ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 90 (2004) 101 – 107

Scaling the Unpleasantness of Sounds According to the BTL Model: Ratio-Scale Representation and Psychoacoustical Analysis

Wolfgang Ellermeier
Department of Acoustics, Fredric Bajers Vej 7 B5, DK-9220 Aalborg Ø, Denmark. we@acoustics.dk

Markus Mader
Institut für Experimentelle Psychologie, Universität Regensburg, 93040 Regensburg, Germany

Peter Daniel
Cortex Instruments GmbH, Erzbischof-Buchberger-Allee 14, 93051 Regensburg, Germany

Summary
The goal of the present study was to determine, (1) whether auditory unpleasantness is jugded consistently across a wide range of acoustic stimuli, and (2) which sound features contribute to that sensation. To that effect, all possible paired comparisons on a heterogeneous set of ten environmental sounds were collected from 60 listeners. The judgments conformed with the highly restrictive *BTL model* [1], thus justifying a ratio-scale representation of perceived unpleasantness. The resulting scale values varied by a factor exceeding 100 (boat diesel vs. jackhammer). While they were not predicted by differences in A-weighted sound pressure level, a linear combination of the psychoacoustic parameters of roughness and sharpness accounted for more than 94% of the variance in perceived unpleasantness.

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

1. Introduction

One and a half century of psychophysics have endowed us with sophisticated methodologies to measure detection thresholds, indices of discriminability, and sensory equivalences. When it comes to quantifying the magnitude of sensations at suprathreshold levels, however, we are often left with unsatisfactory alternatives. Direct scaling methods such as category scaling or Stevens' magnitude estimation provide us with numerical estimates of the stimuli under investigation, but the validity and scale type of these measurements remains doubtful [2, 3].

Procedures derived from axiomatic measurement theory [4, 5] on the other hand, explicitly address these problems by formulating the conditions (axioms) under which measurement is possible and by specifying the scale type of the outcome.

An example for such a well-founded approach is the Bradley-Terry-Luce (BTL) model [6, 1]. It may be derived from the more general *choice axiom* [1], and states that given certain testable conditions, preference probabilities may be related to scale values in the following fashion:

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

where p_{ab} denotes the probability of "prefering" object a over object b (or judging it to sound louder, feel more painful, appear brighter in appropriate paired comparisons), and $v(a)$, $v(b)$ are the scale values of these objects. Note that the v-scale values constitute a ratio-scale, unique up to multiplication with a positive constant[1]. This is more evident, when eq. (1) is rewritten as:

$$p_{ab} = \frac{1}{1 + \frac{v(b)}{v(a)}} = \frac{1}{1 + \frac{k*v(b)}{k*v(a)}}. \tag{2}$$

Obviously, multiplication by a positive constant k neither changes the ratio of two scale values, nor its relationship to the preference probabilities.

It should be emphasized that the BTL model (eq. 1) implies a very strong form of stochastic transitivity: Not only do the preference probabilities have to be rank-ordered systematically, but given two such probabilities (p_{ab} and p_{bc}), a third one (p_{ac}) may be computed precisely. Clearly,

Received 20 December 2001,
accepted 20 September 2003.

[1] For a more general treatment of the uniqueness properties of BTL systems, see [7].

such a restrictive model need not hold for any given data set. In this respect, obtaining a ratio-scale via scale construction is fundamentally different from instructing subjects to produce numerical ratio (or category) judgments, where typically, no such consistency checks are applied.

The advantages of using the approach may be summarized as follows: (1) The BTL model is falsifyable as such, and it specifies the conditions under which an attempt at scaling may *fail*. (2) It separates data collection (which is achieved by obtaining a complete set of paired comparisons) from the enterprise of assigning scale values. (3) If successful, it leads to a ratio-scale representation of the objects studied.

Earlier attempts to apply the model - mostly in economics and sociology - have been summarized by Luce [8]. Selected, more recent applications include studies of attitudes towards politicians [9], of taste qualities of champagne [10], and of facial attractiveness [11, 12].

