noticeable only by critical persons
9	noticeable only by experienced listeners
10	not noticeable even by experienced listeners

Table 1: 10-point scale according to VDI Guideline 2563

In the judgment based on a category scale, various bias effects can occur. For example, a sound that is to be judged directly after a sound that was perceived as very sharp may be assigned to a different category than if it was heard after a sound with little sharpness, i.e. the judgment of a sound may be influenced by the previous sound (context effect). This effect can be avoided by presenting each sound several times in a random order. Furthermore, it is possible that the scale is used to a different extent by different subjects. Subjects often avoid using the extreme categories and choose answers in the medium range of the scale instead (“central tendency of judgment”). That way, the subject wants to make sure not to be surprised by an extreme

sound in the course of the test, for which no suitable category would be available if the extremes were already used. This effect can be avoided by a sufficient training. During this training, the sounds with the most extreme characteristics can be presented to the subject in advance, so they are known and can be judged accordingly. The different utilization of the scale by different subjects can also be compensated by normalizing the results (with the average value as the reference) after the test.

Figure 3 shows an example for a category scale.

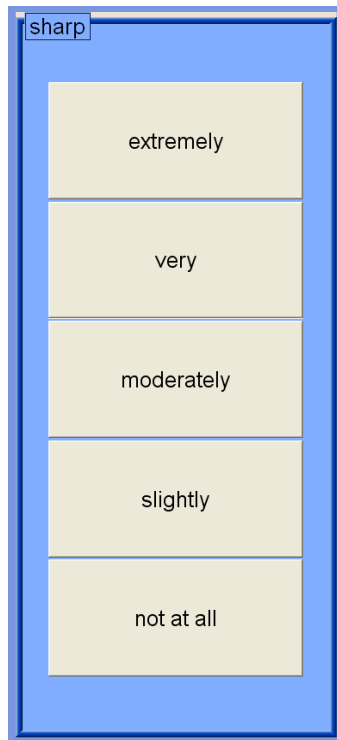


Figure 3: Example of a category scale generated with SQuare

Semantic Differential

Using a semantic differential for judgment allows a differentiated examination of the sound samples. Whereas in the test methods described above the subject concentrates on only one prescribed criterion for judgment, this test method allows several attributes of a sound to be judged. The subject evaluates the presented sounds based on several bipolar scales, whose ends are labeled with an adjective and its antonym (opposite). The scales often have seven or nine points. Figure 4 shows an example for a semantic differential with a seven-point scale and four pairs of antonyms.

This test method delivers a detailed profile of a sound including a lot more information than just the fact that one sound is preferred to another and to what degree. The judgment on several scales makes it easier to find correlations to the results of technical measurements and analyses. This makes it possible to find out why a sound received a bad rating and which aspect of the sound must be changed to improve the sound quality.



Figure 4: Example of a semantic differential using the SQuare interface

Of course, the judgment of a sound with a semantic differential is more time-consuming than other test methods. The number of sound samples and judgment scales should not be too large, otherwise the subject may lose their concentration towards the end of the listening test. Experience has shown that a number of eight to twelve attribute pairs should not be exceeded.

The selection of the attribute pairs requires great care for several reasons. If unsuitable attributes are chosen for a sound, the subjects will often mark the center of the scale, with the result that the test does not yield any useful information except that the wrong attributes were chosen. Furthermore it is important that the attribute pairs refer to different aspects of the sound. If this is not the case, the answers for the different attribute pairs will significantly correlate with each other. Using only one of these attributes (e.g. in a category test) would have been completely sufficient in that case, as the other correlated attribute pairs do not deliver new information. The selection of the antonym influences the judgment as well, as the following example illustrates: For the attribute “old”, the antonym “young” could be used as well as “new”. In many cases, the bipolar scale “old – young” will lead to different results than the scale “old – new”. So the objective of the listening test must be taken into account when choosing the antonym pairs.

When creating an evaluation sheet with several scales, the negative attributes should not all be placed on one side of the scales in order to avoid an adaptation effect of the subjects. In some listening tests, the items are rearranged to a different order for each new sound sample. That way, the subject is forced to fully concentrate on each individual sound without adaptation.

In the EU project OBELICS (Objective Evaluation of Interior Car Sound, BRITE-Euram 96-3727), semantic differentials for the judgment of vehicle interior noise were examined and various antonym pairs were compiled. This compilation contains the English antonym pairs, and also the corresponding pairs in German, French and Italian. This is very important for listening tests involving subjects from different countries, as it is advisable to interview each subject in his or her native language.

