Acoustics Bulletin. Jan/Feb, 26-31. (2008)

ANASE: UNRELIABLE – OWING TO DESIGN-INDUCED BIASES

Peter Brooker

Cranfield University

p.brooker@cranfield.ac.uk

1. Introduction

In November 2007, the ANASE (Attitudes to Noise from Aviation Sources in England) report was published. It claimed that people are increasingly annoyed by aircraft noise, and it estimated how much they would be 'willing to pay' to get rid of it. But its quantitative 'findings were rejected as unreliable by the Department for Transport [DfT]' (BBC webpage). Immediately after the report's release, a DfT Minister (BBC Politics Show) said:

> "The reason why it [ANASE] was delayed was that the scientists – the peers reviewing this major scientific study – said that it isn't up to standard...it isn't good enough for what the Government wanted, ie to formulate Government policy."

About a quarter – *sic* – of the project's duration was spent on expert peer reviews. ANASE's website (http://www.dft.gov.uk/pgr/aviation/environmentalissues/Anase/) includes the report, its technical appendices and several of these critical reviews. In particular, DfT paid two objective and knowledgeable acoustics experts to review the ANASE draft material (Havelock & Turner, 2007). Their comments include:

> "…in the first version of this review it was stated that there were sufficient technical and methodological uncertainties still remaining with the study to mean that reliance on the detailed outcome of ANASE would be misplaced. In view of developments since the review of the July 2007 version of the ANASE main report, the reviewers are even more convinced that their concerns are fully justified…"

The DfT did not refer to these conclusions in its publicity material about ANASE, but this review is a key document.

The following summarises the main ANASE claims, and then examines its design, methodology and statistical analyses as set out in published documents. Neither the history of the project, nor the managerial and professional issues in its conduct, is discussed. Brooker (2004, 2006) provide general background on past research and the technical issues explored here, particularly in the context of the earlier Aircraft Noise Index Study (ANIS – Brooker et al (1985)).

2. ANASE's Objectives and Claims

The 1985 ANIS study concluded that there was no better metric than the noise energy measure LAeq (Leq here) in terms of correlation between aircraft noise and community annoyance. Following consultations, the Government decided to adopt the use of Leq to describe noise, and decided that 57 Leq (16-hour period) marks the approximate onset of significant community annoyance from aircraft noise.

In mid-2001, the DfT announced a major study into aircraft noise:

> "…the new study underlines the Government's commitment to underpin our policy on aircraft noise by substantial research that commands the widest possible confidence";

and that conclusions from the ANIS research have:

> "…been broadly confirmed by other studies here and abroad, and we have no reason to doubt their validity."

Commercial contractors (led by MVA Consultancy Ltd) were commissioned to conduct the ANASE project in late-2001. The ANASE study has two aspects:

> Relationship between aircraft noise and annoyance
>
> Monetary valuation of annoyance by aircraft noise (Stated Preference [SP])

The following does not discuss the SP part of the work, but it does indicate how that component markedly affected the work on annoyance – note that the bulk of the DfT managers' attention was on the SP components.

ANASE adopted several basic ideas from ANIS. Social survey questionnaires were used to elicit respondents' annoyance from aircraft noise as well as socio-economic data. Fifty six survey sites near nine airports were included in the study, with levels of aircraft noise from 36 to 68 Leq. The study report makes a number of aircraft annoyance claims, including (slightly edited):

> *Claim*: "For the same amount of aircraft noise, measured in Leq, people are more annoyed in 2005 than they were in 1982."
>
> *Claim:* "The modelling work also showed that respondents were less sensitive to changes in sound level below 42 Leq and above 59 Leq, adding support to a logistic dose-response form. There was no threshold, or discontinuity, in the relationship between mean annoyance and Leq."
>
> *Claim:* "The results from the attitudinal work and the SP analysis both suggest that Leq gives insufficient weight to aircraft numbers, and a relative weight of 20 appears more supportable from the evidence than a weight of 10, as implied by the Leq formulation."

These are dramatic claims. To meet the DfT criterion 'commands the widest possible confidence', they would need to be robust, technically reliable, and capable of withstanding scrutiny.

3. ANASE Problems: Questionnaire

When carrying out an attitudinal survey, choices must be made about question wording, response scale, question context, and data collection technique. But all these choices can generate errors and biases. The responses to attitudinal questions may easily be affected by the way the issue is posed, the sequencing of questions, the particular wording of a question and its context.

Psychologists interpret attitudes as 'structures in long-term memory', and suggest a four-stage cognitive process needed to answer attitude questions:

(i) Interpret the question ("What is the attitude about?").

(ii) Retrieve relevant beliefs/feelings.

(iii) Apply these beliefs/feelings to generate appropriate judgement.

(iv) Use this judgement to formulate response.

