ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 90 (2004) 1019 – 1028

Using Probabilistic Choice Models to Investigate Auditory Unpleasantness

Karin Zimmer, Wolfgang Ellermeier, Christian Schmid
Department of Acoustics, Fredrik Bajers Vej 7 B5, DK-9220 Aalborg Ø, Denmark. kaz@acoustics.aau.dk

Summary
The potential of probabilistic choice models (the Bradley-Terry-Luce model, or preference trees) in scaling the perceived unpleasantness of sounds was evaluated. To that effect, 74 subjects made pair-wise comparisons of the unpleasantness of twelve binaurally-recorded, environmental sounds presented over headphones. The stimuli varied in their psychoacoustic characteristics and half of them were of technical, half of natural origin. A more sophisticated model than previously tested, namely a preference tree, was identified to account well for the structure underlying the data, indicating (1) that subjects changed criteria, when evaluating different sound pairs, and that (2) these criterion changes combined in a lawful way, so that it was possible to measure unpleasantness on a ratio-scale level across the entire set of sounds investigated. Contrary to expectation, sound origin (technical or natural) did not influence the unpleasantness judgments. Instead, the sounds could be grouped according to their (non-acoustical) intrusiveness, and loudness. A subsequent multiple-regression analysis showed that in the sub-groups of soft and loud sounds, a combination of sharpness (S_{mean}) and roughness (R_{mean}), the latter differing in magnitude for the two groups, explained the unpleasantness-scale values very well ($r^2_{corr} = 0.91$). Direct magnitude estimates of unknown scale type, collected from the same listeners covered a much smaller range of ratios, and were roughly linear with the logarithm of the ratio scale derived from the preference tree. The advantage of the choice-theory modeling in providing information on the structure underlying the judgments is discussed.

PACS no. 43.66.Yw, 43.50.Ba, 43.66.Cb

1. Introduction

In sound-quality evaluation, it is useful to distinguish between the specific *qualities* a sound possesses (which may be captured by elementary psychoacoustic attributes such as loudness, roughness, or tonalness), and the *overall quality* attributed to the sound (which may be construed as auditory pleasantness, product-sound quality, or reproduced-sound quality, depending on the domain studied). Recently, Blauert and Jekosch [1] have emphasized this distinction, recommending to use the term *sound character* for the auditory profile characterizing a sound, and reserving the term *sound quality* for its overall appreciation.

When we attempt to measure (overall) sound quality in this way, we should be able to answer the following general questions with regard to the stimulus set studied:

1. Do we obtain a consistent measure of overall quality by querying our participants?
2. Is this measure uni-dimensional?
3. Is it based on a single decision criterion resp. a common set of criteria, or do different criteria come into play, depending on the sounds under consideration?
4. How does the overall quality measure relate to the *sound character*, i.e. to each sound's profile of elementary psychoacoustic attributes?

Received 8 November 2003,
accepted 19 June 2004.

In the present study, we use the perceived unpleasantness of environmental sounds as a simple paradigm to investigate these issues. In the remainder of the article, we first propose that conventional direct scaling methods using verbal or numeral categories, magnitude estimates, or visual analogue scales constitute inadequate methodologies to provide satisfactory answers at least to the first three questions raised above. In the following, we therefore investigate alternative methodologies based on paired-comparison data, and use probabilistic choice models to represent the perceptual structure underlying the sound-quality judgments. Finally, we explore how well the structure found is predicted by known instrumental metrics of sound quality. We will start out by developing the rationale for using these more sophisticated scaling methodologies, and by specifying the advantages they offer.

1.1. Scaling methodologies

The methodologies used in the majority of studies to collect subjective evaluations of overall sound quality (e.g. in the form of "mean opinion scores", MOS) leave the task of "measuring" it entirely up to the participant. The role

of the scientist in this part of the study is confined to that of a notetaker, who records whatever the participant pronounces, and converts the verbal or numerical responses made into (mathematical) numbers, the means (or medians) of which are taken to represent subjective magnitudes. The problems of this 'direct-scaling' approach (discussed more extensively in [2, 3, 4]) are that (1) no consistency checks (other than retest reliability) may be performed on such data, (2) the scale type (ratio, interval, ordinal[1]) of the data is undetermined, and (3) it remains unclear, whether the subject bases his or her evaluation on one, or several (changing) attributes.

By contrast, methods of indirect scaling, such those derived from axiomatic measurement theory [5, 6] and probabilistic choice theory [7, 8], put the burden of constructing the subjective scale on the scientist. Typically, they require all but very simple *qualitative* (e.g. preference) judgments from participants and uncover a representation of the data in terms of a quantitative scale by modeling the observers' decision strategies. They explicitly formulate the conditions (axioms) under which measurement is possible, and specify the scale type of the outcome.

Yet another alternative for inferring scale values in an indirect fashion is given by multidimensional scaling (MDS; [9]). Here, typically, quantitative (dis)similarity ratings are obtained for all stimulus pairs in a set. These data are then analyzed statistically, extracting orthogonal dimensions by which the pattern of similarity ratings may be described parsimoniously.

MDS and choice-model approaches are complementary, in that the former seeks to provide the sparsest solution, statistically, while the latter aims at modeling the cognitive processes actually involved. The MDS technique is not as rigorously founded, however, in that it will always provide some statistical solution, the applicability of which cannot be properly proved, or disproved. It is typically used as a heuristical method, rather than one suited for scaling a given attribute. Therefore, the focus of this paper shall be on choice models, two of which are elaborated in the following sections.

1.2. Probabilistic choice models

Depending on the decision strategies assumed, different model classes can be investigated. For the present purpose, the Bradley-Terry-Luce (BTL) model and the preference-tree model are specified.

1.2.1. BTL model

One particular approach, the BTL model [10, 11], has – some 40 years after its inception – found its way into occasional applications by psychoacousticians [12, 13, 14, 15, 16, 17] interested in representing paired-comparison data on a metric scale.