AISP (Exploration of Associated Imagination on Sound Perception)

The AISP test method is fundamentally different from the methods described above, because in this method, the subject is not given a list of predefined answers to choose from. In the AISP method, the subjects can freely express their emotions and associations in their own words. That way, the subject can judge a sound in an uninfluenced and unbiased way. This means that during an AISP test, the subjects don’t have rating sheets in front of them, but instead describe their impressions in their own words. The test operator makes an audio recording of these statements,

but otherwise makes sure not to interfere actively with the test procedure, i.e. he does not ask questions and does not comment on the subjects' statements.

In the second part of the test, the test operator asks additional questions, based on methodic interview guidelines, for the purpose of understanding and obtaining a further explanation of the subjects' judgments. This allows additional data to be gathered and the subjects' judgments to be validated communicatively.

The advantage of this method is that the subjects express their judgments in their own words. They are not restricted to predefined judgment scales based on a vocabulary that may not even match their own. By providing the possibility to freely choose the words, it is almost impossible that important aspects of judgment will be overlooked, whereas, for example, a semantic differential may fail to include a certain sound aspect that is essential to a subject.

Of course, the possibility to use one's own words makes the evaluation of the test results more difficult. The effort for the evaluation is considerable, because the subjects' statements first need to be converted into a consistent, comparable form. Since the answers of different subjects may be hard to compare in many cases, the task requires a lot of experience with this kind of listening test.

In the OBELICS project mentioned above, the AISP method was applied and examined in detail. Based on further research, the E³ method* (Explorative Environment Evaluation) was developed, where the AISP method is applied in the field (for example, in a real vehicle) rather than in a laboratory.

Test Procedure

In order to conduct a listening test successfully, i.e. with meaningful results, some basic rules must be observed. Before the actual listening test, the subjects must be instructed sufficiently. These instructions should include all necessary information and explanations about the planned test. Depending on the test objectives and procedures, it may also be useful to explain the purpose of the test to the subjects. Of course, this information should only be given if it is not likely to influence the subjects' judgments. The instructions should be given in both written and oral form. In many cases, it is sufficient to use a short summary of the oral explanations for the written instructions (see figure 5). It is important that the oral instructions are the same for all subjects. There are several possibilities to present the instructions. With SQuare, the written instructions can be displayed on the screen before the test. Furthermore, SQuare provides the possibility to play a video containing a recording of the oral instructions. Of course, the written instructions can also be printed on paper and handed out to the subjects.

* The method was developed by Prof. Dr. Schulte-Fortkamp at the Technical University of Berlin and has already been used successfully in cooperation with HEAD acoustics GmbH.

In the following listening test, you asked to judge the quality of several car door closure sounds.

At the beginning of the test, you will hear all the sounds to be judged, one time. After that, the actual judgement begins, where you should rate each door sound on a scale from “1” to “7”.

In this test “1” stands for a poor quality impression and “7” for an excellent quality impression.

Thank you very much for your cooperation!

Figure 5: Written instructions for a listening test

Only if the subjects have understood the task, will they feel confident during the test and complete the task in a reliable manner. In the briefing, it is important to assure inexperienced subjects that there are no wrong answers in a listening test. The judgment of sounds is based on each person’s individual perception of these sounds. A subject’s perception cannot be wrong – it can only be different of other subjects’ perceptions. In addition, it is useful to inform the subjects about the duration of the test, so they know what to expect.

After the briefing, a training of the subjects can be conducted. Whether such training is necessary depends on the difficulty of the task and the subjects’ experience. The less experienced the subjects are and the more difficult the task is, the more comprehensive the training must be, whereas experienced participants may not require any training at all. In a training session, some or all of the sound samples to be judged can be presented to the subjects in advance. That way, the subjects are prepared for the actual listening test and know what to expect during the test. The training may or may not include a judgment by the subjects. If the subjects are not familiar with the judgment procedure, it is advisable not only to present the sounds, but to ask for a judgment, too. The training should not be too long in order to avoid a loss of concentration during the actual listening test. Furthermore, a too lengthy training will cause boredom of the subjects and make them lose their motivation for the actual listening test.

After the subjects have been sufficiently informed about the test procedure by means of instructions and possibly some training, the actual test can begin. Of course, the subjects should not be disturbed during the test. It may be useful to allow the subjects to write down additional remarks about their judgments. This additional information can help to interpret the results during the evaluation of the test. During the test, the subjects should not feel left alone. This means that the test operator should be available in person or by phone for inquiries (e.g. about technical problems) during the test. After the test, the operator should ask the subjects about their impression regarding the test procedure, the duration etc. It may be necessary to modify the test procedure based on the answers to these questions. In this interview, additional information relevant for the evaluation can be gathered as well, such as the age, sex or profession of the subject. Of course, the privacy of each subject must be respected at all times.

