Scientific committee critique of the EASA research project 'Effectiveness of flight time limitations in commercial air transport (CAT) operators'

Scientific Committee, 31 January 2019: Philippe Cabon, Alexandra Holmes, Steve Hursh, Barbara Stone, Kristjof Tritschler

1. RESEARCH PROJECT OBJECTIVE AND METHOD

The objective of the research project was to evaluate the effectiveness of the European Flight and Duty Time Limitations in controlling fatigue among crew who work for CAT operators.

The research had two main stages: firstly, to select the two areas of highest fatigue, a crew survey (> 15,000 responses) and an analysis of example schedules with three biomathematical models was undertaken. Long nights (duties of more than 10 hours at the less favourable time of day) and disruptive schedules (consecutive early starts, late finishes, night duties, and combinations of these) were identified as being associated with the highest fatigue, and thus the focus of the second stage. In addition, crew schedules (6 airlines) were evaluated using two of the biomathematical models.

The second stage was a field data collection campaign using a crew sourcing approach, exploring long night duties and disruptive schedules. A total of 413 crew (277 pilots and 136 cabin crew) from 24 CAT operators contributed to the final dataset. For a period of 14 days, crew recorded data using an application downloaded online from Apple iTunes for use on an iPhone or iPad, on the timing and duration of sleep, flights and duties. Crew subjectively rated their fatigue, sleepiness, mental effort and identified hassle factors. Crew were asked to rate their sleepiness and fatigue at top of climb (TOC), approximately 15 minutes before top of descent (TOD) and during cruise for longer flights. A small sub-set of crew also wore an Actiwatch to track sleep and wakefulness. Psychomotor vigilance task (PVT) data was collected by pilots, but the data quality was inadequate for analysis.

The research data showed that at TOD of short or long night duties, 33–36% of crew members experienced high fatigue. Inadequate data was collected on disruptive schedules to enable conclusions to be drawn regarding the associated level of fatigue.

2. SCIENTIFIC COMMITTEE (SC) CRITIQUE OF THE RESEARCH

This critique focuses on the final report of the researchers. The SC reviewed supporting documents delivered by the researchers during the research project, and provided comments. This critique does not repeat all the SC's comments, or responses (if any) by researchers, provided throughout the research project. The researchers were free to use the SC's advice or not, often without discussion.

All members of the SC provided unbiased scientific advice to the researchers by giving comments at meetings and conference calls, and by written feedback on each of the deliverables. There is a large degree of agreement among SC, and there has been no need for conflict resolution, nor to bring in other scientists.

The researchers adopted the SC's suggestions about statistical analysis, and took some other comments and suggestions into account. In particular, researchers ignored the SC's suggestions about increasing the data collection phase of the research to ensure that adequate data was collected. In many instances, the researchers did not respond to or incorporate the SC's comments on writing style and terminology. For example, the terms 'state-of-the-art', 'fatigue hot spots', 'benchmarking' and 'fitness for purpose' remain in the documents, despite the SC raising questions about whether the terms were being appropriately used.

The SC often had insufficient time to review drafts of the deliverable documents because drafts were delivered late to the SC. The Sinapse 3 Platform for storage and distribution of documents was not available to the

research project; therefore, keeping track of different versions of documents was difficult. There did not appear to be a clear system of version control in place. The relationship between the deliverables was difficult to understand, and some seemed to overlap. There were even deliverables with the same number but with a long and short version. The structure of the documents did not follow normal scientific reports, and the separation of the discussion of the results in the light of previous research into a 'benchmarking' section was confusing.

The researchers produced minutes after each meeting with the SC, but the minutes were very brief, and did not always adequately represent areas of discussion. Early in the research project, the SC asked to see the data collection tool and instructions to subjects, but these were added as an appendix at the very end of the research project, and the SC has not reviewed them.

