Information on this document

Contents

This document is the second of four application notes on performing jury tests. It describes the most common types of tests used to perform jury tests. A summary lists the advantages and disadvantages of the different test types.

Target group

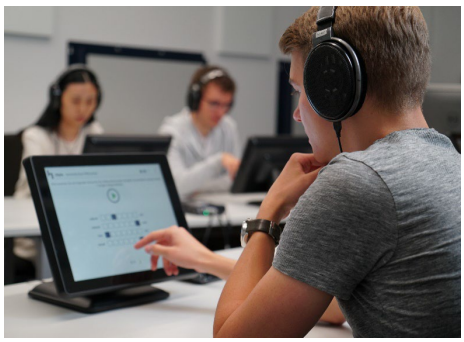
The following text is particularly addressed to (potential) users of the ArtemiS SUITE Jury Test Module SQala who want to be informed about the types of tests available in this software.

Questions?

Do you have questions? Your feedback is appreciated!
For questions on the content of this document: Imke.Hauswirth@head-acoustics.com
For technical questions on our products: SVP-Support@head-acoustics.com

Performing Jury Tests – Part 2

The tests described below represent the most common, established test types in the fields of sound quality and benchmarking.
All these test types can be performed with the ArtemiS SUITE Jury Test Module SQala. To perform other types of tests, the API interface ASX03 can be used.



1. Ranking

Determining a ranking order

In a ranking test, participants are asked to rank a set of N sounds into a ranked order from 1 to N based on a criterion (e.g., annoyance). This task becomes more difficult when there are many sounds to be sorted. For this reason, ranking tests are not recommended when there are many sounds. In most cases, participants can reliably[1] rank up to six sounds.

Figure 1 shows what the interface for a ranking test might look like. By clicking on the play button, the participant can play back the sounds and change their ranking order via drag & drop, for example.

Figure 1: Ranking test in SQala

Ranking order without information on distances

This test method can be used to check the first impression, e.g., customer preference, in a simple, uncomplicated way and can be done in a very short time. Furthermore, the participant can listen to and compare all the sounds directly one after the other. A disadvantage of this test method is that the test supervisor only receives information about the ranking, but not about the distances between the sounds. For example, no statement can be made about whether the distance between the first and second rank is the same as the distance between the fifth and sixth rank.

[1] In this context, „reliably“ means that participants will choose the same ranking order with only minor deviations when repeating the jury test.

2. Paired Comparison

Jury test with paired comparisons

In a paired comparison, participants are presented with two sounds in succession. Each participant is asked to rate these sounds on the basis of a given criterion (e.g., loudness). For this purpose, the participants can choose from two or three judgement options:

A > B	• Example: A is louder than B
B > A	• Example: B is louder than A
A = B	• Example: A and B are equally loud

Figure 2 shows an example of the SQala interface for a paired comparison. The playback of the sounds can be started by pressing the upper buttons. The lower buttons are used for the rating.

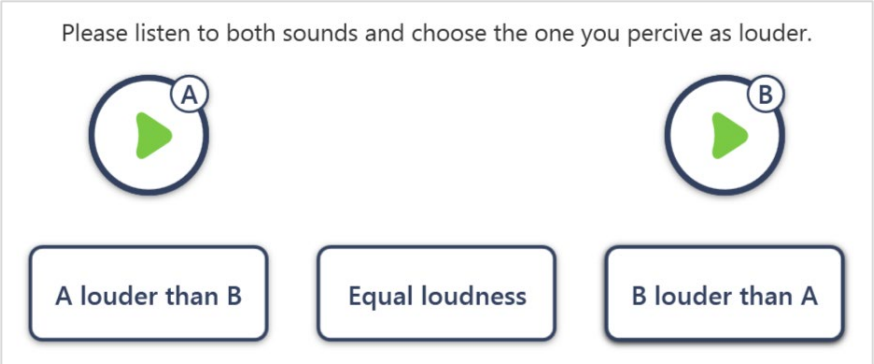


Figure 2: Jury Test „Paired Comparison“ in SQala

Forced Choice

An uncertain participant may tend to avoid making a decision and may frequently choose the A = B response option when offered. This may reduce the validity of the test. To avoid this, the „forced choice“ test variant can be used. In this test variant, a decision is enforced by only offering the response options A > B and B > A. However, it has to be noted that using this test variant puts some pressure on the participant, since a decision is enforced even if the participant does not hear any difference. In this case, the participant has to make a judgement that does not correspond to their perception. Both effects can be reduced by providing a proper test introduction and instruction.