The BTL model may be derived from Luce's [10] *choice axiom*, and postulates a very simple relationship between preference probabilities, as observed in an experiment, and the scale values (with respect to the attribute in question) to be inferred via modeling:

$$p_{ab} = \frac{v(a)}{v(a) + v(b)}, \tag{1}$$

where p_{ab} denotes the probability of "prefering" object a over object b (e.g. judging it to sound louder, more annoying, or more tonal, depending on the task in a paired-comparison experiment), and $v(a)$, $v(b)$ are the scale values of the objects (with respect to loudness, annoyance, tonalness, for example). This model puts strong constraints on the structure in the data: Note that given a set of stimuli a, b, c, etc., the scale value for one stimulus $v(a)$ can be chosen freely. The v-scale values for any other two stimuli are then determined by two preference probabilities (say p_{ab} and p_{bc}), and the remaining preference probability in any triple, p_{ac}, is to be predicted accurately, using the same v-scale values. Furthermore, it can be shown that the v-scale thus defined constitutes a ratio scale, unique up to multiplication by a positive constant.

The criticisms raised earlier with respect to the indeterminacies of a direct-scaling outcome thus do not apply to a properly constructed BTL scale, since (1) the approach allows for prior consistency checks of the raw data, e.g. in terms of the transitivity of paired comparisons [2, 7], (2) it specifies the scale type of the representation, and (3) it requires one-dimensionality of the underlying judgments, otherwise the model will be rejected [7].

The simplicity of the BTL model, however, also entails its major drawback: It is highly restrictive in that it requires *context independence* of the paired-comparison judgments to hold. In a psychoacoustical context that means that the auditory features used when comparing object a with object b must be the same as those used when comparing a and c, for all pairs of objects. For example, if the elementary sound attributes of loudness, roughness, sharpness, and tonalness are used by listeners to determine auditory (un-)pleasantness, they must be applied uniformly across all pairs to be judged. It is not hard to see, how this requirement may break down for a given set of stimuli. Typically, that happens when *similarities* emerge for subsets of the stimuli, and when the decision strategy shifts to disregarding these similarities, and focusing on the distinguishing features. If two stimuli are equal in their perceived sharpness, for example, but differ in loudness, only the latter sound attribute might be used as a basis for the decision.

1.2.2. Preference trees

Fortunately, less restrictive models have been developed to account for these situations, namely preference-tree [18] models, or a further generalization, called *elimination-by-aspects (EBA)* models [19]. The preference-tree model, for example, accounts for the situation sketched above by modifying the model equation in a seemingly slight, but significant way:

$$p_{ab} = \frac{u(a' - b')}{u(a' - b') + u(b' - a')}. \tag{2}$$

[1] Strictly speaking, it may happen that the test objects – unnoticed by the experimenter – cannot even be consistently judged on an ordinal scale.

Now the preference probabilities do not simply depend on the v-scale values (as in eq. 1), but rather on modified scale values that reflect the influence of the "unique" features only, with $a\prime - b\prime$ denoting the set of features of sound a that it does not share with b, and $b\prime - a\prime$ refering to the set of features of sound b that it does not share with a. The consequence, when fitting this model, is that additional parameters will have to be estimated that reflect the common features defining subgroups of the stimuli. Note that these features do not necessarily have to be salient to the observer (or the experimenter, for that matter), their emergence is subject to empirical evaluation, only.

Except for two exploratory studies from our laboratory investigating tonal prominence [20, 21], the less restrictive choice models (preference trees, EBA) have not been used in psychoacoustics. Furthermore, most of the cited work using the BTL model employed it as a scaling heuristic, and did not include rigorous model tests. Therefore, it is conceivable that some of the data sets explored using the BTL approach might have been better accounted for by a preference tree, for example. The unavailability of pertinent statistical software has made it hard, however, to explore these more sophisticated choice models. That situation has improved, since (a) a tutorial-like overview of the steps involved in testing these models [2], and (b) a software module that estimates, and tests BTL, preference tree, and EBA models [22] have recently been published.

1.3. Auditory unpleasantness reconsidered

In the present study, the potential of the more sophisticated choice models is investigated with respect to judgments of overall auditory unpleasantness. Overall unpleasantness may be seen as a high-level sound-quality attribute which may be traced back to multiple, more elementary features of sound character. In terms of the models discussed, it is highly likely that the sound features the listener bases his or her decision on vary, depending on the sounds to be compared.

An earlier study using the BTL-model approach [16] found a heterogeneous set of ten environmental sounds to be well represented by the BTL model; that is, a ratio scale of auditory unpleasantness emerged. Furthermore, the scaling outcome was well predicted by instrumental measures of elementary sound-quality attributes, so-called basic psychoacoustic metrics, namely a combination of roughness, and sharpness. A drawback of the study was, however, that the sound selection was rather arbitrary, and that loudness, and roughness, for example, correlated so highly in the stimulus set that it was impossible to disentangle their contribution.

Therefore, in an attempt to minimize such correlations, a more "balanced" stimulus sample was used in the present study. Such a stimulus set has been made available by Johannsen and Prante [23] in an article published in this journal, that provides a rather complete psychoacoustical analysis of 25 sounds, which were explicitly selected to contain orthogonal, i.e. uncorrelated, combinations of known psychoacoustic metrics.

Furthermore, in order to facilitate the emergence of subgroups of the sounds, for the present study, a set of six "technical sounds" was contrasted with an equal number of "natural sounds" that were similar with respect to the profile of their psychoacoustic indices. It was hypothesized that a preference-tree analysis might reveal these subgroups, if the origin of the sound is relevant to its evaluation. Stimulus subgroups, if they emerged, might also be governed by different "combination metrics", that is different ways in which elementary psychoacoustic attributes combine to yield the overall evaluation. This hypothesis shall be tested using group-wise instrumental metrics of sound character as predictors of overall unpleasantness.

1.4. Goals of the present study

In terms of the general questions raised at the outset, the goals of the present study are to

1. determine whether a consistent measure of overall quality, specifically *auditory unpleasantness* may be obtained,
2. establish the dimensionality and scale type of this measure, considering more sophisticated probabilistic choice models than have been previously applied in psychoacoustics, and
3. relate the overall quality measure to indices of the *sound character*, i.e. to each sound's profile of elementary psychoacoustic attributes, using a stimulus set that has been optimized with respect to the independent variation of these attributes.

   Two more specific research issues related to these hypotheses are
4. to determine, whether sounds of technical vs. natural origin are judged differently with respect to their unpleasantness, and
5. to investigate whether the more economic method of direct magnitude estimation yields similar scale values as the more time-consuming indirect scale construction from paired comparisons.