With the notable exception of a few recent applications to specific stimulus materials [13, 14, 15, 16, 17] in which no rigorous model tests were performed, the BTL model has not been used to investigate substantial problems in auditory perception. Recently, however, the need for more refined suprathreshold measurement in psychoacoustics has arisen in the applied field of "sound quality evaluation," dominated by pragmatic ad-hoc methods of uni- or multidimensional scaling [18]. The present study is thus in line with recent attempts by other investigators [19, 20] to derive tractable subjective representations from the most simple preference or similarity ratings conceivable, gaining sophistication from subsequent modeling or statistical analysis, rather than expecting it to be present in the subjects' semantic or numerical judgments. Furthermore, psychoacousticians attempting to compute auditory attributes such as loudness, annoyance, tonal character, for example, from the signal directly (s. [21]) often use implicit (and untested) assumptions about ratio-scale properties of the measures they employ.

Therefore, the present study was designed with two goals in mind:

1. To determine whether a ratio-scale of the "unpleasantness" of sounds may be derived from paired-comparison data, and
2. to relate the outcome to more elementary auditory sensations which might contribute to the sensation of "unpleasantness."

Unpleasantness, rather than "annoyance" was chosen as the attribute to be judged, since the latter is often conceptualized in reference to the interference with other tasks the listener is trying to accomplish [18]. Since in the present laboratory study, the listener's undivided attention was focused on the stimuli, judging their unpleasantness comes very close to what other authors have called "unbiased annoyance" [22, 23], in an attempt to minimize non-auditory effects in the judgments elicited.

Furthermore, care was taken to include stimuli of maximal heterogeneity, for which previous attempts at fitting the BTL model have been most successful [8]. Further studies in progress in our laboratory will investigate more homogeneous sets of sounds, and perceptual dimensions less complex than "unpleasantness."

2. Method

2.1. Subjects

Sixty subjects, none of whom reported any hearing problems, took part in the experiment. The majority of the sample was drawn from a pool of psychology students participating to fulfil a degree requirement. The sample had a mean age of 24.14 years (range: 19–35 years) and consisted of 40 female and 20 male participants.

2.2. Stimuli

The sounds to be compared were chosen from a database recorded "in the field" by an environmental agency (Medizinisches Institut für Umwelthygiene, Düsseldorf) for noise-evaluation studies. From the 64 sounds which were available to us on digital audio tape, we selected ten for their heterogeneity in source and sound characteristics. These sounds are identified in Table I, and consist of natural, traffic, and industrial noises ranging from "water running from a faucet" to the sound of a "jackhammer".

For presentation in a paired-comparison paradigm, all sounds were recorded in WAVE file format with 16 bit precision, and a 22-kHz sampling rate. They were shortened to a uniform duration of 6 s, including linear rise/decay ramps of approximately 0.2 s. Like the "originals", the resulting sound samples had vastly different (linear) sound pressure levels ranging from 60 to 81 dB SPL. Further acoustical measurements made on these sounds are listed in Table II and discussed in the results section.

2.3. Apparatus

A computer program controlled sound presentation and response collection. The stimuli were stored on the hard disc and played via a 16-bit sound card (Soundblaster AWE PnP). From the output of the sound card, the signal was diotically delivered to Beyerdynamic DT 550 headphones after adequate amplification by a commercially available preamplifier (Linn K 33). Throughout the experiment the subject was seated in a double-walled sound-attenuating chamber.

2.4. Procedure

Each subject was presented with all possible pairs of the 10 sounds selected. However, each subject judged only one of the two orderings (a,b) or (b,a) of a given pair of sounds. Subjects proceeded through different random sequences of the ensueing $n(n-1)/2 = 45$ comparisons with two constraints imposed: (1) To prevent stereotypical responding which might occur when one particular sound was repeatedly presented as the first one in a pair, this was not to occur more than 6 times (out of a maximal 10) for a given sound and subject. (2) To counterbalance

Table I. Cumulative preference matrix ($N = 60$). *Note.* Absolute frequencies are given with which the sound in the row was judged to be more unpleasant than the sound in the column. Sounds: 1 - truck, 2 - brake, 3 - train, 4 - water, 5 - boat, 6 - jackhammer, 7 - mower, 8 - crash, 9 - mixer, 10 - vent.

