# A method for perceptual evaluation of products by naive subjects: Application to car engine sounds.

Emilie Poirson, Jean-François Petiot, Florent Richard

# A method for perceptual evaluation of products by naive subjects: application to car engine sounds

Emilie Poirson[1*], Jean-François Petiot[1], Florent Richard [2]

[1]*Institut de Recherche en Communications et Cybernétique de Nantes (UMR CNRS 6597) Ecole Centrale de Nantes - 1, rue de la Noë, BP 92101, 44321 Nantes, Cedex 3, France*

[2]*PSA Peugeot Citroën, DRIA/SARA/EMSA/PEFH - 2, route de Gisy, 78943 Vélizy Villacoublay, France*

**Abstract**

Vehicles makers are now extremely concerned by the perceived quality of their products. For the design of many parts of vehicles, sensory profiling techniques are used, which are traditionally carried out by experts. These tests incur a significant financial effort and are time consuming. We propose in this paper an evaluation method with naive subjects, based on paired comparisons. Two studies have been carried out on diesel motor sounds: firstly, the panel of experts of a car maker made a conventional sensory profile and an evaluation by paired comparison. Secondly, 30 naive subjects also completed two tests (ratings and paired comparisons). For the experts, we noticed a very good agreement between the sensory profile and the paired comparisons. For the naives, the paired comparisons gave a better agreement between the subjects and were more discriminating than the ratings. The results of these two tests with the naives were then compared with the conventional sensory profile of the experts using Generalized Procrustes Analysis. Results show that the consensus is better with the paired comparison test. As a result, the evaluation method proposed with naives could be an interesting method for the perceptual evaluation of products.

**Relevance to industry**

To stay competitive, a company must take into account customer's perception and react quickly to other competitors. Industry needs efficient techniques for evaluating the perceived quality of industrial products, and to integrate this data into the design process. Using naive subjects in perceptual tests can save the industry significant costs and time reaching customer need.

*Keywords*: Sensory evaluation; paired comparison; perceived quality; Generalized Procrustes Analysis; motor sounds.

## 1. Introduction

In today's highly competitive markets, developing new products that satisfy consumers' needs and preferences is a very important issue especially in the automotive industry. Beyond technical performances, the perceptions of the customer become very influential on their decision to purchase. To be successful, a product should not only satisfy objective requirements, but should also satisfy the customers' tastes which are inherently subjective (MacDonald, 2001; Giannini et al., 2006). Improving the perceived quality of products is then an important challenge in product design. This objective is not simple to achieve because it needs to include, in the design loop, a rather complex entity: the human (McDonagh et al., 2002).

The difficulty for companies is first to understand the dimensions of the perceived quality, through the exploitation of qualitative or quantitative customers surveys. This may be

* Corresponding author. Tel.: +33-2-40-37-69-57; Fax: +33-2-40-37-69-30.
E-mail addresses: Emilie.Poirson@irccyn.ec-nantes.fr (E. Poirson). Petiot@irccyn.ec-nantes.fr (J-F Petiot). Florent.Richard@mpsa.com (F. Richard).

problematic as customers often find difficulty in expressing them, the subjective answers of a subject being generally non reproducible, semantically ambiguous, and depending on cultural and training aspects of the subject.
The second difficulty is to establish links between dimensions of the perceived quality and product characteristics. This necessitates making correspondences between two kinds of data, which are very different in essence: on the one side the emotions or feelings of the user, on the other side product design parameters or physical measurements on the product. Furthermore, designing products to arouse positive customer emotions requires understanding the information gathered through the human senses.

To address this problem, industry and researchers have developed specific design methodologies to take into account customers' feelings and preferences. Three main categories of methods tackle this problem and are subjected to research efforts in engineering design (user oriented design).

- Methods based on the study of products semantic and semiotics intend to understand how we, as human beings, interpret the appearance, the use and the context of a product (Krippendorff and Butter, 1984). The research of "design rules" between the product form and the product semantics are proposed, using mainly tools of artificial intelligence or shape grammars (Hsiao and Chen, 1997; McCormack and Cagan, 2004). These methods aim to connect engineering design to industrial design,
- In Japan, kansei engineering aims to investigate customer feeling and proposes an ergonomic, consumer-oriented technology for product design (Nagamachi, 1995). Kansei engineering proposes to quantify people's perceptions about the product form and to translate the consumer perceptions into the design elements. The principle is to collect subjective evaluations of users on a set of products, and to analyze and interpret the ratings using multivariate statistical techniques. Various modelling methods are developed to provide design rules (linear or non linear model, neural networks, rough set theory) (Hsiao, 2002; Lai et al., 2006),
- Sensory analysis has been used for many years by the food industry to study the links between product characteristics and consumer's perceptions. Tools and methods have been developed (panel of experts, sensory profiling, preference modelling) and can be fruitfully translated into the engineering design domain (Poirson et al., 2007). The applications in the automobile industry are numerous, and they concern all the sensory modalities (design of car horn sounds (Lemaitre et al., 2003), touch of seats fabrics (Giboreau et al., 2001), comfort of seats (Kyunga et al., 2008), engine sound character (Roussarie et al., 2004), steering wheel vibrations (Jeon et al., 2009)).

Restricted in the beginning to the products' quality control phase, sensory analysis is now a competitive method for the industrial development of new products. The sensory approach is based on the construction of a product space (the stimuli) and three types of measurements: (1) preference measurement (hedonic tests), established by a huge panel of consumers; (2) sensory measurements, made by a panel of trained assessors, and (3) instrumental (or physical) measurements, corresponding to a technical characterisations of the products. The principle of the sensory approach for product design (Figure 1) is to explain the preferences by the sensory evaluations (definition of sensory requirements), and after to explain the sensory evaluations by objective measurements, in order to get the technical specifications (Stone and Sidel, 2004).

*insert figure 1 here*

To get a reliable sensory assessment of the products (and eliminate subjectivity as much as possible), the quality of the measurement must be carefully checked (accuracy, exactness, fidelity). For this reason, the sensory evaluation is generally made by a panel of trained experts. The evaluation of the reliability of the sensory measurement by the panel of experts is assessed through four characteristics:

- The representativity of the average of the assessments: check if the average is representative of the scores of the panel (adjustment to a normal distribution),
- The repeatability of the panel: ability of the judges to evaluate in a similar way the products from one session to another (checked with two-way ANOVA),
- The homogeneity: ability of the panellists to grade the products homogeneously (checked with two-way ANOVA),
- The discriminability of attributes: refers to the capacity of the experts to discriminate products according to the descriptors (checked with multiple comparisons tests – Newmann Keuls or Duncan).

Traditionally, the most widespread procedure for sensory assessment is the conventional profile. It is based on the following steps: development a common language by the experts (sensory attributes) – training of the assessors – individual evaluation (in a monadic sequential way, and with several repetitions). Many variants are available to carry out a conventional profile with experts (Quantitative descriptive analysis - QDA® - Spectrum method$^{TM}$) (Stone and al., 1974; Meilgaard et al., 1999).

One main limitation with the conventional profile is that the time necessary to train the experts or to organise the evaluations can be very long. Therefore, due to the competitive environment, a company must react quickly to other competitors, and propose possible modifications in the prototypes. The shortening of innovation cycles is nowadays a key tendency in the current industrial environment. The control of the risks in product innovation and the reduction of the innovation cycles require valid and fast sensory measurements. For this reason, sensory evaluation is often dropped when results are needed quickly. Furthermore, the financial effort supported by companies can also be in many cases prohibitive. We notice particularly that car makers are now seeking faster sensory evaluation methods.

For this reason, alternatives methods have been developed, in order to shorten the duration of the sessions, and to decrease the cost. For example, Free choice profiling (Williams and Langron, 1984) or the flash profile (Delarue and Sieffermann, 2004) can be in certain cases very interesting alternatives to the conventional profile. A study with naives is proposed in (Faye et al., 2004). In the past, we also developed a methodology for the assessment of product semantics with naives, based on paired comparisons (Petiot and Yannou, 2004). This paper is in continuation with this work.