2. Method

2.1. Subjects

Data were collected from a sample of 79 participants, consisting of university students, who were paid hourly wages for participation in the experiment, and a small number of staff members from the Department of Acoustics. With the exception of the first and second author who also participated, the subjects were oblivious to the goals of the investigation, and the nature of the sound sources.[2] As assessed by self-reports, none of the subjects suffered from current hearing impairments. For five participants, the paired-comparison data and the direct magnitude estimates were contradictory, in that sounds judged to be highly unpleasant in one paradigm were rated not at all unpleasant in

[2] Including two of the authors was considered unproblematic, since none of the research questions was directional in nature, i.e. the questions did not lend themselves to clear-cut hypotheses to be retained or disproved, or to expectations that might have guided the participant's answers.

the other. We concluded that these participants may have misunderstood the instructions to one of the experimental tasks, and therefore excluded their entire data from further analysis. The remaining 74 participants were between 20 and 45 years of age (mean age: 25.3 years), and consisted of 30 female, and 44 male listeners.

2.2. Stimuli and Apparatus

Stimuli were taken from a set of 25 natural and industrial sounds, which were systematically collected, binaurally recorded, and extensively documented by Johannsen and Prante [23] to be available for further study. Twelve sounds were selected for the present investigation half of which were of *technical* (denoted T_i in Table I), and half of *natural* origin (denoted N_i in Table I). They were matched in pairs (identified by the indices in Table I) having roughly similar profiles with respect to psychoacoustic loudness ($N5$), sharpness ($S10$), and roughness ($R50$) according to the measurements reported in [23]. The sounds had a uniform duration of approximately 5 s.

All stimuli – including a 94-dB SPL calibration tone – were available in audio format on a compact disc published along with the article [23]. These files were converted to 16-bit, 44.1 kHz wave format, and pre-processed to equalize for the headphone response. That is they were filtered with the inverse of the headphone transfer function as measured on an artificial head [24] using the maximum-length sequence (MLS) method. Subsequently, they were set to be delivered at 6 dB below the recording level, since pilot trials had indicated that reproduction at the original levels would be too annoying for an extended paired-comparison experiment. All stimuli were analysed (using the Brüel & Kjær PULSE Sound Quality Module Type 7698) with respect to a number of psychoacoustic metrics (defined in [25]). A list of the stimuli, and a summary of relevant metrics (always giving the maximum of the left-ear and right-ear values) is found in Table I. It shows that our attempt to de-correlate the psychoacoustic metrics by choosing appropriate stimuli was only partly successful: The Pearson product-moment correlations of the values on these metrics were $r(N,S) = 0.522$, $r(N,R) = -0.184$, and $r(R,S) = -0.11$, with N refering to (mean) loudness, S to sharpness, and R to roughness. Nevertheless, the variance accounted for by these correlations was always less than 28%.

Playback and response collection were controlled by a microcomputer that delivered the stimuli via a sound card (RME Digital 96/8), after appropriate amplification by a headphone amplifier (Behringer HA 903), to Beyerdynamic DT 990 headphones. Listening took place in a double-walled, sound-attenuating chamber.

2.3. Procedure

All listeners participated in two sessions. In the first session, dissimilarity ratings of all pairs of sounds were collected. These data are not part of the current report. The second session focused on the unpleasantness of the sounds. In the beginning, all twelve sounds were played back once more in a fixed order to ensure that the listeners recalled the entire sound set.

Subsequently, they judged all pairs of stimuli once, indicating which of the two stimuli in a pair sounded more unpleasant to them. The sounds in a pair were separated by a 500-ms pause, and subjects responded by pressing one of two buttons labelled '1' (first sound is more unpleasant), or '2' (second sound is more unpleasant) on a computer keyboard. In all, each listener gave $(12 \times 11)/2 = 66$ judgments. Both the order within a pair, and the succession of pairs were randomized separately for each subject. After 33 assessments had been completed, there was a short break. Subjects took up the task again in a self-paced manner, after approx. 1 min. In all, data collection took ca. 20 min on average.

After completing the paired-comparison task, subjects were asked to give direct magnitude-estimates of the unpleasantness of the 12 sounds as compared to a reference sound, i.e. the fan-noise. On every trial, listeners were first presented with the reference, which was assigned an unpleasantness value of 10, displayed on the computer screen. After a 500-ms pause, the second sound was played, and subjects were asked to assess its unpleasantness, as they perceived it, relative to the standard sound. Thus, if the second sound was twice as unpleasant as the fan-noise, it should be given a value of "20", if it was half as unpleasant, its unpleasantness should be rated as "5" etc. Subjects were made aware of the fact that they could use whole numbers and decimals, but that an input of the number '0', or of negative numbers was not accepted. The listeners typed their responses into a field using the computer keyboard. For each subject, the unpleasantness of each sound was assessed three times, thus $12 \times 3 = 36$ sound-pairs were presented in a random sequence. After the subject had given 18 estimates, a short pause was introduced. In all, giving the magnitude estimates took approximately 12 min.

At the end of the session, subjects listened to each of the sounds once more in a random sequence, and were asked to identify the sound source by writing an appropriate label into a chart.

3. Results

In the following, the quality of the paired-comparison data is evaluated, before the choice models are specified, and the model fits to the data are assessed statistically, using likelihood-ratio tests. The scale values of unpleasantness derived from the best-fitting model are then compared to the direct magnitude-estimates obtained. Finally, the potential of instrumental measures of basic sound-quality attributes in predicting the scale-values of unpleasantness is explored.

3.1. Consistency of paired comparisons

Before analyzing the adequacy of the choice models, the data quality was checked by evaluating the ordinal consistency of the judgments separately for each individual.

Table I. Psychoacoustic metrics determined for the sounds. *Note.* The entries are the natural logarithm of the preference-tree values, $ln(PT)$, average loudness (N_{mean}), the 5th loudness percentile (N_5) average sharpness (S_{mean}), roughness (R), fluctuation strength (FS), and the prominence ratio (PR). Analyses were performed separately for each ear; the values in the table are the maximum across ears.

