REPORT ON ANALYSIS OF NAS 2B CONDUCTED BY AIRSERVICES AUSTRALIA

PROFESSOR TERRY O'NEILL

Provided to CEO of CASA

26 AUGUST 2004

Executive Summary

Various changes to Australia's airspace system, collectively known as Stage 2b of the National Airspace System (NAS), were implemented in November 2003. An analysis by Airservices found that certain aspects of these changes associated with the replacement of Class C airspace with Class E airspace had led to a reduced level of safety.

This report was commissioned by the CEO of CASA to comment on the quality of Airservices' analyses.

Airservices' methodology involved two basic stages. First, estimations were made of the likely number of potential conflicts that might occur in the absence of any operating defences. These data were then entered into a 'fault tree', which simulated the likely success of various collision avoidance activities in various classes of airspace. This method of risk assessment appears not to have been used in any other country.

While legitimate questions can be asked about the quality and representativeness of the data which led to the generation of potential conflict estimates, the main deficiencies discovered related to the fault tree methodology. Points of particular concern were:

- The method used to select raters to serve on the panels was not statistically sound and may have led to the selection of inappropriate panel members.
- Estimation of probabilities by experts is well known to be quite prone to error and a search of the literature discovered little published evidence for the validation of the particular probability estimation technique adopted (the 'Shang' method).
- Assuming however that the Shang method is appropriate, the procedures actually employed by Airservices deviated from it in many significant ways. Most of these differences would have been expected to decrease the accuracy of probabilities estimated.
- In some cases, Airservices seems to have arbitrarily decided to estimate certain probabilities themselves rather than referring the task to the panels. In these instances, the entered values were occasionally many orders of magnitude lower than what might have been expected from panel estimates. For this reason, it is possible that these values were particularly influential in determining model outputs.
- Examination of the actual probability estimates provided by panellists revealed cases where results were either clearly in error (e.g. where an individuals 'high' estimate was actually lower than their 'low' estimates) or inconsistent with what might reasonably have been expected to be provided by truly expert raters acting in an independent way.

- Airservices subjected their model to only limited validation. Validation is generally considered an indispensable part of model development and it would be unwise to rely on outputs that had not been subjected to extensive testing and confirmation.

As a result of these findings it was concluded that the analyses conducted by Airservices were insufficient to support a determination that the 2b changes had in fact led to a change in safety or to determine the extent of any such changes. Based on the evidence available, it is not possible to make judgments about the safety effects of NAS 2b.

The report also draws attention to the fact that even if there was complete certainty about the safety effects of the 2b changes, this information alone would not be sufficient to guide what steps should now be taken in relation to airspace reforms.

Introduction

Background

In November 2003 a number of changes to the airspace system were introduced, known collectively as the 2b stage of the National Airspace System (NAS) project. Two changes of particular note were the replacement of areas of enroute Class C airspace with Class E airspace, and the replacement of Class C airspace over Class D towers with Class E airspace.

AirServices Australia (AA) analyses

An analysis of the effects of these changes carried out by AA came to the following conclusions:

Enroute analysis

Results indicated that there were very small differences in the estimated number of fatalities per 100 years between Class C and Class E enroute airspace with radar (1.18 vs 1.2 respectively). In non-radar enroute airspace, Class E was estimated to result in 0.7 fatalities per 100 years compared with 0.64 for Class C. Aside from these general figures, individual estimates were also made for passengers and crew members that took into account their exposure to air travel. The individual risk estimates for aircrew and passengers were found to be within the 'tolerable' range (as defined by the authors) but at the upper end of pre-established control limits.

Class E over D analysis

Four locations were modelled with varying levels of radar capability:

- Albury (predominantly non-radar)
- Hobart (current radar capability, post radar deployment at Launceston)
- Tamworth (predominantly radar)
- Mackay (full radar)

Results indicated that Class E airspace represented the highest overall risk but there were major differences between locations. At Mackay, the difference in the number fatalities per 100 years between Class E and Class C was very small (2.07 vs 2.01). On the other hand, at Hobart the estimated number of fatalities per 100 years for Class E airspace (5.38) was almost 3 ½ times that for Class C airspace (1.55).

The absolute value of estimated risks in Class E airspace were classified as 'intolerable' for at least one aircraft type (VFR/IFR) by Group (Passenger/Crew) combination set at each of the locations.

Terms of reference for this report

The terms of reference for this report were to examine the information in the document entitled 'Safety Assessment Study' prepared by AA in relation to Stage 2b of the National Airspace System (NAS), together with any additional information on which the Safety Assessment Study is based and is available from AA.

Reviews were to be made of all aspects of the analytical work contained in the Safety Assessment Study and advice provided on:

(a) the quality and appropriateness of data used in the analysis in the Study and whether significant relevant data was inappropriately excluded;

(b) the appropriateness of models or analytical techniques used in the Study;

(c) the validity of the outcomes of the analytical techniques; and

(d) the reliability and appropriateness of the interpretation of the analysis results by the authors of the Study.

1.3 Where it was felt that changes should have been made to the analysis or its interpretation, consideration was to be given to the likely practical effect of these changes on the conclusions of the Study.

People interviewed

The people listed in Appendix A provided technical input to this review.

Study limitations

The analysis presented in this report represents only what was possible to achieve under the circumstances of the project. Work began on this project on 13 August 2004 and was required to be finished by 25 August 2004 to enable the report to be finalised in time for the AA Board meeting on 27 August 2004.

In addition to the very tight deadline, the review was hampered by the lack of availability of relevant information at crucial times. Information was sourced directly from AA as it was not in CASA's possession at that time. Some information did not arrive at CASA until 19 August 2004, and on 23 August 2004 further information was offered by AA but was not accepted on the grounds that it would be impossible to consider it in the time frame available.

Further, a large number of individuals referred to under People Interviewed gave unsolicited technical input to the review. This input was of highly variable utility and relevance to the review, but which nevertheless was assimilated and considered, and this process necessarily impacted on the time available to review documentation.

For these reasons, it was not possible to examine all material of interest or relevance, and the material that was examined was not investigated in as much depth as desirable. In addition, there was insufficient time available to check conclusions with relevant experts.

While it is felt that this report makes a significant contribution to understanding the nature of the analysis work performed by AA, it has not been possible to subject it to the normal level of scrutiny and checking. As a result, it is not possible to guarantee that the report is completely free of error or omissions.

Examination of 2b analyses

Considerations prior to the introduction of 2b

Initial assumptions

It could be argued that it would have been reasonable to make an initial assumption that the 2b changes would result in a higher level of risk.

