



COVER PAGE

Document downloaded by @DAEL

Thu Oct 22 18:57:32 2026

For personal use

When automatic English translation is provided, only the original document is authentic.

The EAA cannot be held responsible of any translation error

Bibliographical reference

Measuring Community Annoyance from Aircraft Noise, Peter Brooker, *Acta Acustica* **vol. 95** (Number 3), 2009, pp. 573-577

DOI

<https://doi.org/10.3813/AAA.918182>

Measuring Community Annoyance from Aircraft Noise

Peter Brooker

Aviation Consultant, West Wickham, Kent, United Kingdom.
p_brooker@btopenworld.com

PACS no. 43.50.Qp, 43.50.Rq, 43.50.Lj

Abstract

How should annoyance responses from individuals receiving a similar aircraft noise exposure be combined, in order to measure ‘community annoyance’? Schultz’s metric, the percentage of people who express ‘High annoyance’, is widely used. Estimates of average annoyance, which take into account all the surveyed responses, can be constructed either from Category Ratings (using differences) or from Magnitude Estimates (via ratios). An illustrative measurement theory-based analysis of empirical data gives an example where the Schultz metric and Magnitude Estimates show a better agreement than Category Ratings. If Magnitude Estimates were to be well-established, then this would enable proper statistical comparisons between studies using ANAS and ICBEN scales.

1. Introduction

Measurement Theory is at the heart of acoustics. Measurement is a vital element in acoustics because it is the common ground between different disciplines: engineering, physics, physiology, psychology, etc. In addition to his work in acoustics, Stanley Smith Stevens also wrote one of the fundamental papers on measurement theory [1]. Measurements and the attribute being measured are different kinds of thing. To draw quantitative conclusions about an attribute, there must be an understanding of the nature of the correspondence between the attribute and its measurements. This kind of understanding is crucial if the analytical methods developed for the physical sciences are to be used in a behavioural science (eg. see [2] for some history and key references).

An instance where measurement theory is very important is in the assessment of community annoyance from transport noise sources. The most frequently used method for developing this is to compare people’s responses to social survey questions with their estimated noise exposure. Three important questions are:

- i) What kind of scale best measures an individual’s annoyance?
- ii) How should the annoyance responses from individuals receiving a similar noise exposure be combined, in order to measure ‘community annoyance’?
- iii) With what noise exposure measure should this community annoyance be compared?

The second question turns out to be the most difficult to answer, particularly because of the wide variations between people’s responses at the same noise exposure.

Schultz [3] offered answers to the three questions that have gained a measure of (by no means uncritical) acceptance over the last thirty years (eg. [4, 5]). Schultz concluded:

- i) Different scales could be used to measure an individual’s annoyance, as long as they enabled an estimate to be made that an individual was ‘Highly Annoyed’ (HA)
- ii) The most appropriate metric for community annoyance is the percentage of HA respondents (%HA)
- iii) Community annoyance should be compared with the noise exposure measure Ldn. (Ldn is the Day-Night Average Sound Level used in the USA and several other countries. It is simply the 24-hour Leq, the noise energy measure, but with night (2200 to 0700) flight noise levels artificially increased by 10 decibels.)

Schultz’s answers have been in widespread use for 30 years, so they must have merits. But is it possible to identify a better set of answers? For example, as his %HA metric ignores responses below that level of annoyance, would some kind of average annoyance estimate be a better indicator, as recently suggested in some UK work (for references see [6])? What specific empirical evidence needs to be analysed? The following explores these questions, focusing on the metrics for community annoyance, using simple measurement theory and empirical results from studies on psychometrics and aircraft noise.

2. What were the reasons for Schultz’s answers?

Schultz was careful to set out a number of reasons for his answers to the three questions. On the use of scales, he noted, in reviewing earlier work:

“On one point there seemed to be agreement from the beginning³; namely, that people’s subjective responses could be measured along a scale of annoyance running from (approximately) ‘not at all annoyed’ to (approximately) ‘very much annoyed’.”