Code	Description	ln(PT)	N_{mean}	N_5	S_{mean}	R	FS	PR
T_1	circular saw	5.00	42.50	46.90	6.72	2.53	0.80	5.85
N_1	stadium	2.37	57.60	64.10	3.65	1.57	1.29	3.64
T_2	dentist's drill	3.69	27.40	33.10	5.74	1.70	1.22	19.20
N_2	waterfall	2.12	38.10	42.00	3.52	1.37	1.29	—
T_3	ship's horn	2.51	43.20	69.20	2.32	4.39	1.36	8.73
N_3	stone in well	1.89	10.70	33.10	2.27	3.93	1.29	—
T_4	typewriter	2.24	13.20	25.70	4.07	4.30	1.29	5.49
N_4	hooves	1.58	16.20	26.30	2.54	5.04	2.28	—
T_5	fan	2.30	16.60	17.90	2.05	1.85	1.18	8.16
N_5	howling wind	1.67	7.68	9.75	1.33	1.63	0.73	6.34
T_6	tyre on gravel	1.54	8.89	13.10	2.36	2.35	1.28	—
N_6	wasp	3.96	8.61	19.50	1.53	2.13	1.08	8.96

That was done by counting the number of intransitivities, or *circular triads*, in each individual data set. A circular triad occurs, if $a > b$ and $b > c$, but $a < c$. The participants produced a median number of 4.5 circular triads out of a possible 70, and none of the 74 listeners exceeded the number of intransitivities that may be expected to occur by chance [26] alone (χ^2-test, $\alpha = 0.05$). Thus, all subjects gave sufficiently consistent judgments of unpleasantness to be included in further analyses.

Consequently, the data were pooled over individuals, yielding a 12×12 matrix, given in Table II. In this cumulative matrix, cell entries indicate how many individuals judged the sound given in the row as more unpleasant than the sound in the respective column.

The consistency of the "preference" probabilities evident from these pooled data may again be subjected to a number of transitivity tests [7, 2]. Note that even with consistent individual data, inconsistencies in the pooled data set may occur, indicating the presence of groups of observers, who show different decision behavior, and should therefore be analyzed separately.

Typically, three kinds of stochastic transitivity are distinguished: weak, moderate, and strong. Weak stochastic transitivity (WST) holds, if for all stimulus triples a, b, and c:

$p_{ab} \geq 0.5$ and $p_{bc} \geq 0.5$, then $p_{ac} \geq 0.5$.

For the present data set, the cumulative data matrix did not show any out of 220 possible violations of WST. That implies that it is possible to establish a uni-dimensional ordering of the stimuli with respect to their unpleasantness over all subjects. Furthermore, the data also showed moderate stochastic transitivity (MST), in that

$p_{ab} \geq 0.5$ and $p_{bc} \geq 0.5$, then $p_{ac} \geq min(p_{ab}, p_{bc})$

was only violated in a single instance in 220 tests, indicating that choice-models like the BTL-model, or preference trees, may well be fit to represent the data. Interestingly, the BTL model implies an even stronger form of transitivity: strong stochastic transitivity (SST) which is given by

$p_{ab} \geq 0.5$ and $p_{bc} \geq 0.5$, then $p_{ac} \geq max(p_{ab}, p_{bc})$.

SST was violated in 39 of the 220 tests. Taken together, these diagnostics suggest that the BTL model may not hold, but that less restrictive models such as preference trees, stand a good chance. Since, however, no statistical tests are available to test for the significance of these violations, this conjecture will have to be confirmed by actually evaluating the fit of the various choice models, given in the next section.

3.2. Choice-model representation of the sound set

As suspected, a likelihood-ratio test[3] for the fit of the BTL model (eq. 1) indicated significant departures from the model prediction, $\chi^2(55) = 84.19; p = 0.007$. The failure of the BTL model may be due to a violation of context independence, i.e. to the fact that the criteria used in judging unpleasantness may differ depending on the sounds entering into a paired comparison. That is often the case, if subgroups of sounds emerge in the stimulus set.

In this case, a preference tree might be suited to represent the data. Starting from the a-priori hypothesis that "technical" vs. "non-technical" sounds might be judged according to different criteria, a preference tree branching off into two nodes thus defined was evaluated: Contrary to our expectation, however, neither a tree based on our a-priori classification of technical, and natural, sounds [$\chi^2(53) = 84.18$; $p = 0.004$], nor one based on the a-posteriori labelling[4] by the subjects [$\chi^2(53) = 79.29$; $p = 0.011$] fit the data. Thus it may be concluded that the type of sound source thus defined was not relevant in judging unpleasantness.

Taking a closer look at the preference matrix, the sound of a wasp recorded humming close to the listener's left

[3] All model testing and parameter estimation was done using the MATLAB program published in [22].

[4] When the sounds were categorized as technical or natural *a posteriori*, i.e. based on the labels assigned in the identification task, the only sound that turned out to be ambiguous with respect to its origin and that was classified as being "natural" rather than "technical" in 58% of the cases was the "tyre on gravel" sound.

Table II. Cumulative paired-comparison matrix. *Note.* The cell entries denote the number of subjects (of a total of 74) who judged the sound listed in the row as more unpleasant than the sound in the column.

Sound	saw	drill	fan	hoov.	wind	ship	stad.	well	type.	tyre	wasp	water
circular saw	-	61	72	73	74	72	73	73	71	74	45	73
dentist's drill	13	-	67	72	72	60	59	68	60	72	35	69
fan	2	7	-	64	68	27	34	58	43	68	17	39
hooves	1	2	10	-	29	4	7	20	3	46	6	14
howling wind	0	2	6	45	-	13	18	22	16	57	3	20
ship's horn	2	14	47	70	61	-	43	53	44	64	17	48
stadium	1	15	40	67	56	31	-	53	43	64	17	45
stone in well	1	6	16	54	52	21	21	-	26	59	8	26
typewriter	3	14	31	71	58	30	31	48	-	70	11	42
tyre on gravel	0	2	6	28	17	10	10	15	4	-	4	11
wasp	29	39	57	68	71	57	57	66	63	70	-	63
waterfall	1	5	35	60	54	26	29	48	32	63	11	-

ear was identified as the prime obstacle for fitting a simple choice model. A preference tree with the 'wasp' on a branch separate from all other sounds fared better than the models previously suggested, but still did not provide a satisfactory fit; $\chi^2(54) = 68.72$; $p = 0.086$. The best preference-tree representation found (after excluding a number of other plausible alternatives) assumed the 'non-wasp' sounds to branch out once more, creating two additional groups of stimuli (sounds T_1, N_1, T_2, N_2, T_3, N_3 vs. T_4, N_4, T_5,N_5, T_6 in Table I) which may be characterized as a loud ($N_5 > 27$ sone) vs. a soft ($N_5 < 27$ sone; s. Table I) group. This model, which assumes two additional nodes, one comprising the "non-wasp" sounds, and one the soft sounds, provided a satisfactory fit to the data; $\chi^2(53) = 58.47$, $p = 0.282$. It is schematically depicted in Figure 1. In a graphical preference-tree representation, the path lengths starting from the origin are proportional to the scale values, and the lengths of the "branches" leading to a node are proportional to the degree of similarity of the stimuli connecting to that node. Note, however, that Figure 1 is only schematic, in that it depicts the *structure* of the tree, not the actual scale values. Due to their considerable range, they are awkward to draw to scale in a graph; instead, they are given in Table III, along with the associated standard errors.