This is because the major changes at Stage 2b involved the replacement of Class C airspace with Class E airspace which, in turn, implied a reduction of Air Traffic Control (ATC) involvement. Specifically, while all aircraft are actively separated in Class C airspace, in E airspace this was only true in regard to pairs of Instrument Flight Rules (IFR) aircraft. In the case of IFR-VFR aircraft pairs in Class E, IFR aircraft would be given traffic information about the presence of Visual Flight Rules (VFR) aircraft (of which ATC was aware). No separation or information was provided to avert the risk of collision of VFR-to-VFR aircraft.

Attention was drawn to the following arguments:

- The reclassified airspace had low traffic densities and therefore limited risk of collision no matter what airspace existed;
- Any theoretical differences between Classes C and E would disappear under Instrument Meteorological Conditions (IMC) conditions (i.e. when VFR aircraft would be excluded); and
- Pilots would be likely to adapt their behaviour as a result of the changes, becoming more vigilant than they are at present.

These arguments, if accepted, are only relevant to the *size* of the expected increase in risk and do not invalidate the basic conclusion that risk should increase.

The only argument advanced that the changes might actually *decrease* risk relates to cases where Class C towers (operated by several controllers) were replaced by Class D towers where these might be operated by only a single controller. In this case, the replacement of the C airspace above the D tower with E airspace was said to enhance safety. This is based on an assumption that the need to separate aircraft in Class C airspace could distract the controller's attention from activity in the Class D region where the risk of collisions was higher.

It should be noted that this posited safety advantage only operates given the existence of another 2b change (i.e. the designation of some towers' airspace from C to D). In addition, the potential for distraction would appear to be limited since, as those advancing this argument admit, there was very little traffic in this area of Class C airspace.

Safety case

The NAS 2b safety case conducted prior to the implementation of these changes seems to have consisted primarily of subjecting those aspects of the 2b changes that were not consistent with the system in use in the United States, to examination by a panel of experts who discussed their likely safety acceptability. There is no evidence that there was any attempt to subject proposed changes to formal quantitative analysis.

CASA provided input to the safety case, most notably in a minute of 12 September 2003. Extracts from that document stated:

- "CASA believes the US FAA NAS to be safer than the existing Australian airspace management system based on a simple comparative analysis of mid air collision statistics between the two countries"
- "The detailed evaluation of the presented design safety case evidences that the Stage 2b implementation will bring about an increase in risk beyond that which exists in Australia today."
- "If the NASIG claim that it is necessary to the full implementation of NAS that Australia transition through this higher risk stage is valid, then clearly the sooner NAS Stage Four is implemented and the potential safety benefits realised, the sooner this risk is removed".

In other words, CASA's advice was that it felt that the 2b changes would result in increased risk but that these might be justified (given appropriate risk management measures) if they were necessary as interim measures to bring about a much safer end state.

Australian Transport Safety Bureau (ATSB) evaluation of 2b work

A report by the ATSB of July 2004 examined changes in the incidence of four indicators before and after the introduction of the 2b changes: Airproxs, Breakdowns of separation (BOS), Violations of controlled airspace (VCAs), and Resolution Advisories (RAs).

Analysis of the data found no significant differences in the number of Airproxs, BOSs, or RAs. While there was a significant increase in the number of VCAs after the 2b changes were introduced, ATSB quoted the view of AA that the increase may have been due to an increase in sensitivity of air traffic control to VCAs and a failure of pilots to fully understand their responsibilities under the new system. The former point refers to data quality and if true would undermine the VCA data. The latter point however suggests a real but temporary increase in risk.

At any rate the ATSB concluded that the "data currently available does not enable reliable conclusions to be drawn about NAS 2b safety trends" and that "the results in this report cannot be used to confidently assert that safety has either improved or deteriorated as a result of NAS 2b changes"

General comments on AirServices Australia methodology

AA sought to estimate the risk associated with the NAS 2b airspace model. The procedure used employed two main components. The concept was to simulate air traffic and collect occurrences of conflict pairs and then to assign a collision probability to conflict pairs using a fault tree. A conflict pair in the simulation is a pair that would have come within 1 nautical mile horizontally and 500 feet vertically, in the absence of avoidance strategies of any kind. The simulation component of the exercise was conducted using Total Airspace and Airport Modeller (TAAM)^{TM} software. The output from the simulation was fed into a fault tree model, (developed by AA with consultation with Risk and Reliability Associates), which estimated the likelihood that defenses would operate to prevent a conflict pair from becoming critical i.e. dependant on chance alone to prevent a mid air collision.

Advice was that this approach is unique to Australia and is not used elsewhere in the world in this manner. This was confirmed by several interviewees. Since the method has not been tested in the international arena, it is crucial that the procedure be subjected to rigorous scrutiny.

AA indicated that they had not had sufficient time to engage in extensive validation work but there seems to be no reason why the modeling process could not have begun much earlier, certainly prior to the adoption of NAS 2b, which would have allowed time for the necessary scrutiny.

It is also worthwhile to consider whether there are alternative approaches that might have been used, and that the best possible modeling approach has been used. If modeling is accepted and widely used in the international community, then it would be instructive to compare AA method against a leading international benchmark in a relevant airspace.

One issue that should be addressed is why more empirically based data modeling procedures were not examined in more detail. Thus an examination could have been made of the relationship between traffic flows and types of aircraft in various types of airspace and resultant incidents both in Australia and internationally. While it is by no means clear that this approach would definitely have been successful in producing highly accurate estimates, given the well-known difficulties associated with pure modeling approaches, a significant attempt to explore this option would seem to have been warranted.

It is interesting to note that in relation to proposed NAS 2c changes associated with CTAF procedures, section 12 of the NAS-IG safety case report of 21 May 2004 (Airspace modeling) argues that

"…key input assumptions require largely subjective judgments. These subjective judgments can be very inaccurate because reality can be counter-intuitive in actual practice. If the key input assumptions and relationships are highly subjective and inaccurate, then the modeling results will also be highly subjective and inaccurate" (page 42).

Similarly the report states

"It would not be useful to employ a computer model using highly doubtful subjective input assumptions to generate 'the answer' on airspace reform. This answer would simply NOT be any more rational, objective or scientific than the subjective set of input assumptions. Use of an invalid model, or an invalid set of inputs will only cloud the issue" (page 43).

Thus NAS-IG seems to say (correctly) that the value of models depend on their specifications and the quality of input data, but also asserts that any input data of a subjective nature would inevitably lead to questions about the value of the output. As discussed later in this report, the input to the AA fault tree analysis in fact derives largely from subjective estimates of probabilities.

