APPLICATION NOTE

Information on this document

Contents

This document is the fourth of four application notes on performing jury tests. It provides an introduction to the evaluation of noise judgements obtained by jury tests. The document shows some basic static methods for this purpose, and it also gives hints on what to consider during evaluation.

1. Hints for the evaluation of different test types ____ 2
2. Translation of the jury test results into numerical values ____ 4
3. Statistical analysis of the jury test results ____ 5
4. Further evaluations ____ 7

Target group

The following text is meant particularly to address (potential) users of the ArtemiS SUITE Jury Testing Module SQala who need to get some insight into the different evaluation methods for jury test results.

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 4

Introduction

After the jury tests have been performed, the data obtained are evaluated. A variety of statistical analysis methods are available for this evaluation. On the one hand, these are used to examine and evaluate the data themselves and, on the other hand, these calculations can be used to summarize the data from the jury test and put them into a clear form.

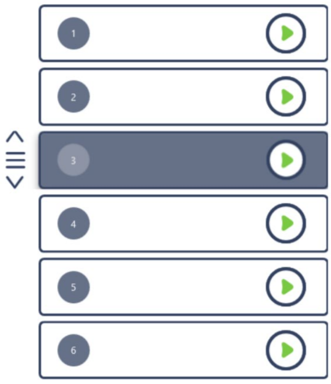
However, before the data can be examined with the help of statistics, they must first be „translated“ into numbers. If the jury test was carried out using appropriate software such as the Jury Testing Module SQala, the test supervisor automatically receives a numerical translation of the judgements at the end of the jury test. The various test methods lead to different evaluations or codings in this case.

1. Hints for the evaluation of different test types

Ranking

Evaluation of the ranking test

In the ranking test method, only rank judgements are given, meaning that it is comparative scaling with no information about the distance between the individual ranks. When evaluating, the test supervisor must not neglect the fact that each sound evaluation largely depends on the evaluation of the other sounds. Averaging the individual judgements of the different participants automatically results in different distances, but it must be decided individually for each jury test whether it makes sense to use this weighting for further evaluation or to convert it again into ranking judgements.



Paired comparison

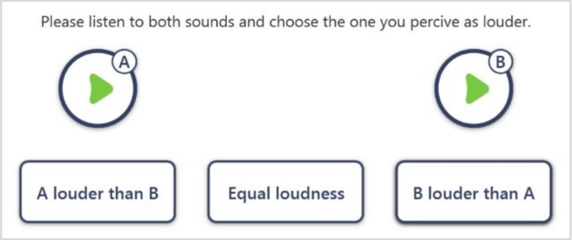
Evaluation of the paired comparison

In a paired comparison, a comparative scaling is performed first (A is better than B). The collected single data can then be easily combined to a ranking order (e.g., the evaluation A>C, C>B results in the order A, C, B). In addition, suitable statistical tools allow the calculation of a scaled order in which the differences between the sounds can also be evaluated. With the help of this scale, correlation studies can be performed. Furthermore, various evaluations can be made for the paired comparison test regarding the judgement reliability of the participants.

Inconsistent triads

This includes, for example, the investigaton of triads. If sound A was evaluated better than sound B and sound B better than sound C, then sound A should also be evaluated better than sound C. If this is not the case, and such inconsistencies occur more frequently, it has to be examined what the reason can be and whether the test design needs to be changed. If inconsistent triads occur with several participants, this may indicate that the participants are overstrained or that the test task was not conveyed properly. Beyond that, the differences between the sounds may be too small for participants to perceive reproducibly.

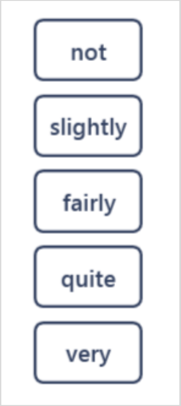
In paired comparison, it is useful to repeat the individual sound pairs several times (also in revese order, i.e., A – B and B – A afterwards). In this way, the repeatability of the evaluation can be checked for each individual participant. This provides additional information about the solvability of the task and the abilities of the participants.


Please listen to both sounds and choose the one you percive as louder.

