Jury-Test Methods for the Measurement of Perceived Quantities

M. CODDA, F. CRENNA, G.B. ROSSI
University of Genova, Dept. of Mechanics and Machine Design
Via opera Pia 15 A, 16145, Genova, Italy
Email: rossi@dimec.unige.it

ABSTRACT

The measurement of perceived quantities is of great interest in nowadays world, where man often interfaces with machines through his senses. This works presents a thorough investigation of the jury-test methods for the construction of reference measurement scales and for the subsequent development of suitable measurement procedures. Cross-validation of complementray tests is addressed, for assuring the reliabilty of the results, as well as the importance of test protocols and of operators' interface. As a support to the proposed methodology, two test cases, concerning automotive noises, are presented in details, showing encouraging results. This study is also intended as to provide support for a standardisation of this kind of methods.

Keywords: measurements of weakly-defined quantities, jury tests, measurement scales, noise quality

1. INTRODUCTION

Ergonomics is an important aspect of the quality of life, since it is aimed at improving the relationship between man and its working or living environment [1-2]. One of the major challenges in ergonomics is the assessment of such relations, which involves measurement of perceived quantities. As far as perception is involved, a high challenging measurement problem arises, including the definition of the measurement scale itself [3-4] and of the related measurement procedures [5].

The basic approach for assessing measurability, constructing a proper measurement scale and finding at least one suitable measurement procedure, is based on jury testing [6]. Such tests may be configured in a wide variety of ways, depending upon features such as the number of stimuli to be compared, the way they are presented, the kind of question which is made, the kind of answer required, the role, active vs. passive, of the subject and so on [7-14].

Now their design is highly critical, for several reasons and a systematic approach to the complex problems involved in the design of jury tests may be, in our opinion, highly beneficial, as a measurement-theory topic.

So in the paper we propose an approach grounded in basic measurement theory, considering the design and implementation of two fundamental kinds of jury tests, namely pair-comparison and magnitude-estimation.

After a brief review of some necessary theoretical basis, we consider both the construction of the reference measurement scale, by selection of a proper series of standards, and the development of a measurement procedure based on that scale. The discussion is supported by two experimental test cases, concerning the measurement of sound quality in the automotive field. Precisely, they concern the measurement of the "pleasantness" of noise due to car-doors closing and to internal vehicle noise.

2. THE THEORETICAL STARTING POINT

According to the "representational" theory, measurement may be defined as an empirical process allowing assignment of numbers to attributes of objects or events, in such a way as to represent relations between them [16-17]. This formally requires three ingredients:

1. an empirical relational system, which is a set of manifestations of the attribute to be measured (i. e. of the measurand) and a class of relations on that set;
2. a numerical relational system, i. e. a set of numbers and a class of relations on it;
3. a function from the set of manifestations to the set of numbers, having the property of mapping the empirical relations to the corresponding numerical relations.

Each suitable configuration of this triple defines a specific *measurement scale* (this term is here used in the most general sense; it will be used in a more specific sense later on). Let us now consider the following three kinds of scales: ordinal, interval and ratio. For these scales we have a formal theory [3-4]. So, for instance, we know in advance which is the appropriate numerical relational system, we know that, as long as the assumed relations actually hold, it is *formally* possible to do the measurement and we know the exact *meaning* of the measurement result. For ease of reference, the main characteristics of the scales are summarised in Table 1

Scale type	Empirical relations	Admissible transformations
Ordinal	Order between manifestations	Monotone increasing
Interval	Order also between distances	Positive linear
Ratio	As above plus empirical addition (positive concatenation)	Similarity

Table 1. Main characteristics of some measurement scales

So let us now consider the problem of measuring a candidate new quantity (in this paper, quantity means "measurable attribute"). The task may be organised in three distinct and sequential steps:

1. *construction of a "reference scale"*[1], which may be understood here as a *series of standards*, with associated numerical values;
2. *definition of (at least) one measurement process*, based on the scale obtained in step 1, which will be the *primary* measurement process for the quantity under investigation;
3. *definition of other (derived) measurement processes*.

Of these steps, only the first two are strictly necessary, whilst the third may be also of great interest for various reasons, both scientifical (better understanding of the way peception takes place) and practical (possible cost reduction). In this paper we concentrate on the first two, to be discussed in detail in the following paragraphs, using the above mentioned experiments as examples.