Total Airspace and Airport Modeller (TAAM)™

General background

TAAM software simulates aircraft movements in airspace. It requires relevant input from real world data on aircraft movements in the airspace under study. The software then uses that data to generate a simulated number and type of conflict pairs. A conflict was defined as when 2 aircraft came within 1 nm horizontal and 500' vertical.

AirServices Australia Airspace Risk Assessment Class D over E Towers Report details the process of obtaining real world data inputs to the TAAM modelling process. Data collection was a complex exercise with different procedures required for IFR and VFR. The data was collected from various sources, including TAATS (RDR & FDR), tower strips, AFTN, tower statistics, field interviews, industry panels and internal panels. In the main, however, data for IFR aircraft was derived primarily from flight plans whereas VFR data came largely from radar observed tracks and level. The data was collected from a period immediately prior to 27 November 2003 and was supplemented by data from June 2004.

Since only 3-4 weeks of actual data was available, this data was used to generate a much larger set by randomising (+/- 30 minutes) take-off time and aircraft performance (appropriate for type), but using existing tracks.

Issues with TAAM

Problems with TAAM could derive from either a) the software itself or b) the quality of the input data.

Software issues

It was noted that interviewees were generally comfortable with the TAAM program. The software was said to be in use throughout the world and it was considered to be a well-regarded tool.

In spite of this, it is important to remember that the TAAM algorithms, like any modelling process, need to be examined and their appropriateness cannot simply be assumed. It is particularly worth noting that its appropriateness might be affected by the special purposes to which its output is being put i.e. input to a fault tree model.

Data issues

There are many aspects of the data which may impact on the accuracy of the modelling.

Data quality

Some of the data is collected by interview. This type of data is inevitably more variable than other sources and may contribute to the variability of the conflict pairs. The method of selecting individuals for interview (other than material presented regarding the industry panel) has not been clarified. It is crucial that the selection is random from a clearly defined population according to documented selection criteria.

One interviewee queried whether the TAAM modelling took account of the NAS recommendation that transiting aircraft avoid the airspace normally occupied by inbound or outbound aircraft and should monitor and avoid traffic associated with the aerodrome. This is a valid point which should be clarified.

Given that the conflict pairs were based on pre-change traffic patterns, it is not clear that the possibility of behavioral modification in response to these changes have been adequately addressed. Various scenarios have been postulated by interviewees, including

- Some IFR traffic may have avoided E airspace which implied an increase in the proportion of VFR in E airspace. (AA interviewee)
- E airspace may attract small aircraft that would not have flown there previously

Data representativeness

AA attempted to assess whether the conditions experienced in the time window of data collection were representative, including seasonal variation and traffic loads. In spite of this, 3 weeks of data which is subsequently randomised may not be sufficient to capture the variety of traffic variation over the whole time window under consideration. Running the simulation for longer periods will not alleviate this issue since all this will do is provide variations on existing patterns. It will not generate qualitative new patterns.

In addition, the effect of trends have not been incorporated into the model. Air traffic is increasing and the effect of varying the overall traffic intensity has not been modeled. The AirServices Australia Airspace Risk Assessment Class D over E Towers Report presents graphs of the load for each of the airports and

there is appreciable variation over the year. It would be instructive to consider the impact of overall traffic level.

Fault tree software and structure

Background and description

The conflict pairs from the TAAM modelling were fed into a fault tree model, to determine whether the conflict would have been resolved by system defences (e.g. ATC intervention). Those that were not resolved were then identified as critical pairs and the probability that a critical pair would collide is modelled by a chance factor. AA used the software package REFLEX^{TM} Fault Tree/Event Tree software package for the fault tree component of the modelling process.

The fault tree structure used by AA was based on previous work by AA and CASA. As noted on page 104 of AirServices Australia Airspace Risk Assessment Class E over D Towers Report 'the full fault tree is very extensive, and so only the first level has been included in this report'. For each conflict the fault tree requires the input of approximately 62 probabilities, of which approximately 12 are determined empirically and the remainder by an expert panel. These probabilities vary across a large number of scenarios, aircraft characteristics etc.

Aside from the conflict pair data discussed previously, the accuracy of the fault tree outputs will depend on:

- The nature of the algorithms incorporated into the fault tree software;
- The logic and structure specified by the particular fault tree configurations;
- The accuracy of the empirically determined probability estimates and
- The quality of the probabilities provided by the expert panel.

The last factor is discussed in more detail later in this report.

Software used

Due to time constraints there was not an opportunity to examine the nature of the algorithms used by the REFLEX software and therefore comment cannot be given on its suitability although it is noted that REFLEX is a major supplier to fault tree modellers around the world.

Fault tree structure

AA indicated that it has self-validated its fault tree model by "years of experience". Although time constraints did not permit a detailed examination of the fault tree structure, issues which did come to light include:

- It was stated at interview that the model contains in excess of 130,000 probabilities, of which approximately 100,000 are derived from expert panels. Clearly this is an extremely complex model. In general the more complex a model, the more prone it is to error and the greater the need for formal validation.

- In practice, the panels estimated only a relatively small number of probabilities. In many cases the same probability was used repeatedly throughout the fault tree model on the assumption that the estimates were, in these cases, independent of context. Without this assumption, it probably would not have been feasible to populate the model. Nevertheless, the notion of equivalence remains an assumption that required testing.
- A prominent factor in the fault tree is visual acquisition of one aircraft by another. The 'Andrews' model is used to predict the probability of visual acquisition in a given situation. An important parameter in the model is T2, the time at which visual acquisition ceases to be useful. In the modelling exercise, T2 was set to 15 seconds, or in other words, if the aircraft are within 15 seconds of collision visual acquisition is no longer of any use in avoiding a collision. In interview, AA mentioned that 12 seconds was often used as the point after which the outcome is determined by chance alone. While it has been suggested by one interviewee that a figure as low as 6 seconds may be appropriate for smaller planes, even setting T2 to 12 seconds may have an effect. This may be quite important since it was suggested that visual acquisition probabilities are inversely related to the square of the distance. The sensitivity of the fault tree to varying the parameter T2 should be tested.
- The issue of whether or not TCAS should be included in the fault tree was very contentious. The AA position was that it would have been inappropriate to model the effect of TCAS given the ruling by ICAO set out in Annex 11 (Air Traffic Services, paragraph 2.4.2), adopted in February 1993.
    - Without commenting directly on the appropriateness of the ICAO ruling, it is noted that the omission of this factor from the fault tree will compromise efforts to accurately model actual prevailing risks[1].
    - Even accepting the ICAO ruling, advice indicates that there would have been no impediment to exploratory work to estimate risk with and without the inclusion of TCAS.
    - It is possible that the inclusion of TCAS would have led to substantially lower risk estimates, perhaps below the pre-established 'intolerable' control limits.
    - It is also possible that the modelled safety effects of TCAS would be stronger in Class E than Class C airspace. If so, this would reduce the estimated differences between these two classes of airspace.

