[redacted]

[redacted]

These two papers supplement my Response.

Could you please acknowledge receipt.

Thank you

[redacted]

[redacted]

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

ANASE: UNRELIABLE – OWING TO DESIGN-INDUCED BIASES

Peter Brooker

Cranfield University

[redacted]

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.

No noise equipment or experiments

Noise playback equipment installed & calibrated in respondents' homes ~ 20 minutes before survey

	ANIS		ANASE
1	General perception of the local area	1	**Annoyed by aircraft/other noises**
2		2	**Annoyed by aircraft noise 10-scale**
3		3	Aircraft at different times, etc
4		4	Airport perceptions, working at airport, etc
5		5	
6	Noise in neighbourhood?	6	
7	Noise annoyance in general?		***Aircraft noise levels played***
8	General noise acceptability	7	Aircraft noise levels questions
9	Noise sensitivity	8	Trade-off and Stated Preference questions
10	**Most bothersome noise**?	9	
11	**Annoyed by aircraft scale**	10	
12	Aircraft at different times, indoors/outdoors, at home, etc	11	
13		12	
14		13	
15			Then Socio-demographic questions
16			
17			
18	**Guttman annoyance scale**		
19	**Aircraft noise acceptable?**		
20	Working at airport, grants, etc		
21			
22			
23			
24			
25			
26			
27			
	Then Socio-demographic questions		

d ANASE Questionnaire context, question order
nces.

ne order given, but the numbering starts at 6
is given for this.
questions to provide 'aircraft disturbance' scales
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.

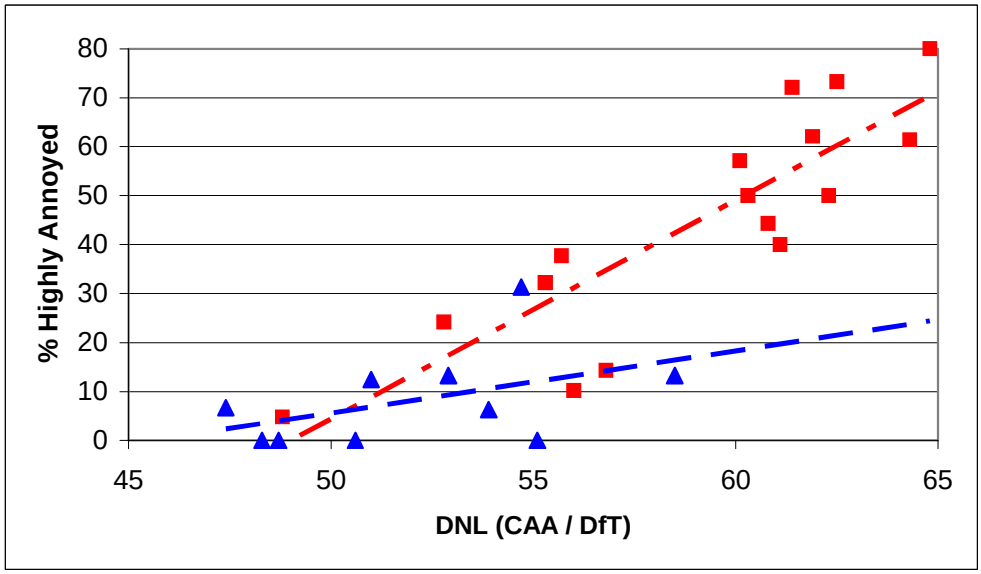

Figure 2. ANASE '% Highly Annoyed' Heathrow results: two distinct data sets. Red squares – Full; Blue triangles – Restricted. Linear trend lines.