This indicates that attitudes are 'evaluative judgements' formed at a particular time, rather than some kind of enduring personal view, waiting to be picked out of someone's mind. Each stage is likely to be influenced by psychological variables dependent on the questionnaire construction and data collection process.

Thus, attitude reports are highly context sensitive. All four stages above can potentially be affected by 'prior items': serious respondents may be building on their earlier thought processes, or they may aim to 'match' the earlier responses, ie be consistent with their answers. They are unlikely to want to mislead about their 'true' attitudes, but they may be motivated to help or 'please' the interviewer, ie provide answers that show that the interviewee is aware of the issues that he or she is to be questioned about. Reputable textbooks (eg Sudman and Bradburn, 1982) warn about context effects, as does UK governmental guidance, eg re question order:

> "Question order can affect the way in which survey respondents interpret survey questions and thus answer them. This is because the wording of preceding questions can help to shape the context in which respondents interpret the current question." (GSRU, 2007)
>
> "Such question-context effects may therefore bias prevalence estimates and invalidate comparisons across surveys where the same questions are asked but not in identical order." (McColl et al., 2001)

Figure 1 shows a schematic comparison of the ANIS and ANASE questionnaire set-ups. Two potential context effects are worth noting:

> The installation of noise playback equipment precedes ANASE, but not ANIS. Thus, ANIS is a social survey and ANASE is a combination of a social survey and a foreshadowed laboratory experiment, as, later in the interview, noises are played to respondents.
>
> ANASE starts immediately with questions on aircraft noise annoyance, but ANIS leads up to them by asking about perceptions of the local area, and thus allows the interviewee to mention aircraft noise spontaneously.

Given the importance of context effects, both of these factors could affect annoyance ratings considerably – discussed later here.

Measured annoyance attitudes also tend to be very variable for other reasons:

Sampling fluctuations: if for a particular noise climate, the true percentage of a proportion is (say) 30%, then a sample of 160 people will produce a range of values purely through sampling variations (95% confidence band is 23%-38%).

Socio-economic variables: few of these produce consistently detectable effects, but working at an airport or having a job dependent on airport activity usually show up as distinct 'confounding factors', and surveys do not consistently include or omit these respondents.

Media attention/trust: there is great deal of research work on attitude measurement showing the importance of recent media attention at the airport in question on respondents' expressed attitudes. Related factors are people's trust in the airport company and national/local government policies.

4. ANASE Problems: Noise

ANASE used noise estimates for common noise areas (CNA) that do not match with official CAA [Civil Aviation Authority] / DfT published values. Table 1 compares the Heathrow site ANASE estimates and CAA /DfT values for Leq (16 hour), adapted from Table 1 of Havelock & Turner (2007)). The Table ranks the Leq data in terms of the ANASE estimate. The fourth column shows the differences between the ANASE estimate and DfT value – the Leq bias. Havelock & Turner explore the technical reasons for the estimation bias.

At the right of Table 1, the average Leq bias is shown for three groups of ANASE Leq estimate: <50, 50-57, and >57. The Leq biases are respectively -2.5, -2.0 and +0.4 dBA. The inference is that ANASE underestimates Leq for CNAs under 57 dBA; thus, when ANASE analyses led to statements about 50.0 Leq estimates, they should be referring to 52.5 Leq on average.

5. ANASE Problems: Annoyance Measure

The ANASE contractors' way of using annoyance scales is odd. First, compare the questions that ask how much a respondent is annoyed:

ANIS	**ANASE**	
Very much?	Extremely?	Highly Annoyed?
	Very?	
Moderately?	Moderately?	
A little?	Slightly?	
Not at all?	Not at all?	

Note that the ANIS version has no middle ranking choice – so the interviewee is not able to take the 'easy way out' by choosing 'in the middle'. For the ANASE version, the combination of 'Very' and 'Extremely' answers is taken as the 'Highly Annoyed' category. The ANASE reports did not offer evidence-based reasons for the change.

There is no perfect recipe for determining 'good' attitude scales, but the key question is the extent to which a possible scale is cardinal in nature (ie corresponding to the properties of integers), rather than just 'ordinal' (ranking responses). If a scale is cardinal, then such results can be manipulated by all the rules of arithmetic, and hence analysed by the standard kinds of statistical testing.

ANIS used the responses above to construct a 'Very Much Annoyed Percentage' scale of annoyance at each survey site. This percentile method actually correlates well with average responses (using non-parametric statistical testing). The ANIS choice of scale is consistent with the great bulk of international research on aircraft disturbance (eg Fidell & Silvati (2004), a recent review paper of international social survey data into aircraft noise annoyance). In contrast, ANASE used the answers to its version of the annoyance question to construct a 'Mean Annoyance'. In its scheme, a rating of 'Not at all' scored 10 points, 'Slightly' scored 30 points', then up to 'Extremely' scoring 90 points; ie each extra level of annoyance added twenty points. The Mean Annoyance estimate for the site was then simply the arithmetic average of the respondents' scores, eg if half the people said 'Not at all' and half the people said 'Extremely', this would be a mean of 50 points.