We propose in this paper to develop a method for a perceptual evaluation of products, made by naive subjects, i.e. subjects who are not particularly trained for sensory evaluation. We focus particularly on the following objectives:

- to show in which extent the evaluations of naive subjects are different of those of experts,
- to study the influence on the results of the type of evaluation test.

To illustrate our approach, the evaluation method was applied to the perception of diesel car engine sounds. The demand for diesel-powered engines and the expectations of customers on vehicle acoustics have both been increasing in the past years. The diesel engine still suffers from a negative image concerning its noise ("it sounds like a tractor"). The diesel knocking (or diesel impulsiveness) is an important issue of vehicle sound quality, which has to be adjusted (if not designed) according to the customer expectations. The aim of car manufacturers is to reduce it, by taking into account the perceptual abilities of customers. The perceptually most critical condition for diesel impulsiveness is usually idle because there is no masking sounds (aerodynamics or rumbling sounds).
In (Parizet et al., 2007), an experimental study on the perceived comfort in diesel car running at idle was presented. A paired comparison task was performed, with stimuli involving sounds and vibrations. Free verbalisations were helpful for the understanding of the perceptual space and for identifying the contribution of noise and vibration stimuli to overall comfort. A comparisons between several listening test methods was presented in (Parizet et al., 2005), on a particular case concerning in-car ventilation noises. The noise pleasantness was assessed with two categories of methods: absolute evaluation and paired comparisons. The results show that paired comparisons provide a good quality of discrimination and are more reliable than an absolute evaluation. In the same way, the method we propose is based on paired comparisons.

We present in section 2 the materials and method used for the experimental approach. This section describes the different characteristics of the tests, and explains the methodology used for the comparisons of the results. A background on paired comparisons and on Generalized Procrustes Analysis (GPA) is exposed. Section 3 is dedicated to the results concerning the tests carried out by the naive subjects and the experts. The results of the different tasks (verbalisation – rating – paired comparisons) are presented. In section 4, a comparison of the results of experts and naives is proposed, using GPA. Conclusions and perspectives are drawn in section 5.

## 2. Materials and method

This section presents the stimuli, the panellists, the different tests carried out and the methods used for data analysis (paired comparison and Generalised Procrustes Analysis).

### *2.1. Stimuli*

The samples used for the tests were recordings of eleven diesel engines at idle ($sI$, $I$=1 to 11), of various brands of cars of the same segment. The engines were all 4-cylinder, except for the sound $s5$ which was a 5-cylinder. The sounds were recorded outdoors, at a distance of 1 metre from the bonnet of the car, in stereo, with an artificial head. For all the recordings, the location of the microphones and the room were identical. The duration of the stimuli was approximately 5 seconds, including an initial and final 200ms fading. The sounds were provided by the research team of the car maker (PSA Peugeot Citroën).
The listening of the sounds by the subjects was made with headphones via the sound card of a computer. User-friendly computer interfaces were designed for all the tests.

### *2.2. Subjects*

Two panels carried out the study:

- a panel of 30 naives subjects (students in our University) (22 males – 8 females), novices in acoustical sensory tests,
- a panel of 10 experts, regularly trained, and employed by the car manufacturer.

### *2.3. Synoptic of the tests and the analysis*

Figure 2 presents the synoptic of the study. From the product space (11 sounds), the panel of naives carried out three tasks (part 1, white boxes in Figure 2):

- A verbalisation task: the objective was to obtain a list of relevant sensory attributes for the description of the sounds (*list 1* of 15 attributes),
- A rating task of the sounds according to the *list 1*. A conventional unstructured monopolar rating scale was used for the assessment. Only one assessment was performed by the subjects (no repetition),
- A paired comparison task, according to a subset of *list 1*, named *list 2* (9 attributes). The pairs of sounds were compared on a 7 levels category scale (<< : very less, < : less, <~ : little less, = : equal, >~ little more, > : more, >> : a lot more). Details on the paired comparison method are given in the next subsection (2.4).

*Insert figure 2 here*

#### *2.3.1. Verbalisation task*

The first part of the study with naives was to generate a list of descriptive attributes of the sounds. As far as possible, these attributes had to be relevant, accurate, discriminating, exhaustive and independent. This part was made of 3 steps, during a first session:

(1)Free generation of descriptive terms (Individual task): generation of descriptive terms, by each subject, without limitation.

This free generation of terms aimed to generate a maximum of words relating to the perception of sounds. Each subject had the possibility to listen to the eleven sounds via a user-friendly interface. The subject wrote on a prepared sheet all the words suggested by listening to the sounds. To stimulate the generation of terms, a paired comparison was suggested, trying to find which word could characterise the differences heard.
The 30 subjects participated in the verbalisation, divided into 5 sessions of 6 participants (duration of a session: 30 minutes).

(2) Collective task: sharing of the descriptive terms, and discussion with the entire group.
After each of these individual sessions, the words generated by the subjects were pooled during a collective summary session, led by the experimenter, in order to eliminate the hedonic terms, to clarify the meaning of ambiguous terms, and to highlight the key dimensions of sounds. The duration of this phase was between 30 minutes and 1 hour. Once all of these terms were generated, a semantic analysis of the terms was carried out in order to select a list of attributes.

(3) Sorting and selection of the main attributes (definition of *list 1*)

The individual and collective part being free verbalisation, the number of terms generated was not controlled. The process used to reduce the database of words was to group words into subsets of synonymous (referring to the same semantic category). The descriptive term chosen as a title of the subset was the most occurent term given by the subjects, or a term suggested by the experimenter during the collective session. No words were generated in this step. This stage led to the definition of *list 1*, made of 15 attributes (table 4).

*2.3.2. Rating task*

During a second session, each subject had to assess the sounds according to the attributes of *list 1*. Via a user-friendly interface, subjects had to click to listen to the sound and to move a cursor on an unstructured scale to give a rating of the specified attribute. The presentation was monadic and sequential (the stimuli were presented for evaluation one by one), and the sounds were blindfolded (coded by different numbers from one attribute to another).

The experimental design was a complete design, without repetition: each of the 30 participants assessed all the $N$=11 sounds. It is well known in sensory analysis that the presentation order of the products may have an influence on the assessments. To prevent order and first-order carry over effects in the evaluation, the presentation order of the sounds was different for each subject. The design of the presentation orders was made by cyclically generated latin squares (MacFie et al., 1989; Wakeling et al., 2001), which are as balanced as possible for order and first-order carry over effects. The row $i$ of the square corresponds to the presentation order given to subject $i$, the column $j$ to the sounds assessed in the $j^{th}$-position. In our case, a 30×11 design was considered to generate the presentation orders of the 30 participants (this corresponds to a partially balanced design). The FIZZ® software (version 2, Biosystèmes, Couternon, France) was used to generate the orders of presentation.

The results of the rating task were analysed, in order to select, among the attributes of *list 1*, the more relevant attributes for the description of the sounds (discriminating – consensual – independent). The discriminating power of the attributes was assessed by Analysis of variance (two-way-ANOVA). Two factors were considered for the ANOVA: the "sound" (product effect) and the "subject" (subject effect). Significant differences between sounds are evaluated by a Duncan multiple comparison test (Petruccelli et al., 1999). Principal component analysis (PCA) of the averaged rating data was used to estimate the independence of the attributes, and to give the contribution of the different attributes on the different factors of the PCA.

For each attribute, the degree of consensus between the subjects was assessed by a normalized PCA on the (sounds*subjects) rating data. The percentage of inertia taken into account by the first principal component is an indicator of the degree of consensus between the subjects (higher the percentage, higher the consensus). After this analysis, 6 attributes, judged as not enough discriminating/consensual/independent, were removed from the *list 1*. This constituted *list 2* of attributes for the paired comparison test.