Given that a valid preference-tree structure was found, the unpleasantness of the twelve environmental sounds can be estimated on a ratio scale, rendering statements about the unpleasantness ratio of two sounds meaningful. Assigning a value of ten to the unpleasantness of the 'fan'-noise, the u-scale values obtained range from 4.66 ('tyre on gravel') to 148.34 ('circular saw'), i.e. they vary by a factor of more than 30.

3.3. Direct magnitude estimation

In the direct-scaling task, listeners judged the unpleasantness of each sound three times. Assessments were given relative to a standard, the 'fan' noise, which had been assigned an unpleasantness value of 10 by instruction. The geometric means of the magnitude estimates over all subjects, and their standard errors, are given in Table III. The

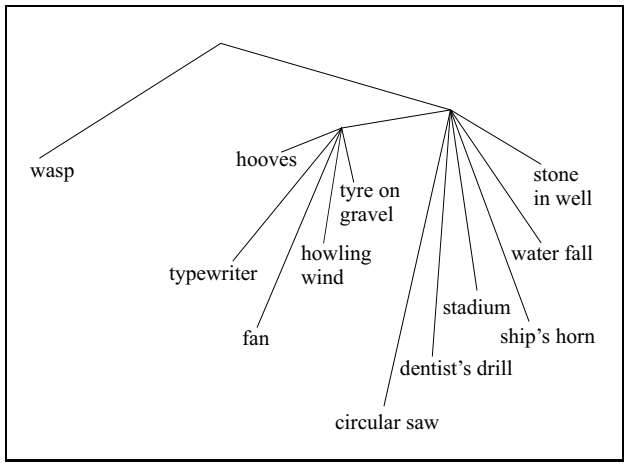

Figure 1. Schematic graph of the preference tree representation.

unpleasantness values thus obtained range from 3.72 for 'tyre on gravel' to 52.47 for the 'circular saw'. Varying by a factor of 14, the magnitude estimation scale is more compressed than the preference-tree ratio scale. Plotting the magnitude estimates over the logarithm of the preference-tree scale values, as in Figure 2, yields an approximately linear relation between the two scales. When the orderings of sounds according to their unpleasantness are compared, a few inconsistencies between preference-tree scale values and magnitude estimates become obvious for the closely spaced stimuli in the vicinity of the standard (see Figure 2 and Table III). Other than that, the agreement is quite high, yielding a product-moment-correlation of $r = 0.93$ (or even $r = 0.98$ when the logarithms of the preference-tree scale values are used as in Figure 2). Thus, the two types of scales have 86%, and 96%, of their variability in common, respectively.

3.4. Unpleasantness and psychoacoustic metrics

It is generally assumed that global unpleasantness judgments may be predicted by a combination of elementary sound characteristics, such as those captured by conventional psychoacoustic metrics. To explore the validity of

Table III. Scale-values of unpleasantness: Preference-tree representation, and magnitude-estimates. *Note.* Scale values (Sv) of the unpleasantness of twelve environmental sounds, and standard errors. The sounds are ordered by magnitude in the ratio-scale, preference-tree representation, in which the 'fan' sound was assigned a value of 10. The magnitude estimates are the geometric means of the numbers given by the 74 listeners. They are based on the 'fan' noise as a reference sound, which was assigned an unpleasantness value of 10 by instruction.

	Preference-tree		Magnitude-estimation		
Sound	Sv	SE	Sv	SE_+	SE_-
tyre on gravel	4.66	1.67	3.72	0.48	0.42
hooves	4.86	1.66	4.03	0.60	0.52
howling wind	5.30	1.63	5.41	0.47	0.43
stone in well	6.62	1.57	9.30	1.11	0.99
waterfall	8.31	1.54	11.99	1.49	1.33
typewriter	9.36	1.54	9.44	1.03	0.93
fan	10	1.53	8.57	0.58	0.55
stadium	10.74	1.55	16.89	2.33	2.05
ship's horn	12.36	1.62	19.41	1.89	1.72
dentist's drill	40.18	4.36	31.01	2.67	2.46
wasp	52.46	8.15	30.20	4.28	3.75
circular saw	148.34	10.05	52.47	4.87	4.46

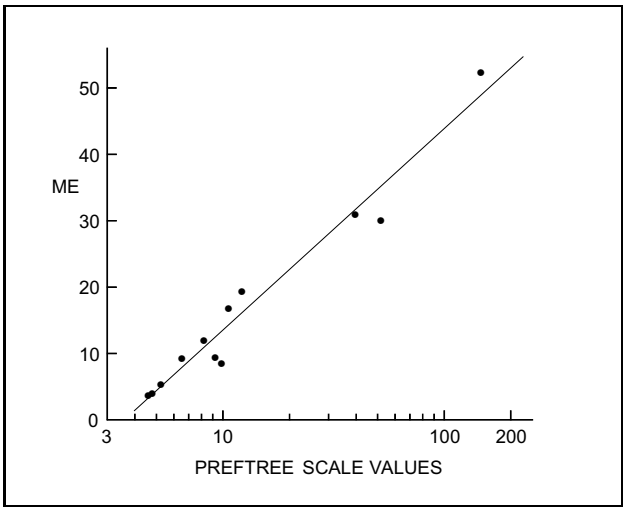

Figure 2. Relationship between preference-tree scale values (abscissa) and direct magnitude estimates (ordinate). The solid line is the best-fitting linear regression of the magnitude estimates onto the logarithms of the preference-tree values.

this assumption for the present data set, a number of psychoacoustic indices, most of which are based on the work of Zwicker and Fastl [25, 27], were computed for the 12 sounds investigated (see method section). Particularly, the contribution of loudness (N)[5], sharpness (S), roughness (R), fluctuation strength (FS), and tonalness (PR) to overall unpleasantness was investigated as has been done in previous studies [28, 29, 30, 16]. The metrics were operationalized as specified in Table I; for the statistical analysis it turned out to be more informative to dichotomize the prominence-ratio values [31] into a "tonal component present" (1) and "not present" (0) score, since for four sounds, no tonal component was detected.[6]

As for the unpleasantness scores to be predicted, both the linear preference-tree scale values, and their logarithms were explored. For the analyses presented below, the (natural) logarithm of the preference-tree scale values was used for two reasons: (1) It generally led to a marginally better prediction, and (2) it is linear with the more conventional magnitude-estimation measure of unpleasantness, as is evident from Figure 2. Further note that most instrumental sound-quality metrics, i.e. the predictors for the overall evaluation, have also been developed based on magnitude-estimation paradigms. When using the logarithm of the estimated choice-model values as has been done by other investigators [17, 32], the ratio scale will of course be transformed into a *difference scale*, on which interval information, only, is meaningful.

