A COMPARISON OF TWO SUBJECTIVE SCALING TECHNIQUES IN THE LABORATORY ASSESSMENT OF NOISE ANNOYANCE

IR Summers & PA Howard-Jones

Medical Physics Group, Dept. of Physics, University of Exeter, UK

1. INTRODUCTION

It is well known that the impulsive and tonal characteristics of a noise can very significantly affect its nuisance potential. Assessments of suitable level penalties for particular types of noise currently rely upon laboratory assessments in which a large group of volunteers are asked to make subjective judgements on a suitable, usually numerical, scale. Results from subjective assessments have been shown to compare well with field trials that sought the opinions of people actually living with the noise (Flindell, 1982). If a standard method of assessment can be arrived at, it is hoped that subjective scaling techniques may have an important role in future surveys of potential noise nuisance. Surveys involving unusual sound sources, such as wind turbines, may particularly benefit from this approach. This paper considers how the scaling methodology itself can influence the results, and investigates differences in the annoyance of noises which are neither strictly tonal or impulsive, but whose general character is likely to influence their potential nuisance.

Perhaps the most popular method of subjective scaling has been categorical scaling. According to this protocol (e.g. Berry, 1987), the subject is asked to make a judgement on a scale of 0 to 9, where the subject is instructed that 0 represents 'not annoying at all' and 9 is 'extremely annoying'. The task is simple and treatment of the results, usually taking the arithmetic mean across subjects, is also straightforward. An inherent problem of this approach, however, is that the subject can only choose from a small set of integer numbers and this introduces quantisation errors. This effect may be reduced by using a 20 point scale, but the increase in choice automatically increases the complexity of the task. The most serious accusation about categorical scaling, however, is that it forces the subject to categorize and to divide a prothetic continuum into parts or segments. This has the effect of making the subject express himself in distances or intervals, rather than ratios and has been shown to introduce a systematic bias (Stevens, 1974). Further problems have been shown to arise from contextual effects (Eisler and Montgomery, 1974; Parducci, 1974; Rossi *et al.*, 1991). For example, a subject may decide that the first signal presented was given too high a score, since he/she now feels restricted in their expression of later judgements. This will result in an unpredictable amount of distortion. Finally, there is considerable evidence that subjects possess an internal absolute scale of judgement; although perhaps counter-intuitive, different groups of listeners usually agree on the absolute value of the numbers assigned to stimuli (Zwisloski and Goodman, 1980). If that is the case, then forcing subjects to restrict themselves to a small set of numbers may only result in them drifting back to their own absolute scale during the test.

The use of free number magnitude estimation, in which the subject uses his own scale to select any positive non-zero number (including fractions) to express his judgement, may theoretically overcome most of the problems discussed. Invariably, when left their own devices in this way, subjects choose to make judgments on a ratio scale. Neither are they restricted by the size of the scale, since there is always a number larger or smaller than the ones they have already used. It also means that subjects are free to use their internal absolute scale, so excluding any possibility of drifting. More importantly, if the presence of an internal absolute scale can be accepted, it makes comparison of studies involving different noises considerably more meaningful and more straightforward. Magnitude estimation has already been used, with apparent success, to measure annoyance (e.g. Berglund *et al.*, 1981; Kryter, 1974, Hellman, 1982).

This paper compares the results from two assessments which differed only with respect to the type of scaling techniques applied.

2. METHODOLOGY

Three types of noise were evaluated in this study. Traffic noise was recorded at a busy junction in London on DAT tape, using the microphone in a Type 1 sound meter fitted with a standard wind shield. Noise emitted from a wind turbine was also recorded, following standard procedures (Henderson, 1987), using a B & K FM recorder. White noise, sampled at 20 kHz, was taken from a CD-ROM database. The recordings of the traffic and turbine noise were then sampled at 20 kHz, through a 9.5 kHz lowpass filter. The digital recordings of the three noises were scaled to the same r.m.s.

Subjects were presented with each of the three noises at -15 dB, 0 dB, and +15 dB relative to the average overall L_{Aeq}, which was set at approximately 50 dB. Spectral differences between the three noises meant that equalisation of the r.m.s. did not guarantee equivalent values of L_{Aeq}, so a type 1 sound level meter was also used to measure the L_{Aeq} of each noise individually.

Forty normally-hearing student volunteers were played 100 seconds of each noise, followed by a 250 ms tone. They were requested, on hearing the tone, to record a numerical judgement of the annoyance arising from the noise. The group of forty subjects was divided into two groups of twenty. The first group was asked to record their response as a categorical estimate on a scale of 0 to 9 (Berry, 1987):

The experiment in which you are participating is to help us understand the reactions of people to certain types of noise. You are going to listen to a number of different sounds and we would like to know how annoying you find each sound.

You will hear a 100 second sample of each noise, followed by a short 'beep'. A 10 second interval of silence will follow, during which we would like you to answer this question:

How annoying would you find the noise you have just heard, if you heard it all the time?

