Synthetic Decision Making for General Aviation

Final research report

Prepared for the Civil Aviation Authority

Steve Jarvis
(Human Factors Department, Cranfield University)
January 2007
## Contents

1. Background . . . . . . . . . 3
   1.1 Introduction . . . . . . . . . 3
   1.2 The Models . . . . . . . . . 3

2. Research Method and Procedure . . . . . 4
   2.1 Introduction . . . . . . . . . 4
   2.2 Methodological Considerations . . . . 5
      2.2.1 The Simulated Environment . . . . 5
      2.2.2 Pilot Workload . . . . . 6
      2.2.3 Use of Models . . . . . 7
      2.2.4 Measures . . . . . 7
      2.2.5 Scenario Considerations. . . . 7
   2.3 Method . . . . . . . . . 8
      2.3.1 Experimental Design . . . . 8
      2.3.2 The Participants . . . . . 8
      2.3.3 Ethical Considerations . . . . 9
      2.3.4 The Procedure . . . . . 9
      2.3.5 The Scenario . . . . . 10
      2.3.6 Radio Calls . . . . . 11
      2.3.7 The ‘Divert Window’ . . . . 12
      2.3.8 Post Experiment Procedure . . . . 12
      2.3.9 Pilot Study . . . . . 13

3. Quantitative Data: Results and Analysis . . 14
   3.1 Types of Data . . . . . . . . . 14
   3.2 Data analysis and results . . . . . 14
      3.2.1 Treatment Of Acronym Recall Sheets . . . . 15
      3.2.2 Treatment Of Divert Times . . . . . 18
      3.2.2.1 Times Across The Whole Sample . . . . 18
      3.2.2.2 Comparison Between Groups . . . . 21
      3.2.3 Treatment Of Subjective Ratings . . . . 24
      3.2.4 Treatment Of Wind Direction Recall . . . . 32
   3.3 Summary Of Key Findings . . . . . 34

4. Qualitative Data: Results and Analysis . . . 35
   4.1 Situational Assessment . . . . . . . . . 35
   4.2 Divert Options . . . . . . . . . 38
   4.3 Decision Processes . . . . . . . . . 40
   4.4 Use Of Tools . . . . . . . . . 46

5. Conclusion . . . . . . . . . . 50
   5.1 Conclusion of Analysis and Findings . . . . 50
   5.2 General Conclusion . . . . . . 52
1 Background

1.1 Introduction

This project was implemented by the CAA due to concern over the number of accidents involving controlled Flight Into Terrain, as well as the number of airspace infringements. ‘In flight decision making’ was recognised as being important to these occurrences. In the past few decades attempts have been made to aid pilots’ decision making by the use of acronyms (models) that attempt to prompt the pilot to include all necessary parts of the decision making process. The first part of the project was a review of these models and look at the scientific literature relating to aviation decision making. Very little academic research has been done to validate any of the models. The main part of the project would be to attempt to evaluate the use of two models in a simulated environment. The two models would be chosen by the project advisory group, made up of representatives from the Civil Aviation Authority, Cranfield University and various general aviation organisations.

1.2 The two models (decision making acronyms)

The chosen models (F.O.R-D.E.C and D.E.S.I.D.E) are summarised below:

FOR-DEC:

- **Facts**
  Collect Facts to assess the situation

- **Options**
  Generate response options

- **Risk & benefits**
  Assess the risks & benefits of each of the options)

- **Do not go further until above 3 steps are done**

- **Decision**
  Decide on the best option

- **Execution**
  Choose how to execute the option

- **Check**
  Check to see if the problem has been solved

DESIDE:

- **Detect**
  Detect Change in the situation

- **Estimate**
  Estimate what will happen if you do not intervene

- **Set safe objectives**
  Assess whether original objectives need changing

- **Identify Options**
  Identify all the possible options

- **Do**
  Choose and Do whichever option is best

- **Evaluate**
  Evaluate if the problem is solved or not
2 Research Method and Procedure

2.1 Introduction

The original request by the client was that two decision-making acronyms be decided upon from the sample identified in the literature. The Project Advisory Group was presented with all the acronyms and selected DESIDE and FOR-DEC as the two to undergo testing. Neither of these models had much scientific background and little research was available to establish their validity.

FOR-DEC was favoured primarily due to its two-part approach. (see above). The advisory group liked the fact that a simple precursor (FOR) was required prior to the decision (DEC) in order to try to avoid problems such as satisfising and confirmation bias. DESIDE (the aviation redesign of DECIDE) was felt by the group to be the most memorable and purposeful acronym.
2. 2 Methodological Considerations

Testing of decision making models in the real world is not a popular scientific endeavour and hence no precedent existed in terms of proven methodology.

2.2.1 The Simulated Environment

The primary problem was establishing a test platform that would provide both a valid a realistic experience for participants while maintaining a reliable scientific environment for collection of consistent data. The most valid environment would be a real aircraft, but apart from the prohibitive cost, there would be no way of ensuring that all participants received a similar experience and therefore comparisons between groups would be meaningless. The two major problems would be the difference in environmental factors between flights (different weather conditions, visibility etc) and the different experiences caused by the decisions of individual participants affecting the flights. The first problem could be easily solved by using a flight simulator instead of a real aeroplane, since the environmental variables could be replicated for each participant. However the second problem would be the same. One could start the participants off in the same synthetic environment but, due to the participants controlling the flight, within a very short time all participants would be facing different issues (altitudes, speeds, location relative to terrain and weather, etc).

Although this is what occurs in the real world, for experimental purposes each participant needs to be faced with the same problems in order for a comparison to be valid. The only way to do this was to take the control of the aircraft/simulator away from the participant. Since handling skills were not relevant and aircraft control not in question, this was felt by the team to be a good compromise; giving us more in terms of scientific reliability than we lost in terms of validity. However it was recognised that the situation being experienced by the participants was now some way from the real experience.

The main difference from reality would be the consequences of actions, stress and motivation. These problems exist whenever an experiment is taken out of its real world setting and so would have been the same even if a simulator had been used. A fundamental motivator in GA decision making en-route is a concept popularly referred to as “get-home-itis” or “press-on-it is”. This urge to continue a flight is often expressed by GA pilots, and is understood anecdotally to be the intense motivation to get home or get to the destination caused by the inconvenience, cost and time implications of diverting or turning back. Clearly an experimental environment such as a simulated flight experience would not naturally recreate this feeling. However without the motivation to reach the destination, the decision making would be wholly unrealistic and therefore the experiment invalid.

In order to address this issue a motivational frame had to be created that came as close as possible to mimicking reality within the realms of the project. In reality pilots have an incentive for getting to the destination as opposed to diverting. Therefore the experiment had to put in place this incentive, and so each pilot was told that if they reach the destination they receive an extra £5 payment. The only control the pilot has over the simulated flight is to decide when to abandon the planned flight (divert, go back, make a field landing etc). This was done by stopping a stopwatch at any one point during the flight when the pilot assessed the situation as being too bad to
continue, but safe to divert. However this £5 incentive in itself cannot work since there is nothing to lose, therefore pilots would simply continue on regardless in the hope of getting there, and therefore never stop the watch. In reality there is a balancing motive, pilots do have a lot to lose by continuing when it is clearly unsafe to do so. Therefore a £5 was introduced for abandoning the flight (stopping the watch) if and when the flight was becoming dangerous. Therefore £5 reward would only be given:

1. If the aircraft reached its destination safely (when the pilot choose to continue)
or,
2. If the pilot ‘diverted’ when the things were too dangerous to continue (which would ultimately end in an accident)

These two payments were clearly mutually exclusive, and the pilot could only get one of them depending on what happened during the simulated flight: either it would be a safe flight and the ‘destination £5’ was up for grabs, or it would become unsafe in which case the ‘divert £5’ was up for grabs.

The experiment would therefore need to recreate the feeling of uncertainty that comes with the true situation. The participant pilot would probably guess by the nature of the study that the flight will run into trouble, and so this needed to be addressed. A well known phenomenon for many pilots is a feeling of ‘invulnerability’, which makes the pilot over-confident in his/her assessment of the situation. It is also well-known that in probabilistic environments human beings have a tendency to over-estimate their influence (for example refer to Langer 1974; ‘The illusion of control’, and subsequent studies). These phenomena add up to the fact that the pilot will over-estimate the odds that he/she can safely get to the destination without the need for chance in a real situation, whereas in the experiment they might well underestimate it as stated. Therefore the experiment would have to arrange the pilots’ probability judgement such that he/she feels that there is a greater likelihood of getting there safely than of needing to divert.