But ANASE's choices of weightings are subjective value judgements. The ANASE contractors did not produce robust evidence to justify the relative numerical scorings (saying the scale is 'standardised' adds no content). Why are nine people saying 'Not at all' equivalent to one person saying 'Extremely', or to three people saying 'Slightly'? Rather than 10, 30, 50, 70, and 90, the analysis could have used any other set of monotonic numbers, with corresponding changes in the inferences made.

Arbitrarily averaged attitude scales, with their unreliable statistical properties, were used very cautiously even before the ANIS work. It is puzzling why ANASE would need to change from the ANIS percentage scales. The following focuses on percentage scale data, as these enable comparisons with ANIS and international work.

6. ANASE Problems: Statistical Analysis

ANASE used two kinds of survey sites. At one ('Full') there was the noise playback equipment of Figure 1, and at the other ('Restricted') there was no equipment. Thus, the context for the two was markedly different. If context effects are crucial in this study, then marked differences would be expected in the data from the two kinds of sites – and they are there.

Figure 2 shows the '% Highly Annoyed' response for the two site types at the 27 ANASE Heathrow sites. The Heathrow sites are selected because CAA / DfT higher accuracy Leq values for these sites are available; because it is simple to approximate internationally-used DNL values (by adding 2.5 dBA to the Leq value); and to avoid airport-dependent factors. [DNL is the Day-Night Average Sound Level used in the USA and several other countries: it is a 24-hour Leq with night flight noise levels artificially increased by 10 decibels.] Simple linear-fit trend lines are also shown for the two sets of data.

Figure 2 indicates that the Full and Restricted scatter plots and trends are very probably different – in particular the trend line slopes differ. In comparing two regression lines, the most basic hypothesis to test is the hypothesis of coincidence, ie if the two underlying relationships are the same. The ANASE contractors carried out statistical testing to compare Heathrow Full and Restricted data – but only at the instigation of the reviewers (Havelock and Turner, 2007: page 20). This rejected the coincidence hypothesis, finding that the differences were statistically significant (t-statistic above the standard 5% level). It is therefore unlikely that the two samples come from the same underlying population. It implies that the introduction of noise equipment changed the aircraft noise annoyance dose-response relationship, by a roughly multiplicative bias here. The ANASE contractors decided to ignore these crucial results.

Only in circumstances when statistical testing accepts coincidence, as examined through (eg) Analysis of Variance techniques, is it permissible to fit a single overall regression line to both relationships. But the ANASE statistical analysis wrongly combines Full and Restricted data sets (eg Figure 3). To ignore the statistical testing results rejecting the coincidence of the data sets is not sound practice. A statistical textbook would offer this kind of thing as an example of 'how to do it incorrectly'. It removes any possible sound foundations for subsequent ANASE modelling claims about (eg) annoyance onsets and the weighting of the number of aircraft.

Why do the Full and Restricted data sets differ? It is not possible to offer precise reasons based on the ANASE documents, simply because the ANASE work did not investigate potential causes. One factor could be confusion between audibility/awareness of noise as compared with suffering a degree of annoyance. The presence, and presumed intended use of the noise playback equipment, is certainly a possible strong factor (would a police officer standing in the corner affect a crime survey?).

An even more telling illustration is a mapping of the Heathrow data in Figure 2 onto the Fidell & Silvati data set – Figure 4. This aircraft annoyance research collated international data from 326 site surveys with an average of ~160 people per site. The Figure shows a scatter plot of all the '% Highly Annoyed' data against DNL. The two trend lines are the linear fits to the Fidell & Silvati data and the ANASE Heathrow Full data. The ANASE Heathrow Restricted data lies roughly on the Fidell & Silvati trend line. The ANASE Heathrow Full data lies markedly above the trend line for the other data: it is hard to believe that it is a sample from the same underlying population.

Figure 5 shows the complete set of Full and Restricted data from ANASE (using wholly ANASE data). This again shows that there are differences between the two data sets: having noise equipment present does make a difference – showing a roughly multiplicative bias at the Full sites. The Figure also shows that ANASE Restricted sites were not wisely selected. The onus was on the ANASE contractors to select sites to be able to test effectively for Full/Restricted differences – Restricted sites at higher Leq values ('control group sites') should therefore have been included.

Figure 6 compares the '% Highly Annoyed' data from all the Restricted sites with a curve fitted to the ANIS results used in policy work (Havelock & Turner, 2007; Fidell

and Silvati (2004) discuss curve-fitting). The ANASE Restricted data points are possibly slightly above the ANIS curve, but this could be a statistical sampling issue (Restricted site ANASE samples were very small, typically 16 people) and/or a context effects-related problem – because of the markedly different questionnaire ordering and a different annoyance question.

