Independent research on Pilot Fatigue Measurement by the Netherlands Aerospace Centre

CAP1756

INTENTIONALLY LEFT BLANK

CAA Preface for NLR/NIN Fatigue Research Report

This independent research study considers potential improvements in the way that pilot fatigue is managed. Fatigue in the aviation industry, particularly for pilots operating across different time zones is a well-recognised issue and is predominantly managed by control of duty hours based on substantial research evidence.

Recent technology developments suggest that it may be possible in the future to consider fatigue management more at an individual level rather than taking a generalised duty-hours approach. This research study considers this potential, mainly from a technological perspective. The results of the study unsurprisingly indicate the complexity of making safety improvements in this area, however as a topic it is likely that there may benefits to be gained and CAA will continue, with others, to explore the science as it develops.

It is recognised that to be effective, ultimately any methodology will need to work within an operational context. This study used complex equipment requiring specialised skills and time-consuming tests however the researchers were aware that to be operationally acceptable, ultimately testing would need to be relatively simple and practicable. Whilst when considering these and similar techniques within the aviation community there has been speculation that a number of fatigue measurement scenarios might be envisaged, ranging from pre-flight (fit-to-fly) tests to continuous flightdeck monitoring, at the time of publication aspirations are limited to supporting the existing fatigue management infrastructure.

Authorship:

This report was compiled for the UK Civil Aviation Authority by:

Netherlands Aerospace Centre NLR (Nederlands Lucht- en Ruimtevaartcentrum)

and

Netherlands Institute of Neurosciences (NIN)

Individual authors are identified by Chapter.

Note

On 1 Oct 2012 the MOR database system was changed to be compatible with the European ECCAIRS format and the use of MOR terminology in this report is used for simplicity, regardless of the date of the incident.

Executive Summary

Introduction

Pilot fatigue is a significant safety concern and has to be effectively managed. A number of procedures are in place to control the associated risks. However, these procedures are based on generalised management techniques, such as controlling hours of work rather than measuring the actual fatigue levels of individual pilots. Being able to objectively measure the fatigue of individuals could potentially offer a significant safety benefit. However, there is no accepted, practical way to do this in an operational context and even in a laboratory setting this remains a challenging matter.

This CAA-UK Pilot Fatigue Measurement study addressed key issues towards a goal of developing potential measurement methodologies that may better manage pilot fatigue. The main research question to be answered was: "Can fatigue in individuals be measured with sufficient reliability in order to make a relationship with changes (deterioration) in flying performance?"

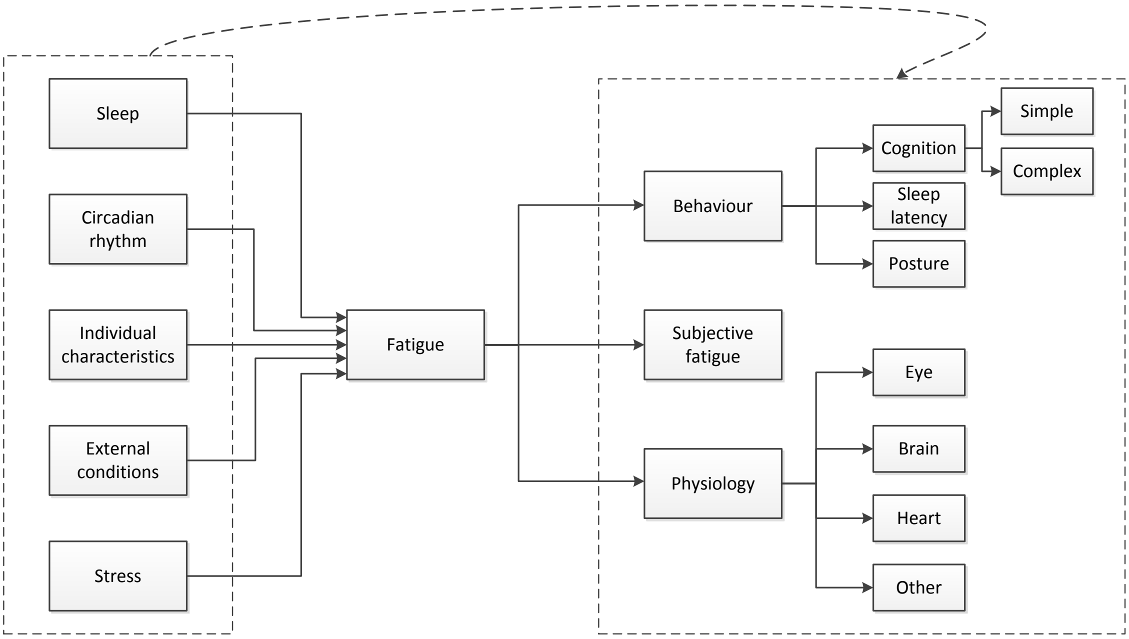
The following tasks were executed:

- Task 1: Available fatigue measurement techniques were considered and an appropriate number were selected for trial.
- Task 2: Experimental tests to investigate the selected techniques for effectiveness in measuring fatigue were defined and carried out.
- Task 3: Significant flying activities were defined that may be sensitive to fatigue.
- Task 4: Fatigue measurement using the techniques derived from Task 2 was related to flying performance in a simulator.
- As Task 1 and 3 provide input for Task 2 and 4, the sequence of the tasks was as follows: Task 1, Task 3, Task 2, and Task 4.

Task 1: A review of fatigue measurement methods

This literature review considered available fatigue measurement techniques in order to establish a comprehensive collection of data leading to a clear position on the most likely practical techniques for assessing pilot fatigue.

Fatigue is a complex phenomenon with many aspects for which there does not exist a single measure. Therefore, several aspects of fatigue were considered to obtain a good view of this multidimensional construct. This included contributors to fatigue as well as expressions of fatigue. Articles were reviewed to obtain a list of relevant measurement techniques and their properties, using the fatigue construct described in the figure below as a classification system.



Criteria for selection of techniques were the following:

- The test should be fast and easy to undertake, ideally self-administered;
- The test results should not require specialist interpretation;
- The test should be robust against falsification of parameters;
- The test results should be unambiguous;
- The test should be socially acceptable and non-invasive;
- Test results should not be subject to learning effects.

These criteria applied to the final (set of) technique(s) to be proposed for application in an operational environment. These criteria did not necessarily apply to the techniques that were used in the process of acquiring sufficient data for the evaluation in Task 2.

Based on the literature search, the following fatigue measurement techniques were selected for the evaluation in Task 2:

Element		Measurement method
Contributors to fatigue	Sleep	Actigraphy
	Circadian rhythm	Actigraphy
	Individual characteristics	Sleep registry Post illumination pupil response
	External conditions	Electrocardiography (ECG) & blood pressure Posture Skin temperature Ambient temperature Light Humidity
	Stress	Ecological Momentary Assessment (EMA)
Expressions of fatigue	Simple cognition	Reaction test Psychomotor Vigilance Task (PVT)
	Complex cognition	Flight simulator 2-back task
	Subjective fatigue	Questionnaire
	Eye	Saccadic velocity
	Brain	Electroencephalography (EEG)
	Heart	ECG & blood pressure
	Other	Skin temperature gradient

Task 3: Analysis of accidents and incidents involving flight crew fatigue

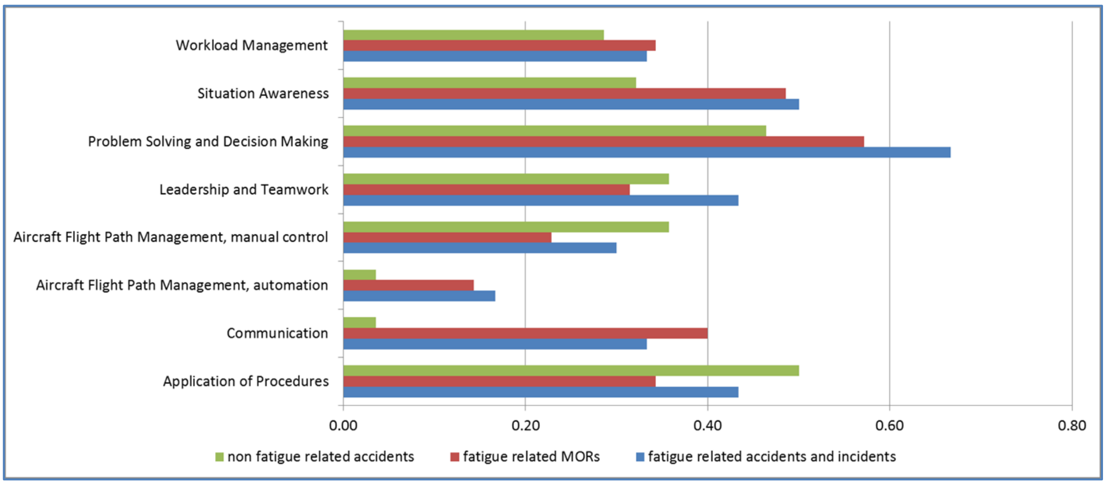
The objective of Task 3 was to define significant flying activities that may be sensitive to fatigue. The result of this task was used to define the scenarios for the flight simulation experiments carried out in Task 4.

The approach for identifying safety-related flying activities that may be sensitive to fatigue was to compare accidents and incidents in which fatigue was a contributing factor with accidents where fatigue was not a factor. Sources of data were official accident investigation reports (mostly from the National

Transportation Safety Board, NTSB[1]) and CAA's Mandatory Occurrence Reports (MORs). For each accident/incident, the core pilot competencies that were affected by fatigue (in the case of fatigue related accidents) or whose substandard performance contributed to the accident (for non-fatigue related accidents) were identified.

The total sample consisted of 30 accidents in which pilot fatigue was a factor, 35 incidents in which pilot fatigue was a factor, and 28 accidents in which pilot fatigue was not a factor. A comparison was made of the relative contribution of the each of the pilot competencies to the fatigue related accidents and incidents and the accidents that did not involve fatigue.

The most frequently mentioned competency for accidents and incidents involving fatigue is 'problem solving and decision making'. The overall picture resulting from the analysis of MORs is remarkably similar to that resulting from the analysis of accidents and incidents as displayed in the figure below.



For accidents that are not fatigue related, 'application of procedures' is the most frequently mentioned competency. Comparing accidents that are fatigue related with those that are not fatigue related showed that 'situation awareness', 'problem solving and decision making', 'aircraft flight path management, automation' and 'communication' contributed more frequently to fatigue related accidents than to accidents that are not associated with fatigue. It was therefore concluded that from the perspective of flight safety, the effect of fatigue on these competencies is most important.

[1] The NTSB was used a primary source of data because of the quality of the investigations and the relatively large number of accidents (resulting from the large volume of US air traffic).

The analysis showed that all ICAO-defined core pilot competencies were mentioned in fatigue related accidents. Therefore, it was recommended that the flight simulation scenarios in the experiments of Task 4 would represent all core pilot competencies, with emphasis on 'problem solving and decision making'.

Task 2: Measuring fatigue

The objective of Task 2 was to investigate the selected techniques (resulting from Task 1) for effectiveness in measuring fatigue through experimental tests; optimisation of fatigue assessment and prediction was pursued using multivariate assessment of physiology, behaviour, performance, environmental exposure, and sleep history. From a complete multivariate assessment, the most discriminative minimal dataset that is feasible in practice yet provides a robust estimate of task-relevant current and near future-projected fatigue in pilots was selected.

Twenty-eight participants with a frozen Air Traffic Pilot Licence (frozen ATPL, age range 20-44 yrs., 4 female) and eight active duty pilots (ATPL, age range 29-46 yrs., 0 female) were enrolled in the study. The experiment for each participant lasted nine consecutive days. Day one and nine took place at the NIN lab while day two through eight consisted of an ambulatory recording period at home.

A data-driven evaluation of the relationship between character traits and individual sensitivity to sleep debt revealed that people with high 'self-directedness' are less sensitive to sleep restriction. Self-directedness, the individual ability to govern behaviour according to situational demands, apparently helps people to be resilient to the effects of restricted sleep. This ability is definitely required in aviation and should warrant future investigations in a sample of active duty airline pilots.

As expected, all fatigue- and sleepiness-related questionnaires that were taken once correlated positively with the average level of sleepiness obtained throughout the week. This implied that questionnaires taken only once can be indicative of average fatigue levels throughout the week and can be considered as a simple means to estimate fatigue, although limited through not meeting the test criteria of Task 1.

The ambulatory study replicated previous laboratory study results, which indicated that subjective sleepiness is more sensitive to sleep(-disruption) than objective measures of sleepiness. However, in an operational context subjective reporting may be sensitive to the outcome desired by the subject who is queried. In such cases subjective reports should be complemented by objective assessments such as the PVT and other methods described in this report.

The multivariate assessment also revealed that recording sleep, time awake, posture, light exposure and skin temperature all provide relevant information for estimating real-time levels of sleepiness. The extent to which these variables affect sleepiness depends on the metric used to measure sleepiness. Although sleepiness, vigilance and fatigue are closely related constructs, all lack formal agreed-upon definitions; therefore here sleepiness refers to the outcome measures of all the tests. Of all these variables, time awake (since waking up that day) was the most potent predictor, followed by posture, light exposure, time spent in bed the previous night and skin temperature. Environmental temperature and humidity did not appear to significantly affect performance. Although fatigue and sleepiness are frequently attributed to sleep disruption, our findings indicated even stronger effects of posture and light exposure than time in bed. These variables should thus not be ignored. Whereas previous laboratory studies have shown that each of these factors individually affects vigilance, the present study was the first to assess their combined effects. The findings indicated that combined measurements are desirable if not essential to optimize estimates of current and predicted fatigue in an operational setting.

The results of the lab study indicated that a supine posture and small Distal-to-Proximal skin temperature Gradient (DPG) can affect sleepiness in as little as 30 minutes, although, again, the results differ per metric. Supine posture affected sleepiness most pronouncedly, followed by a small DPG and a later time of day. Supine posture moreover induces a smaller DPG, thus amplifying its effect on sleepiness. The implication is that fit-to-fly tests should consider or control, at a minimum, the time of day, skin temperature and posture when conducting the test.

It should be noted that although a substantial subset of the features recorded during the study has been analysed for this report, our unprecedented comprehensive multivariate approach has made more data available for future analysis.

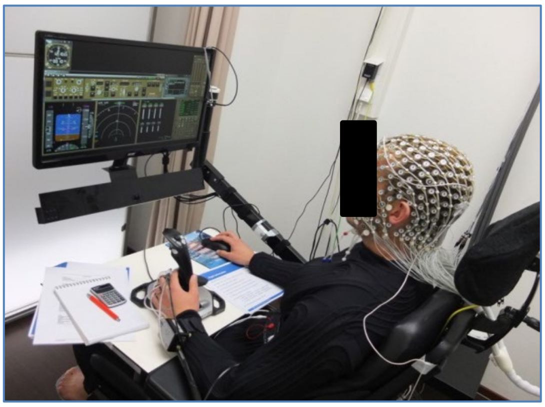
In conclusion, the results from this study clearly showed that optimisation of fatigue assessment and prediction in an operational setting should consist of a multivariate assessment of physiology, behaviour, performance, environmental exposure, and sleep history. Fit-to-fly tests should consider standardization of the recent and current environment of testing to maximize the precision and sensitivity of the estimated levels of fatigue.

Task 4: Fatigue in relation to flying performance

The objective of Task 4 was to relate fatigue to flying performance. Using the fatigue measurement techniques derived from Task 2 flying performance in a simulator was related to fatigue. Flight performance data was used to support analysis of performance of pilots having differing levels of fatigue.

Two flight tests were used to assess performance on pilot competencies. Both flight tests were set-up to replicate a flight in a Boeing 747-400. During the experiments, several events were triggered to test specific pilot competencies.

All participants in the study population (8 active duty pilots and 24 non-active duty pilots) were subjected to a flight test in the AIRSIM desktop flight simulator immediately after completion of the lab protocol (Task 2). The 8 active duty pilots were also subjected to a flight test in the GRACE full-flight simulator immediately after completion of the AIRSIM flight test. In contrast to the AIRSIM test, the GRACE test was a two-pilot operation. The subject pilot was accompanied by a project pilot who was part of the research team.



[redacted]

In order to determine whether the AIRSIM test was a correct way to estimate flight crew performance, the average scores of the active pilots are compared with the average scores of the non-active pilots. The hypothesis was that active pilots have better performance and therefore score higher on the AIRSIM experiment. The hypothesis was confirmed by the results; i.e. AIRSIM was indeed suitable to measure flight crew performance.

To determine if fatigue had an effect on flight crew performance during the AIRSIM and GRACE experiments, PVT and Karolinska Sleepiness Scale (KSS) scores were compared with the performance scores for the various tasks during the AIRSIM and GRACE tests. The results indicated weak correlations between the fatigue measures and performance on the flight simulator experiments. An explanation for this could be that there was a relatively high variation in pilot performance that was not related to fatigue, both between different pilots and within the same pilot.

Final conclusions

The results from this study clearly showed that optimisation of fatigue assessment and prediction in an operational setting should consist of a multivariate assessment of physiology, behaviour, performance, environmental exposure, and sleep history. Fit-to-fly tests should consider standardization of the recent and current environment of testing to maximize the precision and sensitivity of the estimated levels of fatigue.