2.1 The research is not a 'pressure test' of the regulations

The SC understands that sometimes the researchers were constrained by the terms of their contract.The tender for the research project requested a review of the effectiveness of the current regulations within a period of two years. The SC would like to emphasise the complicatated nature of the requested task within limited available time.

The tender requested representative conditions of the EU aviation sector when taken as a whole, which had the unintended consequence of limiting the sample size of the flight duties of interest. In addition, all operators have some additional practices, labour agreements, or fatigue risk mitigations in place. These additional scheduling factors are not identified in this research project; therefore, the the results of this research project provide insight into the fatigue encountered by crew working for 24 operators, rather than an overall assessment of the effectiveness of the regulations.

The research does not constitute a 'pressure test' of the regulations because:

i. No operator operates entirely at all of the regulatory limits, rather the specific limits that are relevant to one business model may not be relevant to another operator.

ii. The 24 participating operators were not specifically selected because they worked schedules that were close to the regulatory limits.

iii. There was no specific 'test', of the effectiveness of the regulations relevant to potentially high-fatigue scheduling, such as a long night duty on the fifth consecutive duty.

iv. In the research dataset, FDP duration on most duties was well below the regulatory limits (see D2.3 figure 23).

v. The research dataset is not large enough to be representative of the schedules worked by crew across Europe, because the 413 crew who participated in the study represent only approximately 0.3% of crew employed by European operators.

2.2 Literature review

The literature review is included in Deliverable D1, 'Definition of the baseline'. The aim of the literature review was to extract some 'lessons learned' from previous studies on fatigue in aviation. The literature review was limited to research published after 2006, to limit the number of papers to be reviewed; research before 2006 was supposed to be covered by the 'Scientific and medical evaluation of flight time limitations', the so-called 'Moebus report' – however, the literature review reference list does not cite the 'Moebus Report', neither does the literature review use specific information from the 'Moebus report'. This is a clear limitation of the literature review, because EASA used the 'Moebus report' to build the existing regulations. No references were

made to key research such as the international collaborative research undertaken by NASA in the 1990s, nor the work funded by the UK CAA that brought a lot of knowledge on the effects of flight and duty time on sleep, fatigue, and human performance.

Another flaw of the literature review is that it primarily focused on methodological aspects, not on the results. A focus on results would have been useful, because these results were used to inform in the current regulation, and could have informed the design of the research hypotheses at the beginning of the research project.

2.3 Research methodology

In the first stage of the research, FDP types were ranked using survey data and results from biomathematical models. This process was well done; however, the SC found the explanation of findings to be somewhat difficult follow, because the models are not referred to by model names but rather by model outputs e.g. Effect (SAFTE-FAST) and CAS (BAM). As with other research data collected, the results from modelling were portrayed in terms of odds ratios (ORs) relative to the base case of all rosters (see report D2.1 'Identification of potential fatigue hot spots'). The key findings were the ORs of occurrence probability of fatigue scores from the models, and the significance tests associated with them. It seemed strange to the SC that lengthy tables were presented of the occurrence probability, while the key findings of significant ORs were not summarised in tables. Further, the final report did not summarise in a table all the significant ORs from modelling across the various comparisons.

The second stage of the research was designed to collect data from a representative group of European CAT operators who work long night duties and/or disruptive schedules. The researchers successfully recruited 24 operators who proportionately represent the geographic distribution and size of operators across Europe.

The researchers selected operators working long night duties, and/or disruptive schedules, but did not consider in-depth the schedules being worked, for example, how much time off was provided before or after long nights, or how early starts or late finishes are sequenced in runs of consecutive duties.

Crew who participated in the research collected all data using a Jeppesen CrewAlert application downloaded online from Apple iTunes for use on an iPhone or iPad. Crew entered data into the app on multiple demographic variables, and for 14 days entered data on their work schedule; timing and duration of sleep; sleepiness (KSS) ratings; Samn Perelli (SP) ratings; mental effort; hassle factors; and completed a five-minute PVT. The application had the advantage of being convenient for both crew and researchers, enabling a crowd-sourcing approach, and the collection of data from the 413 crew who participated in the research.