7. ANASE Problems: International Comparisons over Time

There are comments in the ANASE reports that allude to non-UK studies suggesting that the annoyance dose-response relationship might be moving upwards, ie people are typically more annoyed for a given Leq. This is not a new suggestion (eg see Brooker, 2004). The test of this kind of hypothesis is to examine data.

As already noted, a recent review paper (in the peer-reviewed literature) is Fidell & Silvati (2004). Figure 7 extracts results from the Fidell & Silvati data set. It shows responses in the bands 47.5-52.5, 52.5-57.5, and 57.5-62.5; ie these represent ~50, ~55 and ~60 DNL. The plots cover results after 1980, mainly because the interest is in changes since the early 1980s ANIS work. The Figure plots these responses against the year the survey was published. Simple (unweighted) linear regressions on the data in the Figure – the trend lines – do not show significant changes over time (none of the regression t-statistics is significant at even the 10% level). Thus, there is no strong evidence from this large international data set of a trend over time.

A simple analysis on even this large data set is not statistical proof. To be confident about the magnitude of possible trends over time, it would be necessary to carry out high-quality data collections and statistical analyses, with tight experimental controls on questionnaire context/design, annoyance scales, socio-economic variables, media attention/trust, and of course sampling variations.

8. Summary

DfT was wise to commission the peer reviews and to publish the material rather than be accused of a ‘cover up’. But no reliance can be put on ANASE claims: they cannot ‘command the widest possible confidence’. There are unrepairable major problems with questionnaire design and process, noise estimates, analysis techniques, and selective attempts to compare with international work.

The design of the ANASE questionnaire does not meet the necessary criteria set out in standard textbooks, by the Treasury’s GSRU, or by responsible UK organisations (eg the NHS). This damages the ability to make reliable comparisons with earlier work.

ANASE noise estimates are markedly biased at lower Leq sites compared with official CAA / DfT published values, which distorts several of the analyses.

The analysis techniques used in ANASE do not recognise the problems of using average annoyance scales in parametric statistical analyses. ANASE’s contractors presented no good reasons for changing from earlier, robust scales, *inter alia* preventing proper comparisons.

ANASE fails to meet minimum data analysis requirements for such a study, ie critical examination of raw data to detect potential biases, and always taking proper account of statistical testing results. The regression-based statistical modelling used in ANASE is invalid because it too quickly combines data from Full and Restricted (ie without noise playback equipment) sites samples. This also reveals ANASE's poor design: the onus was on the contractors to test key hypotheses on these effects – there are insufficient Restricted sites at higher Leq values.

ANASE data suggest that the introduction of noise equipment changes the aircraft noise annoyance dose-response relationship by a roughly multiplicative bias factor. ANASE data for Full sites are markedly out of line with the results of reputable international and previous UK work. As data from ANASE's Full sites are unlikely to be representative of people's annoyance attitudes, the SP results that build from these distorted attitudes may similarly be distorted. ANASE Restricted site data are broadly consistent with international and ANIS results.

Thus, a straightforward factual explanation for the ANASE data set is that it has a design-induced multiplicative bias overlaying annoyance responses largely unchanged from past studies. The implication is that the ANASE contractors' claims – eg increased annoyance over time, additional aircraft number effects – are invalid because they mostly derive from the biased data.

References

Brooker, P. (2004). The UK Aircraft Noise Index Study [ANIS]: 20 Years On. Acoustics Bulletin. May/June, 10-16. https://dspace.lib.cranfield.ac.uk/handle/1826/1004

Brooker, P. (2006). Aircraft Noise: Annoyance, House Prices and Valuation. Acoustics Bulletin. May/June, 29-32.

Brooker, P., Critchley, J. B., Monkman, D. J. & Richmond, C. (1985). United Kingdom Aircraft Noise Index Study (ANIS): Main Report DR Report 8402, for CAA on behalf of the Department of Transport, CAA, London.

Fidell, S. & Silvati, L. (2004). Parsimonious alternative to regression analysis for characterizing prevalence rates of aircraft noise annoyance. Noise Control Engineering Journal, 5(2), March/April, 56-68.

GSRU [Government Social Research Unit] (2007). The Magenta Book: Guidance Notes for Policy Evaluation and Analysis. HM Treasury, UK. http://www.policyhub.gov.uk/magenta_book/

Havelock, P. & Turner, S. W. (2007). Attitudes to Noise from Aviation Sources in England: Non SP Peer Review. Environmental Research & Consultancy, CAA; Bureau Veritas. http://www.dft.gov.uk/pgr/aviation/environmentalissues/Anase/nonsppeerreview.pdf

McColl, E., Jacoby, A., Thomas, L., Soutter, J., Bamford, C., Steen, N., et al. (2001). Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Health Technology Assessment [HTA] 5(31). [NHS R&D HTA Programme]. http://www.hta.ac.uk/fullmono/mon531.pdf

Sudman, S. & Bradburn N. M, (1982). Asking Questions: A Practical Guide to Questionnaire Design. San Francisco, Jossey-Bass.