Application area for paired comparison tests

Paired comparison tests are suitable for detecting small differences between very similar sounds. The human hearing is capable, for example, of retaining the sound pressure level in short-term memory, so that even small changes in level can be detected in a paired comparison with sounds presented in quick succession.

However, in everyday life, there is no way to make direct comparisons, for example, when comparing interior noise of a vehicle. These sounds can only be evaluated one after the other and at a certain distance in time from each other, with the human long-term acoustic memory primarily storing sound patterns. For this reason, in the case of sounds that are not presented in direct succession, it is not so much the absolute level that is evaluated, but rather the sound characteristics based on the sound patterns they contain. Therefore, the objectives of a jury test have to be determined before the jury test is performed:

- Is it important to find small differences in sounds?
- Or should sounds be evaluated that cannot be directly compared in practice?

Based on this decision, it can then be decided whether a paired comparison is suitable.

Duration of paired comparison tests

In a jury test with paired comparisons, the test supervisor should not have too many sounds evaluated. If there is a large number of sounds, the test duration increases due to the many possible pair combinations. This is especially the case when both the A – B pairing as well as the B – A pairing are to be tested in order to investigate possible sequence effects. With SQala, it is possible to define in the test properties whether the full matrix (A – B and B – A) or only the half matrix (A – B or B – A) should be evaluated.

Evaluation of results

In a paired comparison, the test supervisor receives an indication for each sound of how often it was preferred to other sounds. When evaluating the results of the paired comparison jury test, a ranking order of the sounds can thus be determined. A statement on how much better one sound is compared to the others cannot be made without further evaluation. There are, however, statistical methods to evaluate the results and thus to make quantitative statements.

3. Semantic Differential

Evaluation with semantic differentials

Using a semantic differential for the evaluation allows a very differentiated examination of the sound samples. Whereas the test methods described above call on participants to focus on a given evaluation criterion, this test method allows several attributes of a sound to be investigated. Participants rate the presented sound on several bipolar scales, the ends of which are labeled with an adjective and its antonym (opposite). The scales used are often seven- or nine-point scales.

Figure 5 shows an example of a semantic differential with a seven-point scale and four attribute pairs.

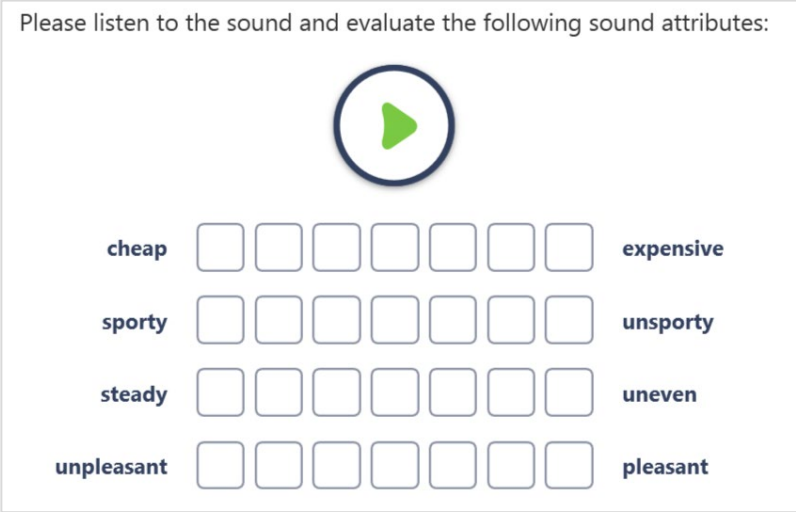


Figure 3: Example of a semantic differential using the SQala user interface

Application area of semantic differential tests

This test method provides the test supervisor with a detailed sound profile that includes much more information than just the statement that one sound is preferred to another and to what degree. Evaluating on multiple scales makes it is easier to find correlations with the results of technical measurement analyses. This allows to find out why a sound was judged negatively and which aspect of a sound needs to be changed in order to improve the sound quality.

The evaluation of a sound with a semantic differential is more time-consuming than other test methods. The number of sounds as well as the number of evaluation criteria should not be too large, otherwise the participants' concentration will decrease towards the end of the jury test. It was found that the number of attribute pairs should not exceed 8-12.