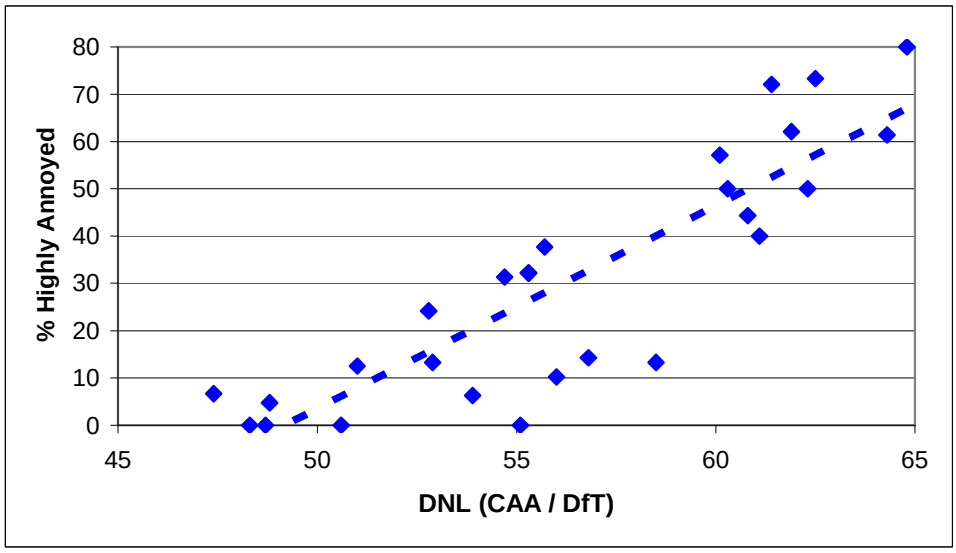

Figure 3. Erroneous ANASE-type fit for Heathrow results – statistical test results disregarded.

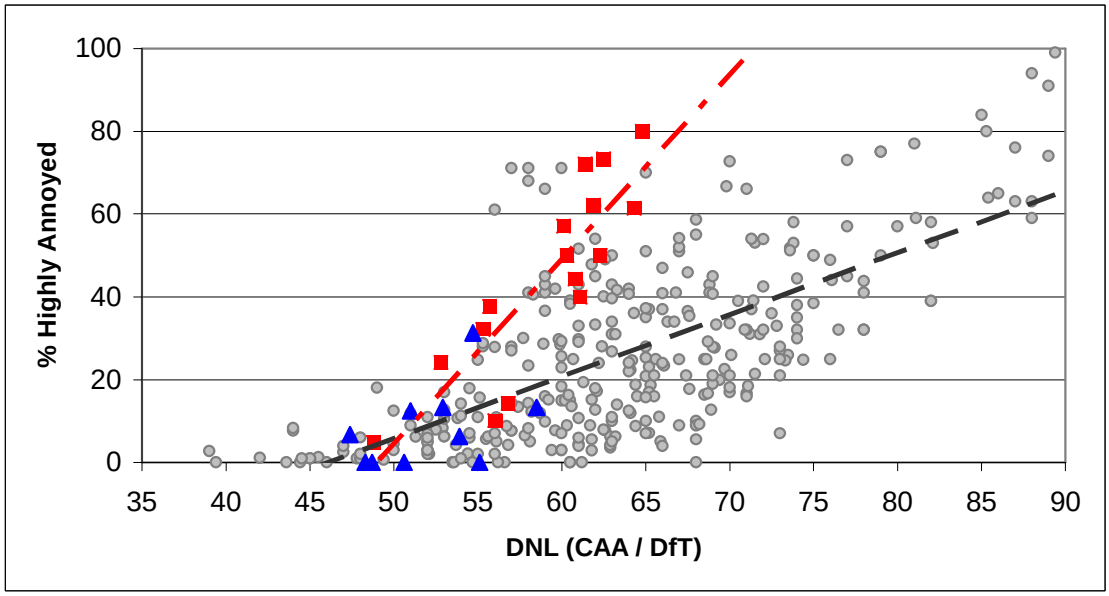

Figure 4. Compares Heathrow ANASE '% Highly Annoyed' with Fidell & Silvati (2004).
Red squares – Heathrow Full; Blue triangles – Heathrow Restricted; Grey blobs – Fidell & Silvati data set. Linear trend lines to Full and Fidell & Silvati data.

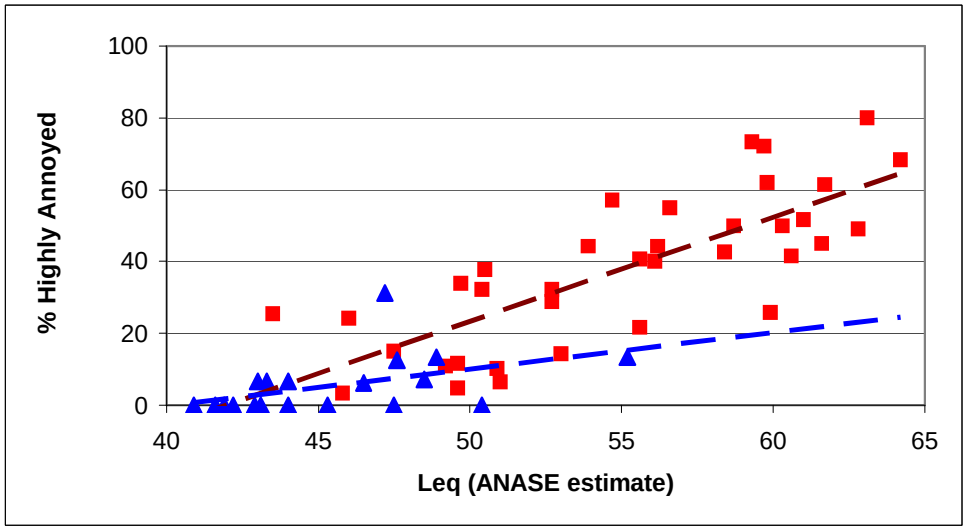

Figure 5. Compares ANASE Full and Restricted sites '% Highly Annoyed'.
Red squares – ANASE Full sites; Blue triangles – ANASE Restricted sites. Linear trend lines. Source Technical Appendices, Table 10 (pages 250/1), Table 6.2 (pages 17/18). Site R17 excluded – as in ANASE analyses.