*2.3.3. Pairwise Comparison task (PC task)*

All the sounds were next assessed by a paired comparison task (PC) according to the attributes of *list 2* (9 attributes). For each attribute, the subject was asked to fill in on a 7-level scale (<<, <, <~, =, >~, >, >>) some comparisons between the sounds in the 11×11 comparison matrix. The interface for the test is presented in Figure 3 (attribute "bass"). To blindfold the test, a randomized number, different for one attribute to another, coded each sound. Sounds were listenable by clicking on their number, and a drop-down menu proposed the 7 comparative choices.

*Insert figure 3 here*

The subjects did not have to fill in all the comparisons in the matrix, but only some of them. So as to get computable data, we demanded to the subject to assess all the pairs

involving the same sound (fill in a complete particular row in the matrix, chosen freely). Next, the subject filled in 12 additional comparisons in the matrix, chosen freely. For the choice of the comparisons, the instruction given to the subject was to choose the pairs of sounds that seemed to be the most convenient and instinctive for evaluation.

The results of the PC test were analysed to asses the discriminating power of the attributes and the degree of consensus inside the panel. Confidence intervals were computed to asses the discriminating power and a normalized PCA on the (sounds*subjects) PC scores (percentage of inertia taken into account by the first principal component) estimated the degree of consensus. These results were compared with those of the rating test.

#### *2.3.4. Evaluations of the experts (Part 2)*

The panel of experts carried out two assessment tasks (part 2, in grey boxes in figure 2):

- A conventional sensory profile, on a list of 7 sensory attributes. These attributes were predefined before the study; they were used by the panel of experts of the company for other sensory tests involving engine sounds. Three repetitions of each sound were proposed, in three different sessions.
- A paired comparison task, on the same list of attributes. Two repetitions were proposed, in two sessions.

The data from the experts was analysed with PCA. For each attribute, correlations between the conventional profile and the paired comparison (PC) were computed.

#### *2.3.5. Comparison of the results*

Finally, averaged results between the naives and the experts were compared. The objective is to study the influence, on the sensory space, of the nature of the subjects (naive or experts) and of the type of task carried out (rating or PC).

The procedure for the comparison is first to assume that the evaluation by the sensory profile (experts) is the reference measurement of the sensory abilities of the sounds. The main argument to justify this assumption is that the sensory profile is the only assessment that provides a reliable and stable measurement.

Secondly, we compared the assessment of the naives (ratings and PC) with the sensory profile. For the comparison, the difficulty is that for naives and experts, the assessments are not made on the same set of attributes. The method used to carry out this comparison is Generalised Procrustes Analyses (GPA), described in section 2.5.

To simplify the comparison, and avoid the use of GPA, we could imagine using the list of attributes of the experts with the naives. But this will be unrealistic because in general, for new products or new projects, there is no list of sensory attributes available. The objective of the method is to use naives to assess the products, and also to create the list of attributes.

The next sub-sections will present the principle of the adaptation of Paired Comparison that we propose and a background on GPA.

### *2.4. Adaptation of Paired Comparison (PC)*

PC methods are particularly suitable for eliciting evaluation between products because they are more instinctive than an absolute assessment on an absolute scale (David, 1988). Furthermore, paired comparisons are known to be easily administrated and to provide a good quality of discrimination. The basic idea is to ask the subject to compare each pair of sounds and to assess the relative evaluation. This leads to a paired comparison matrix,

which can be processed to extract absolute values of the evaluation (evaluation scores). Many methods have been developed for the calculation of scores from paired comparison matrices, for example the well-known eigenvector method (Saaty and Hu, 1998) or probabilistic methods (Bradley and Terry, 1952). We used in this paper the *Least Squares Logarithmic Regression* (LSLR) PC method proposed in (De Graan, 1980; Lootsma, 1982). The main reasons for the choice of this method is that it provides evaluation scores for each subject and it tolerates sparse paired comparison matrices, which is interesting for the relative assessment of numerous sounds (Petiot and Yannou, 2004).

*2.4.1. Computation of scores for each subject*

The principle of the LSLR PC method is as follows. Let's consider a set of $N$ different products. A subject is asked to fill in a PC matrix by assessing a given characteristic, for example the preference, between product $i$ (row $i$) and product $j$ (column $j$) on a 7-level category scale, noted (<<, <, <~, =, >~, >, >>). The subject does not have to fill in all the comparisons in the matrix ($N(N-1)/2$). In order to limit the duration of the test, we limited the number of assessed pairs to $M = 2N$. Nevertheless, so as to have computable data, each product had to be involved in at least one comparison, and all the products must be connected by transitivity. A necessary condition for this is to impose that all the subjects have to assess all the pairs involving the same product (fill in a complete particular row in the matrix).

Next, the category scale is indexed onto a ratio scale. A plausible ratio scale is [1/8, ¼, ½, 1, 2, 4, 8] (see (Lootsma, 1993) for an in depth discussion on the choice of the ratio scale). This leads to a score ratio matrix of generic term $c_{ij}$. Then, $c_{ij}$ is an estimate of the quantity $w_i/w_j$, $w_i$ and $w_j$ standing for the preference scores for product $i$ and $j$.

The problem with the determination of preference scores is to estimate the more reliable values of $w_i$ ($i$ = 1 to $N$) from the ratio matrix. This leads to solving a system of $M$ non linear equation with $N$ variables ($N<M$), for which there is generally no exact solution, only approximate solutions can be found. To solve this problem, the LSLR method proposed by De Graan and Lootsma consists of minimising the cumulated square errors between the logarithmic terms of the estimation of the score ratio $c_{ij}$ and of the actual score ratio $w_i/w_j$, given by equation (1):

$$E = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij.} (\log c_{ij} - (\log w_i - \log w_j))^2 \qquad (1)$$

$N$: number of products

$c_{ij}$ ($i,j$ = 1 to N): score ratio of preference, for the comparison of product $i$ and $j$

$\alpha_{ij}$ : parameter equal to 1 when the subject provided the comparison (product $i$, product $j$) in the matrix, zero otherwise.

The minimisation of $E$ can be done by using the so-called normal equation or, in a similar way, by considering that the error $E$ is the fitting error of a particular multiple linear regression (equation (2)) (Limayen and Yannou, 2004):

$$\alpha_{ij.} \log c_{ij} = \alpha_{ij.} (\log w_i - \log w_j) + \varepsilon_{ij} \quad (2)$$

There is no unique solution to the calculation of the regressions coefficient $\log(w_i)$: an additional relation between the coefficients $w_i$ must be added to solve the system. The coefficients $w_i$ are chosen so that $\sum_{i=1}^{N} \log w_i = 0$.

This can be written as equation (3):

$$Y = X.H + \varepsilon \qquad (3)$$

With: $Y = (\log c_{ij})$: vector of dimension $M+1$, represents the dependent variable

$H = (\log w_i)$: vector of dimension $N$, represents the regression coefficients

$X$: matrix of dimension $((M+1)\times N)$, represents the independent variables (see (Limayen and Yannou, 2007) for a definition of $X$).