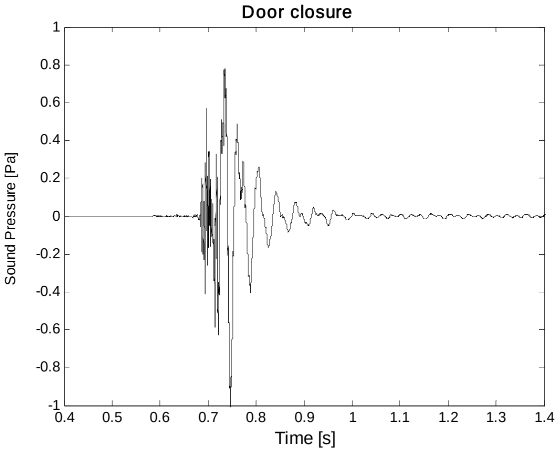

Figure 1: Time history of a car door closure.

3. CONSTRUCTION OF THE MEASUREMENT SCALE

The construction of the measurement scale is based on the concept of *empirical relational system*. Let us then examine it in details. The empirical relational system consists of:

- the set of the possible manifestations of the measurand;
- the class of the empirical relations of interest, which are characterised by their formal properties.

These two points are *extremely important* and a reflection on them provides precious guidelines on how to operate.

3.1 Manifestations (the perception space)

In considering a new quantity, it is necessary to define the set of the possible manifestations and to collect a representative sample of them. Now, when dealing with a traditional physical quantity, a manifestation may be identified with an object[2] carrying a specific value of the quantity. For example, in the case of length, the object may be a gauge block, with a couple of parallel plane surfaces, whose distance is a manifestation of length. In the case of a perceived quantity, we must perhaps be more aware that a manifestation implies not only something which is manifested, but also someone who may receive the manifestation[3]. So in the case of quantities related to perception, we have to consider carefully both the set of objects manifesting the quantity and the set of subjects perceiving it. Such a set may include, in the most general sense, all possible human beings, but in some cases it may be restricted to a suitable subset, for instance considering only people in a given range of ages, or performing a specific kind of work, or habitually using a certain kind of devices, and so on. Moreover it is also important to define a *standard interaction* between objects and subjects, in order to prevent unnecessary dispersion in the observations.

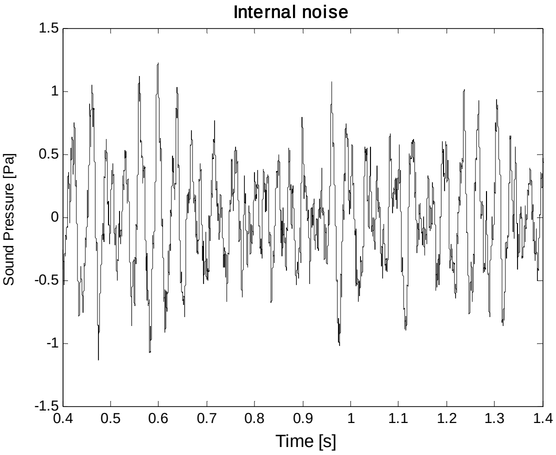

Figure 2. Time history of an internal noise

Let us then consider our reference experiments, as an application example of these ideas. First of all, for the selection of the "objects", which in our case are sound records, we have considered two classes of stimuli quite different from each other. Door-closing signals are transient-impulsive (Fig. 1), while interior car noises, at a fixed regime, are essentially stationary processes (Fig. 2). Then we have to pay attention to the above mentioned *standard interaction* between the manifestations and the subjects perceiving them. This requires *two kind of protocols*, both for recording and for reproducing and presenting signals.

For data acquisition, we have used binaural recording, at a fixed distance from the door in one case, inside the vehicle in the other one. As a further normalisation item, the speed of closure of the door in the first case and the speed of the car in the second one, have been measured For reproduction, the stimuli are presented to the subjects in a chamber with phono-absorbing walls, fixing the

[1] Here we use the term „scale" in a more specific way.

[2] From now on we use the term „object" with the meaning of „carrier" of the manifestation, irrespective from it actually being a material object or an event (such as a sound).