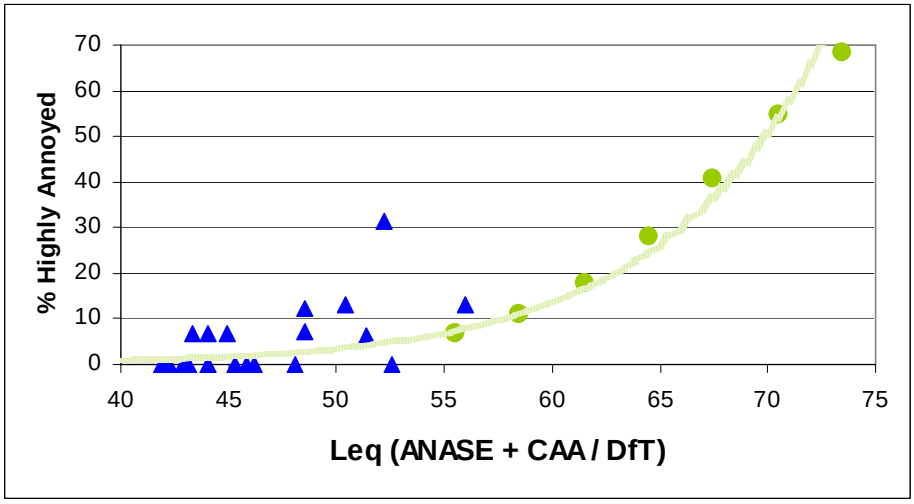

Figure 6. '% Highly annoyed' at ANASE Restricted sites compared with ANIS curve.
Blue triangles – ANASE Restricted sites (source above), sample size typically 16.
X-axis ANASE Leq for non-Heathrow data and CAA / DfT Leq for Heathrow data
Blobs are standard ANIS values from Havelock & Turner Table 2,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

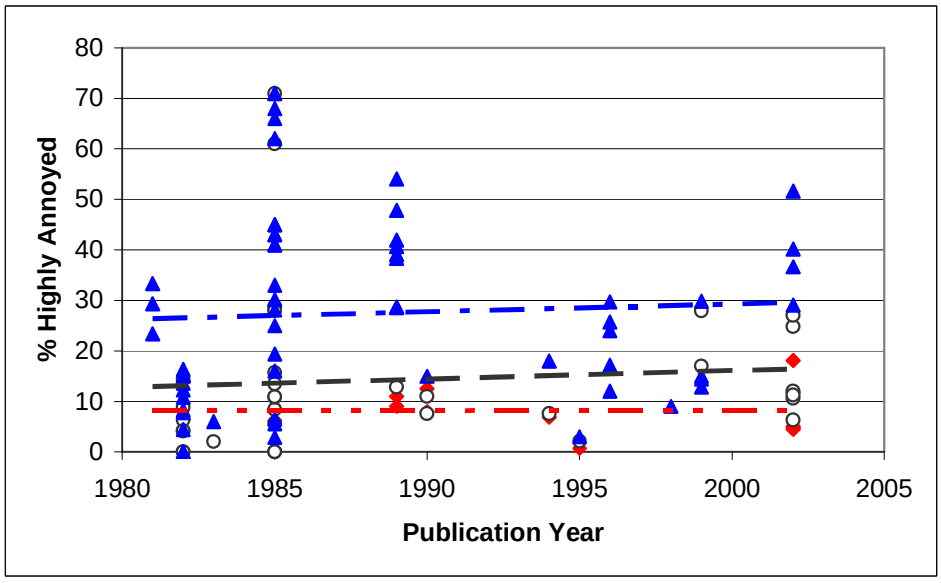

Figure 7. '% Highly Annoyed' from Fidell & Silvati (2004), post 1980 data.
Red lozenges - ~50 DNL; Round open - ~55 DNL; Blue triangles - ~60 DNL. Linear trend lines.

ANASE: LESSONS FROM 'UNRELIABLE FINDINGS'

P Brooker Cranfield University, Cranfield, UK

1 INTRODUCTION

In late 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' (BBC website). The project's managers were warned in its early stages that the work would fail to deliver good value for money and not meet accepted technical/statistical standards. How and why did it fail? What were the methodological and project management failings? What are the lessons for acoustics professionals?