For this to be achieved some deception would be required. The solution was to inform the participants that there were 3 scenarios to choose from at random; two that arrive safely at the destination, and one that does not. A participant was asked to physically select the video tape in front of the other participants to give the illusion that there was a blind choice. In reality, results from the two ‘safe’ scenarios would be worth little to the experiment, and so all three videos contained the same scenario.

2.2.2 Pilot Workload

The participants would not be controlling the flight and therefore would simply be sitting at a desk watching. This would give them more time and greater than normal capacity with which to make the decision. Since all groups had the same advantage this was felt to be acceptable. In order to give them some workload, RT calls were made which required answering on paper. Additionally little planning time was given to participants so they would need to work hard to keep up with the navigation. Comments from participants following the experiment vindicated this approach, as many mentioned that it was a stressful and high workload experience.
2.2.3 Use of the ‘Models’

Clearly the participants could not be forced to use the acronyms, apart from the practical issues, if the participants were forced to use it they would likely to guess the nature of the study and so ‘divert’ early. It was important that experimental groups (those trained in the use of the models) should have the same experience as control groups. By training some participants in the use of an acronym prior to the flight, and encouraging them to make use of it, the data collection afterwards could focus on whether it was used, whether it affected the decision and why.

Because each model would require that participants be trained in very basic decision making, it was important to have a second control group that were trained in exactly the same way but without the acronym. This could be used to clarify any difference in variance between the acronym trained groups and the raw control group, in terms of what extent the decision training itself made a difference.

2.2.4 Measures

The primary performance measure would the time when pilot’s made the decision to abandon the flight (if they did so). Other measures would be drawn from post-experimental questionnaires and group de-briefs. The questionnaires would seek to find out about participant’s option generation, situational assessment and decision processes. The group interviews would assess the same issues in a more open environment.

It was also important to know how useful and memorable the acronyms were.

2.2.5 Scenario considerations

The most obvious element of the scenario was that it had to present a realistic situation of deteriorating weather that required an abort decision to be made. The project advisory group suggested that the rather than a linear deterioration, the weather should get worse then get better, before gradually deteriorating into impossible flight conditions. This would mimic natural occurrences and give a more realistic feel to the experimental situation, showing participants that the artificially generated weather is unpredictable like in real life. As well as worsening en-route weather, the terrain would have to rise towards the destination, in order to provide a dangerous situation representative of many CFIT accidents. The client also wanted considerations of nearby controlled airspace, and so the route would need to be constrained in this respect. The scenario would need to be a relatively short flight for practical reasons but have divert options along the route. Therefore a busy area would have to chosen in terms of aerodromes and airspace.
2.3 Method

2.3.1 Experimental Design

A four-group design was chosen for the reasons given in section 2.3 above. The groups were as follows

A. (Control) This group simply participated in the simulated flight
B. (Control Plus) This group watched a lecture about aviation decision making prior to sitting the simulated flight
C. (FOR-DEC) This group watched a lecture about aviation decision making and the use of the FOR-DEC acronym prior to sitting the simulated flight
D. (DESIDE) This group watched a lecture about aviation decision making and the use of the DESIDE acronym prior to sitting the simulated flight

It was estimated that each group required at least fifteen participants to give sufficient statistical power to the experiment. However without having previous or similar experiments to estimate the effect size, this could only be a rough estimate. Therefore the aim was to get twenty people into each group.

The advantage of the simulated environment was that a number of individuals could sit the experiment at the same time, without any adverse effects. It was found to be practical to have up to six participants sitting the experiment, which meant that group interviews could be conducted post-experiment. The individuals sitting the experiment con-currently were generally those in the same experimental group. From a practical point of view, this arrangement meant that it was practical to have the ambitious targets of four groups and up to eighty participants overall.

2.3.2 The participants

The participants had to have a Private Pilot’s License and be reasonably current (one year was felt to be satisfactory, although three months or less was preferred). Clearly any form of instrument rating would lead to different decisions when flying into IMC and so no participants were recruited with IMC or IR.

With so many participants to recruit in such a short time, incentives were required. £10 was given to each participant (over and above the incentive payments described). In addition each participant was offered a session in the Cranfield University Boeing 747 Simulator. Local flying clubs, organisations and individuals were approached and a web page created explaining the research and the incentives and having a participation link. This proved satisfactory and eighty participants were recruited.
2.3.3 Ethical considerations

The research plan was submitted to the University Ethics Committee for approval. Apart from needing clarification about the ‘deception’ regarding the choice of video tape, the plan was fully accepted.

All participants were volunteers and signed a consent form having been told the details (as far as practicable) before they took part in the experiment. There was no possibility of physical harm, and the ethics committee were satisfied that a very low risk existed in terms of psychological harm (self-confidence, esteem etc).

2.3.4 The procedure

Each participant sat at a desk visually shielded from other participants but with a full and unrestricted view of the screen. Participants were briefed upon the nature of the experiment and the use of the stopwatch to abandon the flight plan at any time that they felt. It was stressed that although it was a simulated environment, the participant should imagine that they were flying the aircraft, and make any decisions as if it were their flight. Participants were also instructed to answer radio calls on paper, and make any notes they wished on the log sheets provided.

The incentive payments were fully explained before the mock selection of the video was made. One participant was randomly assigned to make the selection in full view of all the others, and the video was put straight into the machine. Debriefs indicated that participants were completely taken in by this and fully believed that there were indeed two satisfactory flights and only one ‘accident’ flight.

The participants were then read a brief outlining the procedure and giving limited weather information but no wind information. It was noted that no participant requested the wind information. Participants were either given five minutes to look at the map and prepare, or no time at all. A note was made of which group participants were in. The equipment given to participants was as follows:

1. A half-mil chart of the area,
2. Aerodrome plates for Stapleford and Halton (Departure point and arrival destination)
3. A log sheet (blank sheet to write anything they wish down during the flight)
4. A stopwatch (to indicate an abandon flight decision)
5. An RT sheet to record RT responses
6. The main questionnaire to answer afterwards
7. A fake questionnaire to answer if they arrive at Halton
8. A pen
2.3.5 The scenario

The simulated flight was created on a Flight Simulator computer program with photo realistic scenery of the area. It showed the forward view with instrument panel. However, the view panned around inside and outside the cockpit in a manner similar to that a real pilot. The scenery was viewed periodically as were the instruments and the forward view. A sound track was included, with engine noise and RT calls along route.

Figure 2a Forward view of the simulated flight.

The simulated aircraft was a Cessna 172. This was felt to be the most representative aircraft available on the simulation.

The flight was 19 minutes long and was planned to go from Stapleford to Halton. It began with a take off from Stapleford (rwy 10) and a left turn onto a Westerly heading. The aircraft climbs to 2400ft (never going into the London TMA). A radio call is made to initiate a frequency change to Elstree Information. At approximately 6 minutes forward visibility is momentarily lost as cloud base is encountered, and the aircraft descends slightly and the weather clears again. Clouds build gradually and the cloudbase begins to lower approaching Elstree. The flight passes North of Elstree and turns onto a heading of approximately 300, gradually descending to 1500 to stay below cloud. The cloud gets darker and showers are visible. Cloud begins to form below the aircraft and rain increases. By 16 minutes, the cloud is dark and overcast, there is rain and significant cloud below and around the aircraft, forward visibility is severely compromised. The cloud continues to build and lower, rain increases and forward visibility is lost altogether. In the eighteenth minute the aircraft is flying
below the highest terrain with no visibility. Control is subtly lost, the AI topples and the aircraft crashes.

The wind direction during the flight was 170/15 knots.

2.3.6 Radio Calls

RT calls were regular during the flight. The RT ‘story’ was planned out to be as realistic as possible. Different ‘actors’ were used for different parts. The RT recordings were fed through a telephone line in order to cut off high and low frequencies in order to give realistic sounding transmissions.

The first calls were heard on Stapleford’s frequency, then there was a change to Elstree Information. One aircraft was following about 10 miles behind the experiment flight and was heard receiving FIS from Elstree. Much of the radio activity was circuit traffic at Elstree. There were several clues to the situation in these calls. Firstly the wind direction, strengths and gusts at Elstree, second the fact that Elstree was very busy with aircraft attempting to land, rather than depart or over-fly. This information could have been valuable to the participant, since the wind was not given to them. The participants had to respond in writing to any RT call for their aircraft.
2.3.7 The ‘divert window’

The video tape was watched by a panel of general aviation experts made up from the advisory panel. Each was asked to give the time that they felt was the most sensible to divert. It was agreed that before 12 minutes 30 seconds was premature, but after 16 minutes was too late. Therefore this was used as the divert window. Pilot participants who stopped their watch between 12:30 and 16:00 were awarded the £5 for a sensible divert.