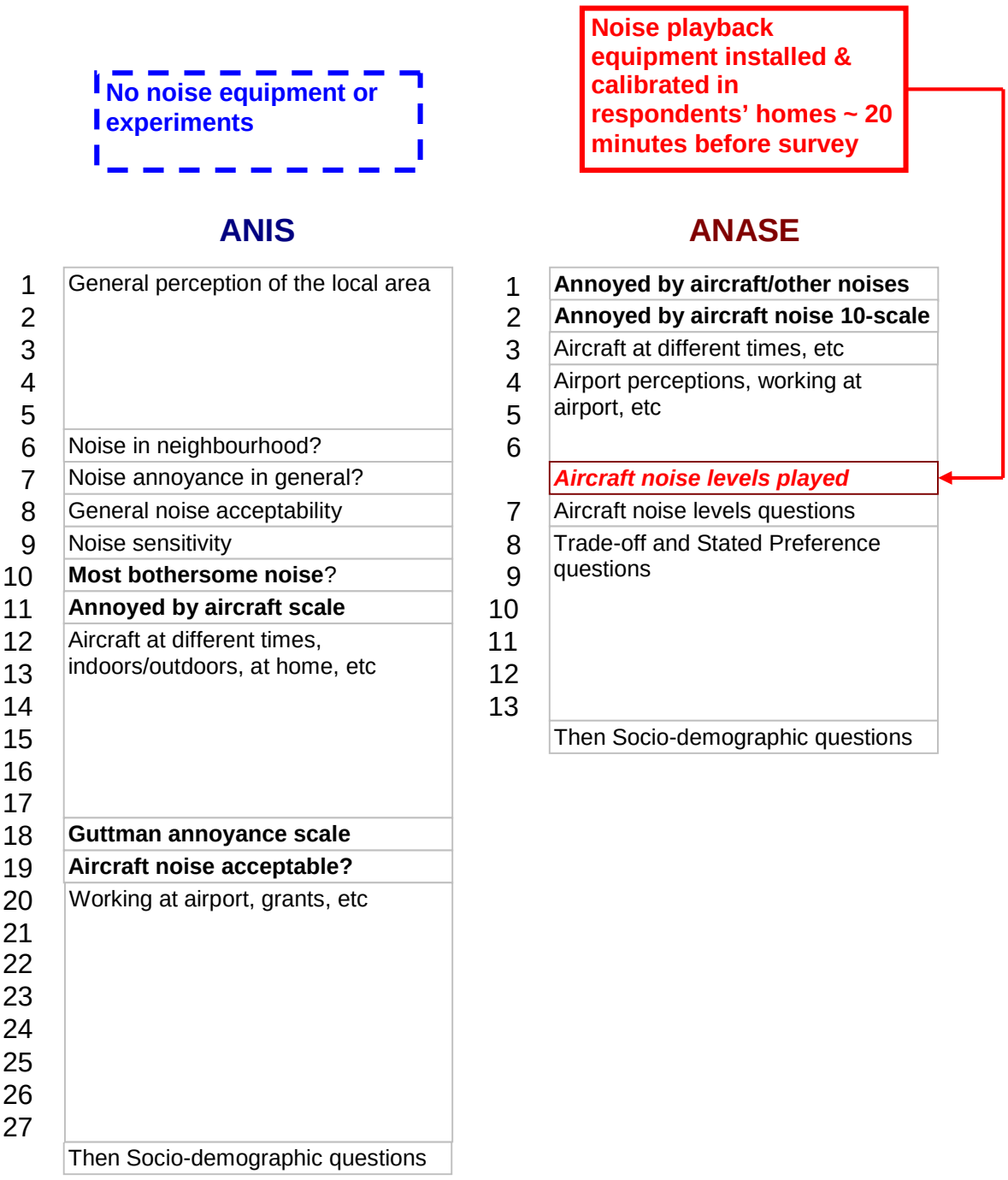

Figure 1. Comparison of key ANIS and ANASE Questionnaire context, question order and noise playback equipment differences.

Notes:

(i) The ANASE questions are in the order given, but the numbering starts at 6 rather than 1 – no explanation is given for this.

(ii) The bold text indicates where questions to provide 'aircraft disturbance' scales used in the statistical analyses were asked.