Table of Contents

CHAPTER 1
Introduction

Background

Pilot fatigue is a major safety concern and has been extensively investigated over many years [1]. As a result a number of procedures are in place in the UK (and elsewhere) to manage the associated risks. However, these procedures are based on generalised management techniques, such as controlling hours of work (through rostering systems) rather than measuring the fatigue levels of individual pilots. Fatigue risk assessments generally assume that safety-critical staff arrive for their work without fatigue accumulated from previous activities. The fatigue situation of individual pilots is managed by self-declaration. However, this is not without problems as individuals may feel under pressure to undertake duties that actually should be declined. Being able to objectively measure the fatigue of individuals could potentially offer a significant safety benefit. However, there is no accepted, practical way to do this in an operational context.

This study addressed three key questions towards a goal of developing methodologies to better manage pilot fatigue. The key research questions[2] to be answered by this research study were the following:

1. Can fatigue in individuals be measured with sufficient reliability in order to make a relationship with changes (deterioration) in flying performance?
2. Can the implications of such deterioration be quantified in safety terms?
3. If so, what would the implications be for operational safety management, if such techniques were to be employed?

It was not expected that these questions would be answered in full by this research. Although question 2 and 3 were addressed somewhat, the main focus of this research was on question 1.

[2] According to the CAA-UK research specification of contract no. 1978.

Measuring fatigue

With respect to the first key research question, although fatigue has been extensively investigated over many years, there is no consensus on a gold standard for the definition and measurement of the actual and predicted level of fatigue [2]. We posit that the lack of successful determination and, more importantly, prediction of fatigue has resulted from an insufficiency of the information required to determine it. Reliable assessment of fatigue requires multivariate assessment and knowledge of the many factors that may (temporarily) mask its effect on performance. For example, fatigue assessment using performance measures only is sensitive to a person's inclination to exert effort and capacity to compensate for the effects of fatigue. Furthermore, physiological measures may appear unaffected at a particular moment in time but may subsequently change to indicate fatigue after a certain interval, the duration of which is determined for example by a person's sleep history. Finally, any measurement may suggest intact vigilance when assessed in a brightly lit cool environment in a person that has recently been physically active [3] and has taken an upright posture [4-6]. Fatigue may soon surface when the person has to be sedentary for hours, maintaining a sitting or supine posture in a dimly lit [7] or warmer environment [8]. Therefore, information about the recent history of environment and behaviour is paramount to reliably estimate fatigue and subsequent deterioration in flying performance.

The primary objective of the present study was to pursue optimisation of fatigue assessment and prediction using multivariate assessment of physiology, behaviour, performance, environmental exposure, and sleep history[3]. Important modulators of fatigue related performance changes as well as subjectively experienced fatigue include individual differences like age; evening or morning chronotype; personality traits; motivation and compensatory effort; individual's circadian time and time awake; individual's sleep-wake history; and behavioural and environmental (notably light) history.

However, the question of measuring fatigue came with the added requirement that the measurement should be operationally practicable. This provided the additional challenge of allowing for only a limited range of assessments. The research aimed to select, from a very complete multivariate assessment, the

[3] This refers to the fundamental measurement techniques, not the commercially available technologies that embody such techniques.

most discriminative minimal dataset that was feasible in practice yet provided a robust estimate of task-relevant present and projected pilot fatigue.

‘Flying’ involves different types of activities that may require different skills that are not necessarily equally affected by fatigue. The question on the relationship between fatigue and deterioration in flying performance therefore required a systematic breakdown of flying into tasks and activities to be able to determine the effect of fatigue on performance of those tasks.

Implications for safety

With respect to the second research question it was noted that the investigation report on the accident with a Connie Kalitta DC-8 at Guantanamo Bay, Cuba on 18 August 1993[4] was one of the first cases where fatigue was mentioned as a causal factor[5]. The investigation report stated that the probable cause of the accident was: “Impaired judgment, decision making, and flying abilities of the captain and flight crew due to the effects of fatigue”. This statement was very significant because it not only identified fatigue as a cause of an aircraft accident, but also stated that fatigue caused a deterioration of several aspects of flying performance (i.e. judgment, decision making, and flying skills).

The National Transportation Safety Board (NTSB) investigation report on the accident at Guantanamo Bay was able to conclude that fatigue was a causal factor because of a systematic scientific analysis of crew fatigue factors. The evidence that fatigue affected performance came from an analysis of the cockpit voice recordings prior to the accident and the testimony of the captain who, although seriously injured, survived the accident.

Since the accident at Guantanamo Bay fatigue has been mentioned more frequently in aircraft accident investigation reports as a causal or contributing factor, see for instance:

- Korean Air, B-747, Guam, 6 August 1997 (NTSB report AAR-00-01) [11];
- American Airlines, MD-82, Little Rock, Arkansas, USA, 1 June 1999 (NTSB report AAR-01-02) [12];
- SI FLY, ATR-42, Pristina, Kosovo, 12 November 1999 (BEA report F-FV991112a) [13];

[4] NTSB (1994) Aircraft Accident Report 94/04.

[5] Note that aircraft accidents usually involve multiple causal and contributing factors.

- FedEx, B727, Tallahassee, Florida, USA, 26 July 2002 (NTSB report AAR-04-02) [14];
- Corporate Airlines, BAe Jetstream, Kirksville, Missouri, USA, 19 October 2004 (NTSB report AAR-06-01) [15];
- Logan Air, BN2B Islander, Campbeltown, UK, 15 March 2005 (AAIB report 2/2006) [16];
- Pinnacle Airlines, Canadair CRJ, Traverse City, Michigan, USA, 12 April 2007 (NTSB report AAR-08-02) [17];
- Colgan Air, DHC-8, Buffalo New York, USA, 12 February 2009 (NTSB report AAR-10-01) [18].

In the Colgan Air investigation report the NTSB concluded: “The pilots' performance was likely impaired because of fatigue, but the extent of their impairment and the degree to which it contributed to the performance deficiencies that occurred during the flight cannot be conclusively determined”. This conclusion is typical for accidents where fatigue is suspected as a causal or contributing factor: There is often sufficient evidence that the pilots were fatigued, but there is insufficient evidence to determine if and to what extent pilot fatigue contributed to the accident. The inability to reach a firm conclusion, despite the rigorous investigation that usually takes place after an aircraft accident, underlines how difficult it will be to quantify the implications of flight crew fatigue in safety terms.

Implications for operational safety management

If a simple and practical, self-administered test to objectively measure the fatigue of pilots existed, this would have significant implications for safety management. The nature and extent of these implications would depend on the characteristics of the test as well as the regulatory framework in which this takes place.

Using such a test for operational safety management would require that a threshold level for maximum allowable fatigue should be set. This means that, aside from scientific insight in the effect of different levels of fatigue on flight crew performance (which was part of this research), a policy decision would be needed (which was not part of this research). Depending on the regulatory context, this policy decision should be made by the airlines and/or by the regulator. Support from the relevant stakeholders, like IATA, IFALPA and regulatory bodies (CAA UK, EASA and ICAO), would be essential.

Modern airline operations are increasingly subject to integrated risk management in which safety risks are part of the risks managed and fatigue risk is part of the safety risks managed. In considering the possibilities of fatigue screening of aircrew, it would be important to also give due consideration to the way in which fatigue screening could fit in the way airlines manage risk. At the more strategic level, these matters would be addressed as part of the Fatigue Risk Management System (FRMS) within the context of the Safety Management System (SMS). In a recent study on the introduction of FRMS for the French regional airlines, Cabon et al. [9] described the link between SMS and FRMS as outlined in Figure 1.1 below.

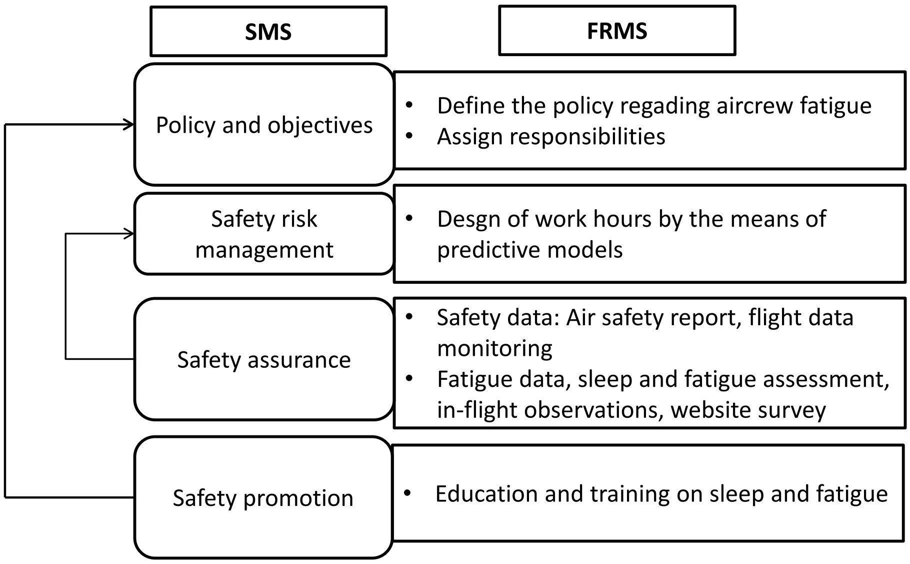

SMS
FRMS
Policy and objectives
• Define the policy regading aircrew fatigue
• Assign responsibilities
Safety risk management
• Desgn of work hours by the means of predictive models
Safety assurance
• Safety data: Air safety report, flight data monitoring
• Fatigue data, sleep and fatigue assessment, in-flight observations, website survey
Safety promotion
• Education and training on sleep and fatigue

Figure 1.1: Link between SMS and FRMS [9]

At the tactical level, fatigue screening would primarily be linked with operational safety management. Fatigue screening could have disruptive effects on flight operations when crews are found to be unfit for duty. It is possible in today's operations that the aircrew reports that they are unfit shortly before the start of their duty, but such occasions are relatively rare. Fatigue screening may well increase the frequency of such events.

Another area of interest with regard to the possibility of integrating fatigue screening into operational risk management would be the use of screening to take operational measures to compensate in case fatigue levels fall within a certain (As Low As Reasonably Practicable, ALARP) range. Recent research into the possibilities of applying so-called fatigue proofing strategies provides some promising insights into the possibilities of managing risk from this perspective. For example Dawson et al. [10] observe that there are multiple layers that precede a fatigue-related incident, for which there are identifiable hazards and control. An effective risk management approach would attempt to manage each layer of risk. In their study, they use Helmreich's threat and error management approach which consists of a series of defensive layers (levels) each providing opportunities to trap and mitigate (the effects of) fatigue. Levels 3 (behavioural symptoms of fatigue) and 4 (assessment and control of fatigue-related error) concern the actual flight operation.

Outline research specification

According to the research specification of contract no. 1978 the following tasks were executed:

- Task 1: Available fatigue measurement techniques were considered and an appropriate number were selected for trial. A comprehensive collection of data was established (in Task 2), leading to a clear position on the most likely practical techniques for any operational use.
- Task 2: Experimental tests to investigate the selected techniques for effectiveness in measuring fatigue were defined and carried out.
- Task 3: Significant flying activities were defined that may be sensitive to fatigue. The definitions were supported by accident and incident data.
- Task 4: Fatigue measurement using the techniques derived from Task 2 was related to flying performance in a simulator. Flight performance data from the simulator was used to support analysis of performance of pilots having differing levels of fatigue in addition to commentary from the simulator observer.

In the chapters below the results of Task 1 to 4 are described. As Task 1 and 3 provide input for Task 2 and 4, the sequence of the tasks is as follows:

- Chapter 2: Task 1;
- Chapter 3: Task 3;
- Chapter 4: Task 2;
- Chapter 5: Task 4.

The overall results and conclusions are described in Chapter 6.

CHAPTER 2

Task 1: A review of fatigue measurement methods

Alfred Roelen[1], PhD

Henk van Dijk[1], PhD

Bart H. W. te Lindert[2], MSc

Eus J.W. Van Someren[2,3], PhD

[1] Netherlands Aerospace Centre NLR (Nederlands Lucht- en Ruimtevaartcentrum), Amsterdam, the Netherlands

[2] Department of Sleep and Cognition, Netherlands Institute for Neuroscience (NIN), an institute of the Royal Netherlands Academy of Arts and Sciences, Amsterdam, the Netherlands

[3] Departments of Medical Psychology and Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research (CNCR), VU University and Medical Center, Amsterdam, the Netherlands

Introduction

This review considered available fatigue measurement techniques in order to establish a comprehensive collection of data leading to a clear position on the most likely practical techniques for assessing pilot fatigue. The review investigated fundamental techniques, not the effectiveness of commercial equipment which may have proprietary processes not suitable for examination.

Structure Chapter 2

Chapter 2 is structured as follows:

- Classification of measurement methods that was used throughout the research;
- List of measurement methods that were originally envisioned;
- Discussion on possible measurement methods to be added to the original list of measurement methods;
- Conclusions and recommendations of Task 1.

Method

A systematic search was conducted of PubMed, using search terms such as fatigue, drowsiness, alertness, review. Additional articles were identified by manually searching bibliographies of retrieved publications. Furthermore, e-mails were sent to relevant scientists, using the authors' network, with a request to mail applicable recent publications on measurement of fatigue and alertness to predict performance.

Review articles such as [1] and [18] were first read to obtain an initial list of relevant measurement techniques and their properties, using the fatigue construct described in Figure 2.1 as a classification system. Other publications were then read to establish if additional techniques (not already on the list) were described and if supplementary information on properties was provided. Publications that did describe measurement techniques, but added no extra information to what was provided by one of more publications that were already read, were set aside and were not included in the list of references, while publications that delivered new information were included in the list of references.

The fatigue construct in Figure 2.1 includes contributors to fatigue for which a review of measurement techniques could easily result in a report of its own. However, this would not contribute to answering the main research question and therefore the literature search in this direction was deliberately restricted.

Selection criteria

Criteria for selection of techniques were the following:

- The test should be fast and easy to undertake, ideally self-administered;
- The test results should not require specialist interpretation;
- The test should be robust against falsification of parameters;
- The test results should be unambiguous;
- The test should be socially acceptable and non-invasive;
- Test results should not be subject to learning effects.

These criteria applied to the final (set of) technique(s) to be proposed for application in an operational environment. These criteria did not necessarily apply to the techniques that were used in the process of acquiring sufficient data for the evaluation in Task 2.

Classification of measurement methods

Fatigue is a complex phenomenon with many aspects that are possibly important contributors to or expressions of fatigue. This is illustrated in Figure 2.1. The left side shows the contributors to fatigue; the right side expressions of fatigue. Note that contributors can also influence behaviour and physiology directly, i.e. without fatigue as a mediator (indicated by the dashed arrow).

Fatigue cannot be measured directly. Therefore the contributors to and expressions of fatigue must be measured in order to obtain information on fatigue. In the next sections, the contributors to and expressions of fatigue are described with respect to possibilities for measuring[6]. Details of specific measurement methods are provided in Appendix A.

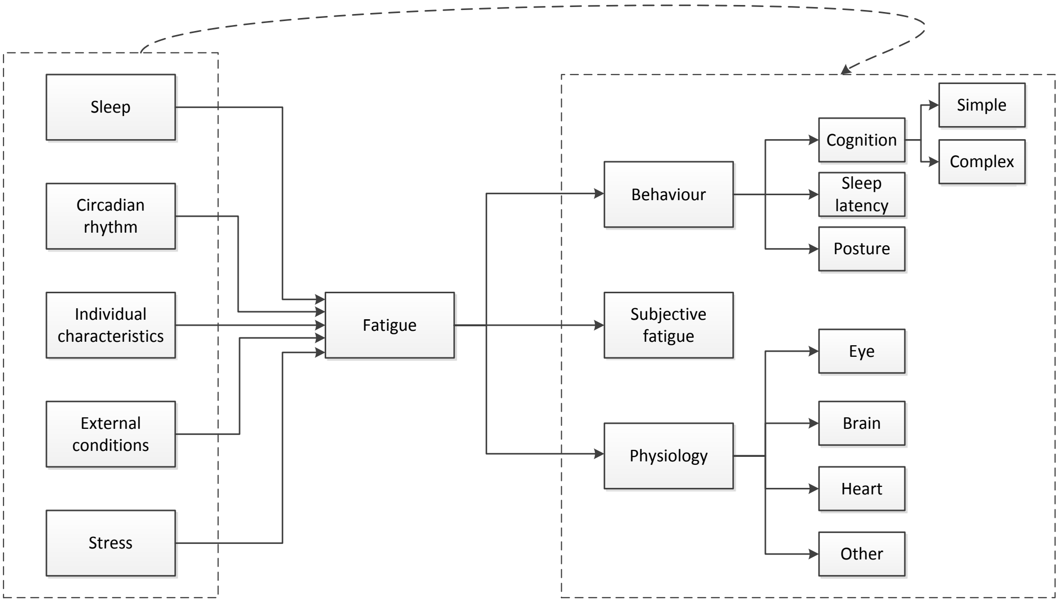


Figure 2.1: Contributors to and expressions of fatigue

Contributors to fatigue

Sleep

The primary physiological method to reduce fatigue is to get sleep. The amount of sleep and the quality of sleep in a relevant period before the actual fatigue

[6] Dependent on the measurement technique different terminology is used, all referring to fatigue; e.g. vigilance, sleepiness.

measurement can be used to make predictions of the level of fatigue and to better interpret results of actual measurements. Because sleep loss can be developed cumulatively, it is important to have information on the amount of sleep and the quality in the days (typically a week) before the measurement.