Applying a standard multiple, linear regression model using the five psychoacoustic metrics as predictors, and the natural logarithm of the preference-tree scores as the criterion, did not produce a very encouraging result: Using a 'step-wise' approach to include or exclude predictors, the only statistically significant predictor found was mean sharpness (S_{mean}) which accounted for but 31.0% of the variance in unpleasantness scores. Subsequent analyses revealed that the psychoacoustic indices did not fare well in explaining the unpleasantness generated by the 'wasp' sound, leading to the conclusion that in this case, other (psycho-)acoustic, or, more likely, non-acoustical factors came into play. When the 'wasp'-sound was excluded from the analysis, mean sharpness remained the only statistically relevant linear contributor to unpleasantness, with its predictive power more than doubled to $R^2_{corr} = 0.771$.

3.5. Preference-tree structure and psychoacoustic metrics

In combining potential predictors of unpleasantness uniformly accross stimuli, it is assumed that the magnitude of a predictor's contribution to perceived unpleasantness is equally strong for all sounds. Thus, the sub-grouping information inherent in the preference-tree structure (see Figure 1) is being ignored. It may well be the case, however, that in the subgroups of sounds found, different psychoacoustic predictors contribute to unpleasantness, or do so to a different degree.

In order to test this more complex hypothesis, a conditional multiple linear regression or, more simply, *moderator analysis* [33, 34] was performed, in which both the simple psychoacoustic metrics as well as those weighted

[5] It turned out that the mean values of these parameters, over time, lead to better predictions than any of the percentile measures often used.

[6] All analyses presented here are based on the maximum values, across the ears, of the psychoacoustic indices. It is worth mentioning that using the minimum or the mean values across the ears led to essentially the same results in all cases, that is, to the same set of predictors, with the predictive power of the regression models decreasing by less than 4% of the variance accounted for.

by a group factor (the so-called interactions with the group) were entered as predictors. Generally, both sets of predictors covary to a large degree, leading to difficulties in applying the regression algorithms. To circumvent this problem of multicollinearity, the predictors were centered around their expected value prior to the regression analysis, as is recommended in the literature [33].

As a result, only two of the ten predictors showed a statistically significant influence on unpleasantness. As with the multiple linear regression reported before, mean sharpness S_{mean} was found to contribute equally strong to unpleasantness in both groups defined by the preference tree, while mean roughness R_{mean} had a larger effect on the unpleasantness of the loud sounds (rightmost branch in Figure 1) than the soft sounds. The resulting regression-model

$$ln(PT) = 0.674 * S_{mean} + 2.518 - \begin{cases} 0.638 * R_{mean} & \text{(loud sounds)} \\ 1.276 * R_{mean} & \text{(soft sounds)} \end{cases} \quad (3)$$

predicted the data quite well, $R^2_{corr} = 0.912$.

Unexpectedly, as seen in equation 3, roughness (the inclusion of which increased the variance accounted for by roughly 14%) contributed *negatively* to unpleasantness; the rougher the sound in the respective stimulus group, the less unpleasant it was judged.

4. Discussion

The results of the present study have implications for three issues which shall be discussed in turn: (1) the usefulness of the choice-model approach taken, (2) modeling auditory unpleasantness, and (3) the relationship between direct and indirect scaling of auditory attributes.

4.1. Advantages of the choice-model approach

The main outcome of the present study is that it was possible to model paired comparisons of the unpleasantness of a set of environmental sounds in such a way that a ratio scale of the stimuli emerged. On such a scale unpleasantness values may be interpreted as mathematical ratios, rendering statements such as "the circular saw sounds more than 14 times as unpleasant as the fan-noise" meaningful. Note that this is not possible, though often claimed, with direct-estimation scales that postulate ratio properties via instructions to the participants [4, 35]. Further note that the scaling model evaluated based on probabilistic choice theory (be it a BTL model or a preference tree) always has a chance to fail, that is, to be rejected on statistical grounds; thus lending its acceptance greater scientific credibility.

For the present experiment, accounting for the data in terms of the BTL model did in fact fail. That requires some explanation, since a previous investigation [16] had derived a BTL scale for the auditory unpleasantness of a set of similarly heterogeneous environmental sounds. The sounds used in the present investigation were better documented, and more carefully selected [23] for eliciting a wide variety of auditory attributes. The more sounds are to be compared, however, and the more attributes available, the more likely is it that subgroups of sounds exhibiting similarities will emerge. This is known to conflict with the context independence required by the BTL model [18], and should lead us to expect its validity to be the exception, rather than the rule. Other investigations from our laboratory, both on the tonalness [20], and the unpleasantness of tyre sounds [32], always required more complex choice models, i.e. preference trees, to account for the data. It may be suspected that in some of the investigations using the BTL approach, it was not put to a severe test, or not evaluated with respect to its alternatives.

What does it mean, that the present data may be modeled by a "preference tree" (s. Figure 1)? As with the BTL model, a ratio-scale is obtained, but a different decision model is shown to form its basis: One according to which the attributes used in the paired comparison *change* depending on the stimuli to be compared. More specifically, subgroups of stimuli are being identified which share certain features that are disregarded in making comparisons within the group, but become relevant if sounds from different groups are compared. Thus, in accordance with Figure 1, it turned out that loudness did not play a role in comparing the unpleasantness of the 'stadium' sound with the 'dentist's drill', for example, while it was used when comparing the unpleasantness of the 'stadium' to that of the 'fan'.

The multiple-regression analysis of the instrumental measures of sound character provides further information: When the group membership obtained from the preference tree is introduced as a *moderator variable*, combination metrics with different weights result for these two groups. The details of how that is conceptualized will be discussed in the next section.