	1	2	3	4	5	6	7	8	9	10
1	-	9	16	45	56	5	29	6	24	33
2	51	-	34	58	58	13	46	30	39	50
3	44	26	-	55	57	9	48	37	38	55
4	15	2	5	-	38	2	17	6	6	20
5	4	2	3	22	-	3	6	3	3	12
6	55	47	51	58	57	-	58	53	55	57
7	31	14	12	43	54	2	-	16	17	41
8	54	30	23	54	57	7	44	-	40	52
9	36	21	22	54	57	5	43	20	-	43
10	27	10	5	40	48	3	19	8	17	-

order effects across the sample, successive subjects were paired, and received the complement of each other's order of presentation. So if subject 21 was presented with the pair (a,b), subject 22 received the pair (b,a), that is the opposite cell in the 10×10 matrix.

A trial consisted of the presentation of a pair of sounds separated by a 2-s pause. In order to facilitate the coupling of observation intervals and response buttons, along with the 6-s duration of each sound, an LED located on the left (first sound) resp. on the right side (second sound) of a panel mounted in front of the subject was illuminated. Following the termination of the second sound, the subject was to decide which of the two stimuli sounded more "unpleasant" by pressing a button either on the left (indicating the first sound) or on the right armrest (indicating the second sound to be more unpleasant). Once the subject had pressed a button, one of two feedback lights located next to (and spatially congruent with) the LEDs marking the observation intervals was briefly flashed. Following a 2-s inter-trial interval, the next pair of sounds was presented. Completing all 45 paired comparisons took approximately 30 min.

3. Results

3.1. Consistency checks

Since each participant performed all possible paired comparisons among the 10 sounds, the individual data sets may be inspected for transitivities before further processing. This is typically done by determining the number of *circular triads* for which $a > b$, $b > c$, but $a < c$. Doing this revealed a median number of $d = 5$ circular triads ($Min = 0$, $Max = 17$) out of the maximal 40 (s. [24], Chap. 9.5.1) inconsistencies to be generated from $\binom{10}{3} = 120$ triads. A χ^2-test [25] performed to evaluate whether the number of circular triads significantly deviates from the number to be expected by chance alone, turned out to be insignificant for each of the 60 participants ($\alpha = 0.05$).

Consequently, the individual paired-comparison matrices may be pooled across the 60 subjects, resulting in the cumulative preference matrix given in Table I. In this matrix, each entry specifies the absolute frequency with which the sound identified by the row of the table was judged as more unpleasant than the sound identified by the columm of the table. It is conceivable that the pooled matrix becomes inconsistent, even though all participants judged consistently, but in different ways. Therefore, the data in Table I were evaluated with respect to *weak stochastic transitivity* (WST) meaning that if $p_{ab} \geq .5$ and $p_{bc} \geq .5$, then $p_{ac} \geq .5$ [26]. There were no violations of this condition in the 120 instances in which the premise held. Since WST - a prerequisite for an ordinal representation of the data - was fulfilled for the pooled paired-comparison matrix, we proceeded to the evaluation of the much more restrictive BTL model.

3.2. Model evaluation

Assuming the validity of the BTL model (as specified in equation 1) those v-scale parameters that best predicted the preference probabilities were numerically estimated using a maximum-likelihood procedure. In order to statistically evaluate the validity of this model, a likelihood-ratio test was performed (s. [27], Chap. 6). In principle, such a test evaluates the likelihood of a restricted model (here: the BTL model) against the likelihood of an unrestricted model (here: the "statistical" model, which assumes independent binomial distributions to generate the entries in each cell). That is done by computing the term $-2\ln(L_{BTL}/L_S)$, where L_{BTL} denotes the likelihood of the BTL model, and L_S denotes that of the statistical model, the whole term being approximately χ^2-distributed. In the present application, a model having only 10 parameters, the v-scale values for the 10 sounds, is compared to a 45-parameter model in which the preference probabilities are simply estimated from the actual relative frequencies in each cell. If the restricted (BTL) model does not fare significantly worse than the unrestricted (statistical) model, the former may be said to hold.

Both estimating the scale values, and evaluating the validity of the BTL model was accomplished using a Matlab function provided by Wickelmaier & Schmid [28]. Performing the likelihood-ratio test showed that the BTL model describes the present paired-comparison data quite well; the null hypothesis assuming its validity may not be rejected, $\chi^2(36) = 38.14; p = 0.373$.