His reference [3] here was in fact the original HMSO [7] UK work – the ‘Wilson Report’. Schultz discusses the variety of scale types used: direct categories – from four to seven (or more) steps, with labels for all steps or at the ends; direct numerical scales from (eg) 1 to 10 or 0 to 10; more complex indirect combinations of scales covering interference with people’s activities – eg. ‘Guttman scales’ used in the Wilson Report.

Schultz’s view was that all the annoyance scales were measuring roughly the same thing, ie individuals’ answers on different scales are statistically highly correlated. This has been tested several times: for example, the sixteen scales compared in [8] were all highly correlated (using non-parametric measures); albeit that scales focusing on aircraft noise alone showed differences from those covering multiple noise disturbance (eg asking about the ‘most bothersome noise’).

Why choose %HA as the metric for community annoyance? There are two issues here: the use of an ‘extreme’ metric; and how to estimate the metric consistently. Schultz has several reasons for adopting an ‘extreme’ metric. In some cases, he compares with the median response, although some kind of average response is similarly a candidate. Some examples (comments in italics – with the benefit of hindsight) are:

“The author suggests that when people are highly annoyed by the noise, the effects of nonacoustical variables are reduced, and the correlation between the noise exposure and the expressed subjective reaction is high. . .” *This approach means that the community annoyance contribution from possibly many (eg) moderately annoyed people is not covered by the measure: does this somehow distort the*

Received 17 March 2008,
accepted 13 January 2009.

'true picture' (compare Olympic goal medals won and the average person's fitness)?

"...the official 'outdoor noise stimulus' may have little or nothing to do with the noise actually heard indoors by the subjects," *This is not a strong argument for airport noise, given aircraft peak noise levels and people's desire to be outside/have windows open in the summer.*

"... 'percent highly annoyed' carries a commonsense import that is clear, even when it is not precisely defined, that 'median response' completely lacks." *People might find average response just as easy to comprehend.*

"Since the median response does not adequately describe that part of the population whose expressed annoyance actually changes with differences in noise exposure, it is too sluggish and insensitive a statistic for regulatory purposes." *What is the evidence for poor statistical performance?*

Schultz's arguments are rational ones, but they do not generate a convincing 'tipping point'. Would a different metric be better – and how would one know if it were better?

One of the practical problems faced by Schultz was how to estimate the %HA:

"Thus, the basic rule adopted was to count as "highly annoyed" the people who responded on the upper 27%–29% of the annoyance scale, if the scale steps were not named; and, in the surveys using annoyance scales with all steps named, so that the respondent could state directly his degree of annoyance, those people were counted as "highly annoyed" who said they were highly annoyed."

Schultz was careful to spell out what he did, but there are issues about the quality of these estimates.

Schultz's third answer was that community annoyance should be compared with the noise exposure measure Ldn. This was probably in line with general thinking at that time. Most of the aircraft noise indices used were Leq-based. Subsequently, most countries using indices attributing a greater weight to the number of aircraft heard (eg see [8]) changed to Leq-based indices. The evidence for the 10dB night weighting is not so strong (eg. see [9]).

3. Scales: Stevens' ideas

The word 'scale' needs some explanation. A verbal scale corresponds to the shaded row of boxes in Figure 1. Respondents like the use of this verbal labelling – it is easy to explain and familiar. In this example, a study's questionnaire respondents are asked how annoyed they are by aircraft ('Aircraft Noise Annoyance Scale' – ANAS) and they have four choices as shown. In measurement theory language, this is an 'Empirical Relationship System'. But words alone are not easy to manipulate statistically, so the words have somehow to be mapped onto real numbers – a corresponding 'Numerical Relationship System'. This is where Stevens' ideas on measurement scales, now standard material in psychology textbooks, come into play [1].

Stevens proposed that there are different levels of measurement, involving different properties of the symbols or numbers constituting the measurements. Each level of measurement has a specific set of 'permissible' transformations. The levels of measurement are, in progressively more demanding order: Ordinal, Interval, Ratio and Absolute. Thus, for example, in an absolute scale, the assigned numbers reflect exactly specific properties of the attribute: a good example is a proportion (or probability).