The subjects' judgments can be written down on prepared forms. But in many cases, the judgments are entered directly into the computer using a suitable on-screen input mask (e.g. using SQuare). This avoids errors that could occur when transferring the data from paper into the computer. Before using a computer for the judgment process, it should be made sure that all subjects are familiar with using a mouse and keyboard. Figure 6 shows an example of such an input mask, which not only allows the subject to enter the judgment, but also to control the playback of the sound. In addition, the subject is informed about the progress and the number of remaining sound samples.

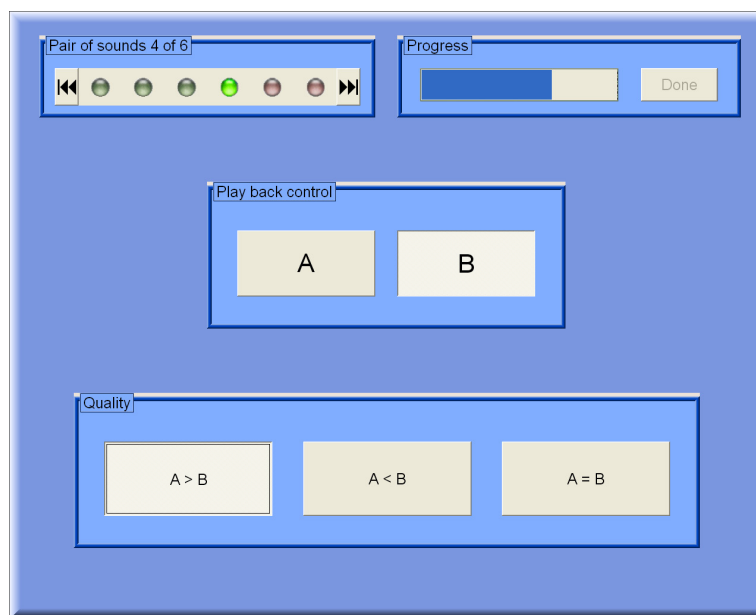


Figure 6: Example of an input mask for sound judgment

Altogether, a listening test should not last longer than 45 minutes to make sure that the subjects' concentration does not weaken (note: tests using the AISP method can last longer). The length and the number of sound samples should be chosen so that this time limit is not exceeded. A test that exposes the subjects to high sound pressure levels must be shorter to make sure that the subjects' health is not impaired. In addition, listening to loud and disturbing noise reduces the subjects' concentration.

Test Environment

The test environment should allow the subjects to feel comfortable. That means that the room should have sufficient ventilation and a convenient temperature. The subjects should not feel as if they were placed in a "broom closet", and they should not be surrounded by too much technology. Depending on the type of person, too much technology will either alienate or distract the subject. The background noise in the test room should be as low as possible. If sound samples with a very low volume are to be judged, the listening test must be conducted in a sound-proof room.

The test environment also includes the other test participants if the listening test is conducted in a group. The influences and interference between one subject and the others should be reduced as much as possible. This can be achieved, for example, by installing dividing walls. This is es-

pecially advisable if low-volume sounds are to be presented. It may be necessary to exclude subjects that are ill from the listening test, so the other test participants are not distracted by coughing or sneezing.

Research has shown that the results of listening tests are more meaningful the closer the conditions of the test environment resemble the real-life conditions the sounds are experienced in. In the EU project OBELICS, the researchers examined how a test setup should be designed in order for the judgment of vehicle interior noise to correspond to that made during a real test drive. It turned out that the correlation is best when the environment during the listening test resembles that of a real test drive. This was the reason for the development of the SoundCar. It consists of a playback system built into a car body, which is capable of presenting not only airborne noise, but also structure-borne vibrations. This means that during the listening test, the subject is sitting in a real car, hears the sound samples via a headphone or loudspeakers and at the same time feels vibrations in the seat and the steering wheel. The vibrations, based on measurements made during a test drive, are generated by shakers on the seat and the steering wheel and match the sound the subject is hearing. When a vehicle interior noise is played in a laboratory instead, it is often perceived as too loud. By including both airborne and structure-borne signals in the SoundCar, the listening test is conducted in the correct environment, allowing the loudness to be judged much more realistically.

Another step towards realism in a listening test is the HEAD 3D Sound Simulation System (H3S). This software allows the sound field to be controlled actively using the accelerator pedal, the brakes, and the shift lever and simulates the corresponding vehicle interior sound. The H3S software can be installed, for example, in a SoundCar, which allows both airborne and structure-borne sound to be played. For mobile use, the H3S simulation system can also be installed in a roadworthy vehicle. During the test drive, the subject feels the vibrations of the real vehicle, but hears a simulated vehicle interior sound corresponding to the actual driving situation. This test setup achieves a maximum level of realism. Figure 7 shows the mobile setup configuration of H3S in action.