2.3.8 Post experiment procedure.

All participants were asked to fill in a post-experimental questionnaire on stopping the watch, crashing, or reaching Halton. Two questionnaires were provided, one for participants who stop the watch and the other being the fake questionnaire for participants who did not stop the watch and arrived safely at Halton. This was provided to further convince participants of the possibility of reaching the destination safely. All participants filled in the real questionnaire. No participant was allowed to leave the room after finishing the questionnaire, until the rest of the post experimental procedure was completed. The questionnaire asked the participant to write words and sentences about a number of issues in the flight such relating to

- What factors were being assessed in decision making
- What decision making strategy was being used
- What options were being considered
- Why the ‘divert decision’ was made when it was

Additionally a section was included to assess (quantitatively) using scales from 1 – 9, where the sources of concern were coming from that eventually led to the divert decision. This will be referred to as the subjective quantitative section of this questionnaire.

When all experimental participants in the session had completed the questionnaire, they were tested on their memory of the decision making acronym. No aids were allowed. The participants were given 90 seconds only to recall the acronym and what each letter stood for (on paper).

The participants were then asked to write down the wind speed and direction of the flight in whatever form they felt was the easiest. Wind information had not been communicated to them directly, no information regarding wind vectors was given prior to the flight, the only information during the flight came from radio calls and cues such as ground speed. It had been communicated a number of times by Elstree information. This was a further test of the participants situation awareness and how well they were gathering important information from the GA environment.

Finally the participants in each session were invited to sit round a table to take part in a group interview about the flight. This was a loosely scheduled interview that intended to further tap those elements asked about in the initial questionnaire such as option generation, reasons for continuing, situation assessment etc.
2.3.9 Pilot Study

The full experimental procedure was tested by a pilot study. The participants in the pilot study were asked to comment on the experiment itself, specifically the suitability of the scenario, its clarity and realism, the workload, the radio traffic, and the motives and incentives. The results were extremely good; none of the participants guessed the true nature of the ‘perceived choice’, and all were convinced of the 3:1 probability. Additionally each commented favourably on the realism of the flight and the incentives.
3. Quantitative Data: Results and Analysis

3.1 Types of data

This experiment generated a vast quantity of data including:

1. Acronym recall sheets
2. Diversion times (minutes and seconds)
3. Subjective ratings in Questionnaire 1 (question 7)
4. Wind direction recall
5. Written questionnaire responses per participant
6. Group Interview transcripts

The first four collections of data above were all treated as quantitative data. Either the data was numerical in nature, or it was converted to be so, so that statistical tests could be performed and conclusions drawn.

3.2 Data analysis and results

The distribution of all quantitative data was looked at and judgements made as to its normality.

There were a total of 70 participants in the research, with a mean age of 39.5 years and a standard deviation of 12.6. The distribution of ages can be seen below (Figure 3a).

![Figure 3a. Distribution of ages](image)

Image shows a histogram with a normal curve overlay, indicating the distribution of ages with a mean of 39.5 and standard deviation of 12.6.
Not all participants attended the sessions and so there was a slight unevenness among the four groups. The least participants belonged to the control group (A). This would not harm the comparisons but would obviously decrease the power of the statistical tests. However this had been predicted to a degree and the groups had been made larger than needed (twenty participants were invited), therefore the effect was minimal.

3.2.1 Treatment of acronym recall sheets

Although not technically quantitative data, the recall of the acronyms could be usefully analysed as such after initial treatment. The recall of the Acronym was a free recall test within one hour of the end of training in the use of the acronym. The only hint the participants were given was that there were 6 letters; otherwise they simply had 90 seconds and a blank sheet of paper. Treatment of this data was performed in order to enable comparisons between the two experimental groups (C and D). The following three scores were generated:

1. The number of correctly remembered letters was noted for each participant (maximum score 6).
2. A score out of 6 was awarded for the order of correct letters (1 mark for each correctly positioned letter, the maximum could only be achieved with 6 letters, and some judgement was required. This was always discussed with team members to be sure that objective treatment was applied).
3. A ‘content score’ was awarded, out of 12. The content was the explanation provided for each letter of the acronym, generally one point was given for a correct word (eg. F – Facts) and a further point for any greater expansion.

Thus 12 points were awarded for the acronym, and 12 for the words or explanations.

The three scores were added together, to give a total out of 24 for acronym recall.
Figure 3b. Tables showing acronym recall for the two experimental groups. The left column shows what the participants recalled, the next column is a score for the acronym recall and the next a score for the recall of content. The far right hand column of each table is simply a total score made by adding together these two scores. DESIDE appears to have an advantage over FOR-DEC in terms of recall, and an independent samples t-test (Figure 3b1) shows that the result tends towards significance (0.077).

<table>
<thead>
<tr>
<th>Acronym as recalled</th>
<th>Recall Score</th>
<th>Content score</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>PRESIDE</td>
<td>8</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>ODC</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>FARDER</td>
<td>8</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>OD</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>FORDEC</td>
<td>12</td>
<td>8</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acronym as recalled</th>
<th>Recall Score</th>
<th>Content score</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>DESIDE</td>
<td>10</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>DESIDE</td>
<td>12</td>
<td>6</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.33082</td>
<td>1.28186</td>
<td>[-4.02373, 26139]</td>
</tr>
</tbody>
</table>

Figure 3b1

Interestingly, this result is based on a two-tailed test. By that it is meant that the premise of the test is that there is no reason to assume that DESIDE is more memorable than FOR-DEC. However when the single reason that the advisory group picked the DESIDE acronym is recalled there is absolute justification to use a one tailed test (in favour of DESIDE) because they hypothesised it would be more easily recalled than other acronyms. When a one-tailed test is applied, the result is now
comfortably significant (p<0.05). It can therefore be safely hypothesised that DESIDE is more memorable than FOR-DEC, in terms of straight acronym recall.

However Figure 3.1 also shows that the DESIDE acronym has a lower score for content recall (4.17 as opposed to 4.78 for FOR-DEC). The latter result suggests that FOR-DEC’s content may be better recalled despite the fact that the acronym itself was less well recalled. This is an unexpected result since recalling the acronym is the first step to recalling the content, and so the FOR-DEC group had a big disadvantage in this test, meaning that if the content recall were similar for both acronyms, one would expect DESIDE to achieve higher score than FOR-DEC (about 20% higher based on the acronym recall result). Thus to achieve a higher score is most unexpected.

To address this issue, and be absolutely objective between content recall for the two groups, it is worth looking at the same tables but without the scores for participants who could not remember the whole acronym. In this way a totally objective content total will be achieved, because, as stated.

In figure 3c above, participants’ scores have been removed unless the whole acronym was remembered (hence the means of 12.00 at the bottom of the acronym recall columns). This gives mean totals for the content recall after the acronym has been successfully recalled (bottom middle score in red). Using this more objective test, the two ‘content recall’ means diverge greatly. It appears that the content (meaning) of the DESIDE acronym can be seen to be harder to recall than the FORDEC acronym.
Additionally, the overall acronym and content means (bottom right corner of tables) have increased and diverged. The increase is due solely to the removal of the poor scores, and the divergence is due to the greater increase in the acronym recall score for the FOR-DEC group over the DESIDE group (the latter had the better straight recall score).

<table>
<thead>
<tr>
<th>Recall</th>
<th>Levene’s Test for Equality of Variances</th>
<th>Hotelling’s Test for Equality of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>Sig</td>
</tr>
<tr>
<td>recall2</td>
<td>2.623</td>
<td>.124</td>
</tr>
<tr>
<td>recall2</td>
<td>2.156</td>
<td>.10</td>
</tr>
</tbody>
</table>

Figure 3d. Independent samples t-Test between groups C and D for the content of the Acronym. Since equal variances could not be assumed (Levene’s F), the result shows a significant difference (p = 040) using a two-tailed test (necessary in this instance).

In summary there is strong evidence that whereas DESIDE is the more easily remembered of the acronyms (as assumed by the advisory group), the content of the FOR-DEC acronym is easier to remember. Both these findings yielded significant results at the p<0.05 level.