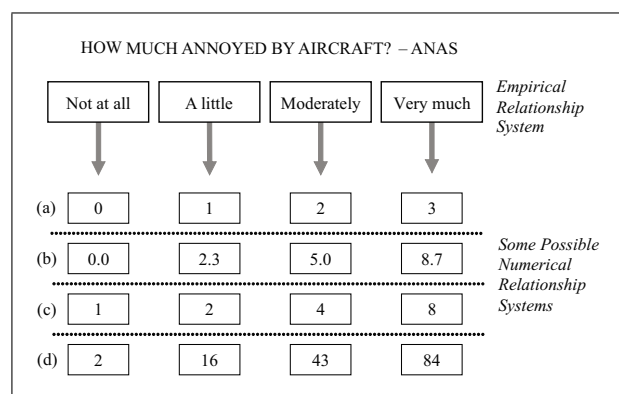


Figure 1. Different kinds of scalings.

The box-shaded row in Figure 1 could correspond to an ordinal scale. Instead of the verbal descriptions, a set of numbers could be used, such as (a). But ordinal numbers are labels rather than ordinary numbers. Adding, multiplying etc them might turn out to have a sensible quantitative meaning – but the labels do not intrinsically guarantee such a property.

If – and the key word is 'if' – the ANAS ratings could reasonably be mapped onto actual numbers, then something like the Numerical Relationship Systems (a) to (d) could be treated as wholly equivalent to ANAS, and be capable of being manipulated by normal arithmetic – and hence be usable in statistical testing/modelling. But writing down ordinal numbers does not turn these numbers into either an interval or a ratio scale. Examples (a) and (b) could be interval scales; the first has equal spacing, the second has unequal spacing: note that the two scales are not equivalent, because there is not a linear transformation that consistently turns the values of one into the other. Examples (c) and (d) are possible ratio-like scales, the first with equal multiplicative steps, and the second with a set of different multiplicative steps: again, the two scales are not equivalent – there is not a simple way of multiplying the values in one scale consistently to get the other one's values.

Which, if any, of these numerical scales would correspond to a true reflection of the intensity of people's annoyance? Could such a 'right' scale be shown to have interval/ratio properties? If there are good answers to both these questions, then the ANAS-equivalent numerical scale might be handled by normal arithmetic.

4. Scales: Progress and psychometrics

There has been patchy progress in developing the ideas in Stevens' measurement paper. Twenty-five years ago, Luce and Narens [10] discussed the reasons for the slow progress. More recent discussions are by Hand [11] (statistical modelling), Michell [12] (scientific justifications for psychometrics), and by Narens and Luce [13] (verifiability and falsifiability of psychosocial measurements). The following sections merely pick out some useful results.

Returning to the numerical scales in Figure 1, what specific kinds of question can be asked about which would be the 'best match' to ANAS responses? One practical step on the way is described in Rohrmann's fascinating work [14, 15]. Words and phrases such as 'Very much' can be thought of as 'verbal qualifiers'. When they act on something like 'annoyed' they generate 'verbal scale point labels' – VSPL – which indicate the intensity of the response. What do these VSPLs correspond to if they are

Table I. Numerical equivalents of verbal qualifiers, ANAS 4-point, ICBEN 5-point, means; CR – Category Rating ('noise'), ME – Magnitude Estimate ('all', GM): see Table 3.1 of Rohrmann [15].

Qualifier	CR	ME
Very much	8.7	84
Moderately	5.0	43
A little	2.3	16
Not at all	0.0	2

compared with numerical scales? The task posed by Rohrmann was the assignment of qualifying words to a point on a continuum, or to a value. Rohrmann distinguished two ways of making these comparisons: Category(-based) Rating and Magnitude Estimation. A sketch of Rohrmann's psychometric methodology is:

Experimental subjects – 'Qualifiers' – are used to grade the degree that a particular attribute is given. Three fundamental judgment dimensions were used by Rohrmann, but the focus here is on *Intensity*, eg. 'Moderately'. The scaling tasks were:

Category Rating: Each VSPL, presented on a card, had to be placed on an 11-point 'equal appearing interval scale' in which '0' was presented as lowest and '10' as the highest level of the respective attribute.