The time required for a listening test with a SoundCar or H3S may be longer than for other test types, because, for example, only one person at a time can make judgments in a SoundCar. However, this expenditure of time is necessary for listening tests where the subject can only make a correct judgment in a realistic test environment.



Figure 7: Mobile setup configuration of H3S installed in a roadworthy car

Test Sounds

The sound samples for a listening test must have a high and consistent quality. To provide the subjects with a spatial sound impression, it is advisable to use artificial head recordings. In combination with suitable playback technology, artificial head technology places the subjects in the original sound field during the playback.

The easiest solution is storing the recording directly on the hard disk of a computer. With the computer, the sound files can then be edited, played back and judged by the subjects. The sounds should be recorded in a way that represents the normal usage of the product to be examined. Furthermore, it is important that all sounds to be used for a listening test should be recorded in the same environment, under the same conditions and, if possible, also with the same recording equipment. This makes sure that the subjects really judge the actual sounds in the listening test and not the different recording conditions. It is advisable to edit the sounds so that they don't contain unnecessary background noise and so that all sound samples have the same length. If the sound samples differ in several aspects (for example different vehicles, different test tracks and different background noise), it is impossible afterwards to find out which of these aspects influenced the judgment.

The signal level and equalization type should be the same for all recordings, too. Otherwise, the playback settings must be adapted accordingly. SQuare allows this adaptation to be made automatically.

In some cases it can be advisable to adapt the signal levels of the sound samples so that they are all perceived as equally loud. This is useful if the sounds are to be judged regarding their sound quality. Inexperienced subjects might be distracted from the actual characteristics of the sound if the samples have different loudness levels.

The length of the sound samples should not be too short. For stationary signals, normally a length between 3 and 5 seconds is sufficient. Non-stationary signals may be longer.

There are two possibilities for the playback of the test sounds. Either the subjects can be given the possibility to control the playback themselves. That way, each subject can decide individually when and how often a sound is played (individual control). The second possibility is to play the sounds according to a predefined playlist and to give the subject a certain time for the judgment. SQuare allows the test operator to choose whether the test is continued automatically after a certain period or the system waits for the judgment from each subject (see figure 8).

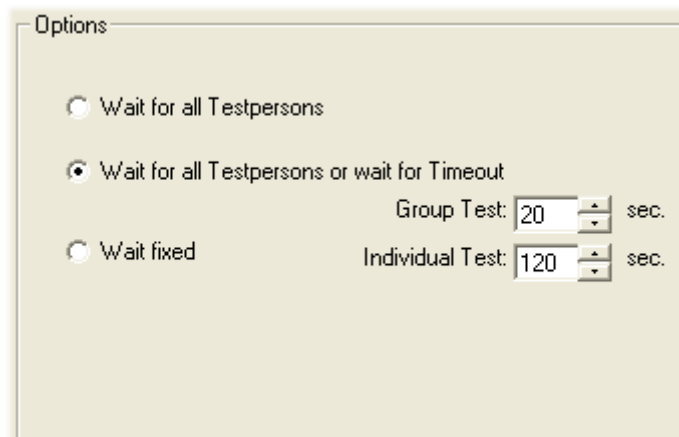


Figure 8: SQuare options for the time control of judgment

The individual control method is especially preferable if the sound samples are very short (e.g. door closing sounds) or have a very low volume. If such a sound is played only once and the subject is distracted or inattentive in that moment, a meaningful judgment is not possible. Furthermore this method offers the possibility to create a different playlist for each subject in order to avoid the context effect mentioned earlier. SQuare supports this by offering a randomizing function for the playlists.

The judgment in a group has the advantage that several persons can participate in the test at the same time, which saves a lot of time and ensures identical test conditions for all subjects. However, this is only possible if the subjects cannot influence or disturb each other. Of course, the selection of the control method depends on several factors: First, it is a question of available time, and second, not every control method can be combined with every test type.

Furthermore, it is necessary to decide whether the sound samples are played back via loudspeakers or via headphones. If loudspeakers are used, the acoustic properties of the room must be taken into account in order to make sure that each subject in the room hears what he or she is supposed to hear. An individual playback control is not possible if loudspeakers are used in a group test. Playback via headphones is an easy way to make sure that each subject hears the same calibrated signal. An inexperienced subject, who is not familiar with listening to artificial head recordings via headphone, may have some difficulties at first. If the characteristics of the recording room differ significantly from those of the playback room, the inconsistency between the visual and the acoustic impression can cause the sound to be perceived, for example, as too loud by an untrained subject. This can be avoided by appropriate instructions. The test operator can ask the subjects to close their eyes and imagine that they are in the room the sound was recorded in. With a little training, the subject will find it easy to imagine the different acoustic environment. If the listening test is conducted in a room with similar acoustic properties as the recording room, this problem does not occur (e.g. playback of vehicle interior sounds in the SoundCar).