3.2.2 Treatment of divert times

3.2.2.1 Divert times across the whole sample

‘Divert times’ are the times that appeared on the stopwatches of the participants after the experiment. Therefore they represent the time that elapsed from the start of the flight until the participant decided that, if they were flying this for real, they would have abandoned the flight in terms of getting to Halton, and looked to divert, turn back or otherwise change the plan. Thus the divert time is a time in minutes and seconds between 0 and 19 minutes. In reality participants had no choice but to accept the flight to begin with, but after take off they could have stopped the watch at any time.

The divert times themselves form a negatively skewed distribution (figure 3e below). This indicates that overall participants were more likely to abandon the flight plan (AFP) later as opposed to earlier.
Figure 3e: Divert times showing a negatively skewed distribution curve.

Figure 3f: Each bar shows the number of participants choosing to divert during that minute. All 19 minutes of the flight are shown.
Figure 3f (above) shows that there was a common point at around 6 minutes where about 10% of participants stopped the watch. This corresponds with the point in the video, requested by the advisory group, where the forward visibility reduces prior to a small descent and subsequent clearance.

5 of the 6 participants in the ‘12 minute’ bar, stopped the watch after 12:30, i.e. they were within the divert window. Figure 3f therefore shows that after the 6 minute blip, the next common divert time corresponds to the first minute and a half of the 3 ½ minute divert window. There is then a lull within the last two minutes of the window. The biggest common divert time occurs just after the end of the divert window, after 16 minutes have elapsed (i.e. within the 16 minute bar). This suggests the anecdotal concept of ‘press-on-it is’!

Figure 3g (below), shows that the percentage of participants diverting during the divert window, and therefore winning £5, was only 26%. Over half (54%) of all participants waited until after the safe divert window had passed before stopping the watch, many of them in the very last minute of the video.

![Percentage diverts in the divert window](image.png)

Figure 3g. The percentage of participants in the three ‘divert categories’ (i.e. who diverted before, during and after the divert window)

Some analysis (particularly from the qualitative part of the study) uses these three groups rather (and as well as) the four experiment groups, to analysis and slice the data. These three groups (containing all 70 participants between them) are referred to as the ‘Pre-window’, ‘In-window’, and ‘Post-Window’ groups. These groups are independent from the definitions of groups A, B, C and D, and all contain participants from each of those groups. Overall these new groups will be known as the ‘divert groups’, or ‘divert time groups’. Figure 3h (below) shows how this ‘divert grouping’ and the experimental grouping compare.
3.2.2.2 Comparison of divert-times between experimental groups A, B, C and D.

Figure 3h. Percentage of participants in the three ‘divert categories’ (pre, in, post window) by experimental group A, B, C and D.

Figure 3h above shows that there is no obvious divert pattern unique to any one experimental group. Taking the pre-window and post-window segments as being ‘safe’ divert times, only group B had more safe decisions than unsafe ones. However the numbers are not statistically significant.

Before directly comparing the divert times between groups, data transformation had to be performed in order to compensate for the negatively skewed group distributions. Once this was done, a one-way ANOVA and post-hoc tests could be performed on the data.
Figure 3i Tukey post hoc analysis of bonus payment by group

To compare the number of participants diverting inside the divert window by group (A, B, C, D), the number of bonus payments per group were compared. Figure 3i shows that no significant differences exist in terms of number of bonus payments awarded. This showed that no group was better than any other in terms of diverting inside the window.

An ANOVA was used to make a straightforward comparison of divert times by experimental group, the results are shown in Figure 3j.

**Divert time**

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>.462</td>
<td>3</td>
<td>.154</td>
<td>.760</td>
<td>.521</td>
</tr>
<tr>
<td>Within Groups</td>
<td>13.381</td>
<td>66</td>
<td>.203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>13.843</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Multiple Comparisons**

<table>
<thead>
<tr>
<th>(I) Trial Group</th>
<th>(J) Trial Group</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>2.38779</td>
<td>1.48705</td>
<td>.383</td>
<td>-1.5317</td>
<td>6.3072</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>.69317</td>
<td>1.46886</td>
<td>.972</td>
<td>-3.2350</td>
<td>4.5133</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>1.73337</td>
<td>1.40450</td>
<td>.608</td>
<td>-1.9685</td>
<td>5.4352</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>-2.38779</td>
<td>1.48705</td>
<td>.383</td>
<td>-6.3072</td>
<td>1.5317</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>-1.74863</td>
<td>1.33388</td>
<td>.559</td>
<td>-5.2643</td>
<td>1.7671</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-.65442</td>
<td>1.26149</td>
<td>.954</td>
<td>-3.9794</td>
<td>2.6705</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>-.63917</td>
<td>1.46986</td>
<td>.972</td>
<td>-4.5133</td>
<td>3.2350</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>1.74863</td>
<td>1.33388</td>
<td>.559</td>
<td>-1.7671</td>
<td>5.2643</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>1.09420</td>
<td>1.24118</td>
<td>.814</td>
<td>-2.1772</td>
<td>4.3656</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>A</td>
<td>-1.73337</td>
<td>1.40450</td>
<td>.608</td>
<td>-5.4352</td>
<td>1.9685</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.65442</td>
<td>1.26149</td>
<td>.954</td>
<td>-2.6705</td>
<td>3.9794</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>-.109420</td>
<td>1.24118</td>
<td>.814</td>
<td>-4.3656</td>
<td>2.1772</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3j. ANOVA and Post-Hoc tests for comparison of divert time between groups.
The one-way ANOVA and Post-hoc results (figure 3j) show that there was no significant difference in divert times between the different groups. This result does not necessarily mean that the models and training were of no use. It is possible that the necessary compromise in ecological validity created too much experimental ‘noise’ that masked any real effect. However without this ecological validity the comparison could not have been made. It is possible that although in this small sample and with these experimental conditions the outcome was not greatly affected, there is still some effect that could increase safety. So many factors weight the decision to stop (divert) that the interaction with all the unknown factors (such as personality, risk utility, previous weather encounters or interpretation of the weather) could have diluted the result.

Figure 3k. The mean divert time by group.

Figure 3k (above) gives a visual representation of the mean divert times in each group
3.2.3 Treatment of Subjective ratings (source of pilot concern)

The quantitative rating scales in the post-experimental questionnaire attempted to quantify sources of concern. The participants rated seven factors on a 1 – 9 scale, 9 being highest concern and 1 being no concern. The factors that participants were asked to rate were:

- Fuel Situation
- Controlled Airspace
- Aircraft Systems
- Weather
- Terrain Clearance
- Traffic
- Radio Calls

After initial data treatment to establish distribution and variance equality between groups, data transformations were performed (square root transformations). This left comparable data sets on which one-way ANOVAs were performed between the four groups for each variable.

Histograms of the means for each group, for each variable, are shown below.

![Fuel Concern](image)

Figure 3L Mean ratings for concern about the fuel state, by group
Figure 3m Mean ratings for concern about controlled airspace, by group

Figure 3n Mean ratings for concern about aircraft systems, by group
Figure 3o Mean ratings for concern about the weather, by group

Figure 3p Mean ratings for concern about the terrain, by group
Figure 3q Mean ratings for concern about other traffic, by group

Figure 3r Mean ratings for concern from radio calls, by group
Figure 3s Mean ratings for all sources of concern, by group
Figure 3t. Single plot of all mean ratings for sources of concern, for four groups

Clearly ‘fuel’ and ‘aircraft systems’ proved to be the least source of concern for the participants, which was to be expected given the brief and the nature of the task. Unsurprisingly weather and terrain were the sources of most concern to the pilots. At a very basic level these findings give a positive indication of the validity of the quantitative part of the questionnaire.

Looking at the results between groups gives less of a clear picture. The ANOVAs show that few significant differences exist. Despite the fact that patterns appear to exist indicating that the concern increases from control groups up to the DESIDE group, Post-Hoc tests show the only significant differences to be the DESIDE group against the control groups (A and B) for ‘Controlled Airspace’ and the Experimental groups (C and D) with the control group (A) for ‘Radio calls’.

Figure 3u (below) shows a Tukey post hoc SPSS output for the concern over ‘Controlled airspace’. It can be seen that the difference between group A (control) and group D (DESIDE) is significant to the 0.001 level (highly significant), whereas the difference between groups B and D are only significant at the 0.05 level.