[3] Actually this is also true, in a sense, in the case of a physical quantity: for example in the case of length, an alternative manifestation may be provided by a couple of straight-line marks on a plane surface, the difference being in the way the manifestation may be detected. Detection is somehow the instrumental counterpart of human perception.

listening distance and normalising the signals with respect to their sound pressure level.

In order to complete the definition of the perception space, it is necessary to *clearly define the perceived characteristic to be judged*. In this sense, before executing the tests, a preliminary verbal description of the stimuli and of the quality to be judged may be useful in order to provide the jury with a uniform vision over the perception space.

In some cases a formal training of the jury is adopted before the test execution. This may be useful for very particular quantities in order to have a uniform way of looking to the stimuli inside the overall jury. On the other hand, training can influence the jury, which then will not be a standard sample of the population but a conditioned one. Another factor to be considered is the behaviour of each subject during the test: a selection of only the good performing subjects, according to their training results, may heavily influence the jury composition.

In our experiment we have made no training and we have made some test for assessing the reliability of each judge, to be used, if necessary, for a post processing of results.

3.2 Empirical relations (the perception mechanism)

This is perhaps the most critical point.

The results of measurement are meaningful as far as they actually reproduce, in a numerical space, relations which actually takes place in the real world. Now, when we measure a quantity we essentially *assume* some kinds of empirical relations *to hold* between the manifestations of the quantity to be measured, and the goal of measurement is to get information about them. It is important to consider which is the foundation of that assumption. For physical quantities, in general, at the present (hystoric) state of scientific and techological development, we may recognise such a foundation in our acceptance of mathematical and physical laws and in the relative derived models. For instance, in the case of gauge blocks, we are usually ready to accept that an empirical relation may be defined through the empirical operation of piling up two or more equal-basis blocks: the resulting block will manifests a length (height) which will be the "sum" of the length manifested by the constituting blocks. Although quite intuitively acceptable, this assumption may be founded on the corresponding geometrical model of a bundle of parallel planes. On the other hand, in the case of perceived quantity, no such foundation is usually possible. So the process of the construction of the scale has in charge *the assessment of measurability* as well.

So the scientifically correct procedure could be: assume an empirical relation first, design and perform a suitable jury test, process the result and check the initial assumption. Only if this validation is fulfilled, it is possible to accept the result and, eventually, move to the assumption of a stronger empirical relation, repeating the same kernel procedure.

This approach is surely not easy to follow, but it is, in our opinion, somehow necessary, if we want to get really reliable results. In order to be more specific, we will discuss this procedure in the following paragraphs, with reference to two fundamental kinds of scales and their related jury tests, namely: order scales and related paired comparison tests and interval scales, with associated magnitude-estimation tests.

Before doing so, we need at least to mention the importance of the test protocol and of the interface between the stimuli and the juries. In general we may say that the perception of the relations takes place in the mind of the subject in a quite confused way, so the test must be designed in order to help the subject in getting proper awareness of its own perception. So details such as the way the question is formulated, the possibility of repeated listening of the sounds, of having a global vision of the perception space and so on, have great importance and will be also touched.

3.3 Order scales and paired comparisons tests

The first empirical relation to be tested should be order. Weak-order is formally characterised by transitivity and strong completeness.

Now transitivity is usually the most critical one. Surely it can not be interpreted in a *deterministic sense*, i. e. as a property applying *to each couple of objects, in any observation, by any subject*. It may be rather assumed to hold in mean, which may be formally expressed by the *weak probabilistic transitivity* [4] assumption.

That is, if a, b and c are objects in the set, and $P(x \precsim y)$ is the probability of x not to preferred to y, then we assume that:

$$P(a \precsim b) \geq 0{,}5 \ \& \ P(b \precsim c) \geq 0{,}5 \rightarrow P(a \precsim c) \geq 0{,}5 \quad (1)$$

This property is very useful, since it may be checked after the test, by obviously replacing the probabilities, with suitable relative frequencies. If it holds, another relation may be defined as:

$$a \precsim_0 b \leftrightarrow P(a \precsim b) \geq 0{,}5 \quad (2)$$

This new relation may be proved to be, for a finite set of manifestations, a weak order [18]. So ordering may be achieved by considering relation $\precsim_0$ instead of the originary relation $\precsim$. The new relation is again an empirical one, since it is based on empirical observation, but it is somehow a mean, average relation.