Consequently, v-scale values may be assigned to the sounds. Since, theoretically, v-scales are ratio scales, that may be done by arbitrarily defining a "unit of measurement." For the present data set, one of the sounds ("a truck passing by"; first entry in Table I) was assigned a scale value of ten; all other scale values were estimated relative to this reference. The resulting "unpleasantness" values are plotted in Figure 1. To provide an indicator of the precision of these estimates, 95% confidence intervals were determined using the method proposed by Bradley [29, 30]. Note that the BTL scale values cover a consid-

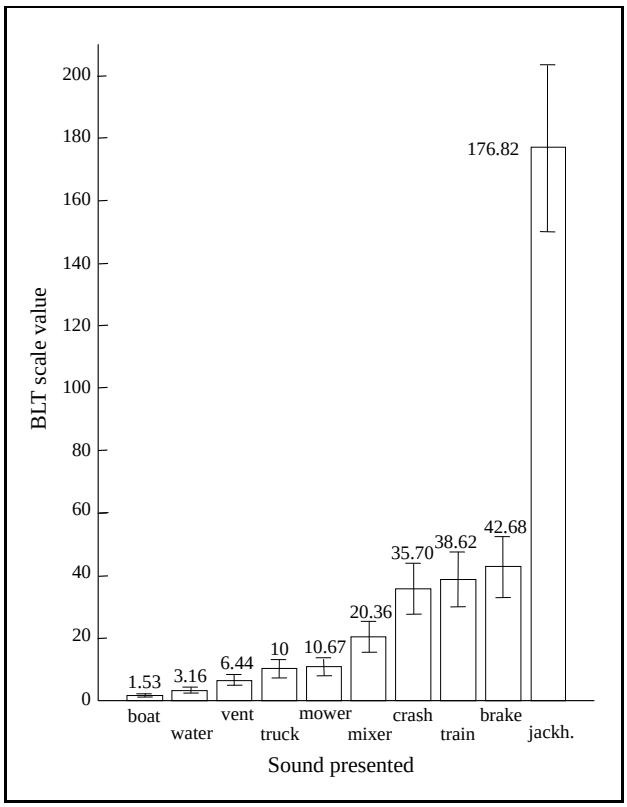

Figure 1. Estimated unpleasantness scale values according to the Bradley-Terry-Luce (BTL) model are plotted for the ten sounds studied along with 95% confidence intervals. The estimates are based on a total of 45 (paired comparisons) × 60 (participants) = 2700 judgments.

erable range, and suggest that the sound of a jackhammer ($v_{10} = 176.82$) is more unpleasant to the ear than the puttering diesel of a boat ($v_1 = 1.53$) by a factor exceeding one hundred.

Given the use of a well-founded methodology we may not only make (ratio-type) statements of this sort, but also appear to achieve greater resolution with respect to the objects we wish to discriminate than is typical of conventional direct rating or magnitude estimation scales.

4. Psychoacoustical analyses

Though it may be of scientific interest to show that an arbitrary sampling of heterogeneous environmental sounds may be represented on a unidimensional ratio scale of "unpleasantness," this finding appears to have little practical relevance for inferences to be made beyond the ten unique sounds studied. Therefore, we decided to investigate, which - more elementary - auditory sensations present in all of these sounds contribute to the impression of their "unpleasantness."

Rather than having subjects estimate each sound's loudness, pitch quality, or temporal structure, with the risk of bias due to the small sample size, and potential inconsistencies due to the multidimensional nature of the stimuli, we decided to use well-established *objective* algorithms to extract psychoacoustical indices (cf. [21]) from the sound samples directly. To that effect, and to include potential distortions in the signal path, the output of the headphones was recorded through a calibrated artificial-head system (Cortex MK1) using a 22-Hz high-pass filter. The recordings thus made were subjected to psychoacoustical analyses as implemented on commerically available sound analysis software (VIPER V2.20; [31]).

In line with previous results (e.g. [32]) predominantly based on the method of direct ratio estimation (s. [33]), we focused on four measures which have been shown to contribute to the unpleasantness of a sound: its loudness, roughness, sharpness, and fluctuation strength. *Loudness* of complex sounds depends on bandwidth, temporal integration, and the spectral distribution of energy, and is consequently not always monotonically related to overall intensity measures [34, 35, 36, 21]. The sensation of *sharpness* [37] depends on the degree to which high-frequency components are present in the sound to be evaluated. *Fluctuation strength* is a sensation elicited by the perception of temporal changes, typically due to slow amplitude or frequency modulations with a peak value around 4 Hz [38, 39]. At high modulation frequencies, when the changes in loudness or pitch are no longer resolvable, it turns into the sensation of *roughness* [40, 15] which has its maximum at a modulation rate of 70 Hz.