The estimate $\Theta$ (generic term $\theta_i$) of the regression coefficients vector H is given by equation (4):

$$\Theta = (X.X^t)^{-1}.X^t.Y \qquad (4)$$

$\theta_i$: estimate of the coefficient $\log(w_i)$, or similarly, $\exp(\theta_i)$: estimate of the coefficient $w_i$

The last stage of the procedure consists in the normalisation of the estimates $\exp(\theta_i)$, to get the scores $W_i$, given by equation (5):

$$W_i = \frac{\exp(\theta_i)}{\sum_{j=1}^{N}\exp(\theta_j)}, \text{ i = 1 to } N \qquad (5)$$

Under the assumption of a normal distribution of the residual $\varepsilon_{ij}$, confidence intervals can be calculated for the coefficients $w_i$ (for a confidence level $(1-\alpha)$). The confidence interval for the coefficient $\log(w_i)$ si given by equation (6):

$$\theta_i - t_{(M+1-N),\alpha/2}.\sigma_{\theta_i} \le \log(w_i) \le \theta_i + t_{(M+1-N),\alpha/2}.\sigma_{\theta_i} \qquad (6)$$

With : $\theta_i$: estimate of the coefficient $\log(w_i)$, given by the regression

$t_{(M+1-N),\alpha/2}$ : Student variable for a confidence level $(1-\alpha)$

$\sigma_{\theta_i}$ : standard deviation of the estimate $\theta_i$, given by the regression

The confidence interval for $w_i$ is given by equation (7):

$$\exp(\theta_i - t_{(M+1-N),\alpha/2}.\sigma_{\theta_i}) \le w_i \le \exp(\theta_i + t_{(M+1-N),\alpha/2}.\sigma_{\theta_i}) \qquad (7)$$

Finally, the confidence interval for the score $W_i$ is given by equation (8):

$$\frac{\exp(\theta_i - t_{(M+1-N),\alpha/2}.\sigma_{\theta_i})}{\sum_{j=1}^{N}\exp(\theta_j)} \le W_i \le \frac{\exp(\theta_i + t_{(M+1-N),\alpha/2}.\sigma_{\theta_i})}{\sum_{j=1}^{N}\exp(\theta_j)} \qquad (8)$$

In addition, the PC methods provide a measure of the subject's judgment inconsistency (Limayen and Yannou, 2007; Yannou, 2002). Indeed, the assessments provided by the subject in the PC matrix are not necessarily consistent, and even not transitive (if A preferred to B is denoted by A>B, the following evaluation is intransitive: A>B>C>A) - (if A strongly preferred to B is denoted A>>B, the following evaluation is transitive but not consistent: A>>B>>C and A>C). These inconsistency phenomena in the paired

comparison matrix may lead to poor results in the calculation of the scores when such inputs are processed.
The determination coefficient $R^2$ of the linear regression (equation 3) is an indicator of the consistency of the assessments of the subject. It represents the fraction of information explained by the regression model ($R^2 = 1$, perfect consistency). Examination of the $R^2$ of the regression for each subject gives an indication of the consistency of the assessments and therefore the credit that we can have in the data. The examination of the $R^2$ may lead to two pieces of important information on the test:

- Identify subjects who provided too inconsistent comparisons (because they did not care about the test, they did not understand the test, or the differences between the sounds were not mono-dimensional). The data from these subjects had to be removed before computing the value of the scores.
- Detect attributes that were systematically subjected to inconsistencies (either because they were not mono-dimensional, or because they were not precise enough). These attributes had to be subjected to additional explanation, or removed from the list for the building of the perceptual space.

To make the decision between consistent and too inconsistent comparisons, we used random drawings of comparison matrices. A set of $N_r$ comparisons matrices ($N_r = 300$), involving $M=2N$ comparisons, were randomly drawn. For each of the matrices, the $R^2$ coefficient was computed using the LSLR method. The sampling distribution of $R^2$ was plotted and the average value $R^2_a$ of $R^2$ was computed. Subjects for who $R^2 < R^2_a$ were ticked as inconsistent (as inconsistent as chance) and their data was removed from the rest of the study.
For the PC task, the examination of the $R^2$ of each subject was done to control the consistency of the evaluation (section 3.1.3). This step was important to verify the coherency of the data and to improve the reliability of the results

*2.4.2. Computation of scores for the group of subjects*
It is possible to compute directly the regression coefficients for a group of $r$ subjects, represented by their score ratio matrices of generic term $c_{ijr}$. It is necessary to assume that the group is homogeneous.
In this case, the cumulated square error between the logarithmic terms of the estimation of the score ratio $c_{ijr}$ and of the actual score ratio $w_i/w_j$ is given by equation (9):

$$E = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{h=1}^{r} \alpha_{ijh.}.(\log c_{ijh} - (\log w_i - \log w_j))^2 \quad (9)$$

The calculation method is similar, the confidence intervals of the score $W_i$ are given by equation (10):

$$\frac{\exp(\theta_i - t_{(rM+1-N),\alpha/2}.\sigma_{\theta_i})}{\sum_{j=1}^{N}\exp(\theta_j)} \le W_i \le \frac{\exp(\theta_i + t_{(rM+1-N),\alpha/2}.\sigma_{\theta_i})}{\sum_{j=1}^{N}\exp(\theta_j)} \quad (10)$$

The interest to compute directly the scores of the whole group of subject is to get more observations for the regression and so to have smaller confidence intervals for the estimates of the scores. These confidence intervals can be used to test the significance of differences between pairs of sounds.

### *2.5. Generalized Procrustes Analysis GPA*

To compare the communality between the averaged data of the different tests and the sensory profile (experts), we used Generalized Procrustes Analysis (GPA).
GPA is a multivariate technique commonly used in sensory evaluation to analyze free-choice profiling data (FCP[†]), to study the consensus among experts in classic sensory analysis, to assess scale use, attribute interpretation, panel performance, monitoring…
It also allows one to compare the proximity between the terms that are used by different experts to describe products. The GPA method was first described by (Gower, 1975), interpretation of GPA can be found in (Dijksterhuis, 1991).
Let's consider the description of *n* products by a set of *K* configurations *Xk*. *Xk* is a (*n*×*mk*) matrix which represents the assessment of expert *k* on the *n* products.
Note that with the Free choice Profiling (FCP), the variables which describe the products are not necessarily the same, the number of variables can also be different for each configuration.
GPA is a method for producing a consensus configuration $\overline{X}$ from the set of *K* different individual data matrices, and to represent the consensus via PCA.
The principle of GPA is to apply transformations (translation, isotropic scaling, rotation/reflection) to the configurations $X_k$ so as to minimise a goodness of fit criterion (the distance between the transformed configuration $X_k$' and the consensus configuration $\overline{X}$). GPA only allows 'rigid-body' transformations to the datasets and respects the relative distances between products. The individual and consensus configurations are typically submitted to PCA and projected onto a lower dimensional space. This space provides a vantage point to compare individual data and to visualise the consensus.
The degree of consensus is assessed by studying the variance of the datasets. The total variance $V_T$ can be partitioned as follows (equation 11):

$$V_T = V_C + V_W + V_R \quad (11)$$

where $V_C$ denotes the variance of the consensus, $V_W$ the within-product variance in the projection space and $V_R$ the residual variance.
By dividing by $V_T$, and sharing the within variance $V_W$ among the *n* products, the equation becomes (equation 12):

$$100\% = R_c + \sum_{j=1}^{n} r_{jW} + R_R \qquad (12)$$

*Rc* corresponds to the consensus ratio: a large *Rc* indicates good consensus.
$r_{jW}$ indicates the within variance of product *j*. A small $r_{jW}$ indicates a bad consensus for this particular product *j*.
In addition to this representation, a significance test of the consensus is available, based on permutation testing (Wakeling et al., 1992). The principle of this test is to verify if the variance ratio $R_c$ is « sufficiently » higher than a variance ratio of a consensual configuration obtained after random permutations of rows in the initial configurations. Several randomisations of the original data set are performed (typically $N_p$= 300), in order to plot the sampling distribution of the statistics $R_c$. The distribution is used to asses the confidence level for the $R_c$ obtained from the original data set: Higher the confidence level, better the consensus. A significance test of the number of dimensions (of PCA) necessary to represent the consensus is also available (Wu et al., 2002).

[†] FCP: Under this type of sensory profiling, each assessor or judge describes a product's characteristics using his/her own list of sensory attributes

## 3. Results

### *3.1. Results Part 1 - Study by the naives*

#### *3.1.1. Verbalisation task*

The verbalisation task generated more than a hundred terms, which were sorted into 3 categories: adjectives, common words and images/evocations. Under the assumption that the clustering of synonyms makes sense, 15 words were finally chosen to describe the different subsets of terms.