[1] This is consistent with the view of Risk and Reliability Associates who distinguished between the tasks of estimating absolute risk and determining the need for air traffic services in an area. It concluded that while: '… TCAS cannot be included in risk modelling when determining the need for air traffic services in an area.', it '…ought to be taken into account in any model which purports to predict absolute risk'. (Review of NAS Class E and Class C, May 2004, Risk Reliability and Associates Pty Ltd., primary recommendation b, page 6 and 7.)

Issues affecting probability inputs to the fault tree

It should be noted that most of those interviewed focused their attention on data provided by the panel which consisted primarily of pilots. Because of limited time, this report also reflects that emphasis. Nevertheless, it should be remembered that there was also an ATC panel which provided input to the model and the quality of its estimates is equally deserving of scrutiny.

Non panel derived inputs to the fault tree

The formula for the probability of collision once system defences have failed has been developed at AA. The formula gives the probability of collision once the aircraft have entered the conflict region. This region is orders of magnitude larger than the volume of aircraft. Since this probability enters at the last stage at the fault tree and multiplies all the other output, it potentially can have a large impact on the absolute levels of estimated risk and the estimated risks may be very sensitive to this formula. The sensitivity to this formula should be assessed.

Collision consequence values are used to determine the number of deaths in a collision. For example if a high capacity aircraft collides with a high capacity aircraft the estimated fatalities per collision is given as 131.8. A full table is presented on page 35 of Airservices Australia Airspace Risk Assessment Class D over E Towers Report. However it is noted in the footnote on page 34 of this report that the collision consequence values date from 1996. This raises the question if load factors and aircraft types in 2004 are sufficiently similar to 1996 for these numbers to be appropriate. Given that the overall risk will depend heavily on the number of occupants of medium and high capacity aircraft, this issue may be important.

General problems with experiential estimates of probabilities

Most of the probabilities in the fault tree derive from estimates provided by expert panels. The psychological literature is replete with research cataloguing the inherent difficulties that people have in estimating and interpreting probabilistic information. These errors and biases are known to operate quite generally, with estimations by subject matter experts not exempt from their influence.

As difficult as normal probability estimation tasks are, those required for the fault tree analysis face special difficulties.

First, the accuracy of probability estimates is particularly problematic when dealing with very rare events that the rater might never have personally experienced or has experienced too rarely to have clear understanding of their frequency.

Another issue with pilot behaviour is the question of generalisation. Even if an experienced pilot had some idea of how often he or she had committed a particular error they may have had little opportunity to directly observe the

behaviour of others. If it is assumed that others will behave as you do this may be misleading if there are differences between individuals in error rates. On the other hand error may also occur if the rater succumbs to the well known psychological tendency to underestimate one's personal errors and failings while overestimating those of others.

Lastly, when IFR pilots are asked to estimate probabilities associated with the behaviour of VFR pilots, they will be generally be required to base their judgements on their own behaviour of several years ago. If these experiences were very distant it will introduce an extra element of unreliability.

Panel member selection

A crucial issue in obtaining useful probability estimations is that raters be both expert in the relevant subject matter and acting in an unbiased and fair manner.

Selection criteria –formal and de facto

AA provided a copy of an email dated 31 May 2004 dealing with the establishment of a NAS Industry Representative Panel. It described the following selection criteria:

- The panel would have representatives from a broad range of the aviation industry[2];
- All members would have significant current experience in aviation as well as experience in one or more aspects of the use of the airspace being studied; and
- Panel members must be prepared to provide practical advice based on personal experience without being influenced by representative interests or responsibilities.

A document summarising panel selection criteria provided by AA on 19 August 2004 restated the last criteria as '*members of the panel are selected as individuals and not as representative of aviations interest groups'*. It was noted that this did not disqualify members or representatives of interest groups, but merely required any such persons not to act in this capacity in making their probability estimates.

While these criteria are all reasonable, two issues are noteworthy:

1. While the final criteria alludes to this, there is no explicit criterion ruling out participation of anyone with strong and fixed views about the value or otherwise of the airspace changes introduced.
2. The criteria are not strongly exclusionary. It is likely that there would be a very large number of experienced pilots who would be acceptable.

[2] The original email suggests that there should be representation from 3 groups that were not listed in the document of 19 August: Defence, Airport owners and RAPAC/NAPAC. In spite of the deletion of the RAPAC group, one of the VFR pilots was described as a RAPAC convenor.

This raises the question of how panellists were selected from among the much larger group of eligible pilots. In discussions with AA it was stated that panellists were chosen by AA staff, perhaps as a result of prior participation in industry consultation processes. It is unlikely that this would constitute a statistically unbiased sample.

An obvious alternative would have been to obtain a list of current pilots and sample at random from this.

Financial considerations would have acted as a de facto selection criterion. The email of 31 May 2004 suggests that not only should panellists be provided with funds to cover travel and accommodation costs, but should also be provided with $750 per day as a sitting fee. Discussions with AA seem to indicate that in fact these fees were not paid. If so, given that panel members had to be available for 2 sessions of 2 days each, this is likely to have affected the representativeness of the sample.

In the case of salaried employees with financial cover from their employers, the possibility exists that permission would only have been granted if the employer expected the employee to argue a case favourable to the employer. In the case of those who were self funded, perhaps only those who were deeply committed and willing to forgo income were selected. This immediately raises the possibility that they had pre-established views and that these views would have affected their estimations.

Vetting of panel members

It might have been prudent to have made some efforts at testing the actually expertise of the potential panel members. For example, it might have been possible to assess potential panellists on topics where empirical data existed to verify the accuracy of estimates.

Composition of selected panel

Table 14.1 of Appendix 5 in the E over D Towers report (page 92) indicates that the panel included two AA staff members – an enroute controller from Brisbane centre and an approach controller from Melbourne centre. The participation of these people would seem to have been inconsistent with the stated selection criteria. Even if it proved to be the case that these controllers were also experienced pilots, their employment by AA would still be inappropriate and it could be viewed as a conflict of interest.