Category Judgement

Evaluation of the category judgement

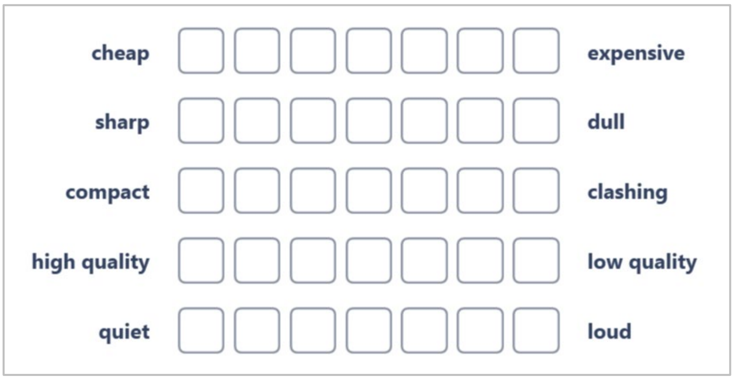
The evaluation of a sound in a jury test with category judgment is more or less independent of the evaluations of the other sounds in the test. If the categories have been chosen in a way that they can be considered equidistant[1], it can be assumed that the data obtained are interval-scaled, and thus the magnitude of the differences can be assessed. This has the advantage that the results of such a jury test can usually be used for a correlation analysis with results from technical measurement analyses. Also, in jury tests with category judgement, participants should evaluate some sounds more than once to minimize context effects and to check for intra-individual differences[2].



Semantic Differential

Evaluation of the semantic differential

In most cases, the results of a jury test with semantic differential are also suitable for correlation studies and thus allow a comprehensive evaluation. The evaluation of a sound with respect to several evaluation items requires more time. For this reason in most jury tests of this kind, not all sounds can be repeated several times, as otherwise the test becomes too extensive and the concentraton of the participants decreases. In most cases, it is reasonable to have at least some sounds evaluated twice in order to check the reliability and intraindividual differenceses of the participants' evaluations.



cheap	□	□	□	□	□	□	□	expensive
sharp	□	□	□	□	□	□	□	dull
compact	□	□	□	□	□	□	□	clashing
high quality	□	□	□	□	□	□	□	low quality
quiet	□	□	□	□	□	□	□	loud

[1] With equidistant categories, the distances between the predefined categories are perceived as approximately equal. So, in the above example for „not", „slightly", „fairly", „quite" und „very", the distance between „not" and „slightly" should be perceived by the participants as the same as the distance between „quite" and „very".

[2] Intra-individual differences are the differences that a participant's evaluations show when the same sound is evaluated repeatedly.

2. Translation of the jury test results into numerical values

Translation of the evaluations into numerical values

In principle, for all test methods, the participants‘ evaluations must be translated into numbers as soon as they are to be subjected to further statistical analysis. The evaluations of a five-point categorical scale are assigned, for example, the numerical values 1 to 5. In the case of a semantic differential with a seven-point bipolar scale, the numerical values 1 to 7 or -3 to 3 can be assigned, for example. Even if the scales on the screen do not always point in the same direction (the negative attributes are sometimes on the left and sometimes on the right), it can be reasonable to assign the numerical values in such a way that the highest value always corresponds to the positive end of the scale and the lowest always to the negative end (or always vice versa). Figure 1 shows an example.

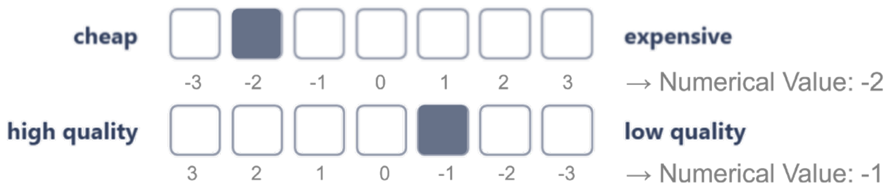


Figure 1: Translation of evaluations into numerical values

With the Jury Testing Module SQala it is possible to define individual values for the individual scale sections of each attribute or attribute pair.

In the following analysis of the evaluations converted into numbers it must be taken into account that these numbers were originally evaluations, e.g., on a categorical scale. The actual evaluations must not be forgotten by the fact that they have been converted into numerical values for statistical analysis purposes.

3. Statistical analysis of the jury test results

Statistical averaging

Once the evaluations of the individual participants are available in numerical values, they can first be plotted and compared graphically. This gives a first impression of the evaluations and helps to decide which statistical tests can be used and whether an averaging of the evaluations of different participants can be carried out.

Histogram and normal distribution

With a histogram of the evaluations given, such an evaluation is usually very easy to make. The histogram shows the number of evaluations for the respective response categories for each sound evaluated.

If the participants‘ evaluations are distributed in a normal way (as shown schematically in Figure 2), averaging can be performed without significant loss of information.