2 BACKGROUND

The ANASE project was initiated by the Department for Transport (DfT) in 2001. Its aim was twofold: to explore the relationship between aircraft noise and annoyance; and a monetary valuation of annoyance by aircraft noise (Stated Preference [SP]). The DfT said that: "…the new study underlines the UK Government's commitment to underpin our policy on aircraft noise by substantial research that commands the widest possible confidence."

The focus here is on the annoyance component of ANASE, which was intended to update the Aircraft Noise Index Study (ANIS) carried out in the early 1980s. (The 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[1].) ANASE's website includes the report, its technical appendices and peer reviews [*http://www.dft.gov.uk/pgr/aviation/environmentalissues/anase/*]. In particular, DfT paid two objective and knowledgeable acoustics experts to review the ANASE draft material on annoyance and noise[2].

3 PREDICTED AND ACTUAL FAILURES OF ANASE

ANASE was in part managed through a Steering Group (SG), covering a variety of interests, including environmental organisations, but largely with a non-technical membership. The author was invited to join the SG as an expert on aircraft noise disturbance, but left it after two meetings because of major concerns about the project. These concerns were set out in a series of letters to the DfT Permanent Secretary and other officials in early 2002. In summary, these concerns were that ANASE would not be robust, nor technically reliable, nor capable of withstanding scrutiny, nor good value for money; and that it would be a source of considerable vulnerability for DfT. Thus, the ANASE outputs would be poor value for the taxpayer, residents near airports, and the aviation industry. DfT did not heed these warnings, nor the specific recommendation to use an independent expert audit team, proposed with the aim of getting the work back on the right track and hence ensuring that the results would command the widest possible confidence.

The ANASE specification said the project should take 'in the order of three years'. The actual duration was from December 2001 to November 2007, including about eighteen months of peer reviews. The time over-run was therefore about 97%.

DfT's Permanent Secretary currently (January 2008) says that the actual spend on ANASE was £1.78 million, compared with the original budget of £0.53 million. These figures do not match responses to parliamentary questions in 2007 because those answers explicitly did not include the 2007/2008 spend. The over-spend was therefore 236%.

ANASE's early results were originally intended to inform the preparation of the Air Transport White Paper, then expected to be published in late 2002. The outputs from the final report should then have had a similar role in the development of the work that has led to the 2007 DfT consultation on 'Adding Capacity at Heathrow Airport'[3]. This is an extremely important document, because it brings forward specific development options for consultation with residents living around the airport. ANASE's outputs failed in three important ways: annoyance, SP valuations, and confusion (its specific claims are examined in the next section).

The ANASE claims on aircraft noise annoyance are disregarded in the Consultation Document. Its analysis simply focuses on existing policy. This continues to use Leq as the annoyance metric, with DfT's standard 57 Leq contours – and the enclosed households and populations – being taken to represent the scale of the noise disturbance generated by the airport.

No use is made in the Consultation Document of the ANASE claims on SP valuations. Its cost benefit analyses instead use results from research on 'hedonic pricing', in which the cash value of disturbance is assessed by examining the reduction in the prices of houses affected by aircraft noise. These kinds of valuations were developed well before ANASE took place.

The Consultation Document discussion of ANASE's claims produces confusion. For example, ANASE's claim of increased annoyance over time is mentioned – a comment guaranteed to cause bewilderment, given that DfT's use of a standard Leq contour explicitly assumes no change in annoyance over time. None of the 'positive spin' on ANASE in the Document is traceable to acousticians or attitude measurement researchers. The outputs have not convinced non-UK researchers, e.g. Dr Rainer Guski, a well-known German researcher on aircraft noise annoyance: "The ANASE statement that the degree of aircraft noise annoyance has changed in comparison with the ANIS study cannot be held, because the methods of data gathering are not comparable."

4 WHAT WENT WRONG? – DESIGN BIASES

What went wrong? The ANASE report generated various claims, but it is easy to demonstrate that the claims are unreliable. There are unrepairable major problems with questionnaire design and process, noise estimates, analysis techniques, and selective attempts to compare with international work. The following summarises an analysis of the problems, most of which are the product of design biases.

The study report makes a number of aircraft annoyance claims:

> *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, particularly given the tough DfT criterion 'commands the widest possible confidence'. Thus, they would need to be robust, technically reliable, and capable of withstanding scrutiny. But there is good evidence that they are the consequence of design biases.

The major design problem stems from 'context effects': the ANIS and ANASE questionnaires are markedly different and are implemented in markedly different circumstances. Figure 1 shows a schematic comparison of the 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.