All the words sorted represented approximately 80% of the vocabulary generated during the verbalisation task. The remaining 20% had a clear hedonic character, were not sufficiently clarified by the subjects, or were too colourful to be consensual.

The list of attributes defined is given in Table 4.

#### *3.1.2. Rating task*

Subjects evaluated the sounds according to the attributes in *list 1* by rating the intensity on an unstructured scale. The raw scores are not given but an example of the average ratings for the attribute ("Sound level") is given in Table 1.

A two-way ANOVA was made on the evaluations. The two factors for the ANOVA are the "sound" and the "subject". The F-ratios are presented in Table 5 (column 1 and 2). All the attributes show a significant "sound effect" (5 % level), with large F-ratios for some of them ("bass", "sound level", "quickness"). The sounds are globally significantly different, showing that the subjects perceived overall differences among the sounds. This is a positive point for the evaluations. But this "sound effect" may be due to only one particular sound, and a significant "sound effect" is not sufficient to conclude that the attribute is discriminating.

Except for "soft", "blowy", "jerky", "perceived power", all the attributes present also a significant "subject effect" (5 % level), indicating differences between the subjects for the use of the rating scale, or differences in the rating of the sounds. Even if a significant "subject effect" is a negative point for relevant evaluations (in particular for a panel of experts), this significant effect is quite acceptable for naives, who are not trained in the use of a rating scale.

The Duncan multiple comparison test was applied to assess the discrimination between pairs of sounds. The results of the Duncan test for the attribute "sound level" is presented in Table 1. Four overlapping groups of sounds can be distinguished. The pairs of sounds which were not significantly different (significance threshold =5%) are linked with a horizontal coloured line (for example, sound *s5* is not significantly different of sound *s6*, but sound *s5* is significantly different of sound *s4*).

*Insert table 1 here*

In total, 31 pairs of sounds were significantly different for this attribute "sound level" (5% level). For each attribute, the number of pairs significantly different is given in table 5 (column 4). For some attributes "resonant", "perceived reliability", "interference", the discriminating power is very weak (less than 8 pairs significantly different).

To analyse the consensus between the subjects, for each attribute, a normalised PCA was made on the (sounds*subjects) rating data. Table 5 (column 3) shows the percentage of

inertia taken into account by the first principal component. It is indicative of the degree of consensus of the subjects: higher the percentage, higher the consensus. The attributes for which the panel of subjects is the least homogeneous are: “perceived reliability” (29.4%), “interference” (28.5%), “enveloping” (27.6%) and “perceived power” (25.5%).

The 6 attributes “resonant”, “steady”, “perceived reliability”, “interference”, “enveloping”, “perceived power” are then the least discriminating /consensual. They are suspected to be not relevant for the assessment of the sound.
The last step of the analysis was to verify that these attributes do not bring a significant level of information to the sensory space. To verify this aspect, a non-normalized PCA of the panel mean rating data was made. The percentage of variance taken into account by the first four dimensions is given in table 2.

*Insert table 2 here*

The contribution of the attributes (in % of variance) on the 4 dimensions is given in Table 3. We verify in table 3 that these six attributes (greyed in Table 3) have particularly low contributions to these four dimensions. We estimate that these attributes could be removed from the list without losing relevant information.

*Insert table 3 here*

The last step of the analysis was to study the semantic world evoked by the attributes. The attribute “muffled” has been considered in the same semantic field as “bass” and “treble”. It was finally removed. The attribute “steady” was considered as representing a “particular” semantic dimension, not present in the other attributes. It has been retained for this reason.

Finally, the characteristics of each attribute and the reasons why it was retained or rejected for the rest of the study are presented in Table 4.

*insert table 4 here*

Six attributes were finally rejected, because they were judged as not being discriminating/consensual/independent enough. For the PC tests, the list was then reduced to 9 attributes (*list 2*)

*3.1.3. The Paired Comparison test (PC)*
To verify the consistency of the subjects, the scores of the eleven sounds and the determination coefficient $R^2$ were calculated for each subject, by the LSLR method (explained in 2.4.1). In parallel, the average value of $R^2$ corresponding to random drawings of the comparison matrices was computed: $R^2_a = 0.4$.
For the $9 \times 30 = 270$ comparison matrices, only 6 matrices, involving 5 particular subjects, had a $R^2 < 0.4$. The data corresponding to these matrices were removed for the rest of the study because they were considered as too inconsistent (as inconsistent as a random filling in of the matrix!). They corresponded certainly to a careless filling in by the subjects.
For each of the 9 attributes, the average value of the $R^2$ remains always greater than 0.75. The assessments were then considered as consistent enough and each attribute was

considered as meaningful for the subjects. In the rest of the study, all the attributed were retained.
After removing the too inconsistent comparisons, the scores of the eleven sounds were calculated for the whole group of subjects (procedure explained in section 2.4.2). By this process, for each attribute, a percentage of 100% of importance is shared among the 11 sounds.
Confidence intervals were computed for the scores, allowing the testing of the significance of differences between pairs of sounds. The number of pairs significantly different (5 % level) for the PC test is given in Table 5 (last column). This number has to be compared to the same result for the rating task (Duncan multiple difference test).

*Insert table 5 here*

From the results of Table 5, it is clear that the PC task provides overall a better differentiation than the rating task. For all the attributes (except bass), the paired comparisons are more discriminating than the ratings (the number of sound pairs significantly different is higher for PC - the higher score is indicated in bold).
Furthermore, according to the subjects, it seems to be easier to compare sounds by pair than on an absolute scale. Finally, a PCA was made on the averaged scores (PC) and it shows that the percentage of inertia on the first component, representative of the degree of consensus between subjects, is globally better with PC than for the evaluation task (the higher score is indicated in bold - Table 5 column 3 and 5). The PC task seems to be more consensual.

### *3.2. Results Part 2 - Study by the experts*

#### *3.2.1. Sensory profiling*

The attributes used by the panel were chosen after several assessment sessions and a rigorous process to check the consensus of the experts (sensory profile). This list is not specific to our study, and is generic for all studies concerning diesel motor sounds in the company. For confidentiality reasons, the name of the attributes are translated and slightly modified. The list is given in Table 6.

*Insert table 6 here*

A normalised PCA of the sensory profile (averaged data) of the 11 sounds on the 7 attributes is given in Figure 4 (two first factors).

*Insert figure 4 here*

Concerning the variables, “intensity”, “quick” “treble” are opposite to “impulsiveness” on the first axis. On the second axis, “purr” is opposed to “blow”, and “knocking” to”treble”.
Concerning the products, *s5* is very different of the other sounds. This is not surprising because sound *s5* is a 5-cylinder engine. This sound has high scores on “intensity” “quick”, “knocking” and low scores on “impulsiveness”.
Groups of similar sounds can be distinguished: *s8 s3 s9* (with low scores on “treble” “blow”, high scores on “Purr”) and *s10 s7* (with low scores on “intensity” and “quick”, and high scores on “impulsiveness”).

#### *a. PC task*

The experts also assessed the sounds with paired comparisons. Similarly to the naives, the experts assessed all the pairs involving the same sound (fill in a complete particular row in the matrix), and filled in freely 12 additional comparisons in the matrix. They made two repetitions of the evaluations (in two different sessions). The scores for the group of experts were computed with the method described in section 2.4.2.

#### *b. Comparison sensory profile/PC task*

In order to quantify the agreement between the results of the two tasks (Sensory Profile and PC), the Pearson and Spearman coefficients between the Sensory profile (SP) and the PC scores were calculated. Results are presented in Table 7.

*Insert table 7 here*

The agreement between the results of the two tests is very good (especially for the ranks). The significance test on the values of R shows that, for all the attributes, R is significantly different of 0 (1% level). The two evaluation methods provide very similar results.