Figure 3v (below) shows a Tukey post hoc SPSS output for the concern over ‘Radio Calls’. It can be seen that the difference between group A (control) and group D (DESIDE) is significant to the 0.001 level (highly significant), whereas the difference between groups A and C are only significant at the 0.05 level.
These results do not necessarily mean that evidence exists to suggest that somehow more consideration was given to the radio calls and the controlled airspace by the DESIDE group than the control group. The ‘radio calls’ finding is further supported by the results from the recall of wind vectors (section 3.2.4). Together these findings suggest that the DESIDE group were taking in the information given over the radio more than the other groups and significantly more than the control group.
Figure 3w. The mean scores for sources of concern, split into the three divert-time
groups: those that diverted before the divert window (pre 12:30), those that were
within the window (12:30 – 16:00) and those that diverted too late (16:00 +). This
demonstrates that there were no overall differences reported as to the sources of
concern, by those diverting at different times during these different periods. Statistical
tests show no significant differences.

These results appear to show very little and the relevance could easily be missed. The
result illustrated in Figure 3w is in fact more interesting than it would have been had
significant results been achieved. One must consider what would be expected. Those
in the blue group (fig 3w) were only flying in non-threatening conditions (up to
12:30), in a safe airspace environment, with non-threatening weather and cloud base,
and with no terrain issues. Therefore the environment they were flying in was far less
concerning than the yellow group (who were making their critical decisions when the
aircraft was in a highly dangerous situation with regards to weather and terrain, as
well as being difficult to orientate with airspace). However the results show that the
yellow participants, despite being in disastrous conditions, rated those conditions as
concerning as the blue group’s benign conditions!

This suggests that the yellow group (post window divert group) were not necessarily
taking greater risks, but were making a quite different situational assessment to the
blue and possibly red groups. The qualitative results for situational assessment
support this (as discussed later). This provides some evidence that situational
assessment is a key driver in the decision process, and therefore greater training in the
assessment of situational elements such as weather recognition may well improve
decision-making.
3.2.4 Treatment of wind direction recall

Participants were asked to estimate wind speed and direction after the flight. These two pieces of data were looked at for each of the four groups. Raw data needed treating to become interval data, since the actual compass direction of the wind would not provide parametric data for comparative testing. Hence the absolute deviation in degrees from the actual direction was calculated for each participant. This gave a total ‘error’ score for each participant from which a mean and standard deviation could be calculated for comparison with other groups.

Many participants felt unable to provide an estimate at all. These blanks are filled grey in the results table and show that the control groups had more than the experimental groups. This interesting finding cannot be submitted to statistical tests since it is not reliable data, being as it was not controlled within the experiment. The statistics for each of the four groups were calculated without taking these missing values into account, since giving a zero score would be inappropriate bearing in mind that the scores are deviations, additionally adding the missing participants to the value of N would bias the mean in favour of that group, wrongly lowering the overall error for each participant who could not give an estimate. This would have bias the results in the opposite direction to the real effect.

<table>
<thead>
<tr>
<th>A - Control</th>
<th>B - Control +</th>
<th>C - FOR-DIE</th>
<th>D - DESIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation (Wind Speed)</td>
<td>Deviation (Wind Direction)</td>
<td>Deviation (Wind Speed)</td>
<td>Deviation (Wind Direction)</td>
</tr>
<tr>
<td>6.0</td>
<td>104.4</td>
<td>6.7</td>
<td>101.7</td>
</tr>
<tr>
<td>5.0</td>
<td>108.8</td>
<td>4.1</td>
<td>54.7</td>
</tr>
</tbody>
</table>

Figure 3x. Results for wind speed and direction recall, for all groups.
As can be seen from Figure 3x, more people in the experimental groups (C and D) felt able to provide an estimate of wind direction (missing values filled grey). This cannot be treated as a result, but it does have implications for data analysis because the effectively lower sample size of n=8 in group ‘A’ results in less statistical power, and therefore a much greater effect size will be needed to find statistical significance. Leaving out these values gave the means as reported along the bottom row. As can be seen, the DESIDE group appear to have made better estimates of wind direction than the other groups, having about half the mean deviation (error). This is shown clearly in Figure 3y below.

![Figure 3y](image)

Figure 3y. The absolute mean deviation of wind direction estimates, by group. As can be seen, group D is clearly different from the others.

However this initial glance at the means must be shown to be valid when statistical tests are applied. A one-way ANOVA yields a result significant only between groups C and D (see figure 3z below).

<table>
<thead>
<tr>
<th>(I) Group</th>
<th>(J) Group</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>-2.708</td>
<td>23.790</td>
<td>.999</td>
<td>-66.12</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>-1.779</td>
<td>23.422</td>
<td>1.000</td>
<td>-64.21</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>49.669</td>
<td>22.347</td>
<td>.132</td>
<td>-9.90</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>4.487</td>
<td>20.866</td>
<td>.996</td>
<td>-5.42</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>-4.487</td>
<td>20.866</td>
<td>.996</td>
<td>-5.42</td>
</tr>
<tr>
<td>B</td>
<td>D</td>
<td>46.961</td>
<td>19.652</td>
<td>.093</td>
<td>-5.42</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>1.779</td>
<td>23.422</td>
<td>1.000</td>
<td>-51.13</td>
</tr>
<tr>
<td>C</td>
<td>B</td>
<td>4.487</td>
<td>20.866</td>
<td>.996</td>
<td>-51.13</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>51.448(*)</td>
<td>19.204</td>
<td>.048</td>
<td>.26</td>
</tr>
<tr>
<td>D</td>
<td>B</td>
<td>-46.961</td>
<td>19.652</td>
<td>.093</td>
<td>-99.34</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>-51.448(*)</td>
<td>19.204</td>
<td>.048</td>
<td>-102.64</td>
</tr>
</tbody>
</table>

Figure 3z. Post Hoc analysis of ANOVA result comparing groups for wind direction recall.
We can be fairly confident that the real picture is statistically stronger than this because the sample sizes of ‘groups A, B and C were much less than that of D because fewer members of group D felt unable to even take a guess. If we therefore take each of these cases in groups A, B and C and give them a score equal to their mean (generous since they did not know the wind direction, and valid since the reason for the missing values was a lack of participant’s knowledge of the wind direction, which is the very variable being tested) then we get a significant result for group D when compared to all other groups.

3.3 Summary of key findings

The quantitative data has produced some interesting results despite the fact that few significant results were found.

1. Groups trained in the use of a decision aid (either FOR-DEC or DESIDE) were significantly more likely to report the radio calls as a source of concern than control groups.
2. Groups trained in the use of DESIDE were significantly more likely to report concern over controlled airspace than either of the control groups.
3. Groups trained in the use of DESIDE were probably more likely to have correctly assessed the situation with regards to wind speed and direction, than the FOR-DEC group, or the control groups.
4. Acronym recall tests showed that DESIDE was significantly more likely to be remembered than FOR-DEC (when a one-tailed test was used). However it appears to be a more difficult model to remember in terms of content than FOR-DEC (test significant to p<0.05).
5. There was no significant difference in the ‘level of concern’ subjective rating scales between participants who had diverted too-early, on-time, or too late.
4. Qualitative Data: Results and Analysis

The recorded interview data were transcribed and content analysed along with some of the qualitative questionnaire data. Unlike the experimental data, no quantitative analysis can be performed on this level of data. This analysis was used to help explain and support the main experimental findings and to uncover concepts and ideas outside the experimental hypothesis. The content analysis performed was assisted by Nvivo data analysis software. Statements were grouped by meaning into categories and each categories further categorised until meaningful categories emerged. This form of analysis (often called template analysis) provides numerous conceptual categories at the bottom level, some containing just a few statements from the original transcripts, others having many statements.

Data coding

The data were coded using the following top-level categories:

i. Situational Assessment
ii. Options
iii. Decision Process
iv. Use of Tool

This categorisation was derived from decision making theory, and basically splits the decision down into components.

Analysis found that the richest categories in terms of understanding and explaining the findings and the general issues were ‘Decision Process’ and ‘Use of Tool’. The other categories are shown in this report for completeness, but are less instructive than those two.