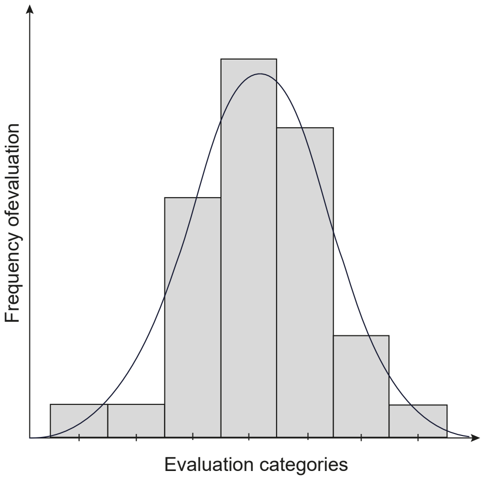


Figure 2: *Histogram with normally distributed evaluations*

Cluster analysis

However, if the distribution shows two or more maxima, it may be useful to divide the partipants into two or more groups in which averaging can then be performed (so-called clustering). This must be decided individually for each jury test on the basis of the data. Various statistics programs provide users with appropriate analysis methods to help them with the evaluation (cluster analysis).

Mean value calculation

Graphical analyses can also be used for other investigations to see if averaging the data is useful. For example, it is possible to check whether the participants‘ scale utilization was comparable by appropriately plotting the evaluations of the different participants.

Median value

In addition to calculating the arithmetic mean value, the median value and the inter-quartile ranges are also frequently determined. The median value is the value that is exceeded by 50% of the evaluations and fallen short of by the remaining 50%. In contrast to the arithmetic mean value, the median value is hardly influenced by extreme values[3], which is why it is usually well suited for the investigation of jury tests in which only a few people participated. In this case, only a few data points are available, which can lead to an outlier strongly distorting the calculation of the arithmetic mean value. In general, the arithmetic mean can be used if the evaluations in the histogram are normally distributed. If this is not the case, the median value should be calculated.

[3] Extreme values are evaluations that differ significantly from the others.

Interquartile range

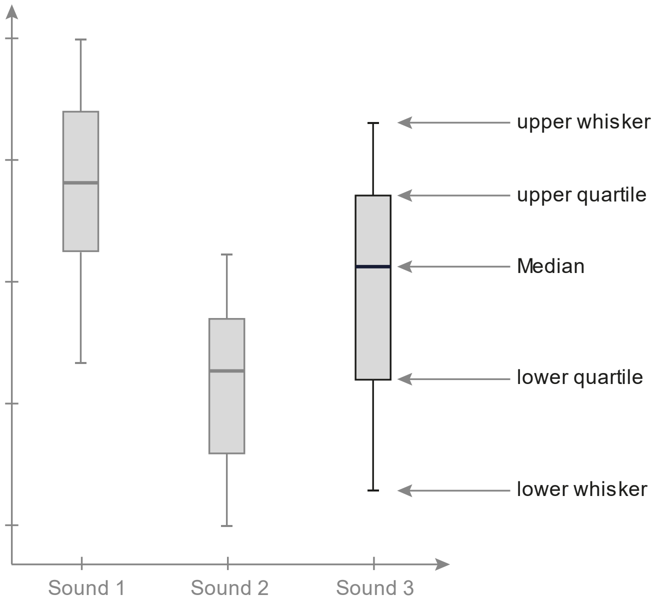


Figure 3: Representation of noise ratings in boxplots

The interquartile range encloses the median value and indicates the range in which the middle 50% of the evaluations are to be found. As for the other 50% of evaluations, 25% are below and 25% are above the interquartile range. The interquartile range thus provides direct information about how widely the individual participants' evaluations scatter. The median value, the interquartile ranges and thus the scattering of the data can be presented very clearly with so-called boxplots (Figure 3). This is very useful to read the data distribution (normally distributed around the median or skewed). In order to visualize the position of the values outside the interquartile range, the boxplot can by supplemented by the display of whiskers. However, the underlying value of these whiskers is not uniformly defined. Often, 1.5 times the interquartile range is plotted, while evaluatons outside this range are usually referred to as outliers. In some cases, the ends of the whiskers also represent the maxima and minima of the ratings.

Standard deviation and confidence interval

Other frequently used statistical quantities are the standard deviation and the confidence interval. When calculating the standard deviation, the mean deviation from the arithmetic mean value is determined. Like the interquartile range, the standard deviation is a measure of the dispersion of the evaluations. The smaller this value is, the more similar the sound was rated by all participants. The confidence interval indicates a range in which the result of a rerun test is expected to lie. For example, the 95% confidence interval shows the range in which the result of an additional test will lie with a probability of 95%.