Site	ANASE estimate	CAA / DfT Published	Bias, ie Difference (ANASE - CAA / DfT)	
R01	40.9	45.8	-4.9	
R02	41.6	46.2	-4.6	
R03	43.0	44.9	-1.9	
H3C	46.0	50.3	-4.3	
R06	46.5	51.4	-4.9	*ANASE: < 50 Leq*
R09	47.2	52.2	-5.0	*Average Bias* ***-2.5 dB***
R05	47.5	48.1	-0.6	
R04	47.6	48.5	-0.9	
R08	48.9	50.4	-1.5	
H5E	49.6	46.3	+3.3	
R10	50.4	52.6	-2.2	
H3A	50.4	52.8	-2.4	
H3B	50.5	53.2	-2.7	
H5A	50.9	53.5	-2.6	
H3D	52.7	52.8	-0.1	*ANASE: 50 – 57 Leq*
H3E	53.0	54.3	-1.3	*Average Bias* ***-2.0 dB***
H1P	54.7	57.6	-2.9	
R07	55.2	56.0	-0.8	
H5B	56.1	58.6	-2.5	
H5F	56.2	58.3	-2.1	
H5D	58.7	57.8	+0.9	
H5C	59.3	60.0	-0.7	
H1L	59.7	58.9	+0.8	*ANASE: > 57 Leq*
H1M	59.8	59.4	+0.4	*Average Bias* ***+0.4 dB***
H1K	60.3	59.8	+0.5	
H1J	61.7	61.8	-0.1	
H1H	63.1	62.3	+0.8	

Table 1. Comparison of published CAA / DfT London Heathrow Summer 2005 Leq (16 hour) with ANASE estimate. Data ranked by ANASE estimate.

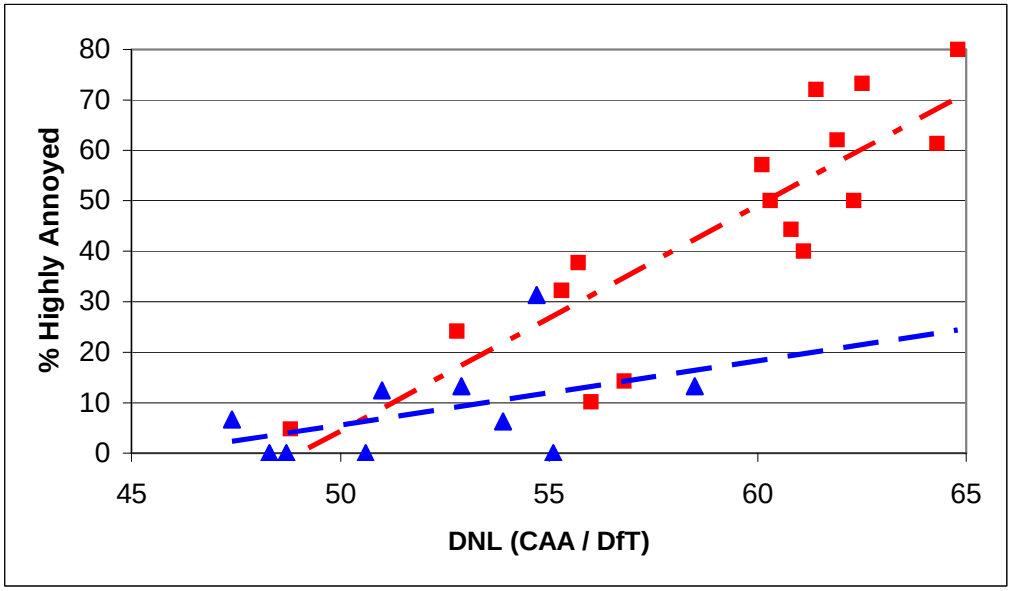

Figure 2. ANASE '% Highly Annoyed' Heathrow results: two distinct data sets. Red squares – Full; Blue triangles – Restricted. Linear trend lines.

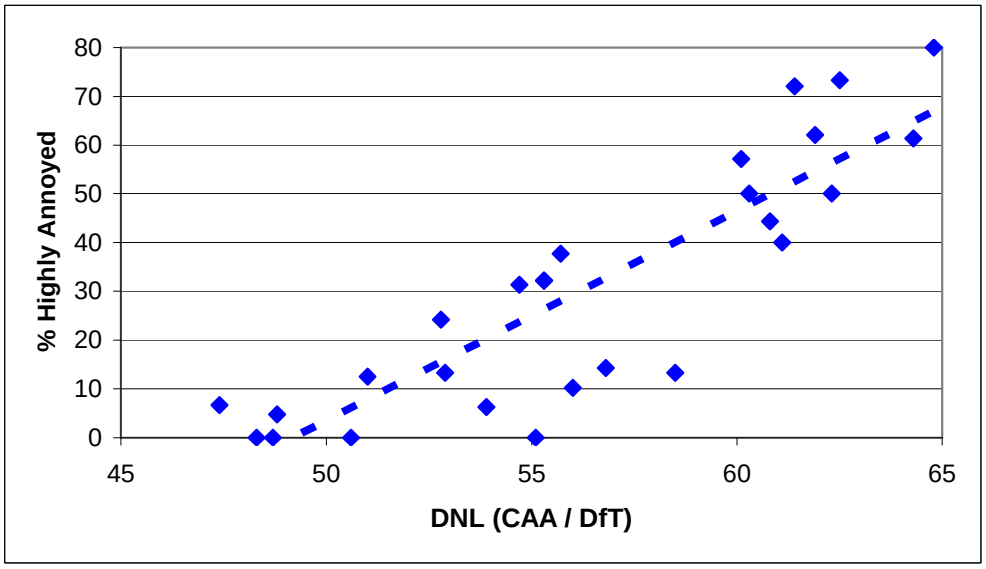

Figure 3. Erroneous ANASE-type fit for Heathrow results – statistical test results disregarded.