4.1 Category i: Situational Assessment.

Situational assessment is often reported in both scientific and aviation literature as being an important skill for a pilot, essential for good decision making. Additionally the quantitative analysis showed that situational assessment was clearly an important concept to be considered, especially in view of the lack of differences between the divert-time groups when reporting their concern over various aspects of the situation (3.2.3). Therefore it was essential that all comments from the transcripts relating to situational assessment be grouped in a top level category (and then decomposed into sub-categories). This was done in order to find out more about what assessments were being made and whether these were substantially different for those that diverted late (as opposed to those who made safe decisions). Also the differences between control groups and experimental groups could be looked at along with the specific acronym groups.

The explanatory power of comments in this category was generally low, participants made statements about their assessments of the situation but few were able to analyse their assessments beyond simple features that they had noticed. The experimental method would need changing and a more in-depth method of interview used if this area were to be focused on further.
A bottom-up approach was taken to the categorisation of the situation assessment data. Seven 2nd level categories emerged but differences between experimental groups or divert times could not be found. Thus, participants that diverted too late were not assessing different elements to those that did not; they were looking at the same issues in general (second level categories - weather, airspace, terrain, radio calls etc).

This backs-up the hypothesis made earlier that situational assessment might be a key issue in decision-making (see end of section 3.2.3).

The categorisation of ‘situational assessments’ is as follows:

**Top level Category: i Situational Assessments, and an example quote from each.**

i.A. Airspace
Any comment that showed that the participant was making an assessment of the airspace around which they were flying.
Example: “I didn't consider diversion too much to either side, it had to be behind me from where I'd just come from because of the airspace at Heathrow and also Luton airspace which we were heading for I think was 3000, 2500 so the north was a practical side because it was underpopulated and less airspace restriction”

i.B. Navigation
Comments about how the participant was navigating.
Example: “fairly new that I was on the way to Elstree because I kept a check on the heading and the time with the airspeed obviously you can work out”

i.C. Terrain and Obstacles
Comments that showed the participant was concerned about terrain, built-up areas or obstacles such as masts.
Example: “You were also dangerously near some masts in that area when it dropped down to about 1200 feet, the aeroplane... some masts at 300 feet in that area.”

i.D. Weather
Any comment relating to assessment of the weather. Comments relating to rain, wind, and cloud/visibility were separated into three categories (below):

i.D1 Cloud (and visibility)
Example “low cloud, we were flying just below the cloud but there were other patches of cloud that were below us”

i.D2 Wind
Example “The main factor from me was that the wind was from behind me, or I perceived it as such”

i.D3 Rain
Example “Just the precipitation to start off with as well, there was a section where only about 8 minutes into the flight we hit the first rain cloud”
i.E. Radio
Any comment regarding information assessed from the radio calls was put into this category.
Example: “Lots of stuff on the radio, other traffic going round Elstree and he other aeroplane behind us going in the same direction which was a bit of a worry turning back”

i.F. Situation Behind
This category was set up to contain a number of comments that specifically related to the assessments being made of the situation (particularly weather) behind the aircraft.
Example: “we were fine but it seems the route, the route seems to be closing in so, if we’d turned around the wedge would have opened”

i.G. Altimeter
Any comments regarding assessments being made based on the altimeter were coded into this category.
Example: “the altimeter wasn’t you know, it got up to about 2000 feet at one point then suddenly it dropped down ‘again”

Figure 4a. Diagram of the categories coded in ‘situational assessment’
4.2 Category ii: Divert Options

This category was set up in order to assess what options were being considered by participants, being as that is a major part of decision making. Groups and divert time groups could also be looked at across this coding to assess whether or not differences could be said to exist that could help explain decision making behaviour.

The categorisation of ‘Divert Options’ is as follows:

**Top level Category: ii Divert Options, and an example quote from each.**

ii.A. Stapleford,    ii,B. Elstree,    ii,C. Wycombe

These three categories contained comments relating to options to divert to Stapleford (turn back, go home), Elstree, and Wycombe respectively.
Example: “Well, I'd just turned past Elstree when I made my decision so I would have probably gone into Elstree”

ii.D. Other Destination.
Comments regarding divert into destinations other than those in categories iiA, iiB or iiC, were put into this category. These destinations were Luton, Panshanger & Denham, but were seldom commented on.
Example: “I was actually going to try and see if I could get into Luton”

ii.E. Emergency Field Landing
Any comment or statement regarding the option of landing in a field was put into this category.
Example: “I think I'd actually try for an emergency descent through the clouds, after then, crash landing”

ii.F. Make radio call
Although not a divert option, this category was included because some participants stated that they would have wished to make a radio call prior to deciding on an option.
“I would have contacted Elstree as regards to their weather...”
& “I certainly would have told someone straight away. I would have talked to, I would have been talking to at this stage, I would have been talking to either Stapleford or Elstree”
Figure 4b. The Divert Options categorisation

This category was the least instructive of the four top-level categories. Apart from being small, it was as expected in terms of options by divert time, and showed no real differences by experimental group. The ‘Post-window’ divert group (those that diverted very early) had many more references to turning back and going to Stapleford than the later diverts (the post-window group had very few). This is as expected and at least shows that participants were making generally sensible divert options. The biggest category was ‘Elstree’, which was again unsurprising since the weather was getting bad when the simulated flight was approximately North abeam Elstree and so option consideration would tend to point to Elstree, especially since the participant was listening to Elstree information. There was no real evidence of increased option generation by those that had been trained in the use of tools, or even those that had had the training, over the relevant control groups.
4.3 Category iii: Decision Process

This category contains all comments that explained why participants decided to continue flying, or decided not to. Thus it is split into two mutually exclusive categories called ‘drivers to continuation’ and ‘barriers to continuation’. The former includes any statement that gives an indication of what factors weighted the participant’s decisions to continue, whereas the latter contain all quotes that suggest factors that made participants decide to stop. An explanation and example quote from each of the bottom level categories is given. These two categories could account for the great majority of comments, but not quite all. Therefore an additional category called ‘Strategy’ was set up to contain any other comments about how the decision was being made.

Top level Category: iv Decision Processes, and an example quote from each.

iii.A. Drivers to continuation

iii.A1. Optimistic weather assessment
Any comment that showed that the participant was making a probability judgement that the weather would not deteriorate any further.
Example: “the cloudbase was changing was going down and up, was going down and up all, well not all the time but a few times and I thought maybe if I continue it will be the same”

iii.A2. Gaming (trying to get through, just following a possibility)
This category contains comments that suggest that the participant was focussing on the possibility of getting further, rather than being concerned with the safety of the situation.
Example: “it was getting lower and you know my first instinct was oh there's a gap”

iii.A3. Inflexibility
A comment that shows that the pilot has made up his/her mind and that is what determines there behaviour (this was a very small category)
Example: “I was just fixed on going into Elstree”

iii.A4. No better option
When the participant stated that they continued because an alternative was no better/safer.
Example: “there was no point in turning back because I'd have to go through the rain cloud again so just keep ploughing on”

iii.A5. Safety net
A relatively large category. This contains comments that show that participants were happy continuing, even into poor weather, because they had a safety net (turn back, divert, etc) in reserve.
Example: “I think I was happy to carry on as long as I did because I knew it was clear behind and it was low ground behind, I could turn and go back”
iii. A6. Not yet hit a barrier
Similar to ‘Safety net’, this contained comments suggesting that participants were happy to continue until one or other criteria was fulfilled (MSA etc), at which point they would turn back.
Example: “I just flew until I thought well that's the limit that gave me the time to get back to diversion.”

iii. A7. Felt Safe to continue because.....