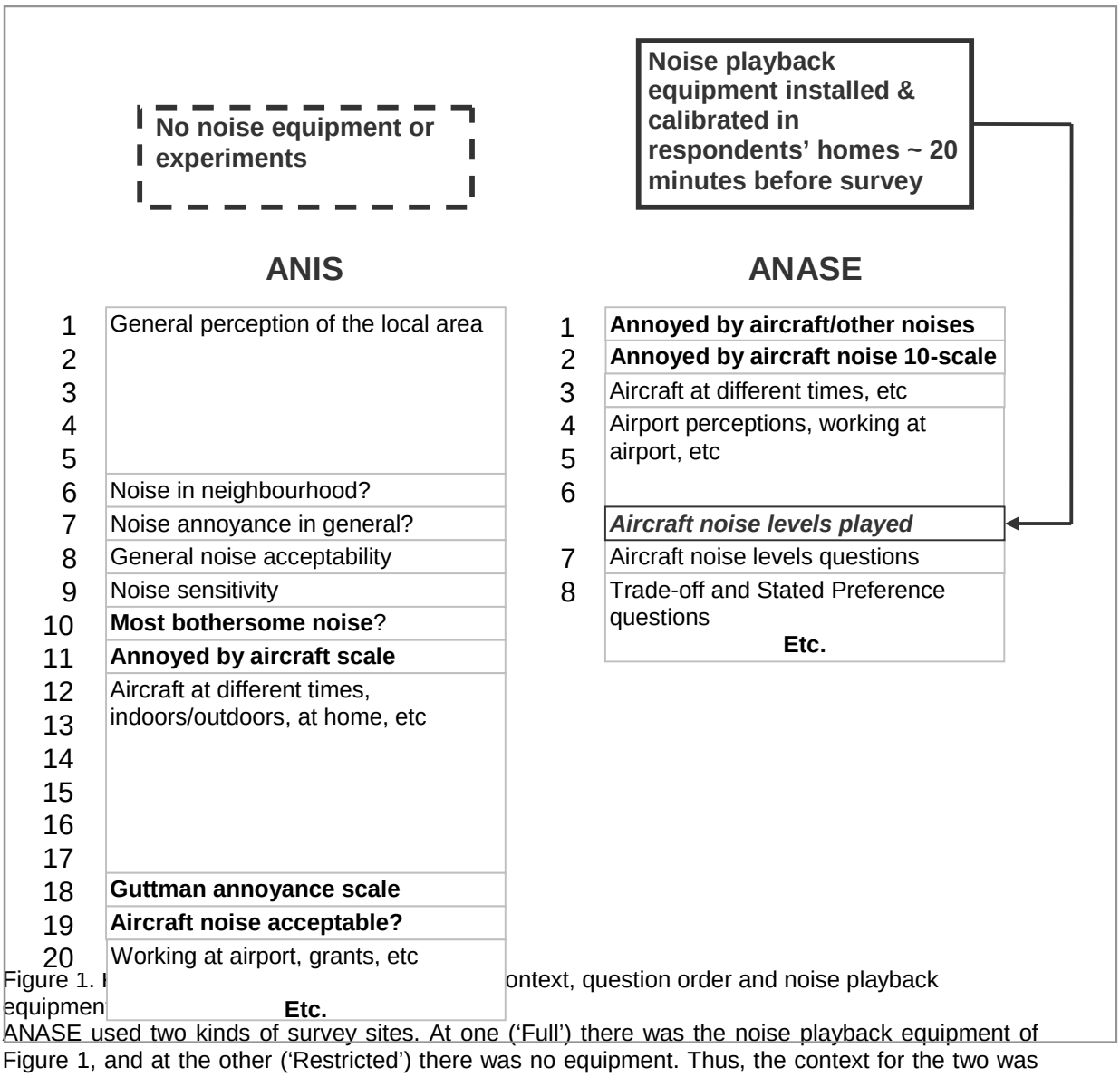

Figure 1. I ontext, question order and noise playback equipmen

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 the data shows that they are there.

A standard international way of matching the annoyance from aircraft to people's noise exposure is to plot the percentage of survey respondents saying they are 'Highly Annoyed' against Leq or a weighted version of Leq[4]. The most common weighted version is Ldn (= DNL, 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. Figure 2 shows the '% Highly Annoyed' response for the two site types at the 27 ANASE Heathrow sites[5]. 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. Simple linear-fit trend lines are also shown for the two sets of data.