The selection criteria, while specifying which areas of the aviation industry should be represented, do not specify their relative proportions. As it turned out, aside from the 2 AA employees, there were 7 IFR/RPT representatives and only 2 VFR representatives. One panel member was listed as being a glider pilot but also as having an IFR rating. Another person from the IFR/RPT sector was invited but did not attend. It is worth noting that if, as was frequently alleged by VFR representatives, IFR pilots tended to unreasonably

underestimate the competence of those in the VFR sector, this imbalance could have led to a significant bias in the result.

In addition to the 2 AA staff on the panel, there appear to have been 5 additional AA employees in the room and 2 other representatives from the firm Risk and Reliability Associates (R2A) acting as advisors. This seems a very high ratio of AA staff to panel members and it raises the possibility of undue (even if unintended) influence, especially if the AA staff participated in topic discussions.

The presence and participation of R2A staff, is also problematic since it would seem to compromise the perceived impartiality of those who would later be called upon to act as independent assessors of the validity of the process. This observation is in no way meant to question the integrity or quality of the work performed by R2A. It is only meant to comment on the impact on their perceived independence, since they were later called on to provide an independent review.

Shang methodology and deviations from the literature

The Shang method appears to have first been described by Ford (1975)[3] as an alternative to the more traditional Dephi methods of averaging expert opinion.

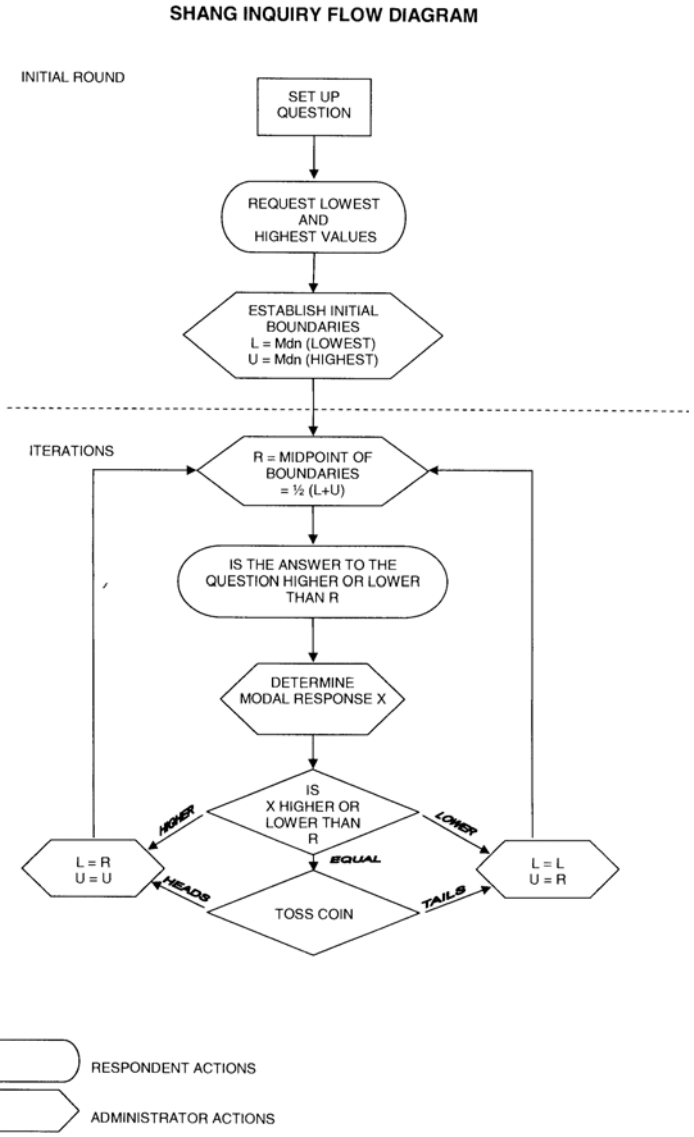
Major characteristics of the approach are that raters are expert in the domain of interest, their ratings are anonymous, and there are no opportunities for them to influence the views of other raters.

In very general terms, the technique consists of asking individual raters to provide both their highest and lowest estimate of a given probability. The median of these estimates is then calculated. The median is then taken of these high and low estimates. Raters are then asked to vote on whether the true probability was higher or lower than this point. If, for example, the majority feels that the true probability is lower than this point, a new point is calculated using as boundaries the former point and the median lower bound estimate. The process is repeated at least three times.

A diagram illustrating the basic features of the model is given in Figure 1 below.

[3] Ford, David, A. Shang inquiry as an alternative to Delphi: Some experimental findings, Technological forecasting and social change 7,139-164 (1975).

Figure 1[4]



One apparent difficulty with this approach is that each expert is asked to estimate the highest and lowest *plausible* value which might be the true answer to the problem. In standard statistical practice such intervals would be specified at a given level of confidence which defined what is meant by 'plausible'. For example, raters might be asked to specify a range which they were, say 90% confident contained the true value. Thus a rater who interpreted 'plausible' to mean 90% confidence might estimate bounds of 1/100 and 1/1000 but someone, *with the same view of the probabilities* but who thought that a 99% confidence interval was being called for, might estimate a much wider range, e.g. 1/10 and 1/10,000.

Ford argues that this technique produces better outcomes than the normal Delphi approach and although it proved impossible within the time frame of

[4] Figure 1 is sourced from page 143, Ford (1975)

this study to locate supporting research, a working assumption could be made that the processes described by Ford produce the best outcome for this method. If so, it becomes important to note that the process used by AA seems to have departed from Ford in several important aspects. These are discussed below:

1. The Ford method presents raters with a single median estimate R and then asks them to vote whether the true value is higher or lower. AA presented raters with *bounded* intervals (i.e. L↔R↔U). This would have provided much more information about the responses of other raters and could conceivably introduced a degree of pressure to conform to the views of the majority[5].
2. It should also be noted that the bounded intervals might not contain the raters initial estimate. In other words it may be that raters were asked to choose between alternatives, neither of which contained what the rater considered to be the true answer.
3. The raters initially wrote down their high and low estimates but were then asked to announce them in front of the rest of the group, one rater after another. This again may have allowed raters to be inappropriately influenced by other panel members.
4. One member of the expert panel indicated that there was some initial discussion of the problem among the panel members prior to writing down their estimates and further discussion after raters had announced their estimates. Although rare, raters were allowed to change the value input after hearing the views of other panelists.
5. Apart from issues due to panel composition, (discussed elsewhere), the AA procedures involved the participation of a number of AA staff members in discussions.
6. Mid point estimates were obtained by calculating the geometric mean where Ford derived mid points from median values. While the use of a geometric mean is not necessarily inappropriate (e.g. if there were strong theoretical reasons for assuming that the judgment errors were log-normally distributed) it should be noted that this estimate of central tendency would generally produce values that were different from the median. The effect of the difference in these estimates would be expected to be greatly magnified given the sequential multiplication of probabilities implied by the fault tree structure.
7. In the case where voting produced a tied result, Ford resolved this by flipping a coin. In the AA study, this was resolved by having one of the AA representatives present cast the deciding vote.
8. Although raters were allowed to absent themselves from particular estimations if they felt that they had insufficient knowledge of the topic being discussed, this rarely occurred. This may have been related to the well documented tendency of raters to be over-confident about the amount of information they possess. This overconfidence may have been assisted by general information provided to raters by AA that