The exact algorithms employed in objectively analyzing the ten test stimuli closely followed the formulae and evidence compiled by Zwicker & Fastl [21]. Note that within this conceptualization all quantities are expressed with respect to some physically-defined unit reference, and are treated as having ratio-scale properties. The resulting psychoacoustic indices are given in Table II, along with the A-weighted sound pressure level averaged over the course of the stimulus, and with the unpleasantness score derived from the paired-comparison experiment.

From a psychophysical viewpoint, it is encouraging to see that the conventional physical description clearly fails in predicting unpleasantness scores: Relating the BTL scale values to the A-weighted sound-pressure levels[2] of the sounds yields a (non-significant) correlation of $r = 0.52$ ($p = 0.12$, two-tailed test). Some of the psychoacoustical parameters, especially median loudness[3] ($N50$) in sones ($r = 0.71$; $p = 0.021$), and average roughness ($r = 0.97$; $p < 0.001$) fare much better in this respect (see Table II).

An attempt was made to improve the variance accounted for by linearly combining the psychoacoustical predictors. Note, however, that given the high correlation between the BTL scale values and psychoacoustical roughness, there is not much room for improvement. Consequently, a "stepwise multiple regression" in which those parameters that

[2] The correlation with SPL improved to yield $r = 0.71$ when it was computed using the logarithm of the BTL scale values, suggesting a nonlinear, i.e. power-function, relationship between perceived unpleasantness and sound pressure.

[3] For the present stimuli, the often preferable 5th loudness percentile ($N5$) did slightly worse in predicting unpleasantness judgments.

Table II. Psychoacoustic parameters computed from the sounds. *Note.* Units of measurement as defined in Zwicker & Fastl [21]. Correlations between the psychoacoustic parameters and BTL unpleasantness scale values are given in the bottom row.

	Sound	**BTL scale value**	Roughness [asper]	Sharpness [acum]	Loudness [sone]	Fluct. Strength [vacil]	SPL [dB(A)]
(1)	boat	1.53	0.25	0.87	10.7	0.11	56.4
(2)	water	3.16	0.32	2.15	14.0	0.14	60.4
(3)	vent	6.44	0.32	1.07	28.0	0.04	70.6
(4)	truck	10	0.35	0.96	31.5	0.15	75.4
(5)	mower	10.67	0.31	0.93	29.3	0.07	72.9
(6)	mixer	20.36	0.32	1.47	15.8	0.06	62.2
(7)	crash	35.70	0.41	1.24	18.3	1.00	74.0
(8)	train	38.62	0.43	1.37	41.5	0.09	81.7
(9)	brake	42.68	0.33	1.68	17.8	0.49	66.7
(10)	jackh.	176.82	1.76	1.51	48.9	0.47	79.8
Correlation:			0.97	0.22	0.71	0.39	0.52

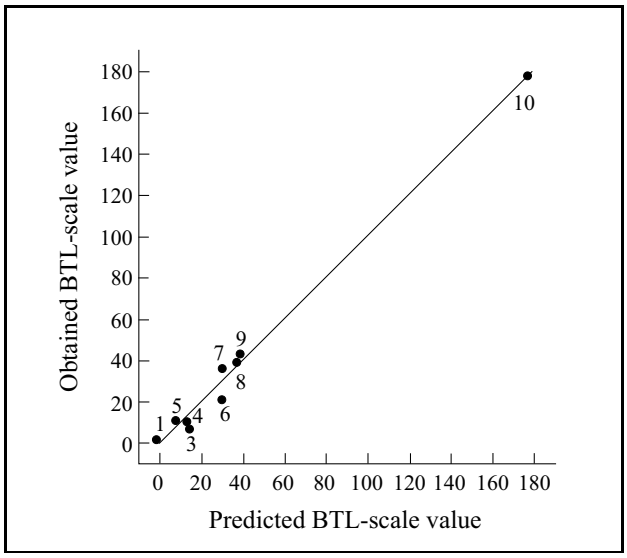

Figure 2. BTL scale values obtained from the paired-comparison experiment (ordinate) plotted over unpleasantness scores predicted from *roughness* and *sharpness* parameters (eq. 3) of the sounds (abscissa). The numbers identifying the sounds correspond to those in Table II. For a perfect fit, all data points would lie on the major diagonal (solid line).