One disadvantage of the CrewAlert app was that it could only be used on an iPhone or iPad, and crew without one of those devices could not participate in the research. While many operators provide iPads to pilots, it is unusual for operators to provide iPads to cabin crew and this may have contributed to the relatively low rate of participation from cabin crew. It is not known how iPhone ownership, for example differences between countries, may have influenced the final dataset.

2.4 Statistical power of the research

The researchers conducted a power analysis for sample sizes of the two main FDP types, and that power analysis guided the size of the data collection effort. What was not considered was the required sample size for sub-categories of FDPs, such as different lengths of night duties or sequences of disruptive schedules. As a result, the sampling method did not ensure that there was sufficient power to address those secondary – but important – questions.

A further complication of the sampling method was the imperative to collect an overall sample that was representative of all flying operations in Europe. That priority may have conflicted with the priority to measure all the various subtypes of FDPs. Unfortunately, the inadequate sample sizes for various research questions was

not identified as a problem until after the data collection had ended, and that leaves several questions unanswered.

Interestingly, the FDPs from rosters used for modelling were categorised by type in Tables 6, 7, 8, and 9 of D2.1. The information from these tables gives an indication of the probability of observing different kinds of FDPs – for example, the odds of observing two consecutive nights were 0.7%. This could have alerted the researchers that a random sample was unlikely to uncover sufficient cases to permit examination of the fatigue effects of consecutive nights.

Data collection period and participation

A related question is whether the duration of data collection with each subject was sufficient to capture data needed to answer the research questions; and whether subjects were adequately motivated to complete the data collection. The researchers decided early in the project that 14 days would be adequate for participants to collect data, and that data collection would stop after a set number of months. The SC questioned these time periods, but apparently, contract milestones worked against extending the data collection period.

Data analysis

Most of the data collection and analysis are based on the frequency of a high level of fatigue (KSS values higher than 7) at the final TOD of the duty. This analysis raises several problems:

1. Depending on the duty schedule, fatigue at TOD could be lower than during earlier phases of the flight, and this could not be captured by the analysis.
2. Most duties had more than one sector, but as the analysis focused only on the final sector of the duty, it misses the effects of consecutive sectors. This is a clear flaw of the research, because the effects of the consecutive sectors are an important aspect of the FTL regulations.
3. The data analysis considers the occurrence of high KSS, rather than on the distribution of all KSS data.

Another limitation of the analysis is the lack of consideration of cumulative fatigue. The analysis is based on the evaluation of isolated FDPs. There is no indication of where the isolated duties occurred in the sequence of duties. This does not allow assessment of the cumulative fatigue resulting from a combination of rosters, which is known to be a key issue, especially in short-range operations. This aspect is also relevant in terms of fatigue management, because operators may more easily and manage sequences of duties, rather than changing isolated FDPs.

2.5 Significance testing

The researchers applied appropriate tests of significance throughout their report, and the conclusions focus on those conditions that were found to show statistically significant differences. It would have been helpful to have included a brief discussion of the difference between 'not finding a significant difference', and 'finding statistical equivalence'. There are several methodological and sampling limitations that can prevent results from two conditions reaching significant statistical difference, but that does not mean that there is no difference, nor that the two conditions are statistically equivalent. In short, a 'non-significant difference' does not imply that conditions are the same or equivalent. The SC discussed this with the researchers, who subsequently avoided drawing any conclusions about conditions that were not significantly different from baseline or from another sub-condition such as night FDPs of varying lengths. This approach is statistically correct.

To their credit, the researchers did re-examine the way the conditions were sub-divided, and discovered that one factor that worked against finding a significant effect of the different FDP types was the very definition of categories, such as night FDPs. In the final report, results shown in Figures 10 and 11 are re-organised by more appropriate categories of FDPs, and clear statistical differences in ORs were found for FDPs of interest, as shown in Table 8.