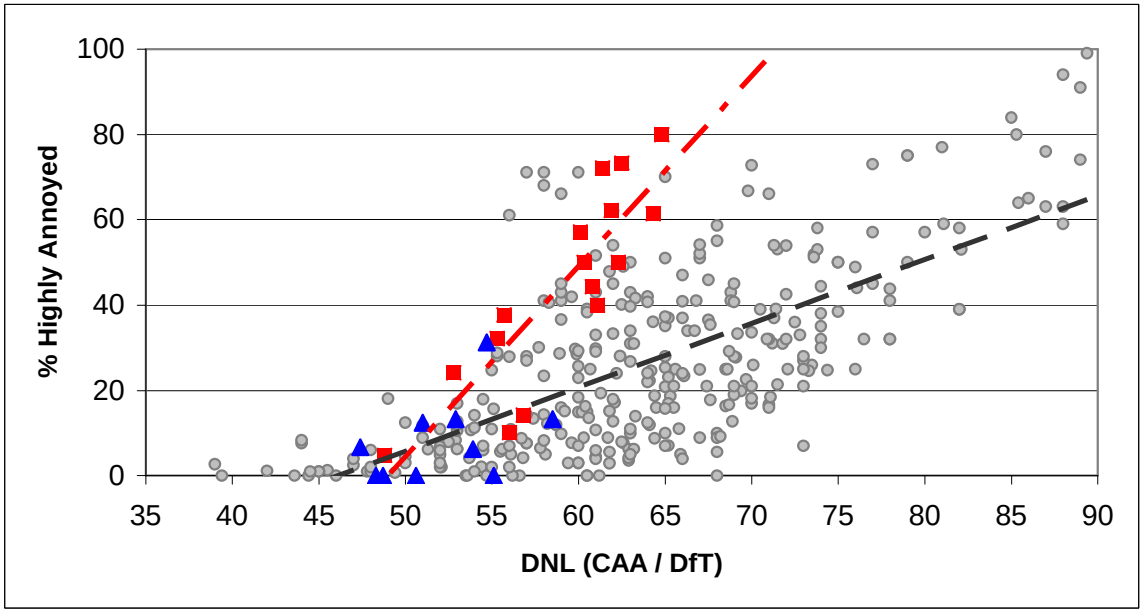

Figure 4. Compares Heathrow ANASE '% Highly Annoyed' with Fidell & Silvati (2004).
Red squares – Heathrow Full; Blue triangles – Heathrow Restricted; Grey blobs – Fidell & Silvati data set. Linear trend lines to Full and Fidell & Silvati data.

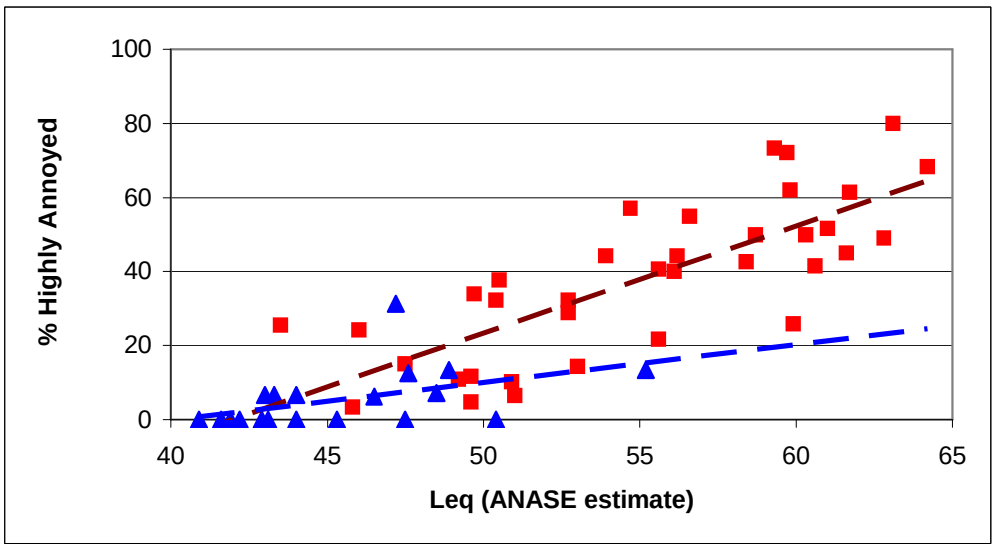

Figure 5. Compares ANASE Full and Restricted sites '% Highly Annoyed'.
Red squares – ANASE Full sites; Blue triangles – ANASE Restricted sites. Linear trend lines. Source Technical Appendices, Table 10 (pages 250/1), Table 6.2 (pages 17/18). Site R17 excluded – as in ANASE analyses.