[5] Appendix 13 of Airspace Risk Assessment, Class E over D tower (pages 89-91) states that only the mid point R was presented. However in conversation with the study's major author, it was indicated that a bounded interval was presented. The use of bounded intervals was confirmed by one of the expert raters.

indicated that skilled individuals had approximately a 1/1000 probability of making an error on a common task within their skill set. In other words, raters who knew nothing about the topic other than this general fact may have convinced themselves that they had sufficient knowledge to participate in the estimation round.

9. The introduction of the 1/1000 figure would also have been expected to operate as an 'anchor'. The psychological literature on estimation heuristics has clearly shown that raters estimates will more closely cluster around such anchor points than would otherwise have been the case[6]. (See Figure 2)
10. Instead of going through a set number of iterations, in the Class E over D Tower study, AA stopped the process when the ratio of the geometric mean high estimate was ≤3 times the geometric mean low estimate. In the vast majority of cases this meant that no 'voting' ever took place. This adoption of the 3:1 rule appears to be totally arbitrary[7]. Moreover this rule allowed the process to stop even when probability estimates of different raters were vastly different. For example, geometric means were calculated when high or low estimates from one rater differed from the comparable estimate of another rater by a factor of 1000 or more (see Figure 2 and Table 1). In the presence of such extreme divergence among raters, questions must be asked about the meaning of any mid point estimate based on this data.

It should be noted that some alleged deviations from the Shang process were not substantiated. For example, correspondence sighted implied that one panel member had, on hearing what they regarded as unreasonable estimates put forward by other panel members, altered their own estimates 'to mitigate the problem'. In discussion with that panel member they made it clear that although they did consider some of the other raters' views to be unreasonable, they made no adjustments to their own ratings as a result, entering only what they regarded as their best estimate of the true probability.

Validation issues

All models are prone to error and the more complex a model is, the more scope there is for various types of errors and distortions to occur. For this reason a thorough validation of outputs is an indispensable part of any modeling work.

On the basis of the material presented, it appears that the internal efforts at validation by AA were very limited. Airservices cited lack of time as the major reason for this. The sensitivity of the process appears not to have been explored in any systematic way, and in some cases was done after the release of the findings of the model.

[6] For example, Judgment under uncertainty: heuristics and biases, Amos Tversky and Daniel Kahneman, in the book of the same title, edited by Daniel Kahneman, Paul Slovic, and Amos Tversksy, Cambridge University Press, 1982.

[7] In fact, during the enroute E analysis, AA adopted a 4:1 ratio rule.

In spite of this, AA expressed confidence that although the absolute values of the risk estimates may not be reliable, the relative results (i.e. between airspace types) were. First, it is extremely significant that AA expressed little confidence in the absolute value of the model outputs given that issues of the 'tolerability' of risk were assessed against these figures. Nevertheless it is difficult to see how there could be confidence even in the relative estimates given the absence of validation work.

Validation could have taken two general forms; comparisons against real world data and examination of the abstract behaviour of the model.

Comparisons against real world data

It is very difficult to compare the predicted collisions against actual Australian data since the number of predicted collisions is extremely low. Nevertheless there are several approaches which may have proved to be very informative about the accuracy of the modeling.

One strategy would to model the occurrence of an intermediate step to collision which occurs more frequently. For example there are many more Airproxs, Breakdown of Separations, Violations of Controlled Airspace and Resolution Advisories than actual collisions. If the approach were successful at modeling these precursors to collisions, then there would be more confidence about the modeling of collisions.

The other approach is to recognize that there is a wealth of relevant data available in other countries, for example the United States. Nevertheless there is considerable disagreement about whether the use of this data would be appropriate[8]. The argument revolves around the comparability of the two countries. For example, Airservices argues that a comparison with the US is not appropriate due to differences in weather, topography, pilot behavior and culture and ATC systems. Some of these factors may increase while others may lower the rate of collisions. The following two points can be made about this controversy:

- It may be possible to estimate and formally characterize the size of any such differences and control for them statistically in the modeling process
- No two situations will ever be exactly similar. What is important is the relative importance of these differences compared with the size of the effect being estimated. The larger the apparent effects noted, the more confidence there would be that these were not entirely due to extraneous factors.

Reasonable agreement of the model with the US data would be reassuring. On the other hand, dramatic differences would suggest that all is not well with the modeling exercise. The comparison would need to allow for the

[8] It is noted that in previous modelling exercises, the accuracy of the ARM was tested when its risk profiles were compared with risk levels predicted by the US FAA's collision probability formulae for uncontrolled and controlled non radar terminal areas – the formula used in air traffic control tower cost-benefit reviews in the US, Canada, Australia and New Zealand.

differences, but the differences should not prevent the comparison being made.

Alternatively, rather than checking the Australian estimated risks against the United States, the whole modeling process could be tested directly on United States system. United States traffic data from comparable locations could be fed into the TAAM software and a United States expert panel could be convened to populate a fault tree with probabilities. If this exercise produced estimated risks in reasonable agreement with observed empirical risks in the United States, then it would supportive of the modeling exercise in Australia.

Abstract behavior of the model

The AA fault tree is very complex and time constraints on Airservices may have prevented a full assessment of its sensitivity. Nevertheless, partial checks could have been performed. The most influential nodes in the tree could have been identified and the effect of errors in the probabilities at those nodes could have been assessed.

The node failure probabilities for the fault tree are estimated quantities which are subject to error. The sensitivity analysis should examine the consequences of those errors on the estimated risks. For the Shang probabilities, the group high and low spread could be used as an indication of the possible bounds for the probabilities[9]. These upper and lower limits could be fed into the fault tree to assess its sensitivity to the probability inputs.

Another approach is that additional expert panels could be convened. The outputs of the model using their inputs would give some indication of the sensitivity of the results to the composition of the expert panel.