## 4. Comparison of the sensory maps and discussion

For the comparison, three sets of averaged data concerning the 11 sounds were considered:

- Evaluations given by experts with the sensory profile, on 7 attributes (Sensory Profile)
- Ratings given by naives with the rating task, on 9 attributes (Ratings)
- Scores given by naives with the PC task, on 9 attributes (PC naive)

### *4.1. Ratings (naives) .vs. Sensory Profile (experts)*

We performed a GPA on the two configurations: ratings (naives) and sensory profile (experts). The confidence level (quartile) for the consensus test was 41%: it signifies that in the permutation test, 59% of the samples (generated by random permutation of rows in the initial configurations) had a variance ratio higher than those obtained with the initial data. The conclusion is that the consensus is in this case very poor (a random re-arrangement of the rows in the configurations has higher likelihood to give a better consensus).

Studying the consensus, the examination of the within variance ratio $r_{jW}$ , for each sound *j*, indicates that the less consensual evaluations concern the sounds *s5* and *s10* (figure 5). The % of variance is greater than 50%, indicating very different assessments for these two sounds. The most consensual sounds are *s6* and *s11*.

*Insert figure5 here*

The significance test of the number of dimensions (of PCA) necessary to represent the consensus indicates that only two factors are significant (p-value<0.05). The PCA of the consensus configuration (2 first factors – map of the attributes) is given in figure 6.

*Insert figure 6 her*

The attributes that are highly correlated with the first factor concern the "intensity" and the "quickness" of the sounds, opposed to "soft" and "impulsiveness".
On the second dimension, the attributes "Blow", "Blowy" and "Treble" are opposed to "Purr".
Even if the consensus is not good, certain proximities of the experts' and naives' attributes are noticeable: (Blow- Blowy), (Quick – Quickness), (Intensity – Sound level). This indicates that the naives have intuitively a correct understanding of these attributes and of the sensations they represent.
The PCA of the consensus configuration (2 first factors – map of the products) is given in figure 7. In this figure, one can see that the main differences in the evaluation concern effectively the sounds *s10* and *s5* (which are far from each other).

*Insert figure 7 here*

In conclusion, this comparison of the ratings of the naives and the sensory profile of the experts indicates important differences in the sensory positioning. Several factors can explain these differences:

- Assessment error of the naives, due to the difficulty of the rating task (and the lack of training)
- Influence of the list of attributes
- Inter-individual differences of the naives

In particular, two sounds (*s5* and *s10*) are subjected to very different assessments. The sound *s5* (which corresponds to a particular motor - a 5-cylinder) was not understood in the same way by naives and experts. The results confirm the fact that naives cannot produce reliable assessments without training, and that the rating task is a difficult test for naives.

### *4.2. PC (naives) .vs. Sensory Profile (experts)*

A GPA was done on the two configurations: PC (naives) and the sensory profile (experts).
The confidence level (quartile) for the consensus test was 69%. The conclusion is that the consensus is in this case better than those with Ratings/Sensory Profile.
The examination of the within variance ratio $r_{jW}$ , for each sound j, indicates that the less consensual evaluations still concern the sounds *s5* and *s10*, the most consensual were sounds *s6* and *s11* (figure 8). For the two tests (rating and PC), the ratios $r_{jW}$ have the same form (figure 5 and figure 8). This is a sign of a certain consistency of the evaluations by the naives between the two tests (rating and PC).
For all the sounds, the % of variance ratio $r_{jW}$ was always lower than those of the comparison ratings/SP. This confirms the fact that the consensus is better with PC.

*Insert figure 8 here*

The significance test of the number of dimensions (of PCA) necessary to represent the consensus indicates that only two factors are significant (p-value<0.05). The PCA of the consensus configuration (2 first factors – map of the attributes) is given in Figure 9.

*Insert figure 9 here*

Similarities can be observed with the previous study: The attributes which are highly correlated with the first factor still concern the "intensity" and the "quickness" of the sounds, opposed to "soft" and "impulsiveness".
On the second dimension, the attributes "Blow", "Blowy" and "Treble" are opposed to "Purr". In Figure 9, we observe proximities as well between "synonyms" attributes from the experts and from the naives (Blow- Blowy), (Quick – Quickness), (Intensity – Sound level).

The PCA of the consensus configuration (2 first factors – map of the products) is given in Figure 10. In this figure, one can see that the main differences in the evaluation concern effectively the sounds *s10* and *s5*.

*Insert figure 10 here*
Results show that the consensus with the sensory profile is better with PC than with ratings, but differences in the evaluation still appear.
The same factors as previously mentioned can explain these differences:

- Assessment error of the naives, due to the difficulty of the PC task
- Influence of the list of attributes
- Inter-individual differences of the naives

Given that the list of attributes and the subjects were the same for the two tests (rating and PC), we conclude that the use of PC test leads to a better consensus. Paired comparisons are more intuitive and provide an assessment closer to the sensory profile, the reference measurement. Subjects also confirmed that they were easier with the PC task than with the rating. With this easier test, naives come nearer the assessments of the experts.

To summarise these conclusions, we plotted the relative positioning of the different tests on a "virtual" sensory space with two arbitrary sensory dimensions (figure 11).

*Insert figure 11 here*

The circle in grey represents the confidence interval of the evaluations for the different tests. According to our results, the confidence interval is greater for the rating test than for the PC test. The arrows represent the relative "distance" (or degree of consensus) between the two tests and the sensory profile. We saw that the consensus is better for the PC test.
To confirm our conclusions, it will be necessary to demonstrate that the confidence interval of the tests (represented by a circle in white), when the tests are replicated, is lower with PC than with ratings. This constitutes the main perspective of this work.
Repetitions of the same PC test will allow the definition of confidence intervals for the PC test. This will be an important stage for the qualification of the test for its use in customer oriented design process.

## 5. Conclusions and perspectives

We presented in this paper a method for the perceptual evaluation of products by naive subjects. It was applied to the assessment of diesel motor sounds, in the general context of customer-oriented design, taking into account the emotional response of the customers.
The method is based on a free verbalisation task of the naives in order to define a first list of attributes, relevant for the characterisation of the sounds. A rating test with the naives

was performed, in order to select a subset of attributes among this list, attributes which are as far as possible discriminating, consensual, and independent. This subset of attributes was used next by naives in a paired comparison task.

The proposed method provided a rapid perceptual positioning of the sounds by the naives. This is one of the advantages of the method on the conventional sensory profile: with two or three sessions, we got the main perceptual differences between the products, and terms (attributes) to explain these differences. At this level, we prefer the term "perceptual positioning" instead of "sensory positioning" because the attributes proposed by the naives do not have the characteristics of sensory attributes (relevant – accurate – discriminating – exhaustive – independent). In particular, the exhaustiveness of the attributes was not checked.

This advantage is important because training experts and carrying out sensory profiles can be an expensive task for companies. A second advantage is that PC provided a more consensual evaluation than the rating method. In consequence, the discrimination of the sounds (assessed with multiple discrimination tests) was better with PC than with ratings. This is an important point in favour of the PC for tests with naive subjects.