Magnitude Estimation: The 'strength' of each VSPL had to be expressed in two kinds of magnitude, numbers and lines (drawn by the Qualifier on a sheet of paper). An item at the lower end of the range (eg 'Not at all' was used as a baseline); then numbers or lines, respectively, had to be allocated, to indicate the Qualifier's perceived ratio between each VSPL and that of the reference stimulus. The Magnitude estimation (ME) was normalized to a 100-point scale. (The magnitude task is obviously similar to the human ear's rating of decibels.)

Rohrmann's work covered both comparison methods; whereas many other studies tend to focus on Category Rating (eg see [16] and its bibliography). A selection from the results of Rohrmann's experiments is shown in Table I, for ANAS and for the 'ICBEN' recommended 5-point scale [16]. Note that the Category Rating (CR) and Magnitude Estimate (ME) columns are averaged numbers, ie it has been assumed that the data from the experiments can be manipulated as real numbers.

The point-estimate equivalences in Table I will produce *approximate* numerical results when used to summarise/combine sampled annoyance responses measured at particular sites. This is because perfect information on the statistical distributions of respondents' *underlying* numerical scores is not available, their scores in effect being allocated to the nearest verbal qualifier. The ICBEN scale has better coverage at the highest CR/ME values, but there may also be relatively few responses for these values because of 'population sorting', ie people with high sensitivity never move to – or preferentially tend to move away from – the highest noise exposure locations.

The immediate question about Table I is which is 'correct' – the Category Rating or the Magnitude Estimate. Both are the product of reasonable experiments. Torgerson [17] posed the

question: "Are there are two distinct subjective operations when subjects are asked to respond in these two ways, or only a single one?" The latter is supported by psychological experiments generally showing that the two scales are roughly logarithmically related – and the logarithmic function converts ratios into differences. If there were two distinct operations, then sometimes experimenters would find different orderings of pairs of stimuli depending on which operation is used ($6/4 < 2/1$ but $6-4 > 2-1$). The current view is that there is in essence a single operation available to experimental subjects, eg see [18] and its references, particularly to Birnbaum and colleagues' work.

If the two measurement methods are somehow equivalent in a person's mental decision processes, then which is a better guide to the *strength* of that individual's attitude? There does not appear to be a conclusive answer in the literature. Rohrmann asserts that "magnitude scaling is principally superior in terms of measurement theory and data quality" [14, 15]. An example of an acoustical study favouring magnitude estimation techniques is [19]. Recent work of transport noise annoyance by De Coensel *et al.* [20] notes benefits from Magnitude Estimates over Category Ratings in terms of scaling bias and reliability. In contrast, note that a great deal of the literature on scales tends to concentrate on Category Rating techniques (eg. [16]).

Table I also has some clear messages about comparing results from studies using different scales. The ANAS 4-point and the ICBEN 5-point scales look quite similar. To estimate %HA in the two cases, the usual practice would be to use the 'Very much annoyed' percentage from the former, and to add together the 'Very annoyed' and the 'Extremely annoyed' from the latter. But the two %HA results would not be exactly comparable. The ICBEN scale results would correspond to a lower threshold for %HA: the Category Rating and Magnitude Estimate numbers are both lower for the ICBEN scale, by about 7% and 15% respectively (8.1 versus 8.7, and 72 versus 84). This means that a study treating ICBEN data in such a way would indicate a markedly higher level of annoyance – *in %HA terms* – than an ANAS-based study of the same (very large number of) respondents in exactly the same noise exposure environment. This could distort the picture of changes in annoyance over time.

5. Combining responses to give community annoyance

Can recent studies on annoyance scales provide help in summarising and/or combining the annoyance results from people in a 'similar noise exposure' community? The Schultz %HA is a summary metric based on just one element of the distribution of responses. As he noted, the median might be used, although he argued that it would be a worse choice, and there would be some issues about interpolation, particularly for 4-point or 5-point scales.