contribute least to the prediction are successively excluded ("backward approach"; [41]), yielded a model according to which *roughness* alone accounts for the variance[4] in unpleasantness scale values quite well ($R^2_{corr} = 0.942$). Including other psychoacoustic parameters (such as fluctuation strenth, loudness or sharpness) only leads to minor, statistically insignificant improvements in prediction. Closer inspection of Table II, however, reveals that the effect of sharpness which typically contributes greatly to the perception of unpleasantness or annoyance is severely attenuated by the sound of "water running from a faucet" (sound 2), the only non-technical sound in the selection, wich shows an untypical combination of great sharpness, and low unpleasantness. If that sound is excluded from the regression analysis, a two-parameter model emerges, according to which *roughness* R and *sharpness* S are sufficient to predict the unpleasantness scores (BTL):

$$BTL = 101.62 * R + 38.60 * S - 60.12. \qquad (3)$$

This model fits the data quite well by accounting for close to 99% of the variance in judgments of the remaining 9 sounds ($R^2_{corr} = 0.988$), as is seen in Figure 2, when the predictions (x-axis) are compared with the BTL scale values (y-axis) actually obtained from the paired-comparison experiment. Even though it may seem that this is largely due to the effect of the jackhammer sound (labelled "10" in Figure 2), analysis of the remaining sounds clustering in the bottom left corner of Figure 2 reveals roughness and sharpness still to account for 84% of the variance ($R^2_{corr} = 0.843$) in unpleasantness scale values. It is obvious, however, that given a small sample of stimuli, the inclusion or exclusion of particular sounds will have great influence on the regression model suggested.

5. Discussion

The main outcome of the present study is to demonstrate that a heterogeneous set of environmental sounds is consistently judged and compared in terms of the unpleasantness of auditory sensations. More specifically: The unpleasantness scale values estimated from subjects' paired comparisons conform to the highly restrictive BTL model.

Note that - in contrast to common direct scaling approaches - this endeavor could have failed, and has often done so in other applications [8]. For the BTL model to hold, the choices made in the paired-comparison experiment must exhibit some form of "context independence" [42, 43]. That is, *all* relevant aspects contributing to a sound's unpleasantness must enter into each decision. By contrast, the model fails, if *different* auditory attributes are

[4] Since the variance accounted for, R^2, is maximized for a given sample, it overestimates the relationship in the population. To compensate for this, the more realistic R^2_{corr}, which adjusts the estimate based on the number of sounds and predictors used, is reported here, and furtheron.

considered (e.g. only those, which distinguish two stimuli most) depending on the context defined by the comparison stimuli. We thus may conclude, that in the present experiment, subjects used relevant indicators of unpleasantness, such as loudness, roughness, or sharpness in a uniform fashion across all of the comparisons they made.

This conclusion also lends greater credibility to the psychoacoustical analyses performed: First, without evidence for context independence it does not make much sense to analyse each sound by itself, as is typically done in applied "sound quality" research [39]. Second, the fact that the BTL model leads to a ratio-scale representation facilitates comparisons with more elementary sensations which have been conceptualized as ratio scales [21], even though that property remains to be demonstrated.

In the present analyses, the elementary sensations of *roughness* and, to a lesser extent, *sharpness* emerged as the main contributors to the complex sensation of (auditory) *unpleasantness*. Of course, such an outcome should not be taken as being of general validity, given that it is only based on a very limited number of sound samples. It serves to demonstrate, however, that ratio-scaled unpleasantness may be predicted from more elementary auditory attributes. Nevertheless, the present outcome is in remarkable agreement with the psychoacoustical literature [32, 15, 44]. Terhardt and Stoll, for example, in a study in which a set of 17 natural and synthetic sounds equated in loudness was assessed using two different psychophysical procedures, found a combination of roughness and sharpness to make a fairly good prediction ($r = 0.672$). Such convergence between studies conducted in different laboratories, despite obvious differences in the choice and calibration of stimuli, and despite the fact that no other study used the present, well-founded ratio-scaling technique, may serve as an indicator of the external validity of results reported here.