However, we showed that the method provided a different perceptual positioning than those of the experts, made by the traditional sensory profile. The consensus, assessed by GPA, was not perfect and differences occurred, due in particular to 2 or 3 specific sounds.
Reasons for the discrepancy between the results are multiple: training of the panel, effect of the list of attributes (non-exhaustiveness of the attributes of the naives), effect of the type of tests.
With PC, a comparative evaluation is made, whereas for conventional sensory profile (or rating), the monadic sequential evaluation implies that each sound is evaluated according to the subject's representation of the product space. The memory abilities of the subject are not involved in the same way, the PC evaluation entails less of the memory of the subject. PC seems to be preferable when non trained subjects are employed.
Nevertheless, the consensus between the sensory profile of the experts was better with the PC than with the ratings. Again, this is an important point in favour of the PC for tests with naive subjects. Furthermore, we verified that with trained experts, PC provides also an excellent consensus with the sensory profile (The results were out of the scope of this paper and not presented here).
Concerning the usage of the method, we think it would be misleading to use it to replace the conventional sensory profile. In particular, the repeatability of the assessments, the homogeneity of the panel, are not checked. The reliability of the measurement is not controlled. In fact, the two approaches do not have the same objectives.
The proposed method could be used in preliminary studies to rough out the product space, in order to select the convenient products. Its interest is also to explain differences between products by a set of attributes. These attributes could be used as guideline for the setting up of the attributes of the sensory profile.
In conclusion, PC with naive subject, following the method proposed in this paper, can be in certain cases a very interesting alternative to the "expensive" sensory profiling task with trained experts.
This work has opened few tracks of research. The first one concerns the Paired Comparison method. In the method used, subjects choose the pairs they compare. The advantage is that the subject can choose a priori the easiest comparisons. But the drawback

is that the same comparisons can be chosen by several subjects. This can lead to a poor efficient experimental design. Future works will consist of the use of optimal criterion (D-optimality) for the definition of the experimental design (the comparisons to fill in).
Concerning the calculation of scores, we matched arbitrarily the category scale on a particular ratio scale. A second perspective will be to optimise the ratio scale according to the consistency of the scores (to find the ratio scale that maximises the $R^2$ of the regression).
Another area of study concerns the setting up of the panel of naives. We propose in this study a panel of 30 naives, but studies could be made to define the "optimal" number of naives that are necessary.
Moreover, concerning the problem of consensus among the panel of naives, a more detailed study of the individual data could allows one to distinguish if the error is generalised or if it is due to some particular individuals.

**References**

Bradley, R.A., Terry, M.E., 1952. Rank analysis of incomplete block designs: the method of paired comparisons. Biometrika 39, 324-345.

David, H.A., 1988. The method of paired comparisons, New York: Oxford University Press.

De Graan, J.G., 1980. Extensions to the multiple criteria analysis of T. L. Saaty. Report National Institute of Water Supply.

Delarue, J., Sieffermann, J-M., 2004. Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. Food Quality and Preference 15, 383–392.

Dijksterhuis, G.B., Gower, J.C., 1991/1992. The interpretation of generalised Procrustes analysis and allied methods. Food Quality and Preference 3(2), 67–87.

Faye, P., Brémaud, D., Durand Daubin, M., Courcoux, P., Giboreau, A., Nicod, H., 2004. Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings. Food Quality and Preference 15 (7-8), 781-791.

Giannini, F., Monti, M., Podehl, G., 2006, "Aesthetic-driven tools for industrial design," Journal of Engineering Design, 17 (3), 193-215.

Giboreau, A., Navarro, S., Faye, P., Dumortier, J., 2001. Sensory evaluation of automotive fabrics: the contribution of categorisation tasks and non verbal information to set-up a descriptive method of tactile properties. Food Quality and Preference 12, 311-322.

Gower, J.C., 1975. Generalized Procrustes Analysis. Psychometrika 50, 33–51.

Hsiao, S.W., 2002. Concurrent design method for developing a new product. International Journal of Industrial Ergonomics 29, 41-55.

Hsiao, S.W., Chen, C.H., 1997. A semantic and shape grammar based approach for product design. Design Studies 18, 275-296.

Jeon, B.H., Ajovalasit, M., Giacomin, J., 2009. Effects of gender differences on the subjective perceived intensity of steering wheel rotational vibration based on a multivariate regression model. International Journal of Industrial Ergonomics 39, 736–743.

Krippendorff, K., Butter, R., 1984. Product semantics: Exploring the symbolic qualities of form. The Journal of the Industrial Designers Society of America, Spring, 4-9.

Kyunga, G., Nussbauma, M.A., Babski-Reevesb, K., 2008. Driver sitting comfort and discomfort (part I): Use of subjective ratings in discriminating car seats and correspondence among ratings. International Journal of Industrial Ergonomics 38, 516–525.

Lai, H.H., Lin, Y.C., Yeh, C.H., Wei, C.H., 2006. User-oriented design for the optimal combination on product design. International Journal of production Economics 100, 253-267.

Lemaitre, G., Susini, P., Winsberg, S., Mc Adams, S., 2003. Perceptively based design of new car horn sounds. Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, July 6-9.

Limayem, F., Yannou, B., 2004. Generalization of the RCGM and LSLR Paired comparison methods. Computers and Mathematics with Applications 48, 539-548.

Limayem, F., Yannou, B., 2007. Selective assessment of judgmental inconsistencies in paired comparisons for group decision rating. Computers & Operations Research 34 (6), 1824-1841.

Lootsma, F.A., 1993. Scale sensitivity in the multiplicative AHP and SMART. Journal of multi-criteria Decision Analysis 2, 87-110.

Lootsma, F.A., 1982. Performance evaluation of nonlinear optimization methods via multi-criteria decision analysis and via linear model analysis. In: M.J.D. Powell, ed. Nonlinear Optimization,Vol. 1981. Academic Press, London, 419–453.
MacDonald, A.S., 2001. Aesthetic intelligence: optimizing user-centered design. Journal of Engineering Design 12 (1), 37-45.
MacFie, H.J., Bratchell, N., Greenhoff, K., Vallis, L.V., 1989. Designs to balance the effect of order of presentation and first-order carry-over effects in hall tests. Journal of Sensory Studies 4, 129-148.
McCormack, J.P., Cagan, J., 2004. Speaking the Buick language: capturing, understanding, and exploring brand identity with shape grammars. Design Studies 25, 1-29.
McDonagh, D., Bruseberg, A., Haslam, C., 2002. Visual product evaluation: exploring users' emotional relationships with products. Applied Ergonomics 33, 231–240.
Meilgaard, M., Civille, G. V., Carr, B. T., 1999. Sensory evaluation techniques (3rd ed.). Boca Raton, FL: CRC Press.
Nagamachi, M.,1995. Kansei engineering: a new ergonomic consumer-oriented technology for product development. International Journal of Industrial Ergonomics 15, 3-11.
Parizet, E., Amari, M., Nosulenko, V., 2007. Vibro-acoustical comfort in car at idle: human perception of simulated sounds and vibrations from 3- and 4-cylinder diesel engines International Journal of Vehicle Noise and Vibration 3 (2), 143 – 156.
Parizet, E., Hamzaoui, N., Sabatié, G., 2005. Comparison of some listening test methods: A case study. Acta Acustica united with Acustica 91, 356-364.
Petiot, J-F., Yannou, B., 2004. « Measuring consumer perceptions for a better comprehension, specification and assessment of product semantics». International Journal of Industrial Ergonomics 33 (6), 507-525.
Petruccelli, J.D., Nandram, B., Chen, M., 1999. Applied Statistics for Scientists and Engineers, Prentice-Hall Inc.
Poirson, E., Petiot, J-F., Gilbert, J., 2007. Integration of user-perceptions in the design process: application to musical instrument optimisation. Journal of Mechanical Design 129 (12), 1206-1214.
Roussarie, V., Richard, F., Bezat, M.C., 2004. Perceptive qualification of engine sound character; validation of auditory attributes using analysis-synthesis method. Proceedings of CFA/DAGA'2004.
Saaty, T.L., Hu, G., 1998. Ranking by the eigen vector versus other methods in the analytical hierarchy process. Applied Mathematical Letter 11 (4), 121-125.
Stone, H., Sidel, J.L., 2004. Sensory Evaluation Practices. Elsevier Academic Press, third edition.
Stone, H., Sidel, J., Oliver, S., Woosley, A., Singleton, R. C., 1974. Sensory evaluation by quantitative descriptive analysis. Food Technology 28, 24–34.
Wakeling, I. N., Hasted, A., Buck, D., 2001. Cyclic presentation order designs for consumer research. Food Quality and Preference 12, 39-46.
Wakeling, I. N., Raats, M. M., MacFie, H. J. H., 1992. A new significance test for consensus in generalized procrustes analysis. Journal of Sensory Studies 7, 91–96.
Williams, A. A., Langron, S. P., 1984. The use of free-choice profiling for the evaluation of commercial ports. Journal of the Science of Food and Agriculture 35, 558–568.
Wu, W., Guo, Q., De Jong, S., Massart, D. L., 2002. Randomisation test for the number of dimensions of the group average space in generalized procrustes analysis. Food Quality and Preference 13, 191–200.
Yannou, B., 2002. Toward a web-based collaborative weighting method in project. Proceedings of IEEE Systems Man and Cybernetics, Hammamet, Tunisia.