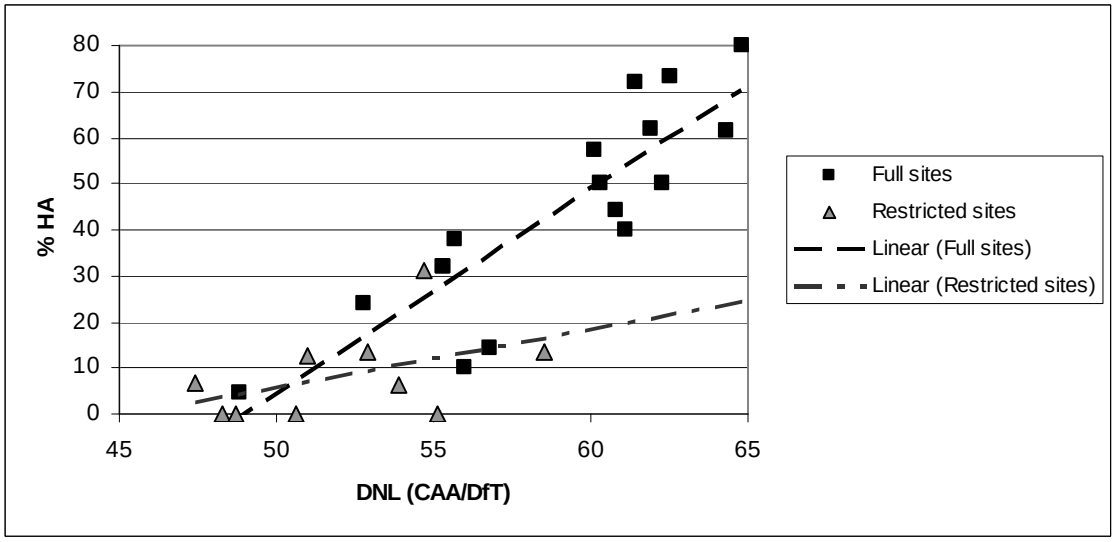

Figure 2 ANASE '% Highly Annoyed' (% HA) Heathrow results.

Figure 2 indicates that the Full and Restricted scatter plots and trend lines 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, i.e. 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[2]. 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 tests.

Only in circumstances when statistical testing accepts coincidence 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. To ignore the statistical testing results rejecting the coincidence of the data sets is not sound practice. It removes any possible sound foundations for subsequent ANASE modelling claims about (e.g.) 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.

Figure 3 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.

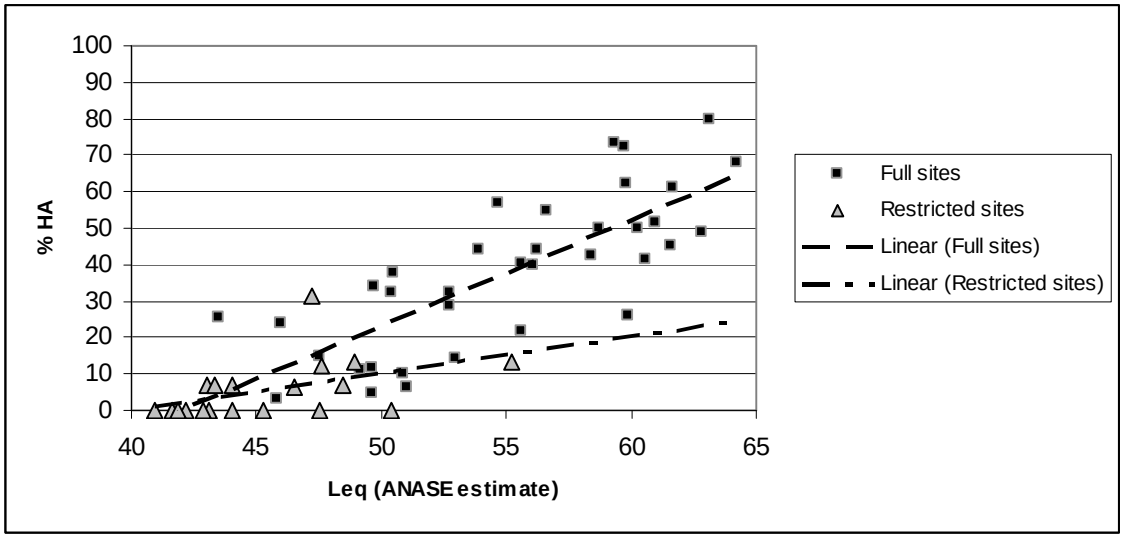

Figure 3. ANASE Full and Restricted sites % Highly Annoyed (%HA).
[Note Site R17 excluded – as in ANASE analyses.]

ANASE data for Full sites are markedly out of line with the results of reputable international and previous UK work[5]. ANASE Restricted site data are broadly consistent with international and ANIS results, although could under-estimate annoyance because of the differences in the experimental/questionnaire contexts – compare Figure 1.

The conclusion is that the introduction of noise equipment changes the aircraft noise annoyance dose-response relationship by a roughly multiplicative bias factor. Thus, the ANASE results are not comparable with previous ANIS work, and the ANASE contractors' claims – e.g. increased annoyance over time, additional aircraft number effects – are invalid because they mostly derive from the biased data.