What kind of *aggregation* of community response would offer the most precise – most reasonable? – picture of the community's disturbance? The obvious answer is an average of some sort. However, it has to be an average of the variabilities between people and across different days. An individual's annoyance response curves versus Ldn are not in practice perfectly knowable. On day 1, A's response curve might be what it usually is, but on day 2 the response curve could be markedly different, perhaps because of atypical kinds of activity that A has engaged in, eg reading, relaxing in the garden. But person B's average response curve could be markedly different from A's average, for probably unknowable hereditary and developmental factors in B's life – B

Table II. ANIS ANAS data and ‘Averages’ CR – Arithmetic Mean, ME – Geometric Mean; NAA – Not at all, AL – A little, M – Moderately, VM – Very much. All data entries rounded to one decimal place, CR values scaled by a factor of 10 to simplify comparisons. ‘weighted’ means ‘Rohrmann weighted’.

Site	Annoyed by Aircraft Noise				CR	ME
	NAA	AL	M	VM	Weighted	
1	34.9	34.9	19.6	10.6	27.0	11.2
2	33.3	24.0	26.0	16.7	33.0	13.7
3	43.9	27.7	17.7	10.6	24.4	9.1
4	40.4	36.5	16.7	6.4	22.3	9.1
5	37.3	32.0	24.0	6.7	25.2	10.4
6	2.3	12.7	32.7	52.3	64.8	50.2
7	5.8	12.2	30.3	51.7	62.9	45.1
8	13.4	13.4	32.9	40.2	54.5	32.6
9	4.2	18.2	33.9	43.7	59.2	42.3
10	14.5	17.0	39.6	28.9	48.9	28.3
11	15.8	21.0	31.6	31.6	48.1	26.6
12	15.5	25.4	33.8	25.4	44.8	24.7
13	55.0	26.8	14.5	3.8	16.7	6.3
14	57.7	24.7	13.5	4.1	16.0	5.9
15	37.0	29.8	23.3	9.9	27.1	11.0
16	44.4	28.2	21.3	6.1	22.4	8.7
17	53.3	18.7	24.7	3.3	19.5	7.1
18	17.8	29.2	35.2	17.8	39.8	21.0
19	38.2	20.8	23.9	17.1	31.6	12.2
20	43.4	37.2	13.3	6.1	20.5	8.2
21	7.5	28.0	39.5	25.0	47.9	30.6
22	13.9	28.3	28.7	29.1	46.2	25.8
23	5.2	26.3	30.9	37.7	54.3	36.5
24	11.7	28.5	30.0	29.9	47.6	27.8
25	19.4	16.7	33.3	30.6	47.1	24.7
26	12.5	18.2	40.2	29.2	49.7	29.9

might simply engage in activities that are less affected by aircraft noise.

People’s annoyance responses are expressions of an attitude using different and varying subjective valuations. The respondents’ *subjective intensities of annoyance* underlying the ratings lie on a variety of personal subjective scales: there is no consensus between individuals about these subjective scales – so there are problems about ‘interpersonal comparability’ [21]. This has major implications if the respondents are being asked to use ratio-like scales, ie for Magnitude Estimation (whereas the arithmetic mean of the ratings can be used for category scales). Aczél and Roberts showed that the only valid way of merging responses for judgements using ratio-like scales must be a geometric mean [22]. The geometric mean for M peoples’ ratings $r_1, r_2, \dots, r_M$ is given by

$$GM(r_1, r_2, \dots, r_M) = (r_1 \cdot r_2 \cdot \dots \cdot r_M)^{1/M}.$$

6. Illustration using empirical annoyance data

To use the ideas from the previous sections, it is necessary to use some empirical data on aircraft noise annoyance. Table II’s source data are from Brooker *et al.* [8, Table C2]. These are now middle-aged data, but the study methodology has withstood challenges, eg. at the Heathrow Terminal 5 Public Inquiry, and all the relevant raw data are published. It uses the ANAS 4-point scale.