Your answer should be recorded on the attached sheet, by circling a number from 0 to 9. For example, if you consider the noise was very annoying, you would circle a high number close to the "extremely annoying" end of the scale; similarly, if you considered the noise to be only slightly annoying, you should circle a number closer to the "not annoying at all" end of the scale. Of course, there are no RIGHT or WRONG answers; we only want to obtain your honest personal evaluation of each sound.

Please refrain from making a judgement until you hear the 'beep' at the end of the 100 second period.

It is not necessary for you to concentrate on the noises while they are being played. You will find a pile of magazines next to chair: please feel free to browse through them or to read them during the test, but remember to make your judgements in the 10 second periods between each 'beep' and the next noise.

The second were group were asked to use magnitude estimation, by recording any positive number (Hellman,1982). They received a similar instruction sheet, except that the CE instruction was replaced by:

We would like you to answer this question by assigning a number to represent your judgement and writing it on the attached sheet. You may use any positive numbers that appear appropriate to you - whole numbers, decimals, or fractions. Do not worry about running out of numbers, there will always be a smaller number than the smallest you use and a larger one than the largest you use. Further, do not worry about consistency. Simply try to match an appropriate number to each noise regardless of how you may have judged some previous noise. Of course, there are no RIGHT or WRONG answers; we only want to obtain your honest personal evaluation of each sound.

3. RESULTS AND DISCUSSION

For the categorical estimation, the arithmetic mean of the responses from twenty subjects was calculated for each of the noise conditions. For the magnitude estimation, the twenty subjects differed significantly in terms of the scale factor which each chose for his/her responses (eg equivalent responses might appear as 1, 2, 3 or as 20, 40, 60). On the assumption that response ratios were consistently maintained between subjects, the geometric mean of the responses was calculated for each of the noise conditions. (Equal ratios in the responses appear as equal increments in the logarithms of the responses. Hence an equally weighted mean response can be obtained by taking the arithmetic mean of the logarithms of the responses, which is equivalent to taking the geometric mean of the responses.) Results are shown in Fig. 1.

Both sets of graphs indicate that white noise is considerably more annoying than either traffic or wind-turbine noise. Differences between the ratings of traffic noise and wind-turbine noise are not, in general, significant. Taking typical values from the range covered by the graphs, the CE and ME experiments give level penalties for white noise with respect to traffic noise of (6 +_ 4) dB and (10 +_ 4) dB, respectively; the corresponding figures for white noise with respect to wind-turbine noise are (8 +_ 4) dB and (12 +_ 4) dB. The calculated errors correspond to the variation within each subject group, and indicate that differences between CE and ME results may simply reflect differences between the two groups of subjects. (An alternative explanation relates to floor and ceiling effects arising in the CE case. Although the average results of the CE experiment appear well distributed on the 0 to 9 scale, many individuals were working close to the top or bottom of the scale. There were, for example, many responses of zero.)

The possibly anomalous point on the ME graph, corresponding to the lowest presentation level of the wind turbine noise, highlights a potentially fatal problem with ME scaling techniques. Magnitude estimates allow individuals to register theoretically infinite ratio judgments, whilst the interval judgements in the CE experiment will always be restricted to between 0 and 9. One individual in an ME experiment must, theoretically, be allowed to completely reverse the tendencies expressed by the rest of the subject group. It would be very difficult and theoretically unsound to apply criteria to reject such subjects. In ME experiments, then, the variance due to individual differences is likely to make an unfortunately large contribution to the overall result. This means that ME results may often be determined by a small number of 'extremists' within the group. The restrictions provided by CE procedures overcome this difficulty, but this control is artificial and may produce a serious underestimation of the differences between noises.

A further indication of the problems of listener variability is given by the fact that 8 out of the 40 subjects gave a lower score to at least one type of noise when it was played 15 dB louder! Seven of these subjects belonged to the CE group, and this may reflect cognitive overload caused partially by the complex task of transforming ratio judgments to interval judgments on a fixed scale. Thus, it would not appear advisable that a scale with more than 10 points should be used since this is likely to add to task complexity and increase the prevalence of this type of response.

The finding, confirmed by both experiments, that white noise is significantly more annoying than traffic or wind turbine noise, emphasises the importance of noise assessment based on regulated subjective assessment. The current British Standard (BS4142) concerned with noise annoyance would not differentiate between the three noises in terms of tonality, since they contain no 'distinguishable, discrete, continuous note'(BS4142, 1990). There are some tonal elements in the traffic noise at high frequencies (due to squealing taxi brakes) but if these were taken account of in the way suggested by BS4142, it would result in a 5 dB level penalty *in favour* of the white noise! This makes planned attempts to update BS4142 with respect to tonality even more important (Porter, 1992). The authors would like to recommend that assessment of level penalties (relative to a standard reference) by CE or ME methods, should be incorporated as an option in future versions of BS4142.