A7a. Slow deterioration
Comments that the participant felt safe continuing because the weather deterioration was very gradually (and couldn’t therefore catch them out)
Example: “I felt that the rate of deterioration was slow enough for me to make that decision later rather than earlier and I think”

iii. A7b. I know the area
Comments that the participant felt safe continuing because they knew the local area.
Example: “I think the fact that we know the area made the decision as late”

iii. A8. Cannot explain
One participant stated that he could not explain why he continued
Example: “It's difficult, to explain why I didn't divert earlier”

iii. B. Barriers to continuation

iii. B1. Tight Barriers

iii. B1a. No Forward Visibility
Comments that showed that when participants could not see ahead, they stopped.
Example: “The fact that I couldn't see was my strategy. I couldn't see in front of me and that was it”

iii. B1b. Bust pre-determined MSA
Comments showing that participants had a ‘pre-determined’ safety altitude, that they used as a go-no go indicator.
Example: “I set myself my own MSA as I said .... 'beyond that I'm going to turn around'. It (his MSA) was just above 1000 feet”

iii. B1c. Very first sign of poor weather
Comments showing participants who stopped (returned home) at the very first sign of poor weather.
Example: “for me, if it starts to get claggy and it's starting to come down then I will turn back and pop back to base”
iii. B2. Loose barriers

Any comment suggesting that participants decided to abandon the flight plan when they were no longer sure of their position.
Example: “I really don't know where I am so I'll bin it.”

iii. B2b. Weather and terrain
General comment suggesting that participants decided to abandon the flight plan due to weather and terrain, although no definition or specifics were given in these comments.
Example: “the large factor was weather and then terrain clearance”

iii. B2c. Out of my depth
Some participants said they stopped due to feeling that they could no longer cope. This included workload, experience and personal limitations.
Example: “I felt I've got too much workload so after about 5 or 6 minutes, I decided enough is enough, I'd just turn around and go back.”

iii. B2d. Gut feeling, stress
General comment suggesting that participants were using on an undefined ‘gut’ feeling to tell them when to stop.
Example: “Well you know, self-preservation and sweaty palms I suppose in reality, inducing a desire to turn back and find somewhere safe”

iii. C. Strategy

‘Strategy’ contained two other types of comments about decision processes that could not easily be placed in the above two categories. These were, firstly, comments relating to the use and evaluation of options in the decision and secondly, comments referring to using no strategy at all.

iii. C.a Option Consideration
This contained all comments relating to the process of options generation or evaluation that did not relate specifically to one of the DM tools. It contained five comments. Comments of this type relating to DESIDE or FOR-DEC were grouped under the ‘Use of Tool’ Node. All five comments are placed below, due to the variety and utility of the comments.

Example 1: “tried to consider all the options in this complete stack and then pick the one from it because I had, several times where I had two or three options all open to me and I kept reviewing them all the time so I suppose I was going through that process but not using an acronym to actually do it”

Example 2: “what I did try and do maybe slightly differently because of what you said was this idea of working with options in a pile rather than options in a line”
Example 3: “I actually went through the three options and picked the safest and easiest...The plan was carry on and go lower which was the first one, which was ??? Stapleford the weather was bad, you know there might have been a time where I'd go on and push but I knew where I was, it was safe and I knew I could get in, take that option.”

Example 4: “List of the general options, the likely options, I could continue, divert or, re-route and just reassess things all the time which of them is still an option it's at the point”

Example 5: “What I found useful was in the decision making process you said about, it's like wiring isn't it, you said parallel or series, I'm very much of the, the person that thinks yes well that's it I'll do that one, tick without actually knowing all the others but where as in parallel yeah I could do that, that, that, that was quite useful for me.”

iii.C.b Non-strategy.
This contained comments suggesting that participants were using no strategy to make decisions
Example: “I didn't use any strategy as such”

---

Figure 4c. Categories and relations for category iii ‘Decision Process’
Those comments contained in the ‘Decision Process’ category were looked at to see if any obvious differences existed between the four groups in the experiment or between the three types of divert group (those that diverted prior to the divert window, those that diverted inside it, and those that left it too late (post window)). No obvious differences existed, and additionally for a number of categories the amount of comments was too small to show any pattern.

The decision processes category data does hint at some interesting possibilities in terms of in-flight decision-making. The major two categories decided upon (drivers and barriers) did end up accounting for most of the comments. Thus it appears that the decision to continue is perhaps partly influenced by the balance between two competing sets of factors; drivers to continuation against drivers to ‘stop’ (referred to as barriers to continuation). Eight drivers to continuation were identified from the data (not including the ‘don’t know’ factor) and 7 barriers to continuation. This can by no means be said to provide quantitative evidence for the reasons to continue flying or to divert, since it is non-experimental, and even if such factors were reinforced by further study, their weightings would be highly complex. However it is possible to hypothesis that at least part of the reason to abandon the flight (divert) is down to the recognition of certain barriers, which may be pre-defined. Some barriers may be ‘non-compensatory’ in nature, ie. MSA (once the MSA is reached the decision is made). However some barriers could combine or be more general in nature. The ‘drivers’ may compensate for the appearance of barriers. For example poor visibility (a barrier) may not cause the flight to be abandoned because it is compensated for in the mind of the pilot by the ‘safety net’ of being able to turn round at any time. Hence a strong set of drivers might lead the pilot further into a situation than his/her safety barriers would have predicted.

Although none of the bottom level categories were particularly large (almost by definition!) the ‘safety net’ category was considerable compared to the others in the ‘drivers to continuation’. This points towards a hypothesis that pilots may be driven to continue into poor weather by the security of having a safety option (turn round at any time). This option will tip the risk in favour of continuation and make them take a greater risk in terms of continuation than perhaps it is safe to do so. Most (but not all) of these comments were referring to a ‘turn back’ option as the safety net, but since the video rarely gave even a small glance behind, it brings up the possibility that the ‘safety net’ might be more of a comfort than a reality. The pilot feels safe to continue and is possibly driven to do so, most of the attention is therefore given to considering the possibilities of getting further and therefore little attention is paid to the safety net. Thus the safety net might easily be eroded or even disappear completely, whilst still acting as a factor that produces a risk imbalance. It should be stressed that the type of data used to generate these possibilities cannot be sufficient to provide scientific evidence for them.

Option generation and evaluation did appear to be something influenced by the decision training, especially the avoidance of ‘serial processing’ or ‘satisfying’ and there is a small quantity of evidence to show that some participants felt that they were helped by the training in this regard. Although these comments were mostly made by participants who diverted within the divert time window, it cannot be taken as scientific evidence to verify the hypothesis, however it does hint towards a hypothesis that those who put this aspect of the training into use were helped to make an
appropriate decision. This hypothesis would require greater evidence before being tested experimentally.
4.4 Category iv: Use of Tools

Any comment relating to the use of decision-making tools was recorded in this category. The top two categories under this heading are related to which model was being described (FOR-DEC or DESIDE). These were further broken down into comments related to either using the models or not using them. The next level down are the categories formed from reasons why participants used or did not use the particular models, including good and bad feedback. Thus the bottom level categories help to identify reasons behind the lack of use of these models (in most cases) as well as constructive comments relating to them.

An explanation and example quote from each of the bottom level categories is given.

Top level Category: iv Use of Tools, and an example quote from each.

iv.A DESIDE

iv.A1 Did not use DESIDE

iv.A1a Couldn’t remember DESIDE
Any comment referring to the inability to use DESIDE due to not being able to remember it.
Example: “I still can’t remember it” & “I think it was too much to take in”
(Only two comments)

iv.A1b Too much workload to use
Any comment referring to the feeling that using DESIDE would make things more difficult.
Example: “I felt it would complicate things”
& “Having to work all of that out in your head like oh what does the acronym stand for and while there's all this cloud-base descending this is the sort of thing that can detract from your flying”

iv.A1c Difficulty of the model
Any comment referring to the difficulty of using the model
Example: “I think it's too long, I think you need to shorten it” & “you get to the S and or was it a C, decide is spelt with a 'C' oh no hang on it's an S”

iv.A2 Used DESIDE

iv.A2a Subconscious use of DESIDE
A comment referring to the use of the components ‘naturally’ without use of the model
Example: “I didn't use the wording of it but I think probably mentally I did”

iv.A2b Partial usage of DESIDE
A comment referring to usage of some aspects of the model.
Example: “Two or three elements of it, the deciding bit and then evaluating and then doing it but not the acronym as such”
Although encouraged to use the model, no participant made any comment suggesting they used the model all the way through. The experiment would not have been realistic had it forced participants to use the model, it was valid to discover whether participants felt able to use it or wished to use it. This is equally as important as finding out if it works.

iv.B FOR-DEC

iv.B1 Did not use FOR-DEC

iv.B1a Couldn’t remember FOR-DEC
Comments that stated that participants could not remember the acronym
Example: “I didn’t actually remember it to be honest. I was thinking of options all the way through but I didn’t have that, the FOR-DEC thing come into my mind”

iv.B2 Used FOR-DEC

iv.B2a Partial use of FORDEC
A comment referring to usage of some aspects of the model.
Example: “I used the facts and the options available and whether I did it all in the right order and at the right time, is another matter”

Figure 4d. Relations of categories in ‘Use of Tools’
The first thing that can be seen is that no participants admitted to successfully using the model in a systematic way, although many tried and some used parts of the models. Interestingly all the participants who commented on trying to use the model, were in the ‘post-window’ divert group. This is not evidence that using the model made participants divert too late! Clearly, had this been the case the control groups would have been better at diverting early, and the quantitative part of the experiment showed this to be not true at all, indeed control group A had the longest mean divert time at nearly 15 minutes (but by no means near statistical significance). Secondly the data is from open interviews, a methodology that cannot be relied upon in any way to produce this sort of evidence. Nevertheless it makes the picture is far from clear.