Now the basic test for assessing order is paired comparison. In this kind of test, a set of couples of stimuli is presented to the subject, who has to decide each time which one he prefers, according to the perceived characteristic established. For example: 'choose which one is less annoying' or 'choose the sound you prefer for this product during this function'.

If the set of stimuli is large, the number of couples will be too large to be judged by a single subject. Besides there are some constraints to be satisfied in order to avoid biasing of the results, such delaying a new presentation of the same stimulus as much as possible and inverting the order in the couple. For these reasons, a random extraction of stimuli or couples may be not satisfying, in terms of speed of convergence and an optimised sequence may be prefereable. We have implemented such a sequence, fulfilling the above requirements and have also introduced in the testr protocol of each subject the possibility of verifying his/her coherence. During the processing it is possible to consider or not this possibility, and in any case it will contribute to the statistical consistence of the results. The following tables presents results of the paired comparison test in the case of noises due to door-closing, for 32 subjects, involving the examination of about 500 couples of signals.

Table 2 presents the probabilities (or relative frequencies) for each signal in a row to be preferred over each signal in the columns. Signals are presented in the raw order due to the formal identification. In order to establish if it is possible to define an order, it is useful to order the matrix from the less preferred to the most preferred one.

	A	B	C	D	E	F
A	-	0.28	0.12	0.16	0.07	0.26
B	0.72	-	0.09	0.47	0.25	0.50
C	0.88	0.91	-	0.77	0.34	0.85
D	0.84	0.53	0.23	-	0.27	0.62
E	0.93	0.75	0.66	0.73	-	0.76
F	0.74	0.50	0.15	0.38	0.24	-

Table 2. Raw probability matrix for noises due to door closing.

Table 2 presents the ordered results. It may be easily checked that the weak probbailistic transitivity actually holds

	A	B	F	D	C	E
A	-	0.28	0.26	0.16	0.12	0.07
B	0.72	-	0.50	0.47	0.09	0.25
F	0.74	0.50	-	0.38	0.15	0.24
D	0.84	0.53	0.62	-	0.23	0.27
C	0.88	0.91	0.85	0.77	-	0.34
E	0.93	0.75	0.76	0.73	0.66	-

Table 2. Ordered probability matrix for noises due to door closing.

In this case signals B and F are equivalent, so only one of the two needs to be included in the series of standards.

In the case of the internal car noises (not reported here for brevity) the ordering was also possible, the transitivity condition was fulfilled and non equivalence condition emerged.

3.4 Interval scales and magnitude estimation tests

Magnitude estimation (ME) is another powerful test, with characteristics somehow complementary to paired comparisons. Depending on the way it is implemented, it may give rise either to ratio or to interval scales. For instance, the test may be implemented by fixing a reference signal, usually called “anchor”, with an arbitrary assigned value, and asking the subjects to assign values to other signals, as a ratio value with respect to the reference one [6]. In this case it may be noted that the empirical relation, i. e. the perceived relation, is actually a ratio, rather then an empirical addition. This makes a difference with physical measurement. Formally this way of constructing a scale based directly on a ratio, may be justified as a kind of indirect measurement [4].

Now this way of performing magnitude estimation may result in a great dispersion. An alternative way is that of providing two or more reference points. In this case the result is rather an interval scale, since the subjects consider intervals rather then ratios.

We followed this second alternative in designing our test. We wanted to fulfil basically two requirements: the possibility to have an active subject with a free interaction with the stimuli set and the need for the consciousness *of the full perception space* before estimating the magnitude associated with each stimulus. The test realised gives the possibility to the subject to interact with a panel on which several buttons correspond to the different stimuli to be investigated. Clicking on a button corresponds to the application of the stimulus, which can be listened. Then the button can be moved according to the perceived quantity to be evaluated. So the subject can freely perceive the various stimuli all the times necessary to define the perceptional space of what is to be judged, and to establish the magnitude of each stimulus in the space.