4.2. Modeling auditory unpleasantness

The outcome of the present preference-tree modeling, though based on a relatively small sample of sounds, pinpoints some of the problems in predicting overall auditory unpleasantness. First of all, the fact that the sound of a wasp required a separate parameter (i.e. branch) in the model, and that its high annoyance was not accounted for by a particular loudness, roughness, or sharpness, illustrates the importance of non-acoustical factors in explaining auditory unpleasantness. Based on comments by participants some of whom reported to instinctively have ducked, or tried to wave off "the bee at their left ear", the excess annoyance of this sound may be tentatively characterized as being due to its *intrusiveness*. It is typically stated in the literature [25] that effects like that cannot be part of psychoacoustical modeling; they are subject to scientific investigation using other methods [36, 37].

But even when we focus on the two major branches of the present preference-tree model, which are quite well acounted for by psychoacoustical parameters, the fact that listeners seem to shift criteria in comparing the sounds argues against generic pleasantness or "unbiased annoy-

ance" metrics as have been proposed in the literature [29, 30, 28]. For the present outcome, it seems that two groups of sounds are formed (s. Figure 1), which are largely defined by different loudness ranges, suggesting that loudness is dominant when comparing *across* these groups. That interpretation is supported by the point-biserial correlation between group membership and loudness being $r = 0.735$, $p = 0.01$, and a stepwise discriminant analysis showing that loudness is the only statistically significant discriminator between the two groups; Wilk's $\lambda = 0.460$, $F(1, 9) = 10.58$, $p = 0.01$. As seen from the moderator analysis, *within each group*, loudness is ignored, and other parameters determine the outcome of the comparison: Overall, the estimated unpleasantness scores are best predicted by a combination of sharpness and roughness, as stated in eq. 3, with roughness being weighted differently in the two groups.

It should be emphasized, however, that the variables entering in our modeling of unpleasantness are the same that have been discussed elsewhere [28, 25, 38], and that a regression model incorporating only these well-established metrics goes a long way in explaining judgments of overall sound quality, leaving not all that much room for non-acoustical factors to play a role.

4.3. Direct vs. indirect scaling

Clearly, in many sound-quality applications, collecting data on the factorial paired-comparison matrix with a large number of subjects as was done here, is not feasable. Therefore, it is worth investigating whether a shortcut to get to the scale via direct estimation methods may be taken. The present data collection using magnitude estimation with a standard indicates that the estimates given by the participants not only reflect the rank order of the sounds when compared to the preference tree scale, but also contain information about the relative distances between the test objects (as is evident in Figure 2).

The magnitude-estimation scale is, however, strongly compressive when compared to the scale values estimated from the paired comparisons, leading to a non-linear relation between the two types of scales. Such non-linear relations are often observed in psychophysics, for example between magnitude, and category scales [39, 40]. Typically, they are due to some constraint on one of the scales. In the present situation, cautious judgments avoiding extreme numerical assignments ("regression bias", [41]) might have produced the compression of the magnitude scale. That is highly speculative, though, and in fact, there is no good reason to believe, that the "raw" outcomes of direct-estimation scales will be linearly related to sensation magnitude [4, 35].

More importantly, the direct estimation techniques have two shortcomings with respect to the choice-model approach advocated here: (1) their outcomes cannot be falsified by subjecting them to rigorous transitivity checks, and (2) they are not suited to reveal the *structure* inherent in the judgments. That is they would not detect the kind of criterion shifts observed in the present data, and thus lead to incomplete modeling of the listeners' behavior. Note that collecting additional ratings on the features supposedly underlying the decisions will not remedy the situation, since such ratings are fraught with the same problem of undetermined dimensionality. Furthermore, the relevant features might not be known or not be accessible to explicit labelling.

4.4. Conclusion

Regarding the general questions raised in the introduction, we may conclude in summary that a consistent, and tractable, measure of overall quality (in this case: unpleasantness) can be obtained. This measure has ratio-scale properties, and is based on multiple attributes with decision criteria shifting dependent on the sounds under consideration. Finally, the auditory unpleasantness thus measured may be predicted quite well by conventional psychoacoustic metrics.

Therefore we hold that the choice-model approach illustrated here is well suited to play a role as a basic-science instrument when it comes to establishing the dimensionality of an attribute, and when the structure of a domain will have to be explored. To this end, the less restrictive models (preference trees, elimination-by-aspects) which have not been previously applied in psychoacoustics, have greater potential than the BTL heuristic previously used.

Acknowledgement

This investigation was carried out while the authors were with the Sound Quality Research Unit (SQRU) at the Department of Acoustics, Aalborg University. The unit receives financial support from Bang and Olufsen, Brüel & Kjær, and Delta Acoustics and Vibration, as well as from the Danish National Agency for Industry and Trade (EFS), and the Danish Technical Research Council (STVF).
We would like to thank Sylvain Choisel and Ville Sivonen for their help with calibrating the equipment, and configuring the measurement of sound-quality metrics, respectively, and two anonymous reviewers for their constructive comments and suggestions, which improved the quality of the paper.

References

[1] J. Blauert, U. Jekosch: Concepts behind sound quality: Some basic considerations. Proceedings of the Inter-Noise, Jeju, Korea, 2003, paper N466 (CD–ROM).

[2] K. Zimmer, W. Ellermeier: Deriving ratio-scale measures of sound quality from paired comparisons. Noise Contr. Eng. J. **51** (2003) 210–215.

[3] L. Narens: A theory of ratio magnitude estimation. J. Math. Psych. **40** (1996) 109–129.

[4] W. Ellermeier, G. Faulhammer: Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. Percept. Psychophys. **62** (2000) 1505–1511.

[5] L. Narens, R. D. Luce: Measurement: The theory of numerical assignments. Psych. Bull. **99** (1986) 166–180.

[6] G. Iverson, R. D. Luce: The representational measurement approach to psychophysical and judgmental problems. – In:

Measurement, judgment, and decision making. M. H. Birnbaum (ed.). Academic Press, San Diego, 1998, 1–79.

[7] J. D. Carroll, G. De Soete: Toward a new paradigm for the study of multiattribute choice behavior: Spatial and discrete modeling of pairwise preferences. Am. Psychologist **46** (1991) 342–351.