It can be seen that only one specific reason was given for not using FOR-DEC; the acronym could not be remembered. This partly accounts for the lack of other reasons in itself since if participants were unable to recall the acronym then they were unable to report any problems they had trying it out. The acronym recall test does show that a number of participants were able to remember it, but it should be remembered that recall under high workload is different from recall after the flight. One participant explained this in the interview:

“I remembered it at the end but I couldn't remember it when I needed it”.

Those who did try to use it only reported using part of it (generally the first part). No participant reported using the whole model, thus there is only one node under the ‘Used FOR-DEC’ category.

All the comments about DESIDE were from the group that ‘diverted’ after the divert window (ie. the unsafe group). The only exception to this is that two comments claiming partial usage with reasons why the whole model was not used were from the pre-window divert group (safe group). The reasons given for not using DESIDE were:

- That it couldn’t be remembered during the flight
- That to use it would require too high a workload, given the situation
- That it was too complex (difficult)

There could be a number of reasons for the fact that these comments all came from those that stopped the watch too late. The comments show that these participants were considering the use of DESIDE or trying to use DESIDE, which means they recognised that the situation was difficult (which for this group it was, being as they were still flying 16 minutes into the video). The comments suggest that they found DESIDE too difficult to use one way or they other (couldn’t recall it, did not have the capacity or found it ‘cumbersome’) and the fact that all the comments came from the late divert group suggests that by the time things have become very bad, pilots will struggle to use DESIDE. This could be equally applicable to FOR-DEC, but since that acronym was harder to remember, the non-usage comments were spread across all ‘divert time groups’. It can be hypothesised that when a situation becomes critical, these types of tools are very difficult to use.

The comments from the ‘use of DESIDE’ category from the Post-window group (ie. those that were still flying after 16 minutes) show that those who did try to use DESIDE were only able to use part of it, or they believed they were using the same
principle without actually going through it, or using it ‘mentally’ (as one participant put it). This strengthens the above hypothesis.

It appears that DESIDE is difficult to use when in a high workload/stressful situation. Of course, training and practice could be solutions to this, as the study groups only received cursory training prior to the experiment.

None of those who diverted on time admitted using either acronym. This could be for several reasons. Firstly these may represent a group of pilots who do not need to use such a model. Their decision-making skills and risk utilities are such that they will judge situations properly before things become too critical. Given the lack of difference in comment between the divert-time groups in the situation assessment category, as well as the subject questionnaire measures, there could be a serious element of ‘improper’ situational assessment involved. In the situational assessment comments category; the ‘post-window’ group were making similar comments to the ‘in-window’ and ‘pre-window’ groups, and yet the eventual situations encountered by these participants were quite different. The post window group should have expressed much greater concern regarding cloud, rain, terrain, airspace etc. It is possible that a poor situational assessment (seeing the conditions as safe when they were not) led to these participants still flying when the conditions deteriorated quickly (after about 16 minutes). At this point they may have begun to think about the model, but been unable to use it (for the reasons that they gave).

The results from this section show something unexpected. That is that the participants who mentioned using the models or trying to use the models almost all ‘diverted’ too late. In other words they were the ones who found themselves in appalling conditions. It appears therefore that participants may have tried to use the models when they realised that their own strategies were leading them into a situation that was unknown and they could not cope. But by the time they got to this stage the model was too difficult to use.

The other participants may not have needed to use an acronym, and indeed the data from the ‘decision processes’ categories (iii) suggest that many facets of the decision to stop are not about comparisons and evaluations between options (as the models try to promote) but about recognising one or more barriers (possibly pre-determined, such as MSA) which, when the barrier(s) overcome the drive to continue, cause the pilot to abandon the flight. Situation assessment would clearly be a big part of this recognition, and so if, as hypothesised the DESIDE model somehow makes pilots consider the situation more, it may well be a good acronym for that group of pilots (if used before the workload became too high).
5. Conclusion

5.1 Conclusions of analysis and findings

The overall experimental hypothesis was that the groups that were trained in decision-making and given a model to use would make safer decisions than those that were not. Experimentally there is very little evidence to support this hypothesis.

Statistical results concerning the decision itself (time to divert) show no significant differences between the groups. Although this result does not necessarily mean that the models and training were of no use, it cannot support the experimental hypothesis. As described there may be a variety of reasons for this and so focus must be put on parts of the decision process.

Of the two models tried, it appears that DESIDE is the most memorable as an acronym and the least successful in terms of content recall. It appears that FOR-DEC may be simpler to remember in terms of content and usage, but harder to remember than DESIDE as an acronym (result significant to 0.05 for a one-tailed test).

The wind vector recall tests did provide a statistically different result between group C and D, and when those participants who felt unable even to take a guess at it were taken into account, D was significantly different from all groups. Therefore the group that had been trained in the DESIDE model were significantly more likely to have made a correct assessment of the wind direction than the other groups. There was little difference between the control groups and the FOR-DEC trained group. This appears to provide evidence that something in the DESIDE training improved this aspect of the participant’s situational assessment.

Additionally the quantitative part of the post-experiment questionnaire provided two significant results, and again these both favoured the DESIDE model in terms of situational assessment. The DESIDE group were significantly more likely to have been concerned about controlled airspace during the flight, and also to have been concerned about the information heard on the radio. The latter finding ties in with the wind direction result because the wind direction was given over the radio. No other differences were found between the four groups.

It is worth noting that the only significant results relating to decision making to be found in the research are supportive of the DESIDE group over the other groups. The reason for these results is far from clear, especially when the results of the acronym recall test and the qualitative data are considered. The acronym recall showed that many participants had forgotten what DESIDE referred to, and the qualitative analysis provided evidence that very few participants actually used DESIDE, and those who tried to found it difficult. Consideration must be given to the possibility that simply having a meaningful word as an acronym acted as a cue to the decision making training to remind participants to use that training, or at least to make regular assessments of the situation.
The idea that using DESIDE improved situational assessment can be backed up by a recent study by Li and Harris (2005).

Qualitative data gave an insight into the decision processes being used. From analysis of the discussions about decision processes and strategy (category iii: Decision Process), it was hypothesised that the decision to abandon the flight may take place when certain barriers, or combinations of barriers are recognised. It was also hypothesised that there are possible ‘drivers’ that may well mean the pilot continues despite recognition of the barrier. The balance between drivers and barriers could go some way to explain the continuation of participants into a dangerous situation, but much more research would be required. This backs up the idea that good situational assessment is critical to the decision process, for without assessing those barriers correctly, the decision cannot be reliable.

It was also found that few of the participants used the acronyms, and those that did, were only able to use a few parts of them. The main reason for non-use of the models was that participants could not remember them at the time, or were under too much workload and felt the model would have increased it. A number of participants who diverted ‘too late’ did attempt to use an acronym (especially the DESIDE group), however none reported being able to do so fully. It was hypothesised that many of those who diverted on time had a safe strategy of their own (based on the findings from group iii of the qualitative data) and would not need to use such a model, hence none did so. However those who continued too long did ‘reach for’ the model, but found it difficult to use. Possibly training and practice would make this possible.

The ‘Decision Process’ categorisation found that one aspect of the training that participants reported finding useful in flight was the training given on ‘satisfying’, the assessment of all options before making a choice, as opposed to choosing an option that is satisfactory and dropping other options as a consequence.
5.2 General conclusion

This study has provided some evidence that, for the general aviation environment and with minimum training, DESIDE has an advantage as a decision aid over FOR-DEC or no acronym, if it is accepted that effective situational assessment improves decision-making.

Much more research is needed in this area in general aviation decision-making. This study has suggested the possibility that the acronyms being used are not efficient because they do not mimic the in-flight decision process for this particular type of flying. Also, participants used aspects of the training that were not connected to the models, while the models were generally not used. Additionally, the ‘divert/abandon’ decision may be based on factors not well triggered by the models (referred to here as drivers and barriers).

On the other hand, there is no evidence here that using a model is counterproductive and may be more likely to cause pilots to continue into dangerous conditions. All the evidence points to the fact that learning the DESIDE acronym particularly, can focus a pilot’s situational assessment, even if he/she does not use the model.