The fact that the contribution of *loudness* turned out to be statistically insignificant may, however, be idiosyncratic to the present choice of stimuli, where loudness and roughness correlated so highly ($r = 0.70$) that using one of them as a predictor was sufficient: Including loudness in the multiple regression equation (eq. 3) serves to increase the variance accounted for by less than one percent. For other stimulus sets, one might certainly have to consider loudness (e.g. [32, 21]) and the prominence of tonal components ("tonalness"; [44, 45]) as additional predictors of unpleasantness.

Having demonstrated that a highly restrictive scaling model holds for the unpleasantness sensations elicited by a heterogeneous set of sounds, in a next step it will have to be shown, how the present approach handles more homogeneous sets of sounds as are typically encountered in sound quality evaluation. A study investigating tire noises [46] suggests that under such circumstances, the BTL model may fail, and reveal subgroups of stimuli, for which different auditory attributes are considered when making paired comparisons. These situations will have to be analysed using more complex choice models like "preference trees" [47]. Note that similarities between stimuli, and the ensuing departures from the assumption of unidimensionality, will remain undetected when using conventional direct scaling approaches. It appears that research on noise evaluation and sound quality engineering might benefit from employing well-founded scaling techniques which have the potential of revealing the structure underlying perceptual judgments [48], and thus offer great benefits over the "pragmatic" scaling approaches currently used.

Acknowledgement

We would like to thank Florian Wickelmaier, Universitet Aalborg, for calculating the confidence intervals for the parameter estimates, and for letting us use a Matlab function [28] to analyze paired-comparison data that he developed together with Christian Schmid. We are also grateful to Fritz Müller (now at Fachhochschule Lüneburg) for supplying us with a DAT recording of the sounds we sampled from, and to three anonymous reviewers who went through considerable pains to double-check our results. Portions of the data have been presented at the 23rd meeting of the German Acoustical Society (DAGA 1997).

References

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

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

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

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

[5] 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.

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

[7] H. Colonius: Representation and uniqueness of the Bradley-Terry-Luce model for pair comparisons. Brit. J. Math. and Statist. Psychology **33** (1980) 99–103.

[8] R. D. Luce: The choice axiom after twenty years. J. Math. Psych. **15** (1977) 215–233.

[9] K.-H. Bäuml: Präferenzen zwischen politischen Kandidaten: Versuch einer Repräsentation durch BTL-Modell, Präferenzbäume und Eliminierung-nach-Aspekten [Preferences for political candidates: Representation by the BTL model, via preference trees, and through "elimination-by-aspects"]. Zeitschr. für Psychologie **199** (1991) 337–351.

[10] J. Lukas: BTL-Skalierung verschiedener Geschmacksqualitäten von Sekt [Scaling different taste qualities of champagne according to the BTL model]. Zeitschr. für experimentelle und angewandte Psychologie **38** (1991) 605–619.

[11] K.-H. Bäuml: Upright vs. upside-down faces: How interface attractiveness varies with orientation. Percept. Psychophys. **56** (1994) 163–172.