The playback via headphones can be supported by an additional subwoofer playback. The subwoofer allows low frequencies to be played that would be missing in a headphone-only playback. Of course, an additional subwoofer playback limits the possibilities in a group test. The playback must be simultaneous for all subjects, so an individual playback control is not possible. The decision about the type of playback is also significantly influenced by the available premises and the available hardware.

The Subjects

The selection of the subjects, too, is influenced by external conditions. The pool from which the subjects can be recruited is normally limited as well as the time scheduled for conducting the listening test. However, since the selection and the number of subjects have an influence on the result of the listening tests, they should be selected carefully.

Before selecting the subjects, the objective and task of the listening must be clearly defined. The following examples illustrate this. An experienced subject (expert) will have no difficulties to solve even complicated "listening tasks". Thanks to his trained hearing, an expert will find it easier to concentrate on a specific aspect in a sound and to judge this very aspect. An untrained subject is not capable of this. On the other hand, an expert may overrate certain aspects of a sound, so a sound may be rated as unacceptable and fail the test, while the same sound would have been considered acceptable by an untrained listener. Also, subjects who never or rarely drive a car

are not suitable for judging vehicle interior noise. Besides the general experience with participating in listening tests, a subject's product experience should be checked as well. Someone who drives a luxury car will probably perceive the interior noise of a sports car as too loud, whereas sports car drivers or enthusiasts would rather accept or even like the dynamic, loud sound of such a car. The subjects' knowledge about the product to be examined and the demographic composition of the group of subjects should correspond to the relevant target customer group. The number of subjects also has an influence on the test results. The more subjects that participate in the test, the better the averaging-out of individual preferences will be in the results. On the other hand, extensive tests requiring an intensive training will likely not allow a large number of subjects to be recruited. In order to determine whether a sufficient number of persons has participated in a test, various statistic examinations can be performed. In general, a subject group is large enough if the average value of the judgments does not change significantly by including the results of another subject in the calculation. The confidence intervals, which can be calculated using statistic formulas also allow a statement to be made about the probability that the average values will or will not change by increasing the number of subjects. That way, the test operator can prove his test results statistically. However, even a large number of subjects cannot compensate for improper selection of subjects (i.e. even 200 drivers of luxury cars will not give you a meaningful judgment of the sound of a sports car). When deciding about the number of subjects, it is important to make sure that the number is still large enough if one or several subjects must be excluded from the evaluation of the results due to a lack of consistency.

Evaluating the Test Results

After conducting the listening test, the obtained results must be evaluated. For this purpose, a wide range of statistic calculation methods is available. These calculations serve two purposes: First, they are used to examine and evaluate the actual result data (e.g. for determining the confidence interval mentioned above), and second, they allow the data gathered in the listening test to be summarized and presented in a concise form. Easy introductions to statistics can be found in books about test methods and evaluations for social scientists.

However, before the data can be evaluated using statistic functions, they must be first "translated" into numbers. If the test was conducted using a PC with SQuare, the test operator is automatically presented the results in a numerically translated format. If the subjects have written down their judgments on paper, they must first be translated into numeric values. Different test methods require different evaluations and numeric coding systems.

In the ranking test method, only ranking judgments are made, i.e. a comparative scale is used containing no information about the distance of the different ranks from each other. For the evaluation it must be taken into account that each judgment of a sound sample highly depends on the judgments of the other sounds. While the averaging of the individual judgments of the subjects automatically yields different distances, it still must be decided individually for each listening test whether it makes sense to use these averaged "metrics" for further evaluation or to convert the judgments back to simple ranking values.

In a paired comparison, too, only a comparative scale is used at first (A is better than B). The individual data gathered can be easily converted into a ranking order (the judgment $A > C$, $C > B$ results in the ranking order A, C, B). With suitable statistic tools, it is also possible to calculate a scaled ranking order, which allows the differences (or distances) between the sounds to be evaluated as well. The resulting scale then allows correlations to be examined. In addition, the pair

comparison test allows various evaluations to be made regarding the judgment capabilities and reliability of the subjects. This includes, for example, the examination of triads. If sound A was judged better than sound B and sound B better than sound C, sound A should also be judged better than sound C. If this is not the case and such inconsistencies occur frequently with a certain subject, it can make sense to exclude the results of this subject from further evaluations. When using SQuare, an analysis of triads is performed automatically when a report is generated with Microsoft Excel®. Figure 9 shows an excerpt of such a report.