There are some modern statistical techniques which are likely to be effective in assessing the sensitivity of the fault tree modelling. The Bootstrap method is a procedure which gives estimates of variation while making minimal assumptions. In simple terms, suppose an estimate is required of the variance of a statistic T which has been obtained from a random sample. However, perhaps because of the complexity of the situation, it may be unwise to make assumptions about the statistical distributions which gave rise to the sample, and so usual procedures for estimating the variance of T cannot be applied. The core of the bootstrap method is to generate a 'bootstrap' sample of the same size as the original sample by selecting with replacement from the existing sample. The statistic of interest is calculated from the bootstrap sample to give a bootstrap observation T_1. This process is repeated a very large number of times, N say, giving rise to a 'sample' of estimators $T_1, \ldots, T_N$ and thereby an estimate of the variance of T.

Another modern statistical technique which is likely to be very useful in assessing the sensitivity of fault trees is Monte Carlo Markov Chains (MCMC).

[9] The node probabilities which are estimated empirically should have also an associated estimated standard error which would give upper and lower confidence limits for the estimated probability.

The core idea behind MCMC is to assign distributions to the inputs in a model and then to simulate from the distributions to obtain an estimated distribution of the quantity of interest. The distributions are specified hierarchically so complex dependencies can be built into the modelling.

Independent Validation

Airservices sought external expert assistance and review. The terms of the tender by Airservices (page 122, Airservices' Airspace Risk Assessment Class D over E Towers Report were to '… provide process assurance and provide expert advice where issues of process are identified' and '… provide a formal report which addresses the validity of the process and verification of the output of the model and its integration with the benefit-cost assessment'. The title of Appendix 10 of the report is 'External Independent Review'.

It is very desirable that the roles of providing expert advice on modeling and providing validation of the final model should be kept separate. If an organization has had an active involvement in the design and conduct of the modeling exercise, then it may be difficult for them to act independently and also to be seen to be acting independently when conducting the validation.

The contract was given to the firm Risk and Reliability Associates Pty Ltd (R2A). Unfortunately their report had not been completed in time to be evaluated in this exercise. Therefore it was not possible to comment on the nature and scope of their validation work.

Nevertheless, without in any way questioning the integrity of R2A or the quality of their work, the close association of this firm with aspects of this project raises a legitimate question about the perception of independence at the validation stage. For example:

- Risk & Reliability Associates Pty Ltd was involved in the conception of the ARM model in 1996.
- Risk & Reliability Associates Pty Ltd provided the estimated fatalities for the nine types of conflict pairs used in the AA Airspace Risk Assessment Class D over E Towers Report (page 34-35).
- Risk & Reliability Associates Pty Ltd was present at the expert panels (AirServices Australia Airspace Risk Assessment Class D over E Towers Report (page 92) and it was confirmed at interview that they participated in the process.
- AA acknowledges that Risk & Reliability Associates Pty Ltd 'has provided useful feedback on improvements to the logic of the fault tree and some aspects of the data validation process.' (AirServices Australia Airspace Risk Assessment Class D over E Towers Report, page 92).

Statistical examination of data quality of fault tree probabilities

Examination of quality of raw data from expert safety panel

The work of this section was conducted under extreme time pressure. As a result it was not possible to subject the analyses to the desired level of checking and it is possible that errors may have crept into the analysis as a result. On the other hand, since only a small subset of the data could be analysed within the timeframe, there may be other problems with the data which remain to be discovered.

An examination of the probabilities estimated by individual panel members[10] uncovered a number of findings which cast doubt on the validity of the figures provided.

First it is likely that no edit or logic checks were performed on the data. Of the 1477 high-low estimate pairs examined, in 28 cases (1.9 %) the 'high' estimate entered was actually lower than the 'low' estimate. In addition there were 39 other cases (2.6%), where the 'high' and 'low' estimates were in fact equal. Obviously if incorrect estimates were entered, these would adversely affect the estimates of the group's high and low geometric means and all subsequent probability estimates within that exercise.

While it is likely that the reversal of high and low estimates were the result of data input errors, it is not clear that the same can be said about cases where high and low estimates were equal. It is possible that the rater actually felt absolutely certain that the entered probability was the correct one. Insofar as this occurred, it casts serious doubt on the suitability of any such rater for a task of this nature.

Figure 2 shows the frequency with which panellists estimated specific probabilities. Truly independent experts may have been expected to advance a very large number of distinct probability estimates. In fact, estimates of 1/1000 (i.e. 10^{-3}) and 1/100 (i.e. 10^{-2}) accounted for over 1/3 of total answers. Two explanations for this concentration of results occur. First, AA advised the panel that a typical error rate for a skilled person is around 1/1000. This information may have biased the panel members' estimates and been especially influential in cases where the rater's previous knowledge was very low.

If a rater had no particular probability in mind, but just wanted to hazard a very rough guess, it seems likely that they might pick a figure such as 1 in 10, 1 in 100, 1 in 1000, 1 in 10,000 etc. In fact these types of estimates accounted for the majority of the total. This suggests that the estimates may have been chosen primarily for ease of mental calculation and may have been extremely imprecise.

[10] Due to time restrictions these analyses were limited to Scenario 1in class E, special E, and C airspace.