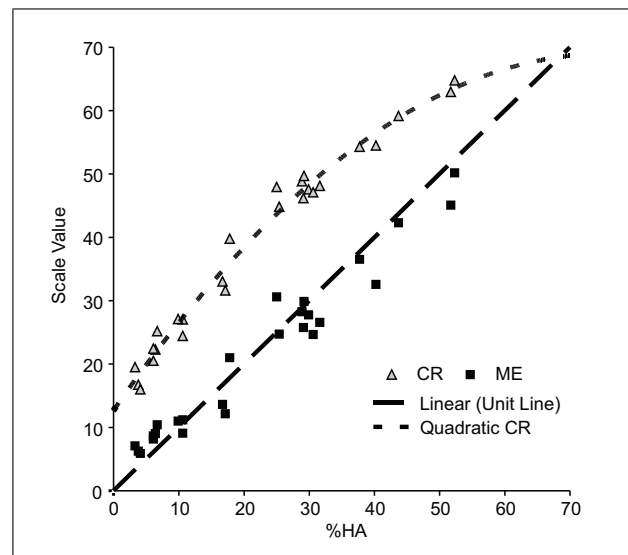


Figure 2. Category Rating (CR) and Magnitude Estimate (ME) versus %HA.

The numbers in the two right hand columns are the averages of the Category Rating (Arithmetic Mean) and Magnitude Estimate (Geometric Mean). These use the Rohrmann weightings in Table I. Thus, for example, the first row’s Geometric mean GM is calculated by

$$GM^{100} = 2^{34.9} \cdot 16^{34.9} \cdot 43^{19.6} \cdot 84^{10.6}.$$

Figure 2 shows the values of the Category Ratings (CR) and Magnitude Estimates (ME) plotted against Schultz’s %HA. Both are well correlated with %HA (noting that sampling variations, mainly from multinomial sampling, affect the datapoint values – particularly when small percentages are involved). The CR points have an underlying curvilinear dependence – the CR trend line shown is a quadratic fit. A step-wise multiple regression for the CR fit shows that the quadratic term is statistically significant at better than the (two-sided) 5% level. In contrast, the line shown through the ME points is simply a unit slope line through the origin. A quadratic fit to the ME points shows that the quadratic term is *not* significant at the 5% level, ie a linear fit is sufficient. Thus, the Magnitude Estimate is very near to a linear equal-scale dependence. The agreement of the %HA and the ME line, remembering that the %HA scale is absolute – as it is a proportion – thus suggests that the ME data can be assumed to be on an approximate ratio scale.

What are the practical benefits of this kind of knowledge, if confirmed by similar analyses on a variety of good datasets? One obvious gain would be the ability to make undistorted, albeit approximate, comparisons between results from studies using ANAS and ICBEN scales. Earlier here it was noted that the two scales could produce markedly different %HA values for identical study methodologies and survey populations. Rohrmann’s results would therefore enable different variants of verbally qualified data – see Tables I and II – to be transformed into comparable ME values with good scaling properties.

7. Conclusions

What kinds of scale should be used to measure an individual’s annoyance from aircraft noise, and how should annoyance responses from individuals receiving a similar noise exposure be combined to measure ‘community annoyance’? This exploratory

analysis uses some simple results from Measurement Theory, Rohrmann's psychometrics, and Aczél and Roberts' work on merging data from scale responses.

Schultz's community annoyance metric, the percentage of people who express 'High annoyance', is widely used. Numerical scales can be constructed from scales that use verbal quantifiers. This process reveals some problems about the interpretation of 'High annoyance', with marked differences being detectable in apparently similar 4- (ANAS) and 5-point (ICBEN) verbal scales.

Estimates of average annoyance – hence taking into account all the surveyed responses – can be constructed either from Category Ratings (using differences) or from Magnitude Estimates (via ratios), using Rohrmann's work. Using for illustration a reasonable quality dataset, Magnitude Estimates *empirically* have a linear variation with Schultz's metric. Given that Schultz's metric is on an absolute scale, this offers an example where the Magnitude Estimates are on an approximate ratio scale, but where the Category Rating dataset is not. If this use of Magnitude Estimates were to be generally applicable, then it would permit statistical comparisons to be made between studies using (eg) ANAS and ICBEN scales.

This technical note has focused on methodological considerations. It is hoped that researchers will carry out analyses of their individual datasets, exploring the relationships between Schultz metric, Magnitude Estimates and Category Ratings, and seek to establish the extent to which the illustrative results here are generally applicable.