Computed by HEAD SQuare, Version: 2.03.200.0, Date: 18.05.2006, Time: 10:41:06

Score matrix

	Door 2 (0,00 - 0,70 s)	Door 1 (0,00 - 0,70 s)	Door 3 (0,00 - 0,70 s)	Door 4 (0,00 - 0,70 s)	Triad(s)!
TP1	2	0	2	2	
Number of stimuli					4
Average Score					1,50
Nbr. of maximal possible circular triads					4,00
Nbr. of circular triads					1,00
Circular error rate [%]					25,00
Coefficient of consistence (Nbr. of S. uneven)					0,60
Coefficient of consistence (Nbr. of S. even)					0,50

Figure 9: Display of inconsistent triads in a SQuare result report

If inconsistent triads occur with several subjects, this indicates that the subjects may be overtaxed or did not correctly understand the test task. In a paired comparison test, it is useful to present the individual sound pairs several times (also in reverse order, i.e. A – B and then B – A). That way, the repeatability of the judgment can be checked individually for each subject, providing additional information about the solvability of the task and the capabilities of the subject.

In a listening test with category judgments, each sound is judged more or less independently of the judgment of the other sounds in the test. Therefore it can be assumed that the results do not represent a simple comparative scale, but a scale that also allows the amount of difference between the sounds to be evaluated (interval scaling). This has the advantage that the results of such a listening test are well suited for a correlation study with the results of a technical analysis. For category scaling, it is also useful to present each sound sample several times for judgment in order to minimize the context effects described above and to check the intra-individual variation (this means the variations within the judgments of one subject).

The results of a listening test with semantic differentials are suitable for correlation analysis as well and thus allow an extensive evaluation. Of course, the judgment of a sound regarding several criteria takes more time. Therefore, in most listening tests of this type, it is not possible to present all sound samples several times, otherwise the test would be too long and the subjects' concentration would weaken. So there is little possibility to verify the reliability of a subject. In some cases it can be useful to present at least some of the sound samples several times in order to do at least a rudimentary reliability check.

The evaluation of listening tests using the AISP method requires a lot of experience. Since the subjects describe their judgments in their own words, it is hard to translate the judgments directly into numbers or to summarize them. However, it is nevertheless possible using suitable methods based on proven techniques of qualitative empiric research, so even these tests allow a further statistic evaluation. Results of listening tests based on the E³ method can be evaluated analogously.

Generally, all test methods require the subjects' judgments to be translated into numeric values if they are to be subjected to further statistical evaluation. For example, the judgments on a five-point category scale are assigned the numbers "1" to "5". For a semantic differential with a seven-point bipolar scale, the values "-3" to "+3" can be used. In this case, it is important to make sure that even if the scales on the result sheet do not always point in the same direction (the negative attributes are sometimes on the left side and sometimes on the right), the numbers must be assigned so that "+3" always represents the positive end of the scale and "-3" represents the negative end. Only that way, a meaningful statistic evaluation is possible. Figure 10 shows an example.

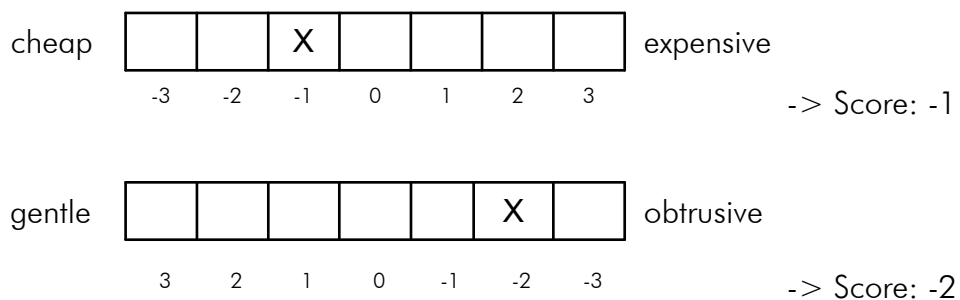


Figure 10: Translating judgments into numeric values

During the evaluation of the numeric values derived from the judgments, it should always be kept in mind that these numbers were originally judgments, for example, on a category scale. The actual qualitative judgments should not be forgotten just because they were converted to numbers for the purpose of statistical evaluation.

Once the judgments of the subjects are available in numeric form, they can be represented graphically and compared to each other. This provides a first impression of the judgments and helps with the decision whether the judgments of different subjects can be averaged. It is possible that the inter-individual variation (the differences between the judgments of different subjects) is too large, so an averaging of the judgment results would reduce the significance of the result. This is the case, for example, if the judgments differ because the subjects have utilized the scales to a different extent. In this case, the judgments can first be adapted by normalization, so that afterwards an averaging is possible without reducing the overall significance of the result. Such a conversion only makes sense, however, if the tendency of the judgments (i.e. the form of the curve and the ranking order) is roughly consistent. Otherwise, normalization and averaging would provide misleading results for the listening test.