The most complete objective measure of sleep is polysomnography, but this has the drawback that it is practically difficult to conduct in an operational setting. Information on recent sleep can also be objectively estimated with actigraphy (sensing and recording body movements) and this is therefore most commonly used to obtain objective sleep data. Subjective sleep information can be obtained from sleep diaries for information on day-by-day sleep history and questionnaires. A standardised sleep diary has been developed [86]. Examples of rating scales for sleep quality are the Pittsburgh Sleep Quality Index [57], the Groningen Sleep Quality Scale [62] and the Insomnia Severity Index [85].

Circadian rhythm

Sleep and wakefulness are controlled by an interaction of an output of the circadian pacemaker and a sleep-wake dependent homeostatic process. The circadian pacemaker and the sleep homeostat contribute about equally to sleep tendency and performance [39]. The human circadian system is normally synchronised with the solar day by exposure to daylight. A remarkably tight association between circadian rhythms of sleep propensity, melatonin and core body temperature has been described in humans [63]. Melatonin, core body temperature and light exposure history can be measured to determine if there has been a shift of the circadian rhythm with respect to the normal wake-sleep pattern (e.g. due to time zone crossing or night work). Melatonin can be measured from saliva samples or blood plasma, but obtaining a blood sample is significantly invasive. Core body temperature can be measured non-invasively. Both melatonin and core body temperature are also influenced by other factors (exposure to light, exercise).

Information on the recent 24-hour activity profile can be objectively measured with actigraphy.

Individual characteristics

No two persons are identical and individual differences will affect the influence of fatigue on one’s performance. Inter-individual variation of test results may be caused by different fatigue levels or by these individual characteristics, which can have origins in recent history, genetics, socio-cultural background or combinations thereof. Comparison of older and young adults for instance indicates that the circadian process has a greater adverse effect (measured as subjective sleepiness, calculation test performance and attention) on younger subjects than older subjects [32] and genotype influences susceptibility to the effect of sleep loss on performance [41].

Information on personality traits such as chronotype, age, gender, medical conditions and recent history (sleep, exposure to light, physical activity, caffeine intake, etc.) can be obtained by self-administered questionnaires. The diversity of individual characteristics necessitates a very large data sample and the 30 participants of phase 2 of this study will not likely be sufficient to draw relevant conclusions. Therefore information resulting from analysis of NIN's web based Netherlands Sleep Registry (a web-based assessment tool for extensive insomnia and good sleep phenotyping has resulted in a growing database of, at the time of writing, 13000 people) will be used as well.

External condition

Any measurement may suggest intact vigilance when assessed in a brightly lit cool environment in a person that has recently been physically active and has taken an upright posture. Fatigue may immediately start to increase when a person has to be sedentary, maintaining a sitting or supine posture in a dimly lit warm environment [82]. External conditions such as ambient temperature, lighting, and noise are known to influence the results of fatigue and performance tests [68] [69] [70] [71]. Environmental conditions during the test as well as the subject's recent history of exposure to environmental conditions can be objectively measured for posture, temperature, light and noise.

Stress

Although acute stress response increases arousal and alertness, after prolonged stress exhaustion may be near. The recent history of subjective stress can be assessed using Ecological Momentary Assessment (EMA) [96], also known as Experience Sampling (ES), using smart phones [84].

Expressions of fatigue

Physiology

The most often used physiological measurements in an operational setting are ocular and cardiographic indices. Percentage of Eyelid Closure (PERCLOS) is often mentioned as a reliable indicator of the onset of sleep, but is more suitable for real-time monitoring than for a (short duration) fit-to-fly test because of compensation effects. Saccadic velocity, pupil size and pupil constriction are not under voluntary control and are therefore robust against compensation effects. Measuring these eye metrics is minimally intrusive.

Of the cardiographic indices the Heart Rate Variability (HRV) is cited as a good indicator of sleepiness but HRV is also influenced by other factors such as exercise and digestion following a large meal.

Sleep deprivation disturbs the co-ordinated thermoregulatory responses between the lower and middle part of the body. Skin temperature gradient (i.e. differences between skin temperatures measured at various locations of the body) can therefore be used as an indicator of fatigue.

Subjective fatigue

Subjective information on fatigue can be obtained from subjective rating scales such as the Stanford Sleepiness Scale (SSS), the Karolinska Sleepiness Scale (KSS) or the Samn-Perelli Scale (SPS).

Behaviour

Cognition

There exists a large variety of tests to measure cognitive performance. As a crude categorisation of these tests a model of human-information processing (Figure 2.2) developed by Wickens and Hollands [64] is applied. According to this model there are two main information processing paths:

- A path involving very simple cognitive operations (sensory processing - perception - response selection - response execution);
- A path involving more complex cognitive operations (sensory processing - perception - working memory/cognition - response selection - response execution).

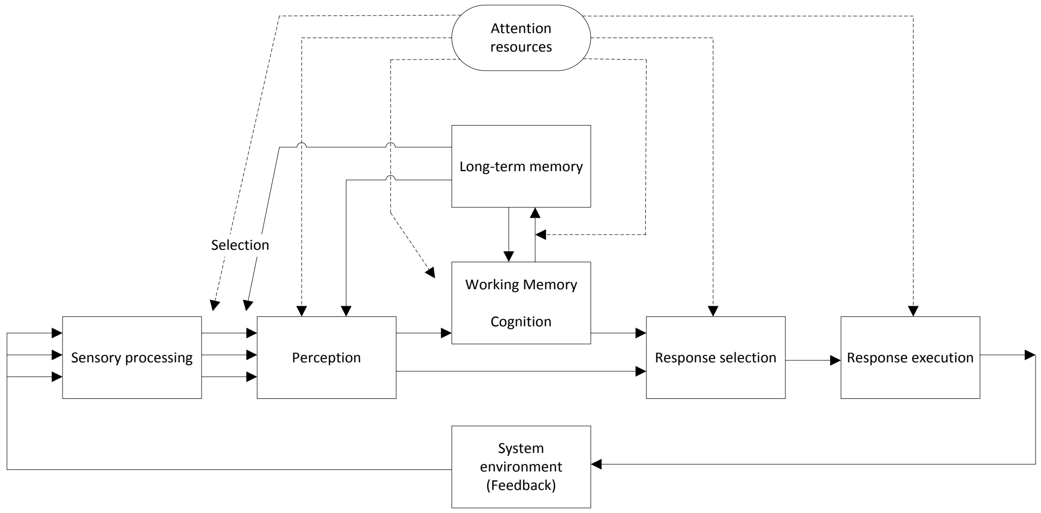


Figure 2.2: Model of human information processing stages [64]

Simple cognitive operations

The most commonly used test for simple cognitive operations is the Psychomotor Vigilance Task (PVT) which is a simple visual reaction time task to test sustained attention. Popularity of the PVT is largely based on its simplicity which allowed early development (i.e. before the introduction of smart phones) of hand-held PVT testers. The Mackworth Clock Vigilance test is also rather frequently used. Due to the more complex visual representation it has primarily been applied in laboratory conditions but with current technology could easily be presented on a smart phone.

Motor skills are tested with tracking tasks. These can be implemented on hand held devices and are therefore popular and quite often used, especially in operational settings. Simulated driving with lane deviation as measure of performance is a similar task but may involve more sophisticated hardware that makes it only practical to perform tests in a laboratory setting. Learning effects can be expected with these types of tests.

Complex cognitive operations

Information processing involving more complex cognitive operations can also be tested in a sophisticated driving simulator (or flight simulator) but the measure of performance is not as straightforward as in a tracking task. The number of threats detected and successfully managed has been used but this requires careful consideration of the scenario such that performance of different individuals can be compared.

A large variety of tests has been developed to assess performance on complex cognitive operations. Several test batteries have been developed that combine a number of tests. The AGARD STRES battery and the Multi Attribute Task (MAT) battery were developed specifically to test pilot performance and are therefore particularly interesting.

Sleep latency

Sleep latency is the time that it takes to fall asleep. The most complete information is obtained with the Multiple Sleep Latency Test (MSLT) or Maintenance of Wakefulness Test (MWT), which uses the same combination of techniques (Electroencephalography (EEG), Electrooculography (EOG), Electrocardiography (ECG) and Electromyography (EMG)) as polysomnography, but similar to polysomnography has the drawback of being difficult to conduct in an operational setting.

Posture

Body posture, head movement and facial expression are often used in real time monitoring systems of driver drowsiness. Similar to PERCLOS, these indices are not considered suitable for a (short duration) fit-to-fly test because of compensation effects.

List of measurement methods originally envisioned

NLR/NIN's proposal for the CAA-UK on pilot fatigue measurement research included a description of a test protocol involving an ambulatory recording period and tests in a human physiology & cognition laboratory. The protocol, as described in detail below, was already approved by the Medical Ethics Committee of the VU University Amsterdam.

Ambulatory period

Participants will be instructed to fill out baseline questionnaires on individual characteristics and vigilance complaints on the website of the Netherlands Sleep Registry.

Using a touch screen mobile device, participants will be queried multiple times with respect to their alertness, fatigue, sleepiness, affect, stress, caffeine intake, and effort to a 3-minute reaction time task. The mobile device will call for responses and task performance at random intervals eight times a day for all days of the ambulatory recording period.

ECG, trunk activity and posture, and chest skin temperature will be obtained from a single recorder on a chest strap with two dry electrodes and a recording device near the plexus. Two small recorders are attached to the upper thigh and the wrist to assess activity, posture and skin temperature. Participants will wear small skin temperature sensors at the finger and foot. A small recorder for ambient temperature, humidity, physical activity and light will be attached to one’s indoor clothing at the level of the chest, a similar recorder to the same location at one’s outdoor coat.

Laboratory tests

The laboratory tests start with a pupil diameter response assessment using infrared Light-Emitting Diode (LED) and digital camera. The pupil contraction response to blue and red light stimulation will be assessed in both a sitting position and a supine position. During light exposure and pupil diameter assessment, participants will perform an auditory reaction time vigilance task, and physiology (skin temperatures, ECG and blood pressure) will be measured.

Subsequently, participants will be subjected to systematic manipulations of posture, skin temperature and light exposure, and will repeatedly perform a 10-minute visual reaction-time vigilance task and answer questions on subjective vigilance (alertness, fatigue, sleepiness, effort required to perform). Other physiological assessments include an ECG, respiratory effort belt, high-density (256 channel) EEG, beat-to-beat finger blood pressure, and skin and core body temperature.

Each participant will be subjected to manipulation blocks in a 2 (light) x 2 (posture) x 2 (temperature) full factorial design. In order to mildly manipulate skin temperature, participants wear a Med-Eng water-perfused suit through which temperature-controlled water is pumped. The light stimulus is provided by four Philips Stratos tube light panels of 1.44 m2 each. The postural manipulations are performed by a computer-controlled commercially available stand-up wheelchair originally designed for physically disabled people.

In NLR/NIN's proposal for the CAA-UK on pilot fatigue measurement research it was mentioned that the psychomotor vigilance task would be complemented by an additional task to be selected after reviewing to quantify more complex cognitive operations. This review was provided in the underlying chapter and the proposed additional task is discussed in the next sections.

Immediately following the laboratory assessments, participants would be subjected to simple computer-based flight simulator investigations using AIRSIM [95] and a subset of participants would be subjected to high-fidelity full-flight simulator investigations using the Generic Research Aircraft Cockpit Environment (GRACE) simulator. Both AIRSIM and GRACE allowed assessment of flight crew performance on a range of divergent skills.

Summary of measurement protocol

Table 2.1 lists the techniques included in the initial measurement protocol.

Table 2.1: Overview of techniques included in the protocol

Element		Measurement technique
Contributors to fatigue	Sleep	Actigraphy
	Circadian rhythm	Actigraphy
	Individual characteristics	Sleep registry Post illumination pupil response
	External conditions	Electrocardiography (ECG) & blood pressure Posture

		Skin temperature Ambient temperature Light Humidity
	Stress	Ecological Momentary Assessment (EMA)
Expressions of fatigue	Simple cognition	Reaction test Psychomotor Vigilance Task (PVT)
	Complex cognition	Flight simulator
	Sleep latency	-
	Posture	-
	Subjective fatigue	Questionnaire
	Eye	Electrooculography (EOG, from EEG)
	Brain	Electroencephalography (EEG)
	Heart	ECG & blood pressure
	Other	Skin temperature gradient

Discussion on possible measurement methods to be added to the list of methods

The proposed measurement protocol included techniques that were candidate for the final set and techniques that were used for the evaluation of the candidates, i.e. it included techniques that met the selection criteria listed as well as techniques that did not meet the selection criteria.

The proposed measurement protocol described in the previous section included many of the techniques that were identified (see Appendix A). However, some of the techniques that were identified were not (yet) included in the protocol. This section reviewed each of the techniques that were not (yet) included in the protocol to determine if they should be added. Techniques that were not included in the lab protocol are listed in Table 2.2.

Table 2.2: Techniques not included in the protocol

Element		Measurement technique
Contributors to fatigue	Sleep	Polysomnography
Expressions of fatigue	Complex cognitive operations	Test battery

	Sleep latency	Multiple Sleep Latency Test (MSLT)
		Maintenance of Wakefulness Test (MWT)
	Posture	Postural sway
		Facial expression
	Eye	Eye closure
		Saccadic velocity
		Pupil constriction latency
	Brain	Functional Magnetic Resonance Imaging (fMRI)
	Other	Skin potential level
		Biomarkers

Sleep

Polysomnography can be applied to determine if subjects are asleep and (in case of sleep) the sleep stage. Because it is not the intention that subjects fall asleep during the lab testing, these techniques were not considered relevant or useful. Information on sleep (quality and quality) in the week before the lab measurements is important but it is not expected that polysomnography would add information to the actigraphy that is already planned.

Complex cognitive operations

The proposed measurement protocol did not yet include a technique to test performance on complex cognitive functions. In setting up the measurement protocol it was envisioned that the PVT would be complemented by a technique to measure complex cognitive functions.

A large quantity of tests has been developed and applied to test subjects' performance on information processing tasks that involve more complex cognitive functions. To determine which (combination of) these tests might be appropriate in the context of the influence of fatigue on flight crew performance it is necessary to have an overview of required flight crew competences. ICAO lists core pilot competencies in the manual of evidence-based training [72]. This overview of competencies and associated behavioural indicators is reproduced in Appendix B. The variety of competencies justifies application of a task battery that comprises relevant tasks. An existing task battery that was specifically developed to assess pilot performance is NASA's MAT battery. There are however a number of arguments against the use of the MAT:

- Flight crew related tasks will already be assessed in the flight simulator experiments (computer-based and high-fidelity research flight simulator);
- The duration of the test is long (30 minutes), a shorter version of the does not produce sufficiently valid results [87];
- Learning effects are expected to influence the results.

Therefore it was proposed to use a simpler task to assess complex cognitive operations. A 3-minute 2-back task is frequently used in research on determinants of genetic and brain-related mechanisms of fatigue. In the 2-back task, subjects repeatedly are shown a letter for a short duration (e.g. half a second) and are required to compare it with the letter presented two trials before. This test has a short duration (3 minutes) and a steep learning gradient (typically only two tests are required to obtain baseline). Therefore we proposed to use a 3-minute 2-back task to assess complex cognitive operations.

Sleep latency

The MSLT and MWT were not considered relevant because it focuses on the ease at which subjects fall asleep. Some people fall asleep very easily even if they are alert and not fatigued. MSLT and MWT were not a candidate for the final set because of the long duration of the test and even for the lab testing the long duration of the tests is a major obstacle.

Posture

Posture was not included because this is under control of the subject and therefore sensitive to compensation effects and is not a candidate for the final set. It is not expected to provide additional information on fatigue and performance that was not measured with techniques already proposed for Task 2.

Facial expression was not included because this is under control of the subject (at least for a short period of time) and therefore sensitive to compensation effects and is not a candidate for the final set. It is not expected to provide additional information on fatigue and performance that was not measured with techniques already proposed for Task 2.

Eye

Eye metrics for assessing effects of fatigue were not originally foreseen in the protocol. However, there is extensive evidence that eye metrics are potential indicators of fatigue, in particular blinks, pupil diameter and constriction latency, pupil saccadic velocity and slow eye movements. Blink rate and pupil size are of limited value for detecting fatigue onset [88]. Regarding the pupil size, there are various influencing factors besides fatigue such as emotion and ambient lighting that play a role, making it difficult to attribute pupil size changes to individual factors [89]. Therefore it is proposed to include saccadic velocity as a measure. Video, infrared and EOG have been used to measure saccadic eye movements [90] [91]. The 256 channel EEG that was part of the protocol included EOG channels so no additional instrument was necessary to be able to measure saccadic velocity.

Brain

Functional Magnetic Resonance Imaging (fMRI) requires a sophisticated scanner and interpretation of results is difficult. Therefore it was not a candidate for the final set. fMRI will only provide information on the location of effects in the brain usually correlated with performance on a task (e.g. PVT [92]). It does not add relevant information that could not already be obtained by EEG.

Other

Skin potential level was not a potential candidate for the final set because of considerable variation in baseline values, both between and within subject. It was not expected to provide additional information on fatigue and performance that was not measured with techniques already proposed for Task 2.

Three types of biomarkers have been identified in literature. The biomarker index that is mentioned in [23] is an indicator of fatigue resulting from exercise and it is unclear if this technique can be used for measuring sleep-related causes of fatigue and time-on-task fatigue. Concentration of Interleukin-6 (IL-6) is influenced by sleep duration and quality, but acute and chronic psychological stress and exercise also increases concentrations of IL-6 [24]. Because of this ambiguity Il-6 is not a candidate for the final set. Melatonin concentration is related to circadian rhythms and fatigue but is also influenced by light exposure and varies significantly between individuals and therefore is too ambiguous to be included in the final set.