2.6 Controlled rest on the flight deck

There is limited published information on the frequency, timing, and duration of in-flight naps, and to our knowledge this is the first large piece of research to have collected data on the incidence of in-flight naps on the flight deck. All the in-flight naps occurred on late finish and night FDPs, and naps were taken during 22.4% (32/144) of duties. In-flight naps were 19.6 ±7.8 minutes in duration, which is within the maximum duration in the EASA guidance material on controlled rest.

Operators, regulators and researchers should better understand the actual practice of napping. Napping is an effective control for fatigue, but because sleep inertia is associated with a reduction in performance, it is essential that crew manage the return of the napping crew member to their active role.

3. RECOMMENDATIONS

The SC acknowledges that this has been the largest crew fatigue study across Europe so far and that some important conclusions can been drawn from this study. The SC supports the recommendations made by the researchers.

The probability of high fatigue during FDPs night is not simply related to the duration of the FDP, as there are other influential factors, such as the amount of sleep a crewmember obtains before the duty and the start and finish time of the duty. Therefore, rather than simply limiting FDP duration, a more comprehensive approach to managing the fatigue associated with night duties is necessary.

The definition of a night FDP in the current regulation is too broad, and the three subcatories for night duties suggested by the researchers should be utilised instead. New definitions are not mitigations, but the SC agrees that this may help operators to tailor fatigue risk management strategies for increased effectiveness.

Further research is necessary to understand in-flight napping and the fatigue associated with disruptive schedules.

Recommendations 5 and 6 explain that crew members should obtain sufficient sleep before all night duties, but don't also recognise that time of day and other factors, for example environmental conditions, determine an individual's ability to obtain sleep. Fatigue risk management is a shared responsibility between the operator, who provides rest times and facilities where sleep is possible, and individual crew who plan and obtain the sleep required to be fit for duty.

Recommendations 5 and 6 therefore open the field for more effective mitigations. A rest period of a few hours close to reporting time may be more efficient than an augmented flight with crew members being awake for a long period before starting their night duty. Corresponding regulations should be adopted to make possible the option of a pre-flight rest period.

4. LESSONS LEARNED

i. The next study could better select operators who are working the rosters of interest.

ii. In retrospect, the risk that there were inadequate data on the FDPs of interest could have been mitigated by daily monitoring of the database to identify when adequate data had been collected. This was not done; therefore, the results database had inadequate women, cabin crew, late finishes, and consecutive FDPs of interest to address important factors. Any future study should anticipate the factors of interest in advance, and monitor the data to ensure that the sample collected on the factor is adequate to observe effects.

iii. The time to review the deliverables should be agreed in advance with the SC, who should make time to discuss and comment on the response to deliverables.

iv. Utilise an e-community or hub (e.g. EU Commission service called Sinapse) to enable storing of documents, tracking of comments and responses to comments.

v. Scientific involvement in writing the next tender.

5. SUGGESTED EDITS TO RESOLVE

i. 'Night duties' and other terms have not been consistently used with one definition. This is the case within the final document, and between the various documents. For example, D2.3 Table 16 is based on the original definition of night duties and Figure 22 onwards uses the new sub-categories of night duties.

ii. Questionable conclusion on long nights. Please check D2.3 Table 16, where sleep is 5h versus 3h and time awake 13h versus 20h. These values seem extreme.

iii. Table 16: please confirm whether nap duration is expressed in minutes.

iv. Final report, Table 9, appears not to be accurate, or needs to be better explained.

v. Chapter 6 indicates that there were 2 cabin crew from cargo operations, but it seems likely that this an error.

vi. Table 9 indicates that adequate data was collected on disruptive schedules, but this is not the case.

vii. The final report includes some references that are not in the reference list, and vice versa.