Using statistics properly

The graphical evaluation already mentioned can give additional indication whether the evaluations of one participant are very different from those of the other participants (i.e., not only the scale utilization but also in the shape of the curve). The evaluations of such participants may then have to be considered separately and must not be included in the calculation of the mean value. The exclusion of participants must not be taken lightly. The test supervisor must not use statistics to alter the data of a test so as to „calculate" the desired result.

4. Further evaluations

Correlation analysis

After the participants‘ evaluations have been consolidated into a mean value or median value, a correlation and regression analysis can be performed. This requires not only the evaluations from the jury test, but also additional data for each noise, e.g., the results of technical measurement analyses. If these are available as single values for each evaluated sound, the similarity of the curve from the results of the jury test and the technical measurement analysis can be determined with the help of the correlation analysis.

Regression analysis

In the regression analysis, the data from the jury test and the data from the technical measurement analysis are plotted in an XY plot, and the mathematical relationship between the axes is calculated. The degree of agreement between this mathematical formula and the actual data is the coefficient of determination R^2. A high coefficient of determination indicates that the results of the jury test can be reproduced very well using the mathematical formula and the results from the technical measurement analysis. Figure 4 shows a simple example of a linear regression analysis. The X-axis of the diagram shows the individual values of a technical measurement analysis calculated for the sounds. The Y-axis represents the mean values of the participants’s evaluations. The participants’ evaluations are well represented by the calculated analysis values in the example shown.

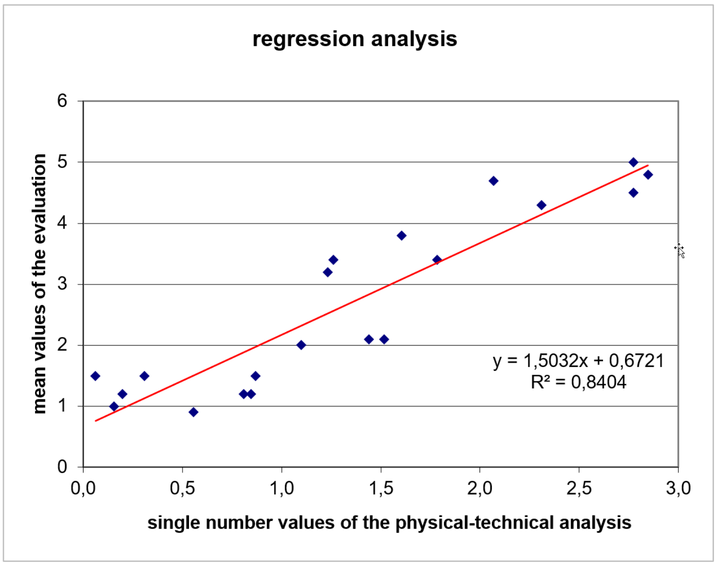


Figure 4: Example of the result of a regression analysis

Using metrics

For a sufficiently high coefficient of determination, the results of several technical measurement analyses may have to be combined. In ArtemiS SUITE, this can be done automatically using the Metrics Project. Details on how to develope a sound quality metric are provided in the application note „Metric development“.

Development of robust metrics

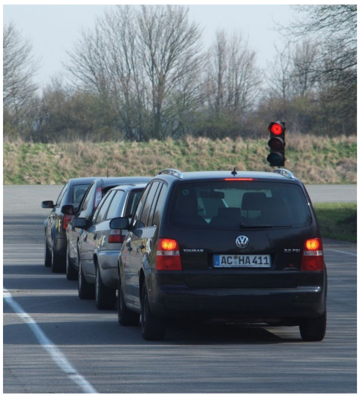
In principle, during metric development, a high correlation between the results of the jury test and the results of the technical measurement analysis is to be achieved. It is not expedient, however, to map the jury test results with a large number of single number values from many different technical measurement analyses. In developing metrics, the influence of each single number value must be systematically examined and selected with respect to its causal relationship with the sound quality of the sounds being evaluated. In most cases, it is more useful for creating a robust metric[4] to use only a small number of analyses to avoid overfitting the model. Provided a robust metric was created, the evaluations of sounds with similar characteristics can then be predicted computationally using the mathematical formula and the results of the technical measurement analysis. In order to confirm the quality of the prediction, additional validation jury tests can be performed.