Conclusions and recommendations of Task 1

Fatigue is a complex phenomenon with many aspects for which there does not exist a single measure. Therefore, the study considered several aspects of fatigue to obtain a good view of the multidimensional construct. This included contributors to fatigue as well as expressions of fatigue.

Based on a literature search an extensive overview of techniques that could be used to measure contributors to and expressions of fatigue was obtained. These techniques were compared with the originally proposed protocol to determine if relevant elements were not yet sufficiently addressed in the protocol. Based on this analysis it was recommended to add two items to the protocol:

- Analysis of EOG parameters to determine saccadic pupil velocity;
- Inclusion of a 2-back task to assess performance on complex cognitive operations.

CHAPTER 3

Task 3: Analysis of accidents and incidents involving flight crew fatigue

Alfred Roelen[1], PhD

Henk van Dijk[1], PhD

[1] Netherlands Aerospace Centre NLR (Nederlands Lucht- en Ruimtevaartcentrum), Amsterdam, the Netherlands

Introduction

The critical factor for safety is not fatigue, or even sleepiness, but accident risk. Therefore fatigue needs to be related to safety by linking fatigue to (reduced) performance and linking performance to safety risk.

The objective of Task 3 was to define significant flying activities that may be sensitive to fatigue. The result of this task was used to define the scenarios that were used in the flight simulation experiments that were carried out as part of Task 4 of the study.

Structure Chapter 3

First, the research approach is described, followed by the results, and finally the conclusions and recommendations. Details of the accidents and incidents included in the analysis are provided in Appendix C.

Research approach

Overall approach

The approach for identifying safety-related flying activities that may be sensitive to fatigue was to compare accidents and incidents in which fatigue was a contributing factor with accidents where fatigue was not a factor. Sources of data were official accident investigation reports (mostly from the NTSB[7]) and CAA's Mandatory Occurrence Reports (MORs). For each accident, the core pilot competencies (as listed in the ICAO manual of evidence-based training,

[7] The NTSB was used a primary source of data because of the quality of the investigations and the relatively large number of accidents (resulting from the large volume of US air traffic).

see Appendix B) that were affected by fatigue (in the case of fatigue related accidents) or whose substandard performance contributed to the accident (for non-fatigue related accidents) were identified. This required some interpretation by the study team as the nomenclature used by the accident investigation boards does not always strictly match with the competency descriptions provided by ICAO). More than one competency could be allocated to each accident or incident.

In nearly all accidents and incidents, some kind of non-adherence to procedures is present. The study team decided to include 'application of procedures' in the competencies contributing to the accident only if this was primary and not if this was a secondary effect as a result of substandard performance on one of the other competencies, such as decision making.

The total sample consisted of 30 accidents in which pilot fatigue was a factor, 35 incidents in which pilot fatigue was a factor, and 28 accidents in which pilot fatigue was not a factor. A comparison was made of the relative contribution of the each of the pilot competencies to the fatigue related accidents and incidents and the accidents that did not involve fatigue.

Method - US accident selection process

Fatigue related accidents in the US

The NTSB aviation accident database[8] was queried for occurrences between 1 January 1980 and 1 January 2013 involving Part 121 (air carrier) operations in which the word ‘fatigue’ was included in the synopsis or full narrative. This resulted in a set of 206 records. Each of those records was read to determine applicability. Of the 206 records, 180 referred to metallurgical fatigue and are therefore not relevant. In six of the remaining 26 cases it was explicitly mentioned that flight crew fatigue was not a factor, and in one case there was insufficient evidence to determine whether fatigue affected flight crew performance. The remaining 19 records described genuine cases where reduced human performance due to fatigue was a factor in the accident or incident:

- Two involved fatigue of a ground crew member;

[8] The NTSB aviation accident database contains information from 1962 and later about civil aviation accidents and selected incidents within the United States, its territories and possessions, and in international waters.

- Two involved fatigue of an air traffic controller;
- One involved cabin crew fatigue;
- 14 involved flight crew fatigue.

All accidents for which a full NTSB aircraft accident investigation report was available were also reviewed which resulted in six accidents that did not meet the initial search criteria that were added to the sample:

- One case prior to 1980 involving a Part 135 (air taxi) operation (the Downeast Airlines accident [9]) was added to the sample (occurrence date 30 May 1979, i.e. before the time period used in the search) because the accident is believed to be the first case in which the NTSB cited flight crew fatigue as a contributing factor in an aircraft accident investigation report;
- One case involving a Part 129 (foreign air carriers) operation (the Korean Air Accident [12]) was added as this is believed to be the only Part 129 accident involving fatigue for which a full NTSB accident investigation report is available;
- One case involving a Part 135 (air taxi) operation (the East Coast Jets accident [19]) was added as this is believed to be the only Part 135 accident (other than the pre-1980 case mentioned above) involving fatigue for which a full NTSB accident investigation report is available;
- One case involving a Part 121 operation (the Continental Airlines accident [11]) was added as the accident investigation report discusses fatigue related aspects in detail. According to the findings the flight crew’s degraded performance is consistent with the effects of fatigue, but there is insufficient information to determine the extent to which it contributed to the accident. Therefore fatigue is not listed as a causal or contributing factor;

- One case involving a Part 121 operation (the Colgan Air accident [18]) was added as the accident investigation report discusses fatigue related aspects in detail. Although fatigue is not listed as a contributing factor, NTSB chairman Hersman submitted a proposal to the Board to amend the probable cause by adding fatigue as a fifth contributing factor, specifically that the flight crew members' fatigue contributed to the accident because they did not obtain adequate rest before reporting to duty. The Board rejected the amendment 2 to 1;
- One case after 1 January 2013 involving a Part 121 operation (the Asiana accident [21]) was added because a full accident investigation report was available.

The resulting set of US accidents and incidents is listed below:

- Downeast Airlines, DHC-6, Rockland, Maine, 30 May 1979 [9];
- US Air, Boeing 737-200, Pittsburgh, Pennsylvania, 6 July 1986[9];
- Ports of Call Air, Boeing 707, Salt Lake City, Utah, 22 January 1989[10];
- American International Airways, Douglas DC-8, Guantanamo Bay, Cuba, 18 August 1993 [10];
- Continental Airlines, Douglas DC-9, Houston, Texas, 19 February 1996 [11];
- Korean Air, Boeing 747, Guam, 6 August 1997 [12];
- American Eagle Airlines, Saab 340B, Jamaica, New York, 8 May 1999[11];
- American Airlines, McDonnell-Douglas MD-82, Little Rock, Arkansas, 1 June 1999 [13];
- Chataqua Airlines, BAe Jetstream 31, Milwaukee, Wisconsin, 21 January 2000[12];

[9] Information about this accident/incident was gathered from the NTSB website www.ntsb.gov.

[10] Information about this accident/incident was gathered from the NTSB website www.ntsb.gov.

[11] Information about this accident/incident was gathered from the NTSB website www.ntsb.gov.

[12] Information about this accident/incident was gathered from the NTSB website www.ntsb.gov.

- FedEx, Boeing 727, Tallahassee Florida, 26 July 2002 [14];
- Corporate Airlines, BAe-J3201, Kirksville, Missouri, 19 October 2004 [15];
- Delta Connection, Embraer ERJ-170, Cleveland, Ohio, 18 February 2007 [16];
- Pinnacle Airlines, Bombardier CL600-2B19, Traverse City, Michigan, 12 April 2007 [17];
- Mesa Airlines, Bombardier CL600, Hilo, Hawaii, 13 February 2008[13];
- East Coast Jets, Hawker Beechcraft Corporation 125-800A, Owatonna, Minnesota, 31 July 2008 [19];
- Empire Airlines, ATR 42-320, Lubbock, Texas, 27 January 2009 [20];
- Colgan Air, Bombardier DHC-8-400, Clarence Center, New York, 12 February 2009 [18];
- World Airways, Boeing DC-10, Baltimore, Maryland, 6 May 2009[14];
- Delta Airlines, Boeing 767, Atlanta, Georgia, 19 October 2009[15];
- Asiana Airlines, Boeing 777-200ER, San Francisco, California, 6 July 2013 [21].

Fatigue related accidents outside the US

The CAA-UK provided a list of non-US accident and incidents involving flight crew fatigue. These non-US accidents and incidents are listed below:

- Air Algerie, Boeing 737-200, Coventry, UK, 21 December 1994 [1];
- Si Fly, ATR 42-300, Pristina, Kosovo, 12 November 1999 [5];
- Crossair, AVRO 146-RJ100, near Zurich, Switzerland, 24 November 2001 [6];

[13] Information about this accident/incident was gathered from the NTSB website www.ntsb.gov.

[14] Information about this accident/incident was gathered from the NTSB website www.ntsb.gov.

[15] Information about this accident/incident was gathered from the NTSB website www.ntsb.gov.

- Epps Air Service, Canadair Challenger 604, Birmingham, UK, 4 January 2002 [2];
- MK Airlines, Boeing 747-200, Halifax International Airport, Canada, 14 October 2004 [23];
- Loganair, Pilatus Britten-Norman BN2B-26 Islander, Scotland, UK, 15 March 2005 [3];
- JetX, Boeing 737-800, Keflavik, Iceland, 28 October 2007 [22];
- Air India Express, Boeing 737-800, Mangalore, India, 22 May 2010 [7];
- Air Canada, Boeing 767-300, North Atlantic Ocean, 14 January 2011 [8];
- Airnorth, Embraer E170, 125 NM north-west of McArthur River Mine, 10 January 2013 [4].

Method - CAA MORs

The events were taken from occurrences from commercial air transport operations[16] reported to the UK CAA MOR scheme between 1 January 2009 and 1 January 2014. The events have all been classified as having a 'human fatigue/alertness' explanatory factor and excluded reports with the lowest severity ratings (D-Low and E-Non reportable), this was to exclude reports where fatigue was cited but did not lead to any further events. The resulting set contained 35 events. Of the events included in this sample there were only two grade B events (high severity, including serious increase in flight crew workload and damage to aircraft) and the remainder were grade C (medium severity).

The events were then allocated pilot competencies based on the ICAO manual of evidence-based training (see Appendix B). More than one competency could be allocated to each event.

Method - Non-fatigue related accidents

All Aircraft Accident Reports (AAR) and Aircraft Accident Briefs (AAB) published by the NTSB of accidents involving fixed wing Part 121 or Part 135 air transport

[16] Defined as: "An aircraft operation involving the transport of passengers, cargo or mail for remuneration or hire". Annex 6 Part 1, Chapter 1.

that occurred between 1992 and 2010 and in which flight crew performance contributed to the accident were included in the sample.

Results and conclusions of Task 3

Following an accident or incident, it is very difficult to determine if and to what extent deficiencies in flight crew competencies are caused by fatigue. Illustrative is the fact that in a few cases the NTSB board members could not come to an agreement on whether or not fatigue contributed to the accident. The Empire Airlines accident is also an example of how shallow the evidence sometimes is: the fatigue factors for both pilots are equal, both pilots make mistakes, but the captain's errors are (partly) contributed to fatigue while the first officer's errors are contributed to distraction and a lack of experience.

An overview of the results of the analysis of accidents and incidents is presented in Figure 3.1, which showed how often each competency was involved on average per accident or incident. In the total set of 30 accidents and 35 MORs involving flight crew fatigue, and 28 accidents not involving fatigue, all ICAO-defined core pilot competencies were mentioned as being substandard.

The most frequently mentioned competency for accidents and incidents involving fatigue is 'problem solving and decision making'. The overall picture resulting from the analysis of MORs was remarkably similar to that resulting from the analysis of accidents.

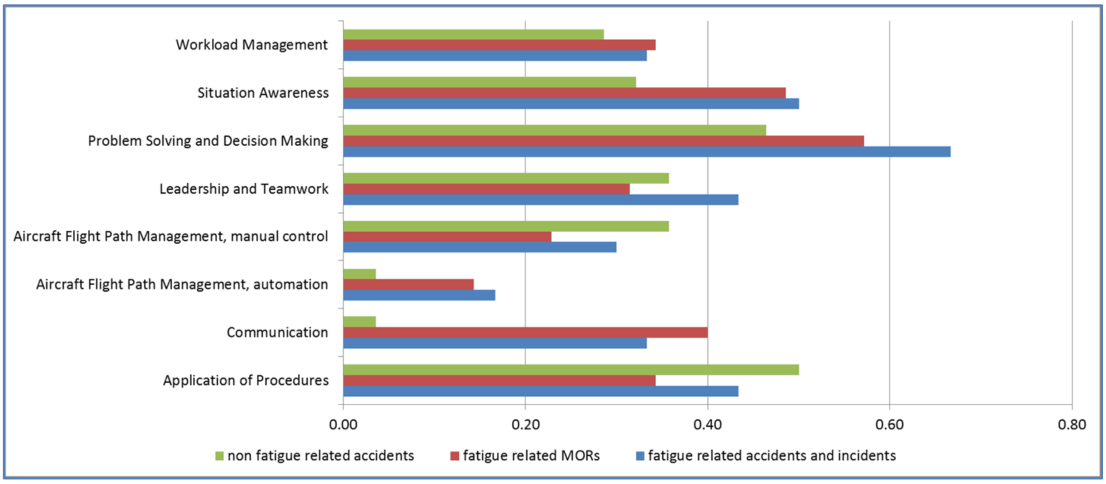


Figure 3.1: Overview of analysis of accidents and incidents

For accidents that were not fatigue related, 'application of procedures' is the most frequently mentioned competency. Comparing accidents that were fatigue related with those that were not showed that 'situation awareness', 'problem solving and decision making', 'aircraft flight path management, automation' and 'communication' contributed more frequently to fatigue related accidents than to accidents that are not associated with fatigue. It was therefore concluded that from the perspective of flight safety, the effect of fatigue on these competencies was most important.

The analysis also showed that all ICAO-defined core pilot competencies were mentioned in fatigue related accidents. Therefore, it was recommended that the scenarios that would be used in the flight simulation experiments of Task 4 represented all core pilot competencies, with emphasis on 'problem solving and decision making'.

CHAPTER 4
Task 2: Measuring fatigue

Bart H. W. te Lindert[1], MSc

Jessica Bruijel[1], MSc

Wisse P. van der Meijden[1], MSc

Eus J.W. Van Someren[1,2], PhD

[1] Department of Sleep and Cognition, Netherlands Institute for Neuroscience (NIN), an institute of the Royal Netherlands Academy of Arts and Sciences, Amsterdam, the Netherlands

[2] Departments of Medical Psychology and Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research (CNCR), VU University and Medical Center, Amsterdam, the Netherlands

Introduction

The objective of Task 2 was to investigate the selected techniques (resulting from Task 1) for effectiveness in measuring fatigue through experimental tests.

Structure Chapter 4

Chapter 4 is structured as follows:

- Methods section, including descriptions of the participants and protocol for the ambulatory study and lab experiment;
- Results of both ambulatory study and lab experiment;
- Summary of the findings.

Research approach

Participants

Study volunteers were recruited by the NLR from an internal database, through emails sent to students of several Dutch flight academies and posters at Dutch airlines. Twenty-eight participants with a frozen Air Traffic Pilot Licence (frozen ATPL, age range 20-44 yrs., 4 female) and eight active duty pilots (ATPL, age range 29-46 yrs., 0 female) were enrolled in the study. All volunteers were nonsmokers and free from medical and psychiatric disorders, ocular pathology and/or colour deficiency, and did not use psychotropic or cardiovascular medication or any medication known to affect thermoregulation, sleep or circadian rhythms. In addition, none of the frozen ATPL participants were extreme morning/evening types, currently on shift work or crossed more than one time zone in the month prior to the onset of the study. Female participants participated between day four and day 12 of the menstrual cycle (mid-follicular phase). All volunteers participated with written informed consent and were compensated financially). The protocol was approved by the medical ethics committee of the Academic Medical Centre of the VU University of Amsterdam.

Participant characteristics

To evaluate if participant characteristics modified the effect of sleep history on sleepiness, participants filled out multiple questionnaires at the online Netherlands Sleep Registry (NSR, www.sleepregistry.org), namely the Duke Structured Interview for Sleep Disorders and Health [1], Action Control Scale [2], Almost Perfect Scale [3], Arousal Predisposition Scale [4], Behavioral Inhibition Scale and Behavioral Activation Scale [5], Hill’s Perfectionism Inventory [6], Hyper Arousal Scale [7], International Personality Item Pool [8, 9], Pre-Sleep Arousal Scale [10], Munich Chronotype Questionnaire [11], Pittsburgh Sleep Quality Index [12], Epworth Sleepiness Scale [13], Fatigue Severity Scale [14], Multidimensional Fatigue Inventory [15], and Insomnia Severity Index [16]. See Appendix F for a detailed questionnaire description.

Experimental protocol

The experiment lasted nine consecutive days. Day one and nine took place at the NIN while day two through eight consisted of an ambulatory recording period at home. On day one the experiment was explained in detail and the experimental task battery to be performed during the experimental lab day (day nine) was practiced twice. The task battery consisted of a 3-minute auditory 2-back working memory task, a 3-minute auditory PVT, a 1-minute saccadic eye movement task and filling out questions on a computer. In addition, two familiarization flights on AIRSIM were performed on a desktop computer. After the practice, participants were given the login credentials to their Dutch sleep registry (www.sleepregistry.org) accounts to fill out questionnaires.

Due to the extent and complexity of the study, the ambulatory recordings and the lab experiment are reported separately and consist of a methods section, followed by the analyses and a brief discussion.