If the judgments of the individual subjects are too different, averaging the results may be inadvisable. In some cases it makes sense to subdivide the subject group into two (or more) smaller groups whose results allow an averaging. This must be decided individually for each listening test based on the results. Statistics software provides suitable analysis methods helping with the evaluation.

Besides the calculation of the arithmetic mean value, other values frequently determined are the median value, the interquartile range and the standard deviation. The median value is the value exceeded by 50 % of all judgments, so the other 50 % are below it. Unlike the arithmetic mean value, the median value is hardly affected by extreme values (judgments that are very far away from the rest). The median value is often used in the evaluation of listening tests involving only a small number of subjects. The interquartile range encloses the median value and shows the range including 50 % of all judgments, i.e. 25 % of the judgments are below the interquartile

range and 25 % are above it. The width of the interquartile range is an indicator of the variation between the judgments of the individual subjects.

The standard deviation is the average deviation from the arithmetic mean value. The standard deviation, too, is an indicator for the variation of the judgments.

The graphical evaluation mentioned above can yield additional information about whether the judgments of one subject differ significantly from those of the other subjects (i.e. not only regarding the utilization of the scale, but also regarding the shape of the curve). It may then be necessary to treat this subject's judgments separately rather than including them in the calculation of the mean value.

Of course, the normalization of the data and the exclusion of subjects should not be done lightly. A test operator should never try to "enforce" a desired result by modifying the test data using statistics.

Once the data from the listening tests have been compiled into a mean value or median value, a correlation or regression analysis can be made. For this purpose, additional data for each sound sample are required in addition to the judgments from the listening test; for example, results from a technical analysis. If these results are available as single number values, the similarity between the result curve of the listening test and that of the technical analysis can be determined with a correlation analysis.

In a regression analysis, the data from the listening test and those from the technical analysis are plotted in an X-Y diagram, and the mathematical relation between the axes is calculated. The degree of conformity of this mathematical formula with the actual data is called the coefficient of determination R^2 . A high coefficient of determination means that the results of the listening tests can be described very well with the mathematic formula and the results from the technical analysis. In such cases, no more extensive listening tests are required for new sounds similar to those already examined, since the relevant information can be derived from the technical analysis alone. To achieve a sufficiently high coefficient of determination, it may be necessary to combine the results from several technical measurements or analyses. It is important that not only a high coefficient of determination is achieved, but the formula found must also be interpretable and meaningful. That means that while searching for the optimal coefficient of determination, the results of the technical analyses should not be combined in some random way, but only such that a meaningful interpretation of the combination is possible.

Figure 11 shows a diagram containing a regression analysis including the coefficient of determination R^2 . The X axis shows the Aures sharpness values determined for the sound samples. The Y axis shows the mean values of the subjects' judgments. The judgments are very well represented by the calculated sharpness values.