**List of figures**

| **Sound level** | *s5* | *s6* | *s4* | *s1* | *s8* | *s11* | *S2* | *s9* | *s3* | *s7* | *s10* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Average rating | 8.56 | 7.33 | 6.48 | 6.31 | 5.75 | 5.54 | 4.69 | 4.59 | 4.51 | 3.57 | 3.2 |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |

**Table 1 : average results for the ratings (attribute « Sound level ») and Duncan groups (5% level)**

| Dimensions | Explained variance (%) | Cumulated explained variance (%) |
|---|---|---|
| 1 | 48.6 | 48.6 |
| 2 | 20.5 | 69.1 |
| 3 | 17.09 | 86.19 |
| 4 | 5.9 | 92.09 |

**Table 2: explained variance of the 4 first dimensions of PCA (rating data - naives)**

| Attribute | % dim 1 | % dim 2 | % dim 3 | % dim 4 | % (dim1, dim2) |
|---|---|---|---|---|---|
| Bass | 25 | 4.4 | 2.6 | 10.3 | 29.4 |
| Sound level | 5 | 25.2 | 1.9 | 0.01 | 30.2 |
| Quickness | 2 | 13.5 | 22.5 | 0.1 | 15.5 |
| Muffled | 14.2 | 0.4 | 3.5 | 28.7 | 14.6 |
| Soft | 14.1 | 10.4 | 0.1 | 1.2 | 24.5 |
| Whistling | 9 | 18.7 | 1.4 | 14.7 | 27.7 |
| Treble | 14.5 | 0.9 | 2.4 | 5.7 | 15.4 |
| Blowy | 2 | 9.9 | 26.9 | 15.2 | 11.9 |
| Jerky | 2.4 | 0.7 | 25.4 | 0.8 | 3.1 |
| Resonant | 0.5 | 5.8 | 0.3 | 0.06 | 6.3 |
| Steady | 1.8 | 0.05 | 2.1 | 5.4 | 1.85 |
| Perceived reliability | 2.6 | 1.7 | 0.01 | 1.9 | 4.3 |
| Interference | 1.6 | 1.7 | 4.5 | 11.5 | 3.3 |
| Enveloping | 3.6 | 2.13 | 6 | 0.07 | 5.73 |
| Perceived power | 1.8 | 4.5 | 0.1 | 2.4 | 6.3 |

**Table 3:** List of attributes, and % of contribution of each attribute on the 4 first dimensions of PCA. Rating data of the naives.

| Attributes of the naives | Characteristics | Decision |
|---|---|---|
| Bass | High contribution on dim 1 and dim 2<br>Good discriminating power | Retained |
| Sound level | High contribution on dim 2<br>Good discriminating power | Retained |
| Quickness | High contribution on dim 2 and dim 3<br>Good discriminating power | Retained |
| Muffled | High contribution on dim 1 and dim 4<br>Good discriminating power<br>in the same semantic field as bass and treble | Rejected |
| Soft | Good discriminating power<br>High contribution on dim 1 and dim 2 | Retained |
| Whistling | Good discriminating power<br>High contribution on dim 1 and dim 2 | Retained |
| Treble | Good discriminating power<br>High contribution on dim 1 | Retained |
| Blowy | High contribution on dim 3 | Retained |
| Jerky | High contribution on dim 3 | Retained |
| Resonant | Very weak discriminating power<br>Weak contribution on dim 1 dim2 dim3 dim 4 | Rejected |
| Steady | Weak discriminating power<br>weak contribution on dim 1 dim2 dim3 dim 4<br>But represent a “particular” semantic dimension, not present in the other attributes | Retained |
| Perceived reliability | Weak discriminating power<br>Weak contribution on dim 1 dim2 dim3 dim 4<br>Too hedonic | Rejected |
| Interference | Weak discriminating power<br>Weak contribution on dim 1 dim2 dim3 | Rejected |
| Enveloping | Weak discriminating power<br>Weak contribution on dim 1 dim2 dim3 dim 4 | Rejected |
| Perceived power | Weak discriminating power<br>Weak contribution on dim 1 dim2 dim3 dim 4<br>Too hedonic | Rejected |

**Table 4**: list of attributes selected from the verbalization task (List 1)

| | Ratings | | | | Pairwise comparison (PC) | |
|---|---|---|---|---|---|---|
| | Anova | | PCA | Difference test (Duncan) | PCA | Difference test (confidence interval) |
| Attribute | Sound effect F-ratio(p-value) | Subject effect F-ratio(p-value) | % of inertia (first factor) | Number of pairs (5% level) | % of inertia (first factor) | Number of pairs (5% level) |
| **Bass** | 43.8 (p<0.01) | 1.9 (p<0.01) | **63.1%** | 37 | 48.2% | 35 |
| **Sound level** | 31.3 (p<0.01) | 2.2 (p<0.01) | 55.6% | 31 | **60.3%** | **35** |
| **Quickness** | 19.1 (p<0.01) | 1.8 (p<0.01) | 47.5% | 26 | **62.8%** | **35** |
| **Muffled** | 15.9 (p<0.01) | 1.7(p<0.01) | 47.4% | 22 | n.a. | n.a. |
| **Soft** | 17.4 (p<0.01) | 1.12 (p<0.3) | **46.3%** | 28 | 42.26% | **31** |
| **Whistling** | 16.2 (p<0.01) | 2.3 (p<0.01) | 44.1% | 18 | **49.31%** | **30** |
| **Treble** | 17.5 (p<0.01) | 1.6 (p<0.03) | 43.8% | 23 | **45.1%** | **36** |
| **Blowy** | 14.3 (p<0.01) | 1.36 (p<0.1) | 40.6% | 20 | **43.23%** | **35** |
| **Jerky** | 9.98 (p<0.01) | 1.33 (p<0.12) | 37.9% | 18 | **41.1%** | **28** |
| **Resonant** | 2.8 (p<0.01) | 1.9(p<0.01) | 33.4% | 0 | n.a. | n.a. |
| **Steady** | 4.7 (p<0.01) | 1.4 (p<0.07) | 30.7% | 7 | **33.7%** | **12** |
| **Perceived reliability** | 3.6 (p<0.01) | 1.6(p<0.03) | 29.4% | 3 | n.a. | n.a. |
| **Interference** | 5.6 (p<0.01) | 1.41(p<0.1) | 28.5% | 8 | n.a. | n.a. |
| **Enveloping** | 7.7 (p<0.01) | 2.1(p<0.01) | 27.6% | 15 | n.a. | n.a. |
| **Perceived power** | 3.8 (p<0.01) | 1.1 (p<0.33) | 25.5% | 5 | n.a. | n.a. |

**Table 5**: results of the two-way ANOVA (F-ratios), PCA and difference test for each sensory attribute

| Attributes of experts |
| --- |
| Intensity |
| Treble |
| Quick |
| Knocking |
| Impulsiveness |
| Purr |
| Blow |

**Table 6**: List of attributes (modified for confidentiality) from the panel of experts

| Attribute | Intensity | Treble | Quick | Knocking | Impulsivness | Purr | Blow |
|---|---|---|---|---|---|---|---|
| $R_{Pearson}$ | 0.93 | 0.95 | 0.86 | 0.92 | 0.79 | 0.88 | 0.82 |
| $R_{Spearman}$ | 0.93 | 0.98 | 0.97 | 0.91 | 0.86 | 0.93 | 0.98 |

**Table 7**: Pearson and Spearman coefficients between the Sensory Profile and PC scores (experts)