Ambulatory study

Methods

For seven consecutive days participants wore ambulatory sensors, 24-hours a day, only to be taken off for water-related activities and contact sports (Figure 4.1). Sensors located at the non-dominant wrist and middle finger, the chest, the left upper thigh and the inner and outer clothing recorded physiological and environmental signals. Some sensors recorded multiple modalities. Three-dimensional accelerometry (64 Hz) and skin temperature (2 Hz) were recorded at the wrist, upper thigh and chest (Movell and ekgMove, Movisens GmbH, Karlsruhe, Germany). ECG (1024 Hz) was recorded using a chest strap with two dry (ekgMove sensor, Movisens GmbH, Karlsruhe, Germany). Skin temperature of the middle finger of the non-dominant hand was recorded with a custom miniature temperature sensor with an infrared sensor (2 Hz for 20 minutes, starting 15 minutes prior to an alarm until 5 minutes after) and gold-plated contact sensor (every minute, Philips, Eindhoven, the Netherlands). Environmental light exposure was recorded every minute (Dimesimeter, Rensselaer Polytechnic Institute, Troy, NY, USA) and ambient temperature and humidity were recorded every 3 minutes (Hydrochron iButton, Maxim, Sunnyvale, CA, USA) at the level of the chest on both the indoor and outdoor clothing using a custom 3D printed broche.

EMA, which uses repeated collection of real-time data on the behavior and experience of participants in their natural environment (Shiffman et al., 2008), was implemented using MovisensXS software (Movisens GmbH, Karlsruhe, Germany) on a touch screen mobile phone (Nexus 4, LG, Seoul, Korea) running an aftermarket version of Android 4.4 (CyanogenMod, www.cyanogenmod.com). Access to the phone was limited to the experience sampling software with SureLock (42Gears Mobility Systems Inc., Santa Clara, USA) to prevent participants from accidently changing the phone settings. Participants were requested to respond to eight randomly generated alarms between 8AM and 10PM every day (Figure 4.2). The alarm triggered a questionnaire followed by a 3-minute visual PVT [19, 20]. The questionnaire consisted of the Daytime Symptoms in Insomnia Scale [21], subjective thermal sensation, thermal comfort, stress, food and fluid intake and effort to perform the PVT using visual analogue scales. Caffeine and alcohol intake were recorded as numeric input. Sleepiness was recorded using a 9 point anchored version of the KSS [22]. To account for differences in bed and wake times, participants manually started a ninth questionnaire just prior to going to bed and a tenth right after waking up. After waking up, the questionnaire started with the consensus sleep diary [23] followed by the questionnaire and the PVT.

The frozen ATPL participants were randomly assigned to one of three manipulation groups to increase sleep debt and fatigue. During the last four nights of the ambulatory week, one group was instructed to sleep one hour less (REDUCTION, n = 8) than their habitual sleep time, a second group to sleep one hour more (EXTENSION, n = 8) and the remaining participants to maintain their habitual sleep schedule (NONE, n = 12). Active duty pilots were instructed to adhere to their flight schedule and not to change their sleep-wake rhythm.

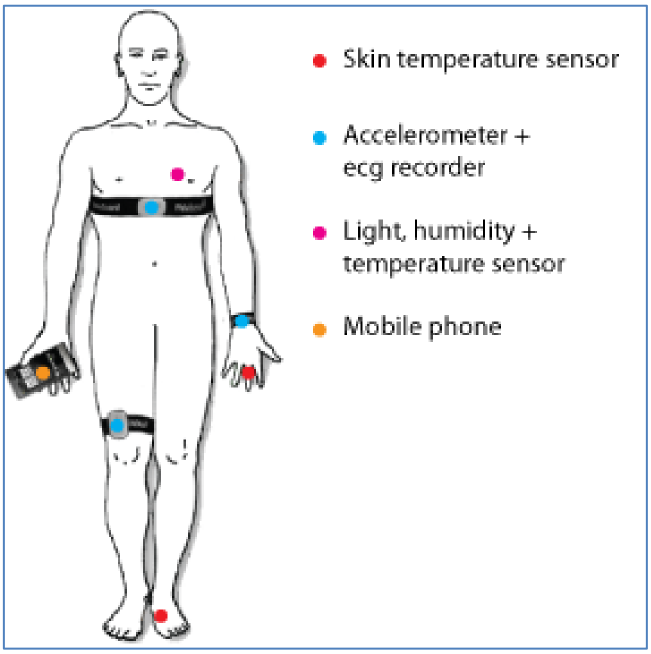

● Skin temperature sensor
● Accelerometer + ecg recorder
● Light, humidity + temperature sensor
● Mobile phone

Figure 4.1: Graphical representation of the sensors worn by the participant. Some sensors record multiple modalities. The phone was used for EMA (see text for details)

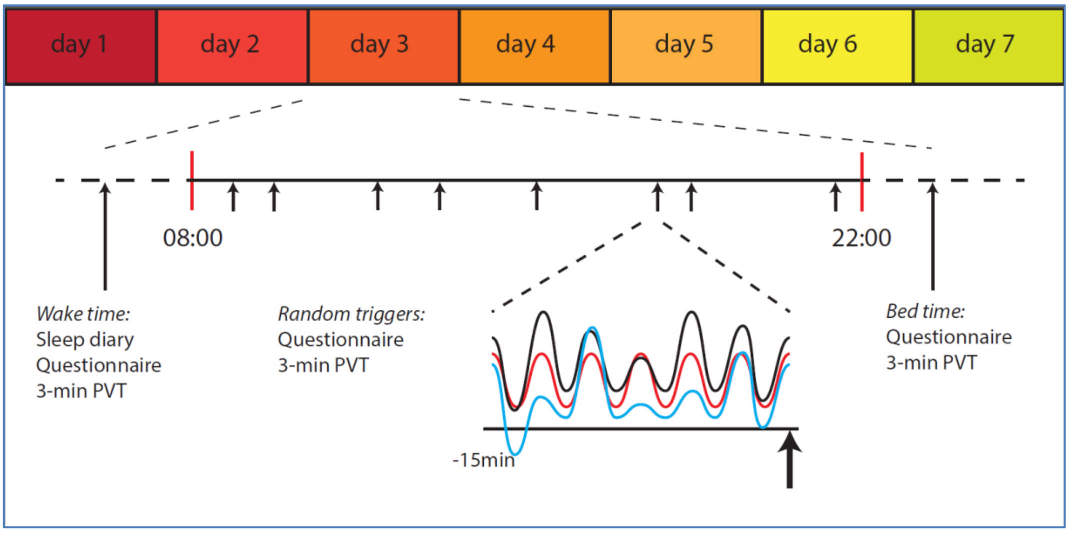


Figure 4.2: Graphical representation of the EMA protocol. Between 08:00 and 22:00 eight random alarms were triggered starting a questionnaire followed by a 3-min visual PVT. In the morning and evening additional questionnaires were manually started by the participant. In the morning a sleep diary was added to request information about the prior night's sleep. The black red and blue time series illustrate how fluctuating features (e.g. skin temperature, activity, ambient light) were extracted from the sensors and averaged over a 15-min interval

Feature extraction

Features were extracted from the sensor data in a 15 minutes window prior to every alarm. For every second in this window posture was classified as either lying, sitting, standing or dynamic movement [24] and the percentages of time spent in every posture were calculated. Sensor selection for data processing (i.e. indoor or outdoor clothing sensor) was based on above noise level activity and/or weighing lux and activity and light exposure in the surrounding 11 minutes of every second. From the selected sensor data, ambient light exposure was calculated as the average of the log10 transformed lux recording across the time window. Ambient temperature, ambient humidity, chest, wrist and thigh skin temperature were all averaged across the entire time window. Gross physical activity was quantified as the average signal vector magnitude of the chest accelerometer [25]. Assumed Time In Bed (TIB) was calculated as the time between lights off and out of bed as reported in the consensus sleep diary. For every alarm the score on the KSS was extracted. Mean speed (SPEED) of the 3-minute visual PVT was calculated as the average of all accurate responses (RT > 100ms). The number of lapses (LAPSES) was calculated as all reaction times greater than 355 ms. Extraction of features from the time series data was performed using custom written Matlab scripts (Matlab 8.3, The Mathworks, Natick, MA).

Data pre-processing

The extracted features were exported to R 3.1.2 [26] for further processing. Prior to statistical analysis measurement errors and outliers (> 3*interquartile range) were removed. Missing data were imputed using predictive mean matching for numerical predictors, proportional odds model for ordinal sleepiness and multi-level normal imputation for PVT features [27] implemented in the mice package for R [28]. Data were then standardized (mean subtracted and divided by the standard deviation) at the group level to allow comparison of regression coefficients across models and dependent variables. Questionnaire data was exported to R and factor scores for each questionnaire were calculated using custom written scripts.

Hypothesis testing was performed using Linear Mixed-Effects regression Models (LMEM) using the lme4 package for R [29]. LMEM is an extension of the linear model, which takes into account the hierarchical structure of the data and dealing with missing data [30]. In addition to the fixed effects covariate terms in the regression, random effects terms can be included, which capture the variance belonging to each level of the hierarchical structure (i.e. participant, day).

To evaluate if participant characteristics modify the effect of sleep history on sleepiness individual LMEM were ran using the KSS [22] obtained during the 1-week EMA protocol as the outcome variable and TIB as a predictor while controlling for time awake (AWAKE). The intercept and slope of TIB coefficients represent the individual baseline sleepiness and the individual sensitivity of sleepiness to TIB, respectively. Both coefficients were compared to scores of the subscales of all questionnaires to explore traits potentially sensitive to sleep disruption.

Effect of time in bed

LMEM analyses were performed to investigate associations between time in bed in the previous night up to 2 nights back (TIB-0, TIB-1, TIB-2 respectively), cumulative time in bed over the last one, two or three nights (TIB-0, TIB-1-0, TIB-2-0 respectively) and each of the dependent variables (KSS, SPEED and LAPSES). Time awake (AWAKE, linear and quadratic) was included as an additional source of variance. LAPSES were included in the first analysis, but preliminary results indicated it to be less sensitive than SPEED in subsequent analyses. The calculation of SPEED includes all the LAPSES and therefore

LAPSES was dropped in the subsequent analyses. For the dependent variables most sensitive to TIB, we explored interactions between TIB and each of the environmental variables (POSTURE, LIGHT, TEMPERATURE, HUMIDITY). TIB, AWAKE as well as interaction terms between TIB and environmental variables were entered as fixed-effects regressors, and random intercepts for each subject and day-within-subject. P-values of the fixed effects regressors were calculated using forward addition [31], i.e. by comparing the model with the added regressor against the model without the regressor. A reference distribution of t-values of the Likelihood Ratio Test statistic (LRT) was created by 1000 bootstrap samples from the fitted distribution under the hypothesis using the pbkrtest package for R [32]. The p-value was calculated by comparing the t-value of the model comparison to the null distribution. The regression coefficients were estimated with maximum likelihood and 95% confidence intervals were estimated using 1000 simulations to compute a likelihood profile and finding the appropriate cutoffs based on the LRT. In LMEM non-significant fixed effects can be due to opposite within- and between-subject effects that cancel each other out [33, 34]. Additional models were therefore run to test whether between- and/or within-subject slopes accounted for the observed effect. The number of observations was kept constant when comparing models.

Results ambulatory study

Data

Two subjects (1 from the REDUCTION group and 1 from NONE group) were removed from the analyses due to an insufficient number of answered alarms (<30%) and unreliable data (e.g. many missed responses on the PVT, unlikely short questionnaire completion time). Preliminary analyses revealed that sleep reduction or extension across the last 3 days was not always achieved and not limited to the experimental groups. Instead of grouping, we therefore analysed the effects of actual time in bed (TIB) during the previous and past two nights. Table 4.1 summarizes the descriptives of the regressors after imputation, but prior to normalization.

Table 4.1: Descriptive statistics for dependent variables and covariates

Variable	Units	Mean	SD	Range
KSS		3.8	2.0	1 – 9
SPEED	s-1	2.9	0.47	0.98 – 4.35
LAPSES		18	12	0 – 53
AWAKE	hrs	7.8	5.3	0 – 22.5

TIB	hrs	7.3	1.3	2.5 – 10.8
LYING	%	23	32	0 – 98
SITTING	%	39	33	0 – 98
STANDING	%	11	16	0 – 89
DYNAMIC	%	26	22	0 – 98
LUX	log10 lux	1.48	0.85	0 – 4.03
TEMP	Degrees C	23.0	4.0	8.7 – 34.8
HUM	%	53	12	20 – 99

KSS, Karolinska Sleepiness Scale; SPEED, mean speed during PVT; LAPSES, number of lapses during PVT; AWAKE, time awake; TIB, time spent in bed; LYING, percentage of time of 15 minutes prior to alarm spent lying; SITTING, time spent sitting; STANDING, time spent standing; DYNAMIC, time spent in dynamic movements; LUX, environmental light exposure; TEMP, ambient temperature; HUM, ambient humidity

To evaluate if participant characteristics modify the effect of time in bed on sleepiness, individual LMEM were ran using KSS as the outcome variable and TIB as a predictor while controlling for time awake (AWAKE). Table 4.2 shows that daytime sleepiness measured throughout the week using EMA is significantly positively correlated with the following fatigue and sleepiness scales: Epworth Sleepiness Scale (ESS), Fatigue Severity Scale (FSS), Insomnia Severity Index (ISI) and the General Fatigue, Reduced activity, Mental Fatigue subscales of the Multidimensional Fatigue Inventory (MFI). Daytime sleepiness is also significantly negatively correlated with Behavioral Inhibition Scale (BIS), Perfectionism Inventory (PI) organization, International Personality Item Pool (IPIP) persistence, IPIP self-directedness and IPIP cooperativeness. This indicates that subjects that show low subjective sleepiness are less prone to anxiety (BIS), show a greater tendency to be neat or orderly (PI organization), have more self-confidence and perseverance (IPIP persistence), are better able to adapt, control and regulate behavior in response to situational changes and in line with personal goals (IPIP self-directedness) and more socially tolerant, empathic, helpful and compassionate (IPIP cooperativeness).

For nearly all subjects the slope of TIB on sleepiness was negative (mean = -0.079, range = -0.204 – 0.008 KSS standard deviation/TIB standard deviation) confirming, as would be generally expected, that sleepiness decreases with increased time in bed. Of all subscales only IPIP self-directedness appears to be significantly positively correlated with individual sensitivity to TIB. Individuals for whom sleepiness is not strongly affected by time spent in bed (flat slopes) reported higher scores on the self-directedness subscale, i.e. people that are more able to adapt, control and regulate behavior in response to situational changes and in line with personal goals [35] are less susceptible to sleep deprivation. Note that the reported p-values are uncorrected for multiple comparisons and should therefore only be used as an exploratory guide to future replication studies in independent samples of airline pilots.

Table 4.2: Descriptive statistics for questionnaires

Questionnaire	Baseline sleepiness (intercept)	Individual sensitivity (slope)
DSISD		
ACS		
Failure vs preoccupation		
Orientation vs hesitation		
APS-R		
Discrepancy		
APS		
BIS/BAS		
BIS	r = 0.48, p = 0.005	
Reward responsiveness		
Drive		
Fun seeking		
PI		
Concern over mistakes		
High standards for others		
Need for approval		
Organization	r = -0.43, p = 0.013	
Perceived parental pressure		
Planfulness		
Rumination		
Striving for excellence		
HAS		
Introspectiveness		
Reactivity		
IPIP		
Novelty seeking		

Harm avoidance		
Reward dependence		
Persistence	r = -0.53, p = 0.002	
Self-directedness	r = -0.54, p = 0.002	r = 0.44, p = 0.01
Cooperativeness	r = -0.41, p = 0.017	
Self-transcendence		
PSAS		
Somatic arousal		
Cognitive arousal		
MCTQ		
PSQI		
ESS	r = 0.43, p = 0.013	
MFI		
General fatigue	r = 0.55, p = 0.001	
Physical fatigue		
Mental fatigue	r = 0.42, p = 0.015	
Reduced motivation		
Reduced activity	r = 0.38, p = 0.031	
ISI	r = 0.40, p = 0.019	
FSS	r = 0.52, p = 0.002	

DSISD, Duke Structured Interview for Sleep Disorders; ACS, Action Control Scale; APS-R, Almost Perfect Scale Revised; APS, Arousal Predisposition Scale; BIS/BAS, Behavioral Inhibition Scale and Behavioral Activation Scale; PI, Hill's Perfectionism Inventory; HAS, Hyper Arousal Scale; IPIP, International Personality Item Pool proxy of the Temperament and Character Inventory; PSAS, Pre-Sleep Arousal Scale; MCTQ, Munich Chronotype Questionnaire; PSQI, Pittsburgh Sleep Quality Index; ESS, Epworth Sleepiness Scale; MFI, Multidimensional Fatigue Inventory; ISI, Insomnia Severity Index; FSS, Fatigue Severity Scale. Only (uncorrected for multiple comparison) significant correlations are shown

Effect of prior night's sleep

LMEM were used to estimate the effect of TIB on subjective (KSS) and objective (SPEED and LAPSES) measures of sleepiness. Time awake (AWAKE, linear and quadratic) was included as an additional source of variance. Figure 4.3 shows the standardized regression coefficient 95% confidence intervals for the effect of cumulative time spent in bed across one, two or three previous nights on KSS, SPEED and LAPSES. The results indicate that KSS and LAPSES decrease and SPEED increases with increased TIB. The magnitude of the TIB coefficient indicates that KSS is the most sensitive to TIB, followed by SPEED and LAPSES. KSS, SPEED and LAPSES are sensitive to TIB during the prior night, the prior 2 nights but not cumulative TIB across the last 3 nights. Since the results were identical for cumulative sleep across one or two nights, TIB of the prior night only was used in subsequent analyses.