In doing so, he/she actually orders the signals and the distances between the signals, which is exactly the empirical relation (order between distances) defining an interval scale. According to the literature examined, this kind of test configuration seems to be rather new. It is important to stress the complementarity of this test with the paired comparison: in the paired comparison the subject examines a couple of signals each time: he will focus on the difference between the two, without having a global vision of the perception space; so, for instance, if the signals are very close, the subject may focus more on the details; on the other hand he will not have a vision of the whole perception space, his response may show inconsistency (violation of the transitivity condition).

So, summarising, we may say that paired comparison is mainly a *local approach*, whilst magnitude-estimation, as implemented in this study, is rather a *global approach*.

So it makes sense to look for a cross-validation of the two tests, checking their agreement.

A comparison of the results from the two test procedures, after judgements from a jury of 32 people, is presented in figures 3 and 4[4].

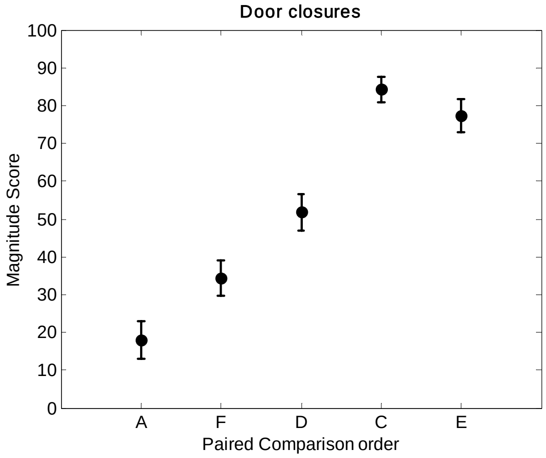

Figure 3. Comparison of the results from the two test methods, noises from door closures. Bars represent mean scores standard deviations

[4] Due to the different approach for each subject to the space available, it is necessary to normalise the set of scores for each subject on a fixed scale, in order to proceed with the evaluation of the mean and dispersion values.

They present the scores according to the magnitude estimation vs the order according to the paired comparison test. For the noises due to door closing there is a slight disagreement in the top of the scale, while for internal noises there is a complete correspondence.

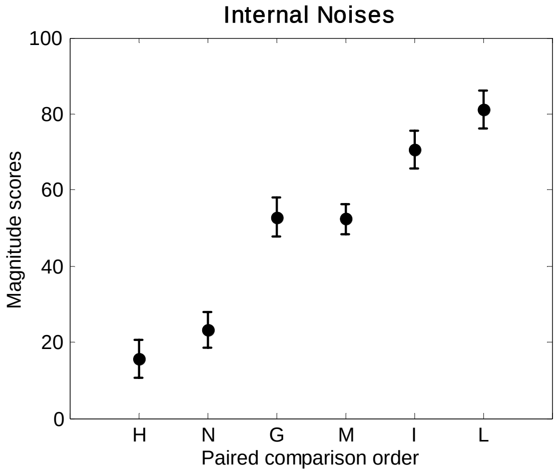

Figure 4. Comparison of the results from the two test methods, internal noises. Bars represent mean scores standard deviations

The disagreement may be explained at the light of a more detailed analysis. As a general trend, "smoothness" of signals grows as long as we proceed towards highest value in the scale. On the other hand, signals C and E, which are both quite smooth as compared to the others, differs in that E is a more "well defined" sound. So in their direct comparison, signal E is preferred, since it appears better characterised. On the other hand, when a subject is asked to order the whole set of signals, he will follow an internal overall criterium (i. e. a global criterium) mainly based on "smoothness", according which signal C actually follows E, since it is a little bit smoother. Now, in our opinion, this example shows how the cross validation is highly sensitive: so on one hand, it may provide precious information for a better understanding of the perception mechanism, as in this case, on the other hand when agreement is found, as in the case of internal noises, it gives much more confidence on the reliability of results.

Finally, what to do when, as in the case of doors closure, there is some disagreement?

Our suggestions are the following:

- first of all, disagreement should be limited to adjacent signals, as in this case, otherwise some measurability problem should be considered;
- then a wise strategy could be to remove one of the two conflicting signals from the scale, so obtaining a lower resolution scale, which is more "robust ";
- resolution of the scale may be improved later, by adding new samples, which prove to be stable; for instance the same rejected signal may be "measured" with respect to the lower resolution scale and its place may be found in this way.