4. CONCLUSIONS

Categorical estimation is more convenient and straightforward to implement, since it is not greatly influenced by the extreme opinions of minorities. It may, however, underestimate level penalties by limiting the judgements of subjects working too closely to the ends of the scale.

The difference in the annoyance stimulated by noises that cannot be differentiated by their tonal or impulsive content, makes it advisable that subjective assessments should be implemented whenever unusual sound sources are being assessed for potential noise nuisance. Rules of thumb (such as the BS4142 5dB penalty) may be wholly inadequate and misleading.

5. REFERENCES

Berglund, B., Berglund, U., Goldstein, M. and Lindvall, T. (1976). "Loudness (or annoyance) summation of combined community noises," J. Acoust. Soc. Am. 70, 1628-1634.

Berry, B.F. "The evaluation of impulsive noise" (1987) National Physical Laboratory report, Ac 111.

BS4142 (1990) "Method for rating industrial noise affecting mixed residential and industrial areas", BSI.

Eisler, H. and Montgomery, H. (1974) "On theoretical and realizable ideal conditions in psychophysics: magnitude and category scales and their relation," Perception and Psychophysics 16, 157-168.

Flindell, I.H. (1982) Community response to multiple noise sources, PhD thesis, ISVR, Univ. Southampton.

Hellman, R.P. (1982) "Loudness, annoyance, and noisiness produced by single-tone-noise complexes" J. Acoust. Soc. Am. 72(1), p62-73.

Henderson, A.R. (1987) "Environmental aspects of wind turbine generators", Report 308/87, National Wind Turbine Centre, NPL.

Kryter, K.D. (1974) "Predictions of paired-comparison and magnitude-estimation judgments of noisiness" in *Sensation and Measurement*, papers in honour of S.S. Stevens, eds. H.R. Moskowitz, B. Scharf., and J.C. Stevens (Reidel, Dordrecht).

Parducci, A. (1974) "Contextual effects: a range-frequency analysis" in *Handbook of perception*, eds. Carterette and Friedman, Academic Press, London.

Porter, N. (1992) "Industrial and Residential Noise - Questions Asked at the BSI Symposium on BS4142", Acoustics Bulletin (IOA), 17(1), p11-13.

Rossi, M., Espesser, E., Pavlovic C. (1991) "The effects of the internal reference system and cross-modality matching on the subjective rating of speech synthesisers" 2nd European Conf. Speech Technology, Genova, Italy, p273-276.

Stevens, S.S. (1974) "Perceptual magnitude and its measurement" in *Handbook of Perception*, eds. Carterette and Friedman, Academic Press, London.

Zwisloski , J.J. and Goodman, D.A. (1980). "Absolute scaling and sensory magnitudes: A variation" Perception Psychophysics 28, p28-38.

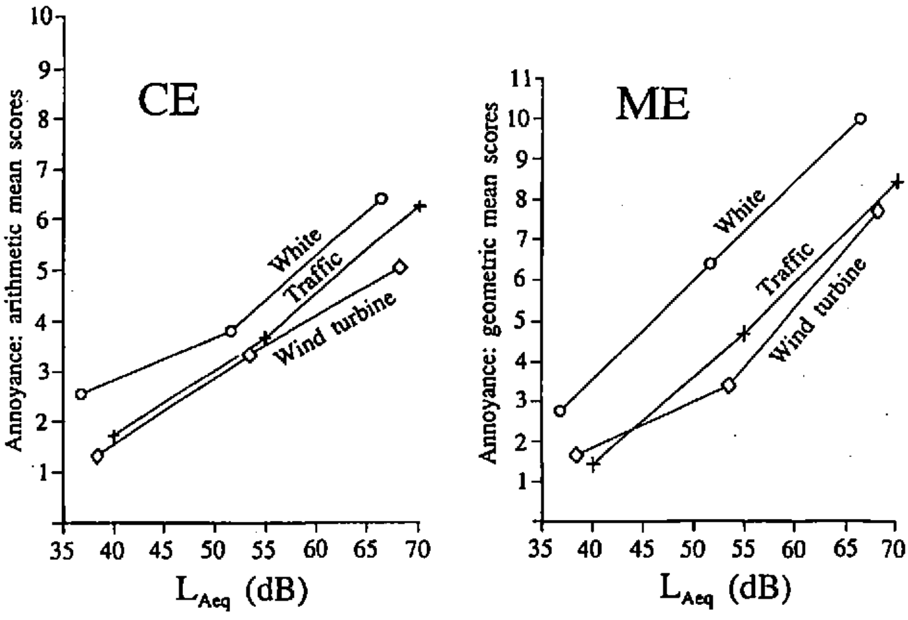

Fig. 1 Results of the two annoyance assessments using categorical estimates (CE) and free number magnitude estimates (ME). Both show that white noise is considerably more annoying than the examples of traffic and wind turbine noise used in the experiment.