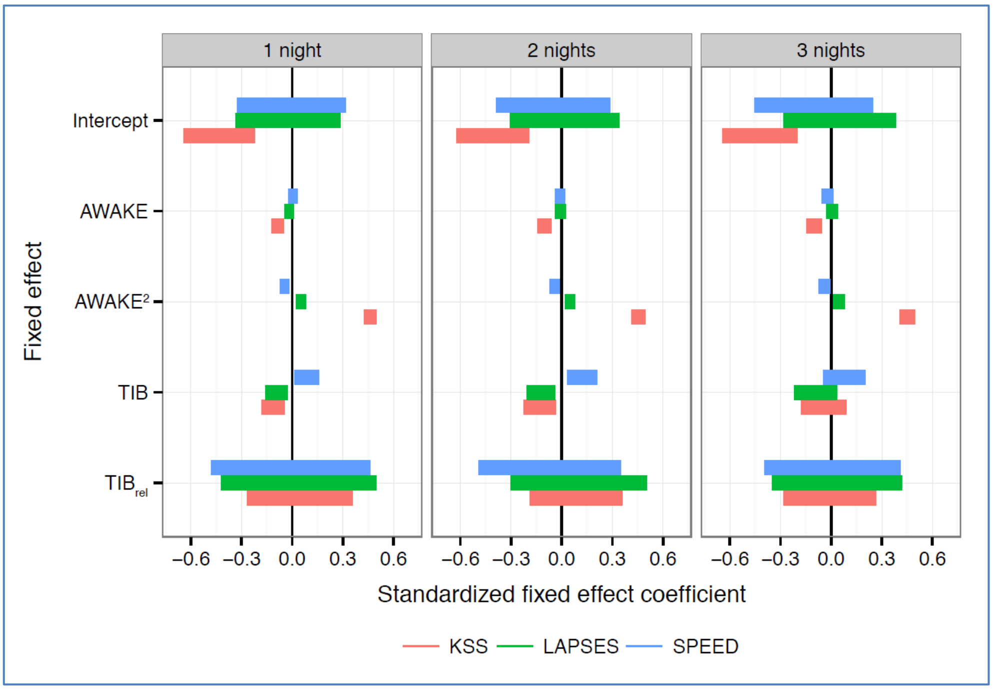


Figure 4.3: 95% CI for standardized regression coefficients estimating the effect of time in bed during the last night (left panel) or the integrated of the previous two (middle) or last three (right) nights on KSS (red), lapses (green) and speed (blue). Horizontal lines indicate 95% CI using 1000 bootstrap simulations. 95% CIs of significant effects do not cross the X = 0 line (black). Coefficients further away from zero indicate greater sensitivity for the corresponding sleepiness measure to variability in TIB. The sign indicates the direction of the effect (i.e. increasing a positive coefficient increases the corresponding sleepiness measure). Smaller ranges indicate more precise predictability. KSS, Karolinska Sleepiness Scale; LAPSES, reaction times > 355 ms; SPEED, mean speed; TIB, within-subject time in bed; TIBrel, between-subject TIB slope relative to TIB; AWAKE, time awake. The slope of the between-subject effect of TIB = TIB + TIBrel. If TIBrel is not significant, the between-subject TIB slope is not significantly different from within-subject slope (i.e. within and between subject effects are similar)

Distal-to-proximal skin temperature gradient

Initial evaluations considered the Distal-to-Proximal skin temperature Gradient (DPG) as a physiological assessment of sleepiness requiring neither a subjective response, nor reaction times. The results revealed that DPG was not sensitive to sleep debt. To investigate if the DPG affected the relationship between sleepiness measures and time in bed, DPG and the interaction between DPG and TIB were added to the model (Table 4.3). The results indicate that subjective sleepiness, but not objective sleepiness, increases with an increases in DPG. DPG does not affect the relationship between TIB and objective or subjective sleepiness.

Table 4.3: Estimated fixed-effect regression coefficients, bootstrapped likelihood ratio test statistics, and random-effect standard deviations for linear mixed-effects models of the distal-to-proximal skin temperature gradient[a]

	KSS		SPEED	
Fixed Effect	Coefficient (MLE ± SE)	Bootstrapped LRT	Coefficient (MLE ± SE)	Bootstrapped LRT
(Intercept)	-0.416 ± 0.105		0.005 ± 0.158	
AWAKE	-0.071 ± 0.019	3.174	0.006 ± 0.014	0.274
AWAKE^2	**0.445 ± 0.020**	485.0[d]	**-0.054 ± 0.014**	11.36[c]
TIB	**-0.108 ± 0.034**	9.216[c]	**0.083 ± 0.036**	5.355[b]
DPG	**0.059 ± 0.024**	6.220[b]	0.028 ± 0.018	2.083
TIB·DPG	0.020 ± 0.036	0.279	-0.047 ± 0.027	3.028
Random-Effect	*SD*		*SD*	
SUBJECT	0.515		0.790	
DAY:SUBJECT	0.202		0.305	

KSS, Karolinska Sleepiness Scale; SPEED, mean speed; AWAKE, time awake; TIB-0, Time In Bed; SITTING, STANDING, DYNAMIC time spent in each posture in the 15 minutes window prior to alarm relative to lying; SUBJECT, participant number; DAY, day of experiment; MLE, Maximum Likelihood Estimator; LRT, Likelihood Ratio Test. The significance of the regressor was tested by comparing the model with the regressor included against the model without the regressor. Model testing was performed using forward addition (i.e. from top to bottom in table). P-values for the fixed effects regressors were calculated using parametric bootstrapping. Significant effects are highlighted in bold font

[a]N = 1476; [b]$p < 0.05$; [c]$p < 0.01$; [d]$p < 0.001$

Posture

We subsequently investigated whether posture or activity during the 15 minutes prior to the measurement would affected the sensitivity of the sleepiness measures to detect sleep debt as indicated by the last night(s) time in bed assessments. To do so, we added posture and its interaction with TIB to the model (Table 4.4). The results indicate that subjective sleepiness decreases the more one spent time sitting, standing or dynamically moving the prior 15 minutes, relative to the reference posture of lying. Sitting and dynamic activity had similar effects on suppressing sleepiness, but more effective than standing. Objective sleepiness increases if the prior 15 minutes are spent sitting, but not standing or moving, at the expense of lying. None of the postures modulates the relationship between TIB and KSS or SPEED. According to these results, the precision of assessing sleep debt and the consequent risk of fatigue-related problems can increase if objective or subjective measures are assessed under standardized postural conditions, but the sensitivity to detect sleep debt does not profit from assessment under specific postural conditions.

Table 4.4: Estimated fixed-effect regression coefficients, bootstrapped likelihood ratio test statistics, and random-effect standard deviations for linear mixed-effects models of posture[a]

	KSS		SPEED	
Fixed Effect	Coefficient (MLE ± SE)	Bootstrapped LRT	Coefficient (MLE ± SE)	Bootstrapped LRT
(Intercept)	-0.393 ± 0.105		-0.011 ± 0.156	
AWAKE	-0.070 ± 0.019	2.428	-0.004 ± 0.014	0.195
AWAKE^2	**0.418 ± 0.019**	487.2[d]	**-0.036 ± 0.014**	11.76[c]
TIB	**-0.108 ± 0.032**	12.79[d]	**0.088 ± 0.036**	6.248[c]
SITTING	**-0.167 ± 0.022**	17.22[d]	**0.048 ± 0.017**	9.113[c]
STANDING	**-0.030 ± 0.021**	7.802[c]	0.002 ± 0.016	0.032
DYNAMIC	**-0.160 ± 0.021**	56.66[d]	0.007 ± 0.016	0.190
TIB·SITTING	0.017 ± 0.030	0.057	-0.011 ± 0.023	0.030
TIB·STANDING	0.049 ± 0.029	3.274	-0.013 ± 0.022	0.424
TIB·DYNAMIC	0.011 ± 0.028	0.145	-0.005 ± 0.021	0.046
Random-Effect	*SD*		*SD*	
SUBJECT	0.515		0.783	
DAY:SUBJECT	0.202		0.313	

KSS, Karolinska Sleepiness Scale; SPEED, mean speed; AWAKE, time awake; TIB-0, Time In Bed; SITTING, STANDING, DYNAMIC time spent in each posture in the 15 minutes window prior to alarm relative to lying; SUBJECT, participant number; DAY, day of experiment; MLE, Maximum Likelihood Estimator; LRT, Likelihood Ratio Test. The significance of the regressor was tested by comparing the model with the regressor included against the model without the regressor. Model testing was performed using forward addition (i.e. from top to bottom in table). P-values for the fixed effects regressors were calculated using parametric bootstrapping (see text). Significant effects are highlighted in bold font

[a]N = 1476; [b]$p < 0.05$; [c]$p < 0.01$; [d]$p < 0.001$

Ambient light

We subsequently investigated whether ambient light exposure (LUX) during the 15 minutes prior to the measurement would affected the sensitivity of the sleepiness measures to detect sleep debt as indicated by the last night(s) time in bed assessments. To do so, LUX and its interaction with TIB were added to the model (Table 4.5). The results indicate that both objective and subjective sleepiness decrease if the prior 15 minutes are spent in bright light. Light exposure did however not modulate the effect of sleep debt (TIB) on sleepiness measures. According to these results, the precision of assessing sleep debt and the consequent risk of fatigue-related problems can increase if objective or subjective measures are assessed under standardized light conditions, although the sensitivity to detect sleep debt does not profit from assessment under specific light conditions.

Table 4.5: Estimated fixed-effect regression coefficients, bootstrapped likelihood ratio test statistics, and random-effect standard deviations for linear mixed-effects models of ambient light[a]

	KSS		SPEED	
Fixed Effect	Coefficient (MLE ± SE)	Bootstrapped LRT	Coefficient (MLE ± SE)	Bootstrapped LRT
(Intercept)	-0.398 ± 0.105		0.015 ± 0.157	
AWAKE	-0.105 ± 0.020	2.428	-0.011 ± 0.014	0195
AWAKE^2	**0.426 ± 0.020**	487.2[d]	**-0.064 ± 0.014**	11.76[c]
TIB-0	**-0.114 ± 0.033**	12.79[d]	**0.086 ± 0.036**	6.248[c]
LUX	**-0.120 ± 0.022**	29.90[d]	**-0.080 ± 0.016**	24.56[c]
TIB-0·LUX	0.014 ± 0.027	1.714	-0.015 ± 0.020	0.551
Random-Effect	*SD*		*SD*	
SUBJECT	0.514		0.785	
DAY:SUBJECT	0.213		0.312	

KSS, Karolinska Sleepiness Scale; SPEED, mean speed; AWAKE, time awake; TIB-0, Time In Bed; LUX, ambient light exposure; SUBJECT, participant number; DAY, day of experiment; MLE, Maximum Likelihood Estimator; LRT, Likelihood Ratio Test. The significance of the regressor was tested by comparing the model with the regressor included against the model without the regressor. Model testing was performed using forward addition (i.e. from top to bottom in table). P-values for the fixed effects regressors were calculated using parametric bootstrapping (see text). Significant effects are highlighted in bold font

[a]N = 1476; [b]$p < 0.05$; [c]$p < 0.01$; [d]$p < 0.001$

Ambient temperature

We subsequently investigated whether ambient temperature (TEMP) during the 15 minutes prior to the measurement would affected the sensitivity of the sleepiness measures to detect sleep debt as indicated by the last night(s) time in bed assessments. To do so, TEMP and its interaction with TIB were added to the model (Table 4.6). The results indicate that both objective and subjective sleepiness are not affected by environmental temperature or its interaction with TIB. Neither the precision nor the sensitivity of assessing sleep debt and the consequent risk of fatigue-related problems increase if objective or subjective measures are assessed under standardized ambient temperature conditions, at least not within the range assessed in the present study (8.7 - 34.8 degrees C).

Table 4.6: Estimated fixed-effect regression coefficients, bootstrapped likelihood ratio test statistics, and random-effect standard deviations for linear mixed-effects models of ambient temperature[a]

	KSS		SPEED	
Fixed Effect	Coefficient (MLE ± SE)	Bootstrapped LRT	Coefficient (MLE ± SE)	Bootstrapped LRT
(Intercept)	-0.422 ± 0.106		0.003 ± 0.157	
AWAKE	-0.070 ± 0.020	2.428	0.010 ± 0.015	0.195
AWAKE^2	**0.450 ± 0.020**	487.2[d]	**-0.050 ± 0.014**	11.76[c]
TIB-0	**-0.122 ± 0.034**	12.79[d]	**0.090 ± 0.036**	6.248[c]
TEMP	-0.033 ± 0.025	1.667	-0.014 ± 0.019	0.499
TIB-0·TEMP	0.004 ± 0.040	0.012	-0.057 ± 0.031	3.403
Random-Effect	*SD*		*SD*	
SUBJECT	0.515		0.785	
DAY:SUBJECT	0.225		0.323	

KSS, Karolinska Sleepiness Scale; SPEED, mean speed; AWAKE, time awake; TIB-0, Time In Bed; TEMP, ambient temperature; SUBJECT, participant number; DAY, day of experiment; MLE, Maximum Likelihood Estimator; LRT, Likelihood Ratio Test. The significance of the regressor was tested by comparing the model with the regressor included against the model without the regressor. Model testing was performed using forward addition (i.e. from top to bottom in table). P-values for the fixed effects regressors were calculated using parametric bootstrapping (see text). Significant effects are highlighted in bold font

[a]N = 1476; [b]p < 0.05; [c]p < 0.01; [d]p < 0.001

Ambient humidity

We subsequently investigated whether ambient humidity (HUM) during the 15 minutes prior to the measurement would affected the sensitivity of the sleepiness measures to detect sleep debt as indicated by the last night(s) time in bed assessments. To do so, HUM and its interaction with TIB were added to the model (Table 4.7). The results indicate that both objective and subjective sleepiness are not affected by environmental humidity or its interaction with TIB. Neither the precision nor the sensitivity of assessing sleep debt and the consequent risk of fatigue-related problems increase if objective or subjective measures are assessed under standardized ambient humidity conditions, at least not within the range assessed in the present study (20 – 99%).

Table 4.7: Estimated fixed-effect regression coefficients, bootstrapped likelihood ratio test statistics, and random-effect standard deviations for linear mixed-effects models of ambient humidity[a]

	KSS		SPEED	
Fixed Effect	Coefficient (MLE ± SE)	Bootstrapped LRT	Coefficient (MLE ± SE)	Bootstrapped LRT
(Intercept)	-0.423 ± 0.106		0.000 ± 0.157	
AWAKE	-0.076 ± 0.020	2.428	0.008 ± 0.014	0.195
AWAKE^2	**0.454 ± 0.020**	487.2[d]	**-0.048 ± 0.014**	11.76[c]
TIB-0	**-0.121 ± 0.034**	12.79[d]	**0.090 ± 0.036**	6.248[c]
HUM	0.015 ± 0.025	0.276	0.009 ± 0.018	0.342
TIB-0·HUM	0.048 ± 0.037	1.694	0.017 ± 0.027	0.377
Random-Effect	*SD*		*SD*	

SUBJECT	0.514		0.784	
DAY:SUBJECT	0.223		0.313	

KSS, Karolinska Sleepiness Scale; SPEED, mean speed; AWAKE, time awake; TIB-0, Time In Bed; HUM, ambient humidity; SUBJECT, participant number; DAY, day of experiment; MLE, maximum likelihood estimator; LRT, likelihood ratio test. The significance of the regressor was tested by comparing the model with the regressor included against the model without the regressor. Model testing was performed using forward addition (i.e. from top to bottom in table). P-values for the fixed effects regressors were calculated using parametric bootstrapping (see text). Significant effects are highlighted in bold font

[a]N = 1476; [b]$p < 0.05$; [c]$p < 0.01$; [d]$p < 0.001$

Active-duty pilots

To test if TIB, DPG, LUX, SITTING, STANDING and DYNAMIC remained effective predictors of objective and subjective sleepiness in an independent sample of active-duty airline pilots (n = 8), the model was run using the aforementioned predictors (Figure 4.4). The results indicate that time awake is a strong predictor of subjective sleepiness and that sitting will reduce subjective fatigue estimates relative to assessment during lying. For objective sleepiness, none of the predictors turned out to be significant.