Principal component analysis

The results of a jury test with semantic differential are very extensive, as the participants make their evaluations on several attribute scales. In order to reduce the amount of data, the results achieved with this test method are often subjected to a principal component analysis. Such analysis can be used to determine which attribute pairs can be grouped together and how much influence they have on the evaluation. Once some attribute pairs can be grouped together, the regression analysis only needs to be performed for the higher-level principal components and not for each attribute pair individually. In addition, the components that are crucial for the overall evaluation can be found. If further jury tests with similar sounds are to be performed, it is possible to omit some of the attribute pairs that could be grouped together. This allows to query new attribute pairs that provide additional information.

Examination of non-stationalry signals

Another special feature is the evaluation of jury tests in which non-stationary sounds were evaluated. Non-stationary sounds change as a function of time (e.g., the sound of driving when starting at a traffic light or when passing a vehicle). When participants are asked to evaluate such a signal, they must combine their sound impression, which, like the signal, may change over time, into one evaluation. This „internal" averaging on the part of the participants will generally not correspond to the arithmetic mean value of the individual evaluations. Also, the mean value of a technical measurement analysis will in many cases not reflect the impression of the participants.



Percentile values

In the case of non-stationary signals, the calculation of percentile values has proven to be useful. The calculation of percentile values is a statistical evaluation that examines the value distribution of the technical measurement analysis. Thus, if the value *10* is entered in the percentile value table on the properties page of an analysis in ArtemiS SUITE, the single number value that is exceeded by 10% of the analysis results is determined, and so on. For time-dependent analyses in combination with entry *5*, ArtemiS SUITE determines the value that is exceeded in 5% of the time during the evaluated period. Figure 5 shows an example of a time-dependent loudness analysis. In addition to the time-dependent loudness curve, the 5%, 10% and 50% percentile values are marked. With ArtemiS SUITE, the single number values can be displayed either in the diagram or in a single number value table.

[4] Guidance on creating robust metrics is provided in the following publication: Fiebig, Kamp; "Development of metrics for characterizing product sound quality", Proceedings Aachen Acoustics Colloquium 2015, 123-133.

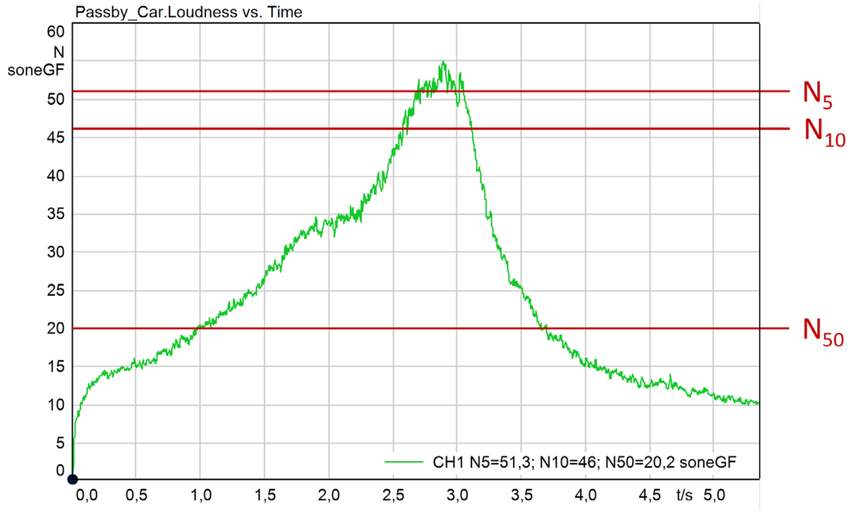


Figure 5: Example of percentile values

N5 loudness

In many cases, percentile values correlate significantly better with the results from jury tests than the arithmetic mean value. Studies of the annoyance of traffic noise showed that the 5% percentile value of loudness (N5) correlates very well with the evaluation of noise in a jury test. This value is higher than the average loudness value, but the loud components of traffic noise are also much more prominent in the evaluation by the participants. ISO532-1 therefore stipulate the calculation of N5 loudness as a single number value for time-varying sounds.

For the correlation test, different percentile values should always be determined to learn more about the weighting performed by the participants and to find the appropriate percentile value.

In summary, the following should be kept in mind during evaluation: each mathematical operation (averaging, exclusion of a participant, etc.) must be chosen and performed with care. In addition, each operation performed must be carefully documented to record the basis on which the results were obtained. Only in this way can a meaningful interpretation of the results be made. A more comprehensive introtuction to the statistical evaluation of test evaluations can be found, for example, in books on test methods and their evaluation for human and social scientists.