Choosing attribute pairs

For several reasons, great care must be taken when selecting attribute pairs. If the same evaluations are made for all sounds within a jury test, it is possible that inappropriate attributes were chosen for the test.

It must be ensured that the attribute pairs refer to different aspects of the sound. If this is not the case, the answers of the different attribute pairs can correlate too strongly with one another. In such a case, it might have been sufficient to refer to only one of these attributes (e.g., in a category judgement), since referring to the other strongly correlated attribute pairs yields only little new information. In addition, choosing the antonym will have an impact on the evaluation. The following example illustrates this: the

attribute „old" can be assigned to the antonyms "young" and "new". In many cases, the bipolar scale „old – young" will yield other results than the scale „old – new". The objective of the jury test must therefore be taken into account when selecting the antonyms.

International jury tests

When conducting jury tests involving semantic differentials with participants from different countries, each participant should be addressed in their native language. In order to compare the different results of the participants, it is important to ensure that the various attribute pairs are actually translated to mean the same thing.

4. Category Judgement

Jury test with category judgement

In a category judgement, participants are presented with one sound at a time that is to be rated according to a given criterion on a multi-level scale. To assess the intensity of a sound, for example, the DEGA compendium for conducting jury tests recommends the following five-point scale:

not – slightly – fairly – quite – very

As an example, figure 3 shows the rating of sharpness with a category scale.

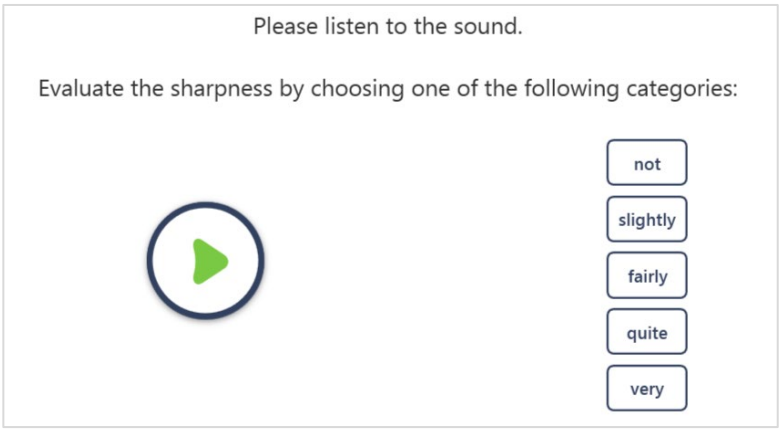


Figure 4: Example of a category scale in SQala

Interval-scaled data

When creating a category scale, care must be taken that the individual categories can be understood as equidistant. Only in this way will the test instructor obtain interval-scaled data that enable them to assign corresponding numerical values to the individual categories. The transfer of the participants' ratings into numerical values allows, for example, a correlation analysis with the results of technical measurement.

Distortion effects and countermeasures

Various distortion effects may occur when rating sounds on a category scale:

- A sound to be rated right after a sound that was perceived as particularly sharp may be rated differently than if it is rated directly after a sound that was not perceived as sharp. Thus, the rating of a sound may be influenced by the preceding sound. This effect is called context effect and can be avoided by repeated rating of the sounds in randomized order.
- Participants often refrain from using extremes. Instead, they tend to give answers in the central range of the scale. In this way, participants want to avoid being surprised by extreme sounds during the test, for which no suitable

category is available. This effect can be avoided by adequate training. During this training, participants can be presented with the most extreme sounds, so that they know what to expect and are in a position to rate them accordingly.

- The scale may be used differently by the participants. This can be compensated by a subsequent normalization (relating the data to the mean value).

5. Simultaneous Category Judgement with Reference

Jury test with simultaneous category judgement

In a simultaneous category judgement test, participants can listen to and compare several sounds. The sounds are then rated individually on a separate category scale. This type of jury test allows the participants to be provided with a reference sound that exhibits a certain characteristic of the sound attribute to be evaluated.

Figure 4 shows an example of a simultaneous category judgement of the loudness of five sounds. In addition, participants have the opportunity to compare the sounds with a reference sound that has the highest loudness compared to the other sounds.