[12] J. Kissler, K.-H. Bäuml: Effects of the beholder's age on the perception of facial attractiveness. Acta Psychologica **104** (2000) 145–166.
[13] 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.
[14] N. Chouard: Loudness and unpleasantness perception in dichotic conditions. Dissertation, University of Le Mans, France, 1997.
[15] P. Daniel, R. Weber: Psychoacoustical roughness: Implementation of an optimized model. Acta acustica - Acustica **83** (1997) 113–123.
[16] D. Pressnitzer, S. McAdams: Two phase effects on roughness perception. J. Acoust. Soc. Am. **105** (1999) 2773–2782.
[17] 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.
[18] R. Guski: Psychological evaluation of sound quality. Acta acustica - Acustica **83** (1997) 765–774.
[19] S. McAdams, P. Susini, N. Misdariis, S. Winsberg: Multidimensional characterization of perceptual preference judgments of vehicle and environmental noises. EuroNoise, Munich, 1998.
[20] P. Susini, S. McAdams, S. Winsberg: A multidimensional technique for sound quality assessment. Acta acustica - Acustica **85** (1999) 650–656.
[21] E. Zwicker, H. Fastl: Psychoacoustics. Facts and models. 2nd ed. Springer, Berlin, 1999.
[22] 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.
[23] R. Guski, H. G. Bosshardt: Gibt es eine unbeeinflusste Lästigkeit? [Is there such a thing as unbiased annoyance?]. Zeitschr. für Lärmbekämpfung **39** (1992) 67–74.
[24] J. Bortz, G. A. Lienert, K. Boehnke: Verteilungsfreie Methoden in der Biostatistik. 2nd ed. Springer, Berlin, 2000.
[25] M. G. Kendall: Rank correlation methods. 3rd ed. Griffin, London, 1962.
[26] P. Suppes, D. H. Krantz, R. D. Luce, A. Tversky: Foundations of measurement. Vol. 2. Academic Press, San Diego, 1989.
[27] T. D. Wickens: Models for behavior. Stochastic processes in psychology. Freeman, San Francisco, 1982.
[28] F. Wickelmaier, C. Schmid: A matlab function to estimate choice-model parameters from paired-comparison data. Beh. Res. Meth. Instr. & Computers (2003) in press.
[29] R. A. Bradley: Rank analysis of incomplete block designs. III. Some large-sample results on estimation and power for a method of paired comparisons. Biometrika **42** (1955) 324–345.
[30] R. A. Bradley: Paired comparisons: some basic procedures and examples. – In: Handbook of Statistics. Vol. 4. P. R. Krishnaiah, P. K. Sen (eds.). Elsevier, Amsterdam, 1984, 299–326.
[31] P. Daniel, H. Mundt: Visualization and auralization of sound quality. J. Acoust. Soc. Am. **108** (2000) 2642.
[32] W. Aures: Der sensorische Wohklang als Funktion psychoakustischer Empfindungsgrößen [Sensory pleasantness as a function of psychoacoustical parameters.]. Acustica **58** (1985) 282–290.
[33] G. A. Gescheider: Psychophysics - the fundamentals. 3rd. ed. Erlbaum, Mahwah, N.J., 1997.
[34] H. Fastl: Loudness of running speech measured by a loudness meter. Acustica **71** (1990) 156–158.
[35] B. C. J. Moore, B. R. Glasberg: A revision of Zwicker's loudness model. Acta acustica - Acustica **82** (1996) 335–345.
[36] B. C. J. Moore, B. R. Glasberg: A model of loudness perception applied to cochlear hearing loss. Auditory Neuroscience **3** (1997) 289–311.
[37] G. von Bismarck: Sharpness as an attribute of the timbre of steady sounds. Acustica **30** (1974) 159–172.
[38] H. Fastl: Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise. Hearing Research **8** (1982) 59–69.
[39] H. Fastl: The psychoacoustics of sound-quality evaluation. Acta acustica - Acustica **83** (1997) 754–764.
[40] E. Terhardt: Über akustische Rauhigkeit und Schwankungsstärke [On acoustical roughness and fluctuation strength]. Acustica **20** (1968) 210–214.
[41] I. H. Bernstein, C. P. Garbin, G. K. Teng: Applied multivariate analysis. Springer, New York, 1988.
[42] A. Tversky: Elimination by aspects: A theory of choice. Psychological Review **79** 281–299.
[43] P. Slovic, S. Lichtenstein, B. Fischoff: Decision making. – In: Stevens' Handbook of Experimental Psychology. 2nd ed. Vol. 2. R. C. Atkinson, R. J. Herrnstein, G. Lindzey, R. D. Luce (eds.). Wiley, New York, 1988, 673–738.
[44] 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.
[45] G. R. Bienvenue, M. N. Nobile: Quantifying subjective responses to discrete tones in noise from computer and business equipment. – In: Proceedings inter-noise '91. A. Lawrence (ed.). Australian Acoustical Society, Sydney, 1991, 53–56.
[46] 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.
[47] A. Tversky, S. Sattath: Preference trees. Psychological Review **86** (1979) 542–573.
[48] K. Zimmer, W. Ellermeier: Deriving ratio-scale measures of sound quality from paired comparisons. Noise Contr. Eng. J. (submitted).