Finally, we may note that in our example, if we consider the response of each member in the jury, and we eliminate the responses of judges which have an incoherent behaviour, we obtain an agreement with E better then C, as predicted by the paired comparison test.

4. THE MEASUREMENT PROCESS

After a scale has been constructed, it is possible to define a measurement procedure on it. In general we may say that this may be implemented *by allowing a subject to compare the measurand with the scale*. In order to do this properly, it is necessary that the subject may have a global view of the perception space as spanned by the scale (the series of standards). So the *interface* is of utmost importance. We designed a graphical user interface similar to the one used for the magnitude estimation test, as depicted in figure 5. The width of the window represents the full scale available for measurement. In the upper part of the window a series of buttons, representing the reference samples constituting the scale, are positioned according to their scores, from the lowest on the left, to the highest on the right. The subject can listen all the times needed to all the samples by clicking on the corresponding buttons, but he can not move them in any case. In the centre of the lower part of the window there is a button corresponding to the signal to be measured. The subject can listen to the signal by clicking on the button or he can move it until he reaches a position satisfying his perception. In such a way the measurement procedure is something like a measurement by comparison of the measurand with each sample on the scale.

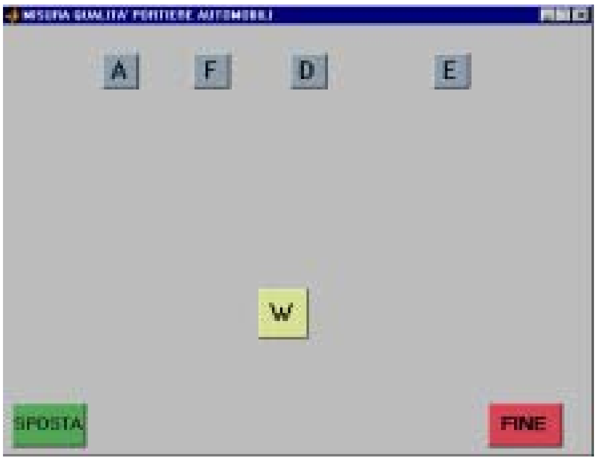

Figure 5. Graphical interface for the measurement test

The measurement procedure was applied to three new noises from door closing. They were measured on a scale constituted by only the four "robust" samples, over which there is full agreement between the two tests. Results from a jury of 21 people are presented in figure 6. Bars indicate the standard deviations of the mean score of each measured signal and of each scale sample. From the subject point of view the measurement procedure requires a judgement duty much easier than the construction of the scale by the magnitude estimation method. During measurement the perception space is clearly marked by the reference samples, which are in some way imposed to the subject. On the other hand during the construction of the scale the subject himself has to establish the subdivision of this space, since when starting the test it is completely void and only its extremes are defined not by perception but by numbers.

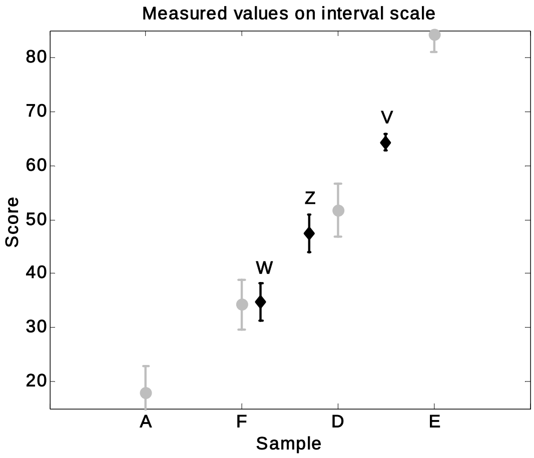

Figure 6. Measurement results. Bars indicate the standard deviation of the mean values.

This explains why the dispersion of the mesaurement is lower than that of the reference standard. Of course the uncertainty of the mesaurement will include both effects. We may use the following model:

$$\hat{x} = x_S + d \qquad (3)$$

where x_S is the value of the nearest element of the standard series, and d is the difference between the value assigned to the measurand, after averaging over the n responses from the jury, and the value of the nearest standard. So the combined standard or expanded uncertainty of the measurement result may be easily obtained.