References

- [1] S. S. Stevens: On the theory of scales of measurement. *Science* **103** (1946) 677–680.
- [2] L. Finkelstein: Widely, strongly and weakly defined measurement. *Measurement* **34** (2003) 39–48.
- [3] T. J. Schultz: Synthesis of social surveys on noise annoyance. *The Journal of the Acoustical Society of America* **64** (1978) 377–405.
- [4] S. Fidell: The Schultz curve 25 years later: A research perspective. *The Journal of the Acoustical Society of America* **4** (2003) 3007–3015.
- [5] R. Guski: How to forecast community annoyance in planning noisy facilities. *Noise and Health* **6** (2004) 59–64.
- [6] P. Brooker: ANASE: Unreliable - Owing to Design - induced biases. *Acoustics Bulletin* (2008) 26–31.
- [7] HMSO: Noise - Final report. 'Wilson Report' Cmnd.2056. Her Majesty's Stationery Office, London, [Appendix XI], July 1963.
- [8] P. Brooker, J. B. Critchley, D. J. Monkman, C. Richmond: United kingdom aircraft noise index study (anis). Main Report DR Report 8402, for CAA on behalf of the Department of Transport, CAA, London, 1985.
- [9] P. Brooker: The UK Aircraft Noise Index Study [anis]: 20 years on. *Acoustics Bulletin* (2004) 10–16.
- [10] R. D. Luce, L. Narens: Comments on the "non-revolution" in the representational theory of measurement. *Psychological Science* **4** (1983) 127–130.
- [11] D. J. Hand: Statistics and the theory of measurement. *Journal of the Royal Statistical Society, Series A* **159** (1996) 445–492.
- [12] J. Michell: Measurement in psychology: Critical history of a methodological concept. Cambridge University Press, United Kingdom, 1999.
- [13] L. Narens, R. D. Luce: Meaningfulness and invariance. – In: Palgrave Dictionary of Economics. L. Blume, S. N. Durlauf (eds.). <http://www.imbs.uci.edu/personnel/luce/Submitted/06-09-05-meaningful.pdf>, 2008.
- [14] B. Rohrmann: The use of verbal scale point labels in annoyance scales. – In: 7th International Congress on Noise as a Public Health Problem. C. Norman, S. R. F. Job (eds.). Noise Effects Pty Ltd., Sydney, 1998, Vol. 2, pp. 523–527.
- [15] B. Rohrmann: Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. Project Report, University of Melbourne/Australia. <http://www.psych.unimelb.edu.au/staff/br/vqs-report.pdf>, April 2003.
- [16] J. M. Fields, R. G. De Jong, T. Gjestland, I. H. Flindell, R. F. S. Job, S. Kurra, P. Lercher, M. Vallet, T. Yano, R. Guski, U. Felscher-Suhr: Standardized general-purpose noise reaction questions for community noise surveys: Research and recommendation. *Journal of Sound and Vibration* **242** (2001) 641–679.
- [17] W. S. Torgerson: Distances and ratios in psychological scaling. *Acta Psychologica* **19** (1961) 201–205.
- [18] L. Narens: Symmetry, direct measurement, and Torgerson's conjecture. *Journal of Mathematical Psychology* **50** (2006) 290–301.
- [19] T. L. Whitehill, A. S. Y. Lee, J. C. Chun: Direct magnitude estimation and interval scaling of hypernasality. *Journal of Speech, Language, and Hearing Research* **45** (2002) 80–88.
- [20] B. De Coensel, D. Botteldooren, B. Berglund, M. E. Nilsson, T. De Muerand, P. Lercher: Experimental investigation of noise annoyance caused by high-speed trains. *Acta Acustica united with Acustica* **93** (2007) 589–601.
- [21] L. Narens, R. D. Luce: How we may have been misled into believing in the interpersonal comparability of utility. *Theory and Decision* **15** (1983) 247–260.
- [22] J. Aczél, F. S. Roberts: On the possible merging functions. *Mathematical Social Science* **17** (1989) 205–243.