These ANASE failures would lead directly to the many problems with the SP valuations in ANASE identified by the peer reviewers. To quote one peer reviewer (Bateman): "the absolute value of such reductions *[i.e. SP valuations]* is, by the authors admission, implausibly high." The high valuations are simply explained by reference to the Figures above. Monetary valuations of noise via SP or other pricing techniques are intended to 'crystallise' people's annoyance or disturbance. All other things being equal, a bias in measuring annoyance would therefore imply a similar degree of bias in subsequent monetary valuations. The multiplicative bias on annoyance found here – from the slopes of trend lines on '% Highly Annoyed' – is roughly a factor of three. Thus, ANASE SP valuations would be expected to be about three times the values that would be obtained from aircraft noise SP studies without such biases. This can only be a rough figure – in practice, weightings and non-linear mappings probably complicate matters. But this evidence of design bias straightforwardly explains why the SP valuations are implausible (note that that the peer reviewers did not offer simple economic explanations to resolve ANASE's problems).

5 LESSONS ABOUT CONTRACTS

The comments here about contracts have to be general ones, because of the legal position about public comments on a specific Government contract. Contracts to carry out research studies often go wrong to some degree. To have a good chance of succeeding, and hence to provide assurance that public funds are not being mismanaged, a contract generally needs to meet some obvious positive criteria. Obvious key ingredients are:

Competent project managers

Competent contractors

Good quality technical advice

The word 'competent' means the appropriate combination of intelligence, integrity and track record. 'Good quality technical advice' means people who know the technical subject, who probably have carried out similar projects in the past themselves, and who persevere in getting answers to key project questions.

Looked at from the negative viewpoint, it is vital for the project managers to detect preventable failures. The contract may not deliver according to the specification; the work may be of poor quality; the outputs may be late and/or over-budget. Examples of the nature and causes of problems, summarised from a variety of governmental contracting guidance material, are

Improper award of contract – e.g. de-scoping specification for a particular contractor, biased contract evaluation process;

Contract not delivered properly – e.g. ineffective monitoring, no effective auditing programme;

Contract cost over-runs – e.g. 'lowballing' by contractor, Ineffective monitoring.

'Lowballing' means the contractor underquotes when bidding for the contract, hoping that the quality of supervision and monitoring will offer potentially large claims for extras and over-runs.

6 METHODOLOGICAL LESSONS

6.1 Get independent expert technical advice

There is a right and a wrong time to get independent technical advice about a project. The wrong time is after the fieldwork and analyses have been done, i.e. to have an ANASE-type peer review by experts. The right time for expert input is very early in the project, when pilot studies are being designed and analysed, so that the work goes in the right direction and avoids obvious traps. It is vital to find genuine experts and take their advice seriously: they may well not be right all the time, but their questions have to be answered. [This is the 'Frobisher lesson'. Martin Frobisher, the 16th Century English explorer, went on three voyages to northern Canada, bringing back increasingly large amounts of gold ore – 1100 tons on the last trip. But it was Fool's Gold – iron pyrites. Frobisher should have taken advice from a reliable metallurgist at the outset.]

6.2 Do the right annoyance study

ANASE's specification for assessing annoyance was largely a reasonable one, but failed to address the key policy decisions required for proposed Heathrow developments relating to 'mixed mode' operations[3]. Heathrow currently operates in 'segregated mode', with landings on one parallel east-west runway and departures on the other one. About 75% of the time the airport is operated westerly, i.e. with flights going towards the west, and 25% easterly. On westerly operations, there is runway alternation: each runway is typically used for takeoffs about half the time, morning or afternoon. There is no alternation for easterly operations, as takeoffs only take place on the

southern runway; northern runway takeoffs are not generally allowed because of the 'Cranford agreement', named after an area to the east of the northern runway.

The proposal to expand Heathrow's operations eliminates segregated mode operations, alternation, and the Cranford agreement. Instead, Heathrow would operate in mixed mode for both runways, for both easterly and westerly operations. Mixed mode means mixing departures (D) and arrivals (A) on the same runway, i.e. at peak hours a sequence A-D-A-D-A-D-… This would change both the total number of annual flights and the noise exposure patterns around the airport. For example, people in Stanwell Moor, living underneath westerly departure routeings from the southern runway, are currently exposed to a westerly mode Leq of about 74, contributed almost entirely by an Leq of about 77 from southern runway operations, and an easterly mode Leq of less than 60. With mixed mode, the westerly and easterly Leq values in projected scenarios (e.g. for 2015[6]) might both be about 74 [2015 traffic would have reduced average noise levels but increased numbers of flights]; but without periods of respite, either from day to day or during each day. How would these changes in noise exposure at different times affect people's disturbance?