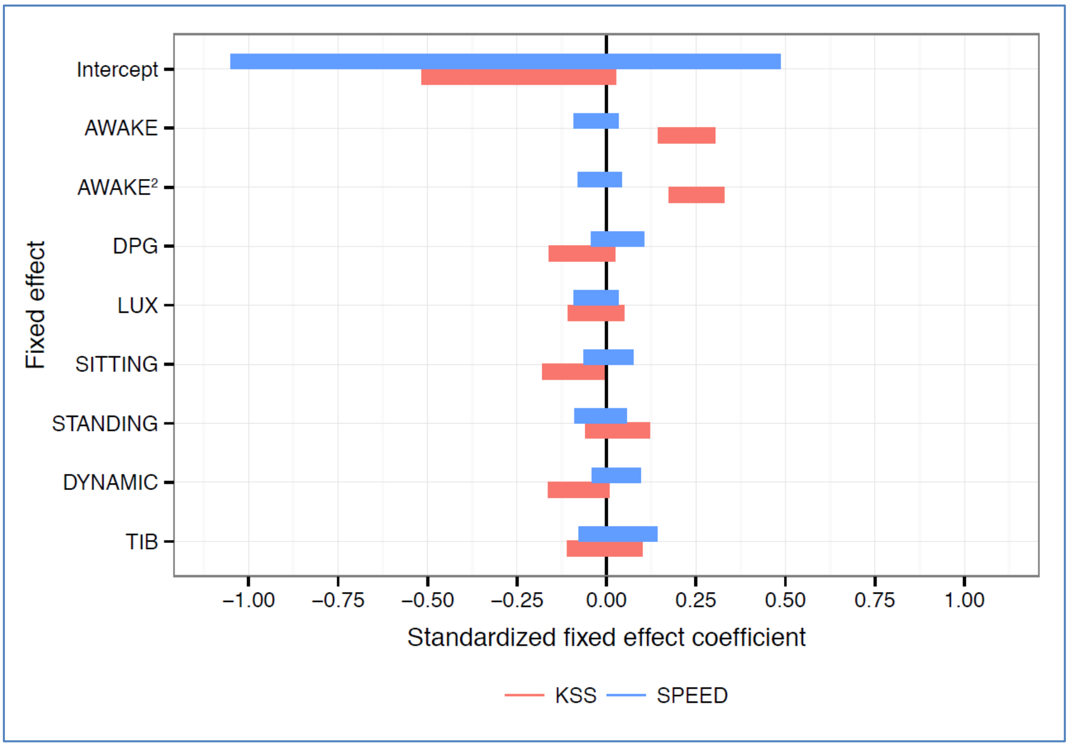


Figure 4.4: 95% CI for standardized regression coefficients for the complete model applied to active duty airline pilot data (N = 363 alarms).

Horizontal lines indicate 95% CI using 1000 bootstrap simulations. The 95% CIs of significant effects do not cross the X = 0 line (black). Larger coefficients indicate greater sensitivity of sleepiness measures to the variable. KSS, Karolinska Sleepiness Scale; SPEED, mean speed; AWAKE, time awake; DPG, Distal-to-Proximal skin Gradient; LUX, ambient light exposure; SITTING, STANDING, DYNAMIC, percentage of 15 minute window prior to alarm spent in each posture; TIB, time in bed

Summary of findings from the ambulatory study

The results indicated that subjective sleepiness is a more sensitive indicator of the time spent in bed the previous night than objective sleepiness. Information about the last one or last two nights provided equally good estimates, which means that in a minimal set of recordings, only the prior night can be considered. Environmental and physiological factors like ambient light, temperature, humidity, skin temperature or posture do not change the sensitivity of subjective and objective sleepiness to detect sleep debt. The precision of subjective and objective estimates of fatigue versus well-restedness due to restricted versus extended TIB, can be boosted by taking into account or standardizing time of day, recent posture, light exposure and the distal-to-proximal skin temperature gradients. Replication in a small sample of active duty pilots results in markedly different results, and warrants future studies in a larger sample of active-duty pilots in an operational setting. A possible explanation could be that these pilots have a semi-chronic exposure to time-zone and sleep-wake schedule shift. This may (1) alter the associations, (2) result in ceiling- or floor-effects, or (3) result in self-selection if such schedules can only be tolerated by persons that experience little effect of time in bed on subsequent fatigue.

Laboratory study

Methods

The aim of this study was to examine the effects of posture, light and skin temperature on vigilance in a controlled laboratory environment. Standardization of the environment in which a future fit-to-fly test is administered can maximize the sensitivity of the test to sleep debt and fatigue. The experiment consisted of one full day at the NIN in Amsterdam. Participants were instructed to refrain from alcohol and caffeine the evening prior to the experimental lab day until the end of the experiment. Participants arrived at the lab at 09:00 (Figure 4.5). The morning session consisted of a 2x2 full factorial design of four 30-minute experimental blocks. Light (<2 lux and >1000 lux) and posture (supine and sitting) were manipulated while skin temperature was allowed to vary freely. The afternoon session consisted of a 2x2x2 full factorial design of eight 30-minute manipulation blocks. In addition to posture and light, skin temperature was manipulated to induce a small and large DPG. To control for first order carry-over effects, the sequence of manipulations was counterbalanced across subjects using a balanced Latin square design [36]. For every eight participants each manipulation was given exactly once in each block, and each manipulation was preceded exactly once by every other manipulation.

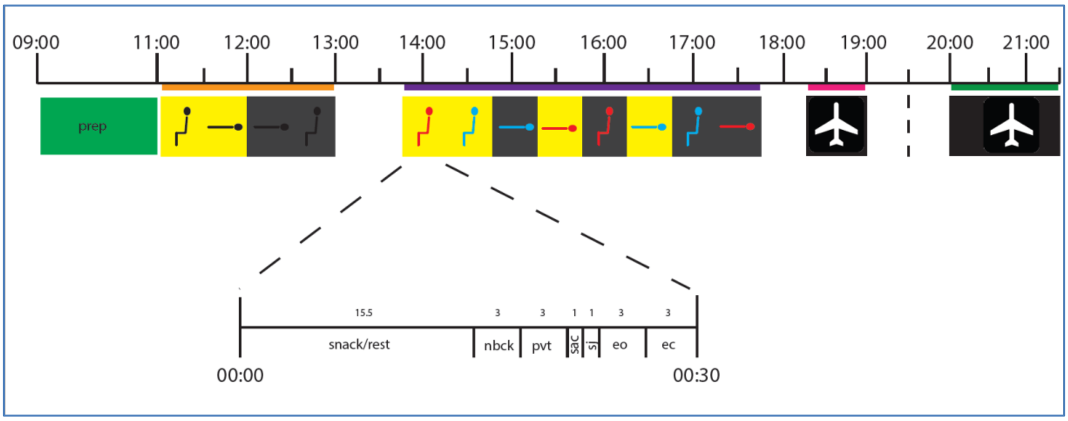


Figure 4.5: Graphical representation of the laboratory experiment. Participants were prepared for the experiment between 09:00 and 11:00. The morning session consisted of bright (yellow) and dim (grey) light exposure and supine and sitting posture (stick figures). In the afternoon, temperature manipulations with a small (red) and large (blue) distal-to-proximal skin temperature gradient were added to the light and posture manipulations. The day ended with an AIRSIM scenario flown on a desktop pc. Only the active duty pilots were then transferred to the NLR for an additional scenario in the GRACE simulator. Each manipulation block lasted 30 minutes and consisted of a baseline period (snack/rest), followed by a 3 minute auditory 2 back task (nbck), a 3 minute auditory PVT, a 1 minute horizontal saccade test (sac), answering some subjective questions (sj) and a 3 minute eyes-open (eo) and eyes-closed resting state EEG recording

Posture, light and skin temperature manipulations

Posture, light and temperature manipulations were pre-programmed and running autonomously using custom-written LabView software (National Instruments Corp., Austin, USA). A computer-controlled wheelchair (C500VS, Permobil AB, Timra, Sweden) was used to alternate posture between sitting and lying.

Light exposure was manipulated using four large (1.44 m2) light panels consisting of 12 28W fluorescent tubes (Philips Lighting, Eindhoven, the Netherlands). Two panels were placed horizontally at the ceiling above and two vertically in front of the participant. During the experiment, only the two panels providing indirect light were used (i.e. the vertical panels during the supine posture, and the horizontal panels during sitting posture). The bright light condition was ~1000 lux and the dim light condition was < 2 lux as measured at the eye level.

Skin temperature was manipulated using a water-perfused thermo suit, consisting of a long sleeved shirt, pants, gloves and socks (Allen Vanguard, Ottawa, Canada). Proximal (shirt and pants) and distal (gloves and socks) skin temperature was manipulated by two separate temperature controlled water-baths (K6KP, Lauda, Lauda-Köningshofen, Germany). The temperature of the water was controlled based on PT-100 thermistors at the entry to the suit (Omega Engineering Ltd., Manchester, UK), to avoid fluctuations in water temperature due to heat loss in the tubing between the baths and the suit. Proximal skin temperature was kept stable by controlling the water entering the suit at 30 degrees Celsius. Distal skin temperature was manipulated by alternating the temperature entering the thermo-suit between 25 and 30 degrees Celsius. The skin temperature manipulations were within the thermo-neutral zone and did not elicit heat or cold defence mechanisms such as sweating or shivering. In the morning session, when skin temperature was not manipulated, participants wore the suit but it was not connected to the water baths.

Measurements

Resting-state high-definition EEG was recorded using a 256-channel HydroCel EEG net (Electrical Geodesic Inc., Eugene, OR) connected to a Net Amps 300 amplifier (input impedance: 200 MΩ, A/D converter: 24 bits) and referenced at the vertex. Physiology was recorded simultaneously from a Polygraphic Input Box (PIB; Electrical Geodesic Inc.). ECG was recorded using Ag/AgCl electrodes (Ambu Neuroline 700, Ambu A/S, Ballerup, Denmark) placed in accordance with the standard lead II configuration [37]. Electrode impedances were kept below 100 kΩ throughout the recording session. Signals were online band-pass filtered between 0.1-100 Hz and digitized at 1000 Hz.

A self-inserted (~10 cm into the rectum) rectal thermistor (D-RA4, Exacon Scientific, Roskilde, Denmark) measured core body temperature. Skin temperature was recorded with 10 neonatal skin thermistors (D-S06A, Exacon Scientific, Roskilde, Denmark) attached to the fingertip of non-dominant (ND) index finger, the palm of both hands; the ND lower arm (dorsal, halfway between elbow and wrist); ND infraclavicular region; at the stomach two cm above the navel stomach, the ND mid thigh at the musculus rectus femoris; the ND mid calf; the palm of both feet. The thermistors were covered with adhesive insulated thermal probe covers (ConMed Corp., Utica, USA) and fixed using medical tape (Fixomull, BSN Medical, Hamburg, Germany).

Room temperature was recorded from a thermistor hanging in the air at the back of the chair. All temperature signals were multiplexed at 2 Hz into a single channel of the PIB (TempMux-1012, Braintronics BV, Almere, the Netherlands). Breathing rate was measured with two inductive plethysmography effort sensors around the chest and stomach (SleepSense, SLP Inc., Tel Aviv, Israel).

Finger blood pressure was recorded continuously from the left ring and middle finger using the Portapres system (Finapres Medical Systems, Amsterdam, the Netherlands) alternating between the two fingers every 30 minutes. Light exposure during the experiment was verified with a custom built photodiode with an opaque filter attached to the EEG net at the forehead of the participant. Pupil diameter was tracked continuously using a custom build eye-tracker sampling at 25 Hz using custom written Matlab scripts and stored for offline analyses in mp4 format.

Task battery

Each block lasted 30 minutes (Figure 4.5). During the first 15.5 minutes of each block, participants were allowed to use the toilet, given a 60kcal snack and 100ml of water and instructed to relax and adjust to the new manipulation. All tasks were implemented in E-Prime 2.0.10.242 (Psychology Software Tools Inc., Sharpsburg, USA) running on a powerful desktop pc. Visual stimuli were presented on an LCD display (G246HL Bbid, Acer, New Taipei City, Taiwan) 80cm from the participant and auditory stimuli were presented using small stereo speakers (Z120, Logitech International S.A., Lausanne, Swiss). The task-battery commenced with a 3-minute auditory 2-back working memory task [38]. A sequence of letters was presented at a 2 second interval. Participants had to press the left mouse button as fast as possible if the letter presented was the same as two letters before (i.e. a target), or the right mouse button if not (i.e. no target).

The 2-back was followed by a 3-minute PVT-B measuring sustained attention [39]. Participants were instructed to respond as fast as possible to an auditory stimulus by pressing a mouse button with the index finger of the dominant hand. The auditory stimuli were presented randomly with an inter-stimulus interval of 1-4 seconds. After the PVT-B, saccadic velocity was measured which is an indicator of fatigue levels and drowsiness [40, 41]. Participants were instructed to follow a target that jumped horizontally at random intervals (800-1400 ms) and angles (10, 15 and 20 degrees left or right. The EOG channels of the EEG were used to measure saccadic velocity. After the saccadic eye movement task participants rated their subjective sleepiness on the 9-point KSS and thermal comfort, sensation and effort required to perform the preceding tasks on 5-point Likert scales. Each block concluded with a 3-minute eyes-open and 3-minute eyes-closed resting state EEG recording. During the eyes open resting state participants were instructed to look at a small white fixation cross on a black screen.

To obtain optimal response timing accuracy, the response mouse was modified to send a digital marker directly to the Net Amps 300 amplifier digitized at 1000 Hz. To obtain optimally timed event markers of visual stimulus presentation, a photodiode attached to the screen recorded the appearance of a white square in the top right corner of the screen at the onset of each stimulus. To obtain optimally timed event markers of auditory stimuli, an AV-tester (Electrical Geodesic Inc.) was attached to the speaker cable and sent a marker to the EEG amplifier at the onset of each auditory stimulus.

Data pre-processing

Preprocessing of physiological data was conducted separately for each participant using the MEEGPIPE toolbox (https://github.com/meegpipe/meegpipe). The full-length recordings were first split into manipulation blocks of 30 minutes using event markers. Average skin temperature at every site was calculated by de-multiplexing the skin temperature signal and averaging across 1 minute intervals. A weighted average was calculated for proximal (0.383 x mid thigh + 0.293 x infraclavicular + 0324 x abdomen [42]) and distal (average of palms of both hands and both feet) skin temperature across the last 6 minutes of each block. From these the distal-to-proximal skin temperature gradient was calculated. The blocks were then further split into the experimental sessions (baseline, 2BACK, PVT, saccade, questionnaire, eyes-open EEG, eyes-closed EEG) using event markers. From the PVT and 2BACK data files the digital stimulus onset and response markers were extracted to obtain millisecond accurate reaction times.

Response times < 100 ms were removed from the analyses and mean speed (1/RT*1000) and lapses (RT > 355ms) was calculated from the correct responses.

Statistical analyses

LMEM were used to estimate the sensitivity of subjective (KSS) and objective (SPEED and LAPSES) measures of sleepiness to controlled changes in distal-to-proximal skin temperature gradients, body posture and light intensity. To account for systematic variability across the day, the model included linear and quadratic terms for the manipulation block number (BLOCK). Study group (CONTROL) was added a covariate to evaluate if ATPLs differed in sleepiness levels from controls (frozen ATPL).

Results of the laboratory study

Data

The control group (frozen ATPL) reported to be less sleepy than the ATPLs, but both groups performed equally well on both PVT measures (Table 4.8, Figure 4.6). PVT performance, but not subjective sleepiness, improved in upright posture compared to supine posture. Larger negative DPGs (those associated with sitting, standing) improve subjective sleepiness and SPEED, but not lapses. Interestingly, light exposure did not affect subjective or objective measures of sleepiness.

Table 4.8: Estimated fixed-effect regression coefficients, bootstrapped likelihood ratio test statistics, and random-effect standard deviations for linear mixed-effects models[a]

	KSS		LAPSES		SPEED	
Fixed Effect	Coefficient (MLE ± SE)	LRT	Coefficient (MLE ± SE)	LRT	Coefficient (MLE ± SE)	LRT
(Intercept)	0.885 ± 0.297		0.585 ± 0.306		-0.569 ± 0.333	
BLOCK	0.059 ± 0.019	2.323	-0.029 ± 0.020	3.975	**0.028 ± 0.015**	7.965c
BLOCK^2	**-0.023 ± 0.007**	8.267[c]	-0.014 ± 0.008	2.648	**0.016 ± 0.006**	5.726b
DPG	**0.176 ± 0.085**	2.589	-0.107 ± 0.093	3.275	**0.067 ± 0.068**	4.113b
BRIGHT	-0.095 ± 0.072	1.726	-0.023 ± 0.079	0.080	0.060 ± 0.058	0.967
UPRIGHT	-0.135 ± 0.073	3.428	**-0.235 ± 0.080**	8.402[c]	**0.278 ± 0.059**	21.14[d]
CONTROL	**-0.887 ± 0.331**	6.43[b]	-0.478 ± 0.341	1.906	0.393 ± 0.376	1.071
Random-Effect	*SD*		*SD*		*SD*	

SUBJECT	0.742		0.760		0.856	

KSS, Karolinska Sleepiness Scale; SPEED, mean speed; LAPSES, number of lapses; BLOCK, block number; DPG; Distal-to-Proximal skin temperature Gradient; BRIGHT, light condition, 0 = dim, 1 = bright; SITTING, posture condition, 0 = supine, 1 = sitting; CONTROL, study group, 0 = ATPL, 1 = frozen ATPL; SUBJECT, participant number; MLE, Maximum Likelihood Estimator; LRT, bootstrapped Likelihood Ratio Test. The significance of the regressor was tested by comparing the model with the regressor included against the model without the regressor. Model testing was performed using forward addition (i.e. from top to bottom in table). P-values for the fixed effects regressors were calculated using parametric bootstrapping (see text for details). Significant effects are highlighted in bold font

$^{a}N = 225$; $^{b}p < 0.05$; $^{c}p < 0.01$; $^{d}p < 0.001$

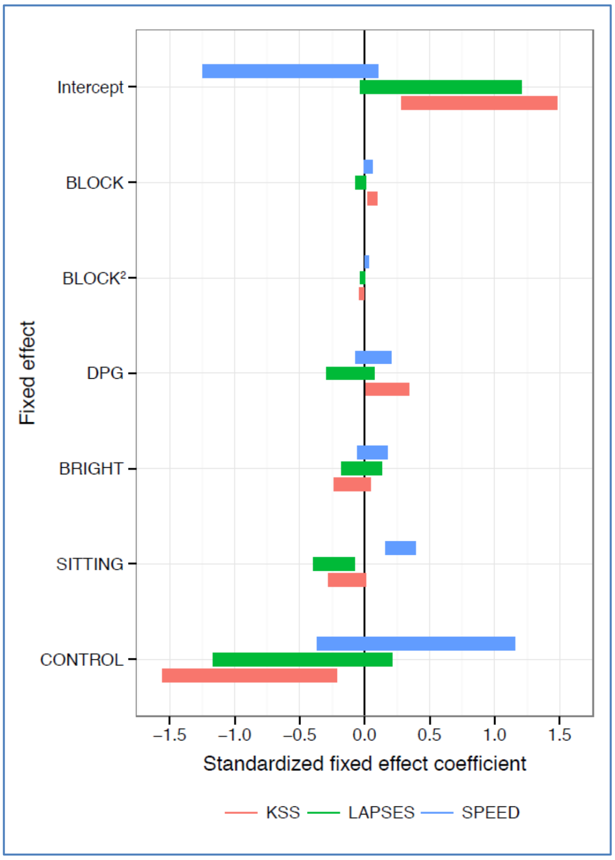


Figure 4.6: 95% CI for standardized regression coefficients for the model presented in Table 4.8. Horizontal lines indicate 95% CI using 1000 bootstrap simulations. The 95% CIs of significant effects do not cross the X = 0 line (black). Larger coefficients indicate greater sensitivity of sleepiness measures to the variable. KSS, Karolinska Sleepiness Scale; SPEED, mean speed; LAPSES, number of lapses; BLOCK, block number; DPG; Distal-to-Proximal skin temperature Gradient; BRIGHT, light condition, 0 = dim, 1 = bright; SITTING, posture condition, 0 = supine, 1 = sitting; CONTROL, experimental group, 0 = ATPL, 1 = frozen ATPL

Discussion on Task 2

The primary objective of the present study was to pursue optimisation of fatigue assessment and prediction using multivariate assessment of physiology, behaviour, performance, environmental exposure, and sleep history. The second objective was to select, from a complete multivariate assessment, the most discriminative minimal dataset that is feasible in practice yet provides a robust estimate of task-relevant current and near future-projected fatigue in pilots.