To give an idea of the resulting uncertainty, in round figures, we may say that with a jury of n=20 subjects, we obtain a standard uncertainty of about 6% of the full range of the scale. The overall measurement process requires about two hours.

8. CONCLUSIONS

The measurement of perceived quantities based on jury tests has been investigated. We have considered both the complex process of constructing a reference measurement scale and the subsequent task of defining at least one complete measurement procedure. For the construction of the scale, we have considered two fundamental kinds of tests, paired comparisons and magnitude estimation, pointing out their complementary characteristics. We have proposed a step by step procedure, including careful checking of assumptions and cross validation of the two tests. As far as a reliable measurement scale is obtained, it is possible to develop an efficient measurement procedure, based on that scale. The overall procedure has been implemented and checked on two classes of sound fenomena, quite different from each other (transient-impulsive vs. stationary), of industrial interest. The importance of testing protocols and of operators' interfaces has also been stressed, and example solutions have been documented. The results seems to be encouraging, since the validation procedure has proved to be sensitive and informative and the final measurement procedure seems to be acceptable in terms of efficiency and uncertainty, for the kind of measurement considered.

So, in our opinion this study may provide useful information for the standardisation of this kind of methods, and, as a subsequent step, for the development of accepted metrological standards and methods.

Should this be the case, considerable benefit could come from the diffusion of these methods, both on the scientic side, for a better understanding of human perception and on the application side, for the assessment of the "perceived quality" of products and environment.

ACKNOWLEDGEMENTS

The authors are grateful to Centro Ricerche Fiat, especially for the experimental set up. Special thanks to Dr. Giorgio Irato for precious advice and encouragement.

9. REFERENCES

[1] Meister D The history of human factors and ergonomics, Lawrence Erlbaum Ass, 1999

[2] Salvendy Handbook of human factors and ergonomics Wiley II, 1997

[3] Krantz D R et al., 1971-1990 Foundations of Measurement, 3 vols. Academic Press, New York

[4] Roberts F S 1979 Measurement theory, Addison Wesley, Reading MA

[5] Muravyov S, Savolainen V Special interpretation of formal measurement scales for the case of multiple heterogeneous properties, Measurement, 29, 2001, 209-224

[6] Purghé F. Methods of psycho-physics and uni-dimensional scaling (in Italian), Bollati Boringhieri, 1999

[7] N. Otto, S. Amman, C. Eaton, S. Lake, "Guidelines for jury evaluations of automotive sounds", Soun and vibration, 2001, pp. 1-14

[8] Engen T., Psychophysics. I Discrimination and detection, II Scaling, in Woodworth & Schlosberg's

[9] M.G. Kendall, *et alii*, On the method of paired comparison, Biometrica, vol. 31, 324-345 (1939)

[10] Kling J K, Riggs L A (ed.), Experimental Psychology, pp. 11-86, Methuen, London 1971

[11] Thurstone L.L., A law of comparative judgments, Psychological Review, vol. 34, 273-86 (1927)

[12] Thurstone L.L., Psychophysical Analysis, American Journal of Psychology, July (1927)

[13] Wherry R.J., Orders for the presentation of pairs in the method of paired comparisons, Journal of Experimental Psychology, vol. 23, 651-60 (1938)

[14] Crenna F Ferrari B Rossi G B Weber R Measurement of the noise quality of post-sorting machines by jury testing ICA Congress Rome 2-4 Sept. 2001

[15] Zwicker E and Fastl H Psycho-acoustics, Springer Verlag, 1999

[16] L. Finkelstein , M. S. Leaning, „A review of the fundamental concepts of measurement", *Measurement*, Vol. 2, No. 1, 1984, pp. 25-34

[17] M. S. Leaning, L. Finkelstein, "A probabilistic treatment of measurement uncertainty in the formal theory of measurement", *Acta IMEKO 1979*, G. Striker ed, Elsevier, Amsterdam, pp. 73-81

[18] R. C. Michelini, G. B. Rossi, Measurement uncertainty: a probabilistic theory for intensive entities", Measurement, Vol. 15, No. 3, 1995, pp. 143-157