The ANASE specification included no examinations of mixed mode/alternation/Cranford agreement. But these elements play a crucial role in the consultation on Heathrow's development: more than half of the Consultation Document's Response form[2] deals with these issues. The inference is that research specifications should try to meet the likely policy needs. Trying to understand the probable annoyance effects of a change to mixed mode would no doubt have been a difficult task. This is because of the very different diurnal and day-to-day noise exposure patterns around the airport, and the presence of confounding factors (in particular, work connections with the airport strongly affect responses – these connections are much more likely to the east of Heathrow because of the better road/underground communication links). However, it would have been a very worthwhile study if it ensured that policy decisions were based on an accurate reflection of potential disturbance.

6.3 Professionally attitude testing/noise estimation/statistical modelling

A professional approach is vital. This does not mean anything very esoteric or technically complex. The need is to do the basic things sensibly, whilst focusing on the goal of meeting the project specification. There are several reputable textbooks and accessible governmental guidance documents about attitude testing and accepted social science methodology[5]. Equally, there are established multiple regression and basic statistical testing techniques[7].

In designing social survey based studies, the aim must be to eliminate or reduce the potential for serious technical/statistical biases, and then in database modelling and analyses to try to uncover possible biases. Proper account must be taken of statistical testing results: they must not be disregarded. It is not good scientific practice to generate complex models – 'rewriting attitudinal acoustics' – without first taking proper care to check if straightforward reasons, in this case design bias, explain the observed data. If ANASE had been properly comparable with previous studies, the devastating laboratory experiment problems noted above would have been detected quickly. The alarm bells would have sounded that there was something very dubious in the study design.

Aircraft noise estimation from computer models is an inherently complex process, particularly because of variability in atmospheric conditions along the propagation path, so estimates need to be matched against appropriate field data collections (e.g. the large-scale programme in ANIS) and current best practice[1,2,8].

6.4 Produce the right specification

The ANASE specification might have generated worthwhile project outputs if efforts had been made to prevent the SP element of the study distorting the annoyance survey, which in turn distorted the SP results. Acoustics professionals know that there are major methodological problems in combining social survey and laboratory experiments. The noise research literature has examples

showing that laboratory experiments – noise playback equipment in this case – change people's disturbance reactions.

It is puzzling why DfT focused so much on SP valuations in the specification, given that hedonic pricing methods are generally used in valuing transport noise[9,10] – and indeed are used in the current Heathrow Consultation document. To quote from the ANASE Executive Summary:

> "Overall, therefore, we do not think that the valuations from either *[ANASE SP]* method are safe, and it will probably be necessary to rely on sources based on Hedonic Pricing.

SP's methodological uncertainties were known to respected researchers in this field, e.g. the peer reviewers, well before ANASE was commissioned.

The implications are that the two kinds of exercise, annoyance and SP, should not be combined unless there is confidence that bias/distortion effects are eliminated or controlled; and that very large-scale SP studies to aid policy decisions should not be carried out until there is good agreement amongst expert researchers that they will produce good quality results.

7 CONCLUSION

ANASE's problems were predicted, and its failure to produce cost-effective outputs to help policymakers was preventable. The study was intended to be 'substantial research that commands the widest possible confidence', but it failed to achieve that central goal or to have any worthwhile impact on the current Heathrow Consultation process – ANASE's claims added nothing but confusion. Nevertheless, there are some valuable lessons to be learned from ANASE's failings regarding aircraft noise study specifications, contracts, design methodology and data analysis.

8 REFERENCES

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

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

3 DfT. Adding capacity at Heathrow airport - Public consultation. UK. (2007). *http://www.dft.gov.uk/consultations/open/heathrowconsultation/*

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

5 P. Brooker, ANASE: Unreliable – Owing to Design-Induced Biases. Acoustics Bulletin. Jan/Feb, 26-31. (2008).

6 D. P. Rhodes & Beaton, D., Revised Future Aircraft Noise Exposure Estimates for Heathrow Airport. ERCD Report 0705 Prepared by the Civil Aviation Authority on behalf of the Department for Transport, London. (2007).

7 N. R. Draper & Smith, H., Applied Regression Analysis. Wiley-Interscience; 3rd edition (1998).

8 P. Brooker, Critchley, J. B., Monkman, D. J. & Richmond, C., 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, Aircraft Noise: Annoyance, House Prices and Valuation. Acoustics Bulletin. May/June, 29-32. (2006).

10 B. Pearce & Pearce, D., Setting Environmental Taxes for Aircraft: A Case Study of the UK. CSERGE Working Paper GEC 2000-26. *http://www.uea.ac.uk/env/cserge/pub/wp/gec/gec_2000_26.pdf*