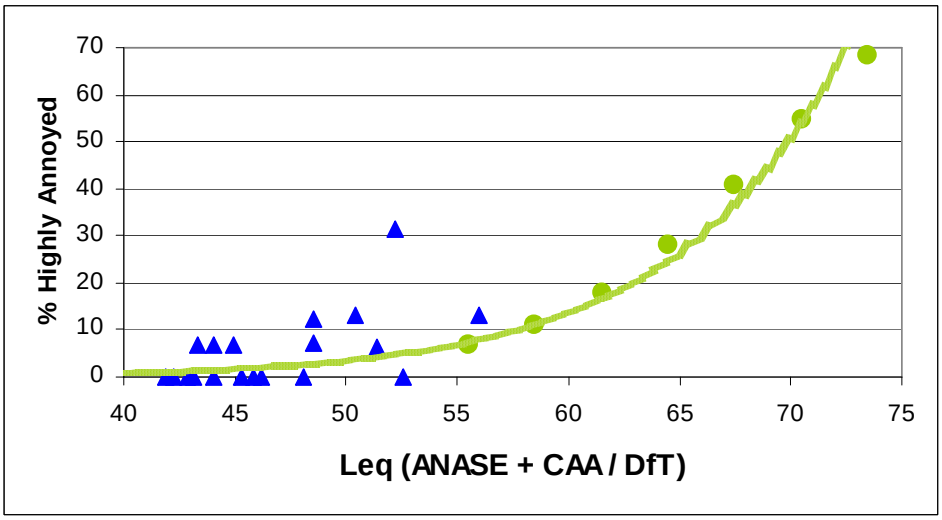

Figure 6. '% Highly annoyed' at ANASE Restricted sites compared with ANIS curve.
Blue triangles – ANASE Restricted sites (source above), sample size typically 16.
X-axis ANASE Leq for non-Heathrow data and CAA / DfT Leq for Heathrow data
Blobs are standard ANIS values from Havelock & Turner Table 2, plus exponential fit.

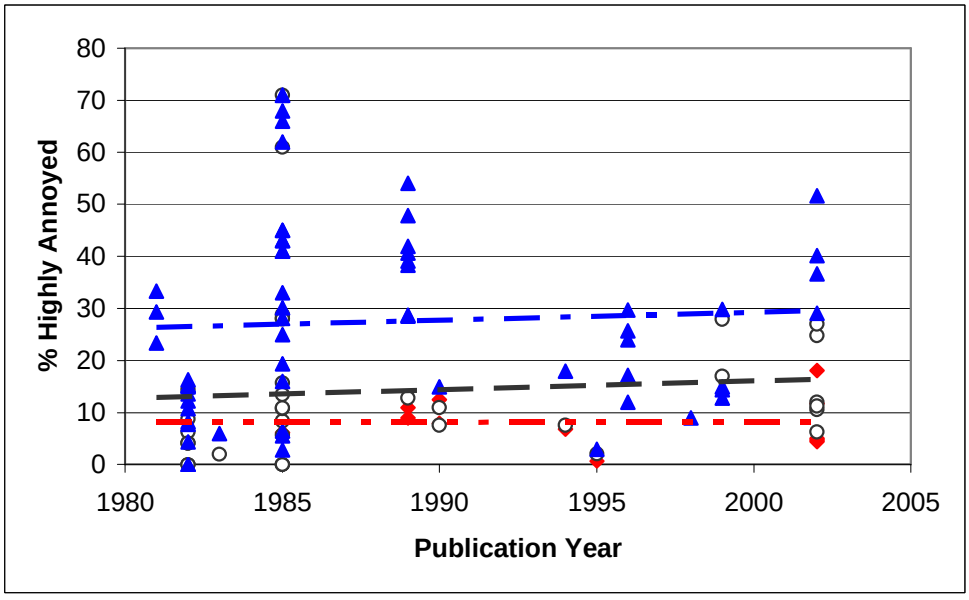

Figure 7. '% Highly Annoyed' from Fidell & Silvati (2004), post 1980 data.
Red lozenges - ~50 DNL; Round open - ~55 DNL; Blue triangles - ~60 DNL. Linear trend lines.

ANASE: Unreliable - owing to design-induced biases

Brooker, Peter

2008-01

Brooker P. (2008) ANASE: Unreliable - owing to design-induced biases. Acoustics Bulletin. Jan/Feb 2008, pp. 26-31

http://hdl.handle.net/1826/2242