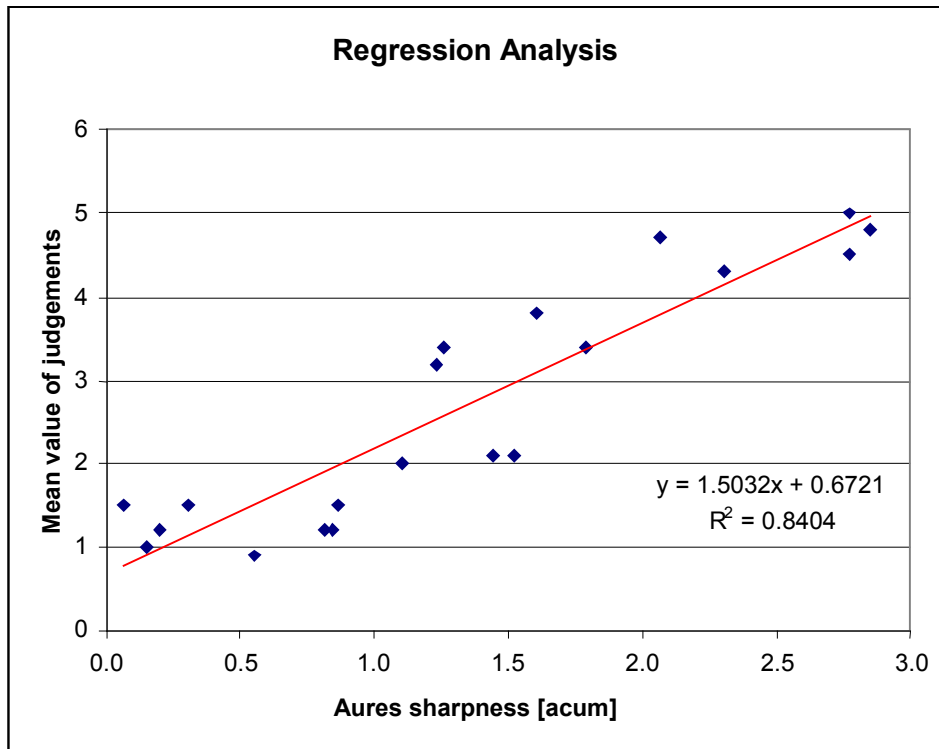


Figure 11: Example for the result of a regression analysis

The results of a listening test with a semantic differential are very extensive, because the subjects enter their judgments on several scales. To reduce the amount of data, the results of this type of test are often subjected to a principal component analysis (or factor analysis). Such an analysis allows a decision as to which judgment items can be combined and how big their influence is on the judgment. Once some factors have been combined, the regression analysis only has to be made for the group factor instead of each single judgment item. Furthermore, it is possible to find the decisive factor crucial for the overall judgment. In future listening tests with similar sounds, it may then be possible to abandon some of the attributes that can be combined into a factor and to examine some new attributes instead that may yield additional information.

Another special situation is the evaluation of listening tests in which non-stationary sounds are judged. When subjects are asked to make a single judgment for a signal that changes over time (e.g. the interior noise of a vehicle starting at a traffic light), they have to summarize their acoustic impression, which also changes over time with the signal. This "internal" averaging by the subject will normally not correspond to the arithmetic mean value of the individual judgments. There will also be little match between the mean value of the results from a technical analysis to the impressions of the subjects. In the case of non-stationary signals, the calculation of percentile values has proven to be a good solution. The calculation of percentile values is a statistical evaluation of the temporal change of an analysis. The percentile values are always shown in combination with a percentage value. The 10 % percentile value is the value that was exceeded only during 10 % of the time period in question. The 50 % percentile value is the value exceeded during half of the time period. Figure 12 shows an example for the 10 % and 50 % percentile values of a level curve. The calculation of percentile values can be done with the ArtemiS analysis software for all types of 2D analysis.

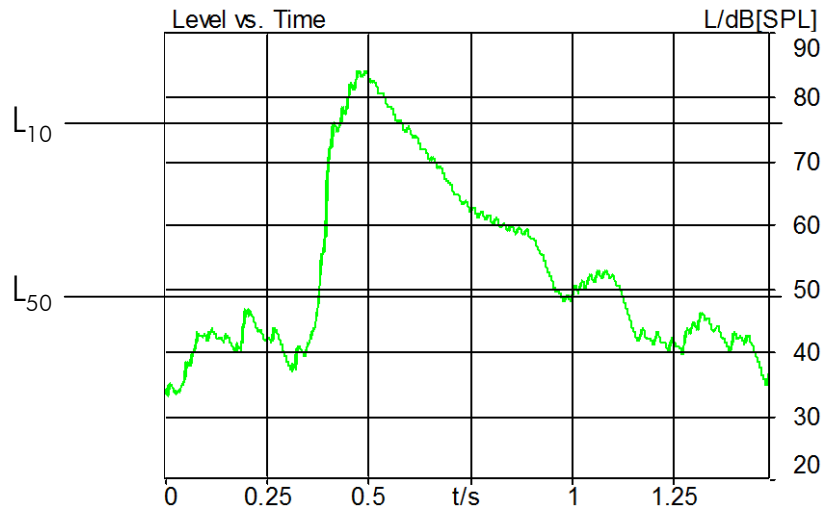


Figure 12: Example for percentile values

In an examination of the annoyance caused by traffic noise, Zwicker found that the 4 % percentile value of the loudness N_4 matches the judgment of the noise by subjects very well. The N_4 value of the loudness is higher than the average loudness value, but indeed the very loud components of the traffic noise are more relevant for the subjects' judgment than the low-level components. This weighting is represented very well by the N_4 value. In the new DIN standard proposal on time-variant loudness (an addendum to DIN 45631), using the N_5 loudness value is suggested for the evaluation of noise emissions. Research has shown that the N_5 loudness value has a good correlation with subjects' judgments of various types of noise (road, rail and air traffic).

Using percentile values allows a statistic evaluation of time-variant technical analyses, whose results match those of a listening test much better than the arithmetic mean value in most cases. For the examination, several percentile values should be determined in order to find out more about the weighting applied by the subjects and to find the optimal percentile value matching their impression.

Generally, the following should be kept in mind in all evaluations: Any mathematic or statistical operation (averaging, excluding subjects etc.) should be carefully considered before applying it. Furthermore, each applied operation should be well documented so it is possible to reproduce the results and know how they were obtained. Only that way, a meaningful interpretation of the results is possible.

Do you have questions for the author? Contact us at: imke.hauswirth@head-acoustics.de.

We look forward to your feedback!