[8] P. Slovic, S. Lichtenstein, B. Fischoff: Decision making. – In: Stevens' Handbook of Experimental Psychology. 2nd ed. Vol. 2. R. Atkinson, R. J. Herrnstein, G. Lindzey, R. D. Luce (eds.). Wiley, New York, 1988, 673–738.

[9] I. Borg, P. Groenen: Modern multidimensional scaling: Theory and applications. Springer, New York, 1997.

[10] R. D. Luce: Individual choice behavior. Wiley, New York, 1959.

[11] R. A. Bradley, M. E. Terry: Rank analysis of incomplete block designs. I. The method of pair comparisons. Biometrika **39** (1952) 324–345.

[12] P. C. Laux, P. Davies, G. R. Long: The correlation of subjective response data with measured noise indices of low-frequency modulated noise. Noise Control Eng. J. **40** (1993) 241–255.

[13] P. Daniel, R. Weber: Psychoacoustical roughness: Implementation of an optimized model. Acta acustica - Acustica **83** (1997) 113–123.

[14] D. Pressnitzer, S. McAdams: Two phase effects on roughness perception. J. Acoust. Soc. Am. **105** (1999) 2773–2782.

[15] D. Pressnitzer, S. McAdams, S. Winsberg, J. Fineberg: Perception of musical tension for non-tonal orchestral timbres and its relation to psychoacoustic roughness. Percept. Psychophys. **62** (2000) 66–80.

[16] W. Ellermeier, M. Mader, P. Daniel: Scaling the unpleasantness of sounds according to the BTL model: Ratio-scale representation and psychoacoustical analysis. Acta acustica - Acustica **90** (2004) 101–107.

[17] M. Marzinzik, B. Kollmeier: Predicting the subjective quality of noise reduction algorithms for hearing aids. Acta acustica - Acustica **89** (2003) 521–529.

[18] A. Tversky, S. Sattath: Preference trees. Psychological Review **86** (1979) 542–573.

[19] A. Tversky: Elimination by aspects: A theory of choice. Psychological Review **79** (1972) 281–299.

[20] W. Ellermeier, P. Daniel: Tonal components in tire sounds: Refined subjective and computational procedures. – In: Proceedings of the Sound Quality Symposium (SQS2002) at Inter-Noise 2002. G. Ebbitt, P. Davies (eds.). Institute of Noise Control Engineering (INCE-USA), Iowa State University, Ames, IA, 2002.

[21] W. Ellermeier, P. Daniel: Effekt des Signal-Rausch-Abstands und der Bandbreite tonaler Komponenten auf die wahrgenommene Tonhaltigkeit. Fortschritte der Akustik, DAGA 2003. Deutsche Gesellschaft für Akustik, Oldenburg, 2003.

[22] F. Wickelmaier, C. Schmid: A matlab function to estimate choice-model parameters from paired-comparison data. Beh. Res. Meth. Instr. & Computers **36** (2004) 29–40.

[23] K. Johannsen, H. U. Prante: Environmental sounds for psychoacoustic testing. Acustica - acta acustica **87** (2001) 290–293.

[24] F. Christensen, H. Møller: The design of VALDEMAR - an artificial head for binaural recording purposes. Proceedings of the Audio Engineering Society's 109th Convention, Los Angeles, California, USA, September 22-25, 2000, preprint No. 5253.

[25] E. Zwicker, H. Fastl: Psychoacoustics. Facts and models. 2nd ed. Springer, Berlin, 1999.

[26] M. G. Kendall: Rank correlation methods. 3rd ed. Griffin, London, 1962.

[27] H. Fastl: The psychoacoustics of sound-quality evaluation. Acta acustica - Acustica **83** (1997) 754–764.

[28] W. Aures: Der sensorische Wohklang als Funktion psychoakustischer Empfindungsgrößen [Sensory pleasantness as a function of psychoacoustical parameters]. Acustica **58** (1985) 282–290.

[29] E. Zwicker: Ein Vorschlag zur Definition und zur Berechnung der unbeeinflußten Lästigkeit. Zeitschrift für Lärmbekämpfung **38** (1991) 91–97.

[30] E. Zwicker: A proposal for defining and calculating the unbiased annoyance. – In: Contributions to Psychological Acoustics. A. Schick, J. Hellbrück, R. Weber (eds.). BIS, Oldenburg, Germany, 1991, 187–202.

[31] G. R. Bienvenue, M. A. Nobile: Prominence ratio for noise spectra with discrete tones: A procedure based on Zwicker's critical band research. Proceedings of the Internoise, 1991, 53–55.

[32] P. Daniel, W. Ellermeier, P. Leclerc: Tonalness and unpleasantness of tire sounds: Methods of assessment and psychoacoustical modelling. Proc. Euro-Noise 98, Munich, Germany, 1998, 627–632.

[33] L. S. Aiken, S. G. West: Multiple regression: Testing and interpreting interactions. Sage, Newbury Park, CA, 1991.

[34] R. Steyer: Wahrscheinlichkeit und Regression. Springer, Berlin, 2003.

[35] K. Zimmer, O. Baumann: Direct scaling of sensations: Are subjects able to produce consistent, and meaningful, numerical ratios? First ISCA Tutorial & Research Workshop on Auditory Quality of Systems, MtCenis, Germany, 2003, 85–90.

[36] H. Fastl: Neutralizing the meaning of sound for sound quality evaluation. Proc. Intl. Congr. on Acoustics (ICA), Rome, 2001, paper 4KN1.03 (CD–ROM).

[37] W. Ellermeier, A. Zeitler, H. Fastl: Impact of source identifiability on perceived loudness. Proc. Intl. Congr. on Acoustics (ICA), Kyoto, 2004.

[38] E. Terhardt, G. Stoll: Skalierung des Wohlklangs (der sensorischen Konsonanz) von 17 Umweltschallen und Untersuchung der beteiligten Hörparameter [Scaling the pleasantness (sensory consonance) of 17 environmental sounds and investigation of the contributing hearing sensations]. Acustica **48** (1981) 247–253.

[39] S. S. Stevens, E. H. Galanter: Ratio scales and category scales for a dozen perceptual continua. J. Exp. Psychol. **54** (1957) 377–411.

[40] G. A. Gescheider: Psychophysics - the fundamentals (3rd. ed). Erlbaum, Mahwah, NJ, 1997.

[41] S. S. Stevens, H. B. Greenbaum: Regression effect in psychophysical judgment. Percept. Psychophys. **1** (1966) 439–446.