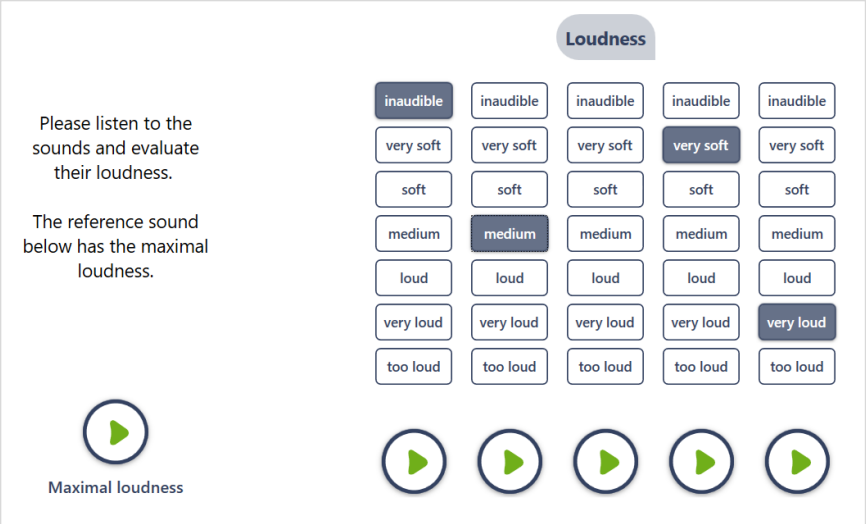


Figure 5: Example of a jury test with simultaneous category judgement in SQala

Evaluation of the jury test results

If a category scale with equidistant categories was used for the rating, the jury test results can be converted into numerical values and used, for example, to correlate jury results with a technical measurement analysis.

Application area for simultaneous category judgement

The advantage of simultaneous category judgement is the possibility to directly compare sounds with a reference sound and other sounds to be evaluated before rating. In this way, even very specific sound phenomena, such as booming or gearbox rattling, can be examined. The possibility of listening to a reference sound again and again allows even non-experts to evaluate a sound phenomenon they are not actually familiar with.

As with the paired comparison, however, the direct comparison creates a test situation that is not necessarily comparable with the everyday situation. In everyday life, a direct comparison is often not possible. Therefore, before conducting a jury test, the

objective must be determined. Only then is it possible to assess whether simultaneous category judgement is the most suitable method for the sounds to be examined and for the participating persons.

6. Summary of the Advantages and Disadvantages of the Different Test Types

Advantages and disadvantages of ranking

Ranking

Advantages:
- Intuitive, uncomplicated judgement
- Judgement is very reliable in most cases
- Quick test method

Disadvantages:
- Ranking only; no statement can be made regarding the distances between the individual positions

Advantages and disadvantages of paired comparison

Paired Comparison

Advantages:
- Intuitive judgement
- Jugement is very reliable in most cases
- Small differences between similar sounds are found

Disadvantages:
- Only feasible with a small number of sounds, otherwise significant increase in test duration
- Ranking only; a statement on distances is only possible by means of statistical calculations

Advantages and disadvantages of semantic differential

Semantic Differential

Advantages:
- Interval-scaled ratings
- Results can be used, for example, for a correlation analysis with technical measurement analyses
- Detailed information on sound characteristics is obtained

Disadvantages:
- Various distortion effects must be considered to receive valid test results
- Time-consuming; only feasible with a manageable number of sounds

Advantages and disadvantages of category judgement

Category Judgement

Advantages:
- Ratings are interval-scaled if a corresponding scale is used
- Results can be used, for example, for a correlation analysis with technical measurement analyses
- Relatively short test duration, even with a larger number of sounds

Disadvantages:
- Various distortion effects must be considered to receive valid test results
- May not be intuitive for unexperienced participants

Advantages and disadvantages of simultaneous category judgement

Simultaneous Category Judgement

Advantages:
- Ratings are interval-scaled if a corresponding scale is used
- Ratings of very special sound phenomena can also be obtained
- Relatively short test duration, even with a larger number of sounds

Disadvantages:
- The rating situation is complex for unexperienced participants
- The test design is more suitable for experienced jury test participants

7. Realization of Jury Tests with Individual Test Designs

Creating individual test designs

The API interface ASX03 in SQala allows the creation of individual test designs beyond the test types presented in this document. This makes SQala individually customizable and enables the implementation of test designs that are not available in the standard software.
Creating extensions with the API interface requires programming knowledge in C#.
For the time being, the realization of jury tests with individual test designs is limited to the local mode.

→ Proceed to the *third application note on jury tests providing basic information on the test procedure.*