Figure 2

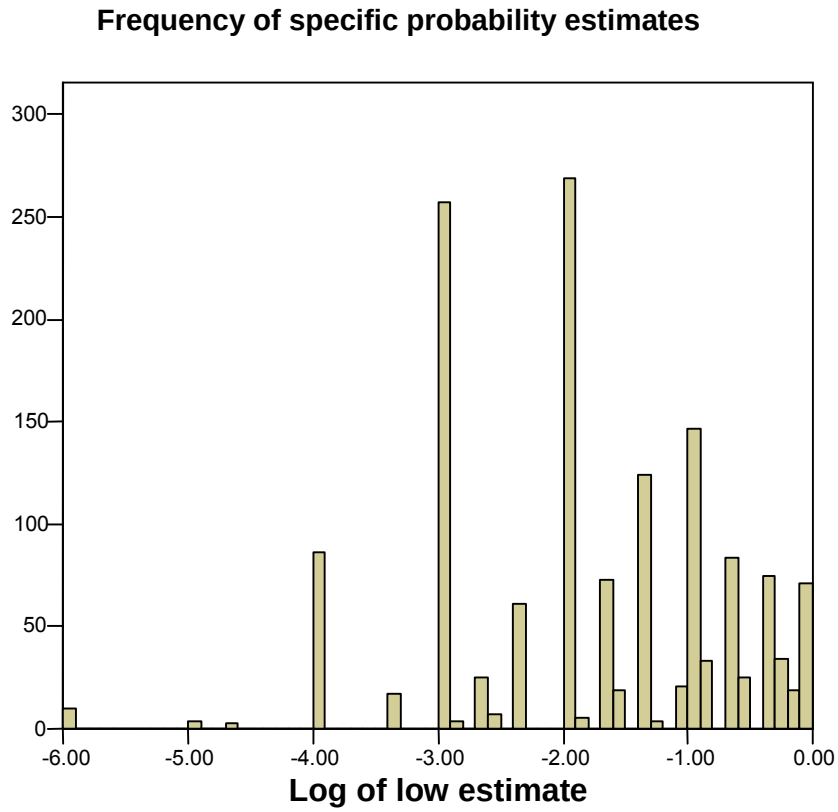


Table 1 shows data from 20 questions relating to Scenario 1 (VFR perspective) in Class E airspace. In total 120 estimations were made since each of the 20 questions involved both a high and low estimation, and estimations were made for three separate contexts (radar 45, radar 85, and non radar).

In each of these 120 cases, a ratio was calculated between the highest and lowest probabilities estimated by the individual members of the panel.

Table 1

Range of probability estimates and method of resolution

Range of rater estimates	**Number of estimates**	**Number resolved by vote**
At least 100 times	41 (34%)	6
At least 1000 times	7 (6%)	1

As shown, in 41 cases (34 %) one panel member put forward an estimate that was at least 100 times as large as another panel member. In 6% of cases the ratio was at least 1000:1.

It is also worth noting that even when there were extreme differences in raters' estimates, the process rarely proceeded to the voting stage to achieve convergence. In other words, the final group point estimate was arrived at by just taking a mid point of estimates that differed by a factor of hundreds or thousands. Averaging would only be justified if it could be assumed that each of the raters was expert and that they all estimating the same underlying quantity. It is extremely doubtful that such extreme ratios could have been achieved if these assumptions were true.

Examination of grouped data

Examination of the set of fault tree probabilities revealed that only questions dealing with communication failures by VFR pilots showed substantial differences between Class C and E airspace. Table 2 below shows data relating to these questions in 2 types of scenarios (designated as 1 and 3 by AA) under the Radar 45 context.

Table 2

Error estimates for 2 VFR scenarios in E and C airspaces under Radar 45 conditions[11]

Question	Scenario type	VFR E	VFR C	Ratio E:C estimates
1. What is the likelihood that you will not choose the ATS frequency responsible for the airspace as the appropriate one to monitor?	1	.085	.000001	**84,900:1**
	3	.087	.000001	**87,000:1**
2. What is the likelihood that you do not have the information available to choose the correct ATS frequency? (e.g., No current charts, frequency information not available, etc).	1	.061	.000001	**61,400:1**
	3	.065	.036	**1.8:1**
3. What is the likelihood that you make an error in determining what the correct ATS frequency is?	1	.028	.000001	**28,400:1**
	3	.024	.024	**1:1**
4. What is the likelihood that ATC or another aircraft does not alert you that you are transmitting on the incorrect frequency?	1	.639	.008	**82:1**
	3	.807	.006	**129:1**
5. What is the likelihood that ATC does not call to find out about your position/intentions?[12]	1	.924	.0002	**4,552:1**
	3	.849	.0002	**4,182:1**
6. What is the likelihood that you choose not to announce your position/intentions?	1	.598	.000001	**59800:1**
	3	.654	.0012	**545:1**

Generally the probabilities appear to show convincingly that VFR pilots are much less likely to make errors in Class C airspace. In fact the table seems to show that VFR pilot errors are in some instances almost 85,000 times more likely in Class E airspace, a difference that is likely to have had very

[11] Some figures have been rounded for ease of presentation
[12] Asked only of the ATC panel.

significant effects on the estimates of the relative safety of the two classes of airspace.

The problem with this conclusion is that the probabilities in the shaded cells, although entered into the fault tree, were not in fact estimated by the panel but directly assigned by AA staff. In other words, in these cases AA merely decided that these probabilities were remote and arbitrarily assigned them a value of 1 in 1,000,000.

There does not appear to be any obvious reason by why AA felt that these particular cells did not need to be estimated by the panel. Put another way, if this procedure was legitimate, why could AA not simply have decided to fill in all cells themselves rather than ask the views of expert panels?

It is interesting to note that in the question dealing with determining the correct ATS frequency (question 3), in scenario 1, the panel estimated the error rate in E airspace as .028 but AA considered the likelihood of this error in Class C to be so remote that it was not worth putting to the panel to estimate. But when the same event was required to be estimated in Scenario 3, it appears that AA was happy to count the probability (.024) as being equal in both classes of airspace.

Decisional issues

A distinction needs to be made between a) the accuracy of the risk estimates made by AA and b) decisions made about airspace changes after consideration of these estimates.

The analysis results from AA appeared to indicate that the 2b changes were associated with an increase in risk. As this report has made clear, legitimate doubts can be raised about the validity of this conclusion. Nevertheless, it is important to realise that even if the conclusions advanced by the AA analysis were indisputable, this information would not be sufficient to uniquely determine any particular course of action. For example, while reverting to the pre 2b situation might be a reasonable means of managing any increase in risk identified by the analysis, the analysis results would not necessary rule out the appropriateness of other risk management strategies. Such alternative strategies might include an acceleration of the NAS timetable, moving on more rapidly to the next stage of the project.

While a number of factors would need to be considered in determining the most appropriate course of action, one very important piece of information would be the likely safety level associated with the project end state and all interim project steps on the way to that end state. Thus decision makers might come to very different conclusions about what was appropriate to do now if it was known that subsequent stages would result in a marked improvement in safety both relative to the 2b stage and the original (i.e. pre 2b) situation, as opposed to knowing that subsequent stages would all result in a further reduction in safety.

Conclusions

The following general conclusions can be reached:

1. Problems with the way in which AA analyses were conducted and the lack of adequate validation work raises considerable doubts about the accuracy of the conclusions presented. The analyses conducted are insufficient to support a determination that the 2b changes have in fact led to a change in safety or to determine the extent of any such changes;
2. Based on the available evidence, this report makes no judgments about the relative safety of Class E and C airspace; and
3. Obtaining more accurate estimates of the nature and size of changes resulting from the 2b changes would not, by themselves, be adequate to determine the most appropriate course of action to take next.

Appendix A: People interviewed

The list of individuals interviewed for the purposes of this report has been deleted as it would disclose personal information contrary to the Privacy Act 1988