A data-driven evaluation of the relationship between character traits and individual sensitivity to sleep debt revealed that people with high 'self-directedness' are less sensitive to sleep restriction. Self-directedness, the individual ability to govern behaviour according to situational demands, apparently helps people to be resilient to the effects of restricted sleep. Interestingly, people suffering from chronic insomnia score lower on self-directedness than matched controls without sleep complaints [43]. Individuals with high self-directedness might not be aware of their fatigue, but still manifest performance decrements on objective tasks. A similar analysis using objective measures of sleepiness (lapses, mean speed) should be performed to evaluate this hypothesis. It would be interesting to evaluate whether self-directedness could be used to screen individuals for operational control jobs where it the ability to sustain attention even after sleep restriction is crucial. The ability to govern behaviour according to situational demands is required in aviation and should warrant future investigations in a sample of active duty airline pilots.

As expected, all fatigue- and sleepiness-related questionnaires that were taken once (MFI, ESS, ISI, FSS) correlated positively with the average level of sleepiness obtained throughout the week using EMA. This implied that questionnaires taken only once can be indicative of average fatigue levels throughout the week and can be considered as a simple means to estimate fatigue.

The ambulatory study replicated previous laboratory study results, which indicated that subjective sleepiness is more sensitive to sleep(-disruption) than objective measures of sleepiness [44-46]. However, in an operational context subjective reporting may be sensitive to the outcome desired by the subject who is queried. In such cases subjective reports should be complemented by objective assessments. If the brief PVT is used for this purpose, the average speed is a better alternative than the number of lapses. This could be due to the fact that mean speed also incorporates the response times classified as lapses (i.e. RT > 355 ms), but in addition integrates reaction time shifts across the whole range from the fastest to the slowest.

The multivariate assessment also revealed that recording sleep, time awake, posture, light exposure and skin temperature all provide relevant information for estimating momentary levels of sleepiness. The extent to which these variables affect sleepiness depends on the metric used to measure sleepiness. Of all these variables, time awake was the most potent predictor, followed by posture, light exposure, time in bed and skin temperature. Environmental temperature and humidity did not appear to significantly affect performance in spite of the fact that a wide range of humidity was covered in the assessments. Previous studies have reported increased sleepiness with increasing environmental humidity rather than temperature [47]. Although fatigue and sleepiness are frequently attributed to sleep disruption, our findings indicated even stronger effects of posture and light exposure than time in bed. These variables should thus not be ignored. Whereas previous laboratory studies have shown that each of these factors individually affects vigilance [6, 9, 10], the present study was the first to assess their combined effects. The findings indicated that combined measurements are desirable if not essential to optimize estimates of current and predicted fatigue in an operational setting.

The results of the lab study indicated that supine posture and small DPG can affect sleepiness in as little as 30 minutes, although, again, the results differ per metric. Supine posture affected sleepiness most pronouncedly, followed by a small DPG and a later time of day. Supine posture moreover induces a smaller DPG, thus amplifying its effect on sleepiness. In upright posture cerebral perfusion and cardiac output is maintained by vasoconstriction of the extremities to avoid blood pooling due to gravity. In a supine posture, the gravitational gradient across the body is removed and leads to vasodilation of the extremities, increasing blood flow and thus skin temperature. The increase in distal skin temperature decreases its difference with the usually higher proximal skin temperature and thus reduces the DPG. The implication is that fit-to-fly tests should consider or control, at a minimum, the time of day, skin temperature and posture when conducting the test.

It should be noted that although a substantial subset of the features recorded during the study has been analysed at this stage, our unprecedented comprehensive multivariate approach has made more data available. Table 4.9 summarizes the foreseen analyses on additional variables.

Table 4.9: Additional variables

Variable
2-back task
Blood pressure
HRV
Pupillometry (e.g. diameter, diameter oscillations)
Saccadic eye-movements (e.g. response time, movement speed)
High-density EEG features (e.g. event-related potentials, spectral properties, granger causality, network connectivity)

In conclusion, the results from this study clearly showed that optimisation of fatigue assessment and prediction in an operational setting should consist of a multivariate assessment of physiology, behaviour, performance, environmental exposure, and sleep history. Fit-to-fly tests should consider standardization of the recent and current environment of testing to maximize the precision and sensitivity of the estimated levels of fatigue.

CHAPTER 5
Task 4: Fatigue in relation to flying performance

Alfred Roelen[1], PhD

Henk van Dijk[1], PhD

Frederik Mohrmann[1], MSc

Edzard Boland[1], MSc

[1] Netherlands Aerospace Centre NLR (Nederlands Lucht- en Ruimtevaartcentrum), Amsterdam, the Netherlands

Introduction

The objective of Task 4 was to relate fatigue to flying performance. Using the fatigue measurement techniques derived from Task 2 flying performance in a simulator was related to fatigue. Flight performance data was used to support analysis of performance of pilots having differing levels of fatigue.

Structure Chapter 5

Chapter 5 is structured as follows:

- Research approach of both AIRSIM and GRACE experiments, including scenarios, events and performance indicators;
- Data recording plan of both AIRSIM and GRACE experiments, including pilot performance calculations and results;
- Results and conclusions of the Task 4 experiments.

Research approach

Two flight tests were used to assess performance on pilot competencies. Both flight tests were set-up to replicate a flight in a Boeing 747-400. During the experiments, several events were triggered to test specific pilot competencies.

All participants in the study population (8 active duty pilots and 24 non-active duty pilots) were subjected to a flight test in the AIRSIM desktop flight simulator immediately after completion of the lab protocol (Task 2, see Chapter 4). The test started with a short familiarisation session. The familiarisation session took the participant through all elements of the controls, instruments and systems displayed on the computer screen. There was also the opportunity for the subject pilots to ask questions. In addition to the familiarisation, the participants had received a preparation briefing (Appendix D) one week prior to the experiment date. The flight test itself took 30 to 45 minutes to complete. The test was completed by a debriefing.

The 8 active duty pilots were also subjected to a flight test in the GRACE simulator immediately after completion of the AIRSIM flight test. The test started with a familiarisation that allowed the participant to become familiar with the controls, general behaviour of the simulator and to ask questions. In contrast to the AIRSIM test, the GRACE test was a two-pilot operation. The subject pilot was accompanied by a project pilot who was part of the research team. The task distribution between the subject pilot and the project pilot was explained during the pre-flight briefing. The test was completed by a final debriefing.

Method - AIRSIM

AIRSIM is a desktop research simulator that in this experiment represents a Boeing 747-400. As shown in Figure 5.1, the simulator consists of a computer screen representing a selection of instruments (primary flight display, navigation display, clock, main engine instruments and Flight Management System (FMS), control levers (throttles, flaps, landing gear and speed brakes) and a Controller Pilot Datalink Communication (CPDLC) window for receiving Air Traffic Control (ATC) instructions. Flight control inputs are given via a joystick and a mouse is used for manipulating the buttons, dials and switches. Figure 5.2 shows the entire simulator set-up.

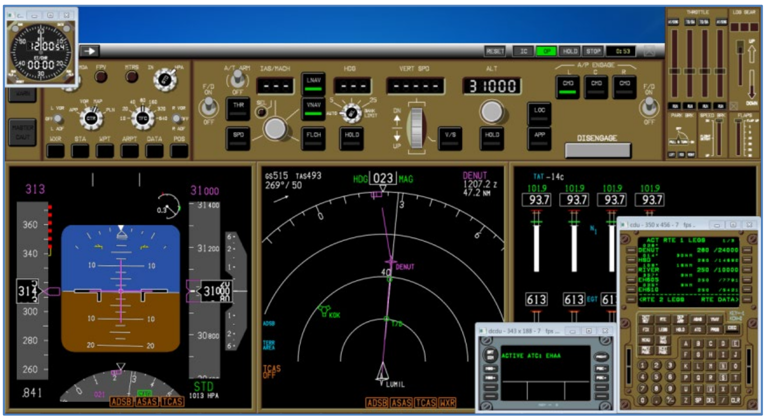


Figure 5.1: Screenshot of the AIRSIM user interface

The scenario used for this experiment represented the last phase of a flight from Rio de Janeiro to runway 06 at Amsterdam Airport Schiphol. The experiment started approximately 20 minutes prior to the expected landing. At the start of the experiment the route was already programmed in the FMS and all relevant autopilot and flight director modes were operational.

The scenario started overhead of waypoint DENUT southwest of Amsterdam Airport Schiphol. The flight continued along the RIVER2A-transition towards SOKSI, after which the aircraft was guided by ATC radar vectors.

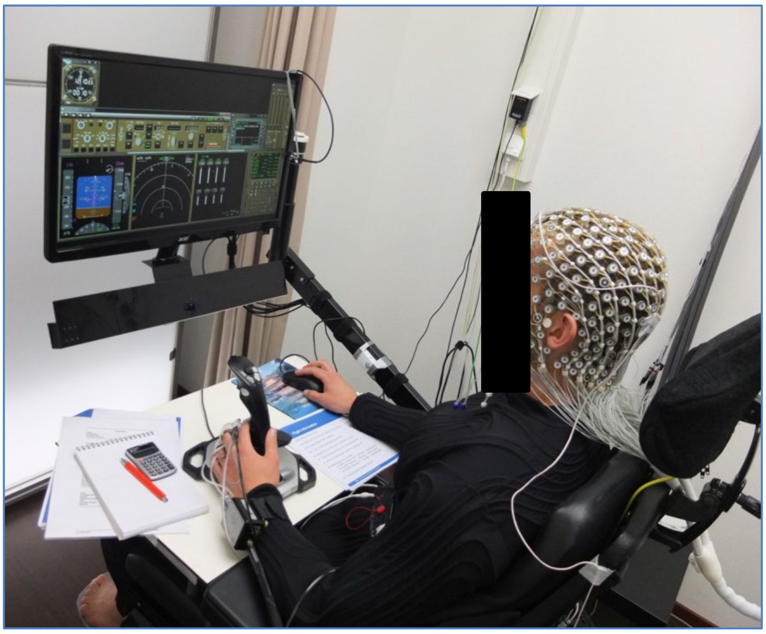

[redacted]

Figure 5.2: Participant flying the familiarisation scenario in AIRSIM

Several events and flight phases in the scenario were used to measure performance on specific pilot competencies. A summary of the AIRSIM and pilot performance measurements is provided in Figure 5.3 and in Table 5.1.

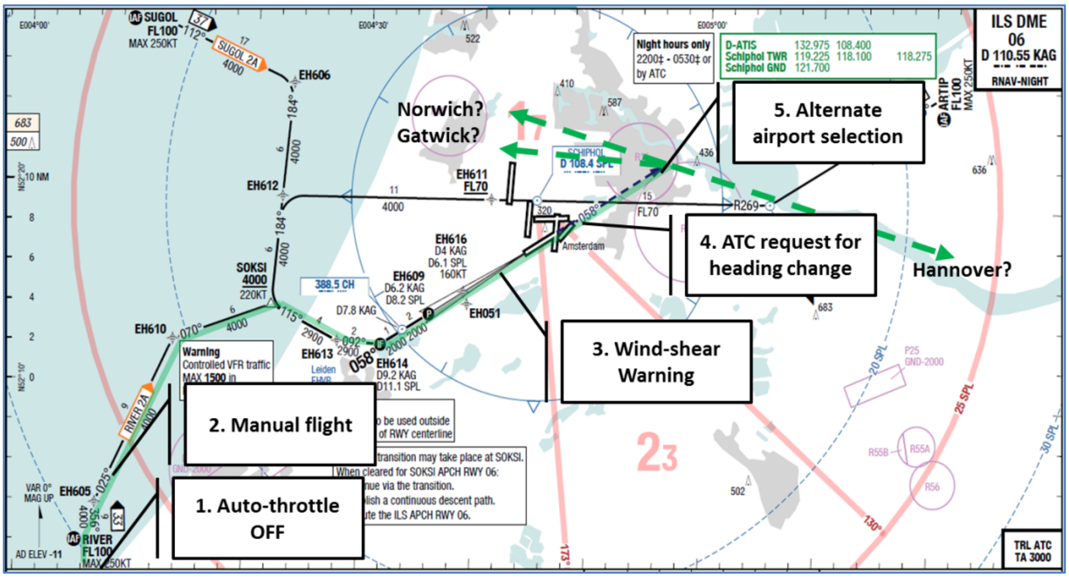


Figure 5.3: AIRSIM flight path and events

Table 5.1: AIRSIM events

Event 1 - Auto-throttle 'OFF' detection	**Timing**	**Performance indicator**
At the start of the experiment the Auto-Throttle (A/T) is off	Initial Condition (IC)	Time to A/T on
Desired behaviour		**Competency**
The participant scans the instruments immediately upon start of the experiment and detects that the A/T is off and should be on		SA; aircraft flight path management, automation

Event 2 - Manual flight	**Timing**	**Performance indicator**
Manual approach from waypoint RIVER towards runway 06	Aircraft is overhead of RIVER	Lateral and longitudinal deviation from FMS flight path; deviation from FMS target speed
Desired behaviour		**Competency**
The participant disconnects the autopilot and A/T overhead of waypoint RIVER and follows the flight director instructions smoothly and accurately. Throttle is adjusted to match the FMS target speeds for the given section		Aircraft flight path management, manual control

Event 3 - Wind-shear	**Timing**	**Performance indicator**
Sudden shift in wind (direction and speed) triggers the wind-shear warning: aural via a spoken "WIND-SHEAR" (3x) and visually via the text "WIND-SHEAR" on the Primary Flight Display (PFD)	Aircraft descents below 400 ft [≡ t(W/S)]	(1) thrust > 80% in 4.5 s, (2) pitch angle > 12 degrees in 12.5 s
Desired behaviour		**Competency**
The participant recognises the situation and immediately executes a wind-shear escape manoeuvre. More than 80% thrust is applied within 4.5 s and the aircraft is rotated to 12 degrees pitch within 12 s		Problem solving and decision making

Event 4 - ATC request for heading change	**Timing**	**Performance indicator**
7 s after activation of the wind-shear	t(W/S) + 0:07	Aircraft heading angle

warning, ATC issues a heading change instruction to 150 degrees and, after a pause of 33 s, to 360 degrees	t(W/S) + 0:40	

Desired behaviour	**Competency**
The participant ignores the heading change instruction until the wind-shear escape manoeuvre has been completed. After the manoeuvre is complete at 2,000 ft altitude, the participant is free to execute the heading change. The second heading change to 360 degrees is executed without delay	Problem solving and decision making; workload management

Event 5 - Alternate airport selection	**Timing**	**Performance indicator**
ATC issues a message that Amsterdam Airport is closed. The subject pilots have to decide to which of three possible alternate airports Norwich, Gatwick and Hannover) the flight will divert	t(W/S) + 1:40	Choice of alternate

Desired behaviour	**Competency**
Participant takes into account the relative distances to the alternates and the given wind conditions and calculates that Hannover is the only viable alternate	Problem solving and decision making

Method - GRACE

GRACE is a research simulator featuring a 6-axis moving platform, a 120 degree visual system and a fully enclosed generic cockpit representing a modern airliner (Figure 5.4). The flight control hardware can be swapped in order to represent different types of aircraft.

Figure 5.4: GRACE generic research flight simulator

In this experiment, GRACE was configured to represent a Boeing 747-400. The simulated flight was executed with two pilots: the subject pilot and a confederate pilot who was part of the research team. ATC communication was provided from the control room by one of the researchers. The experiment set-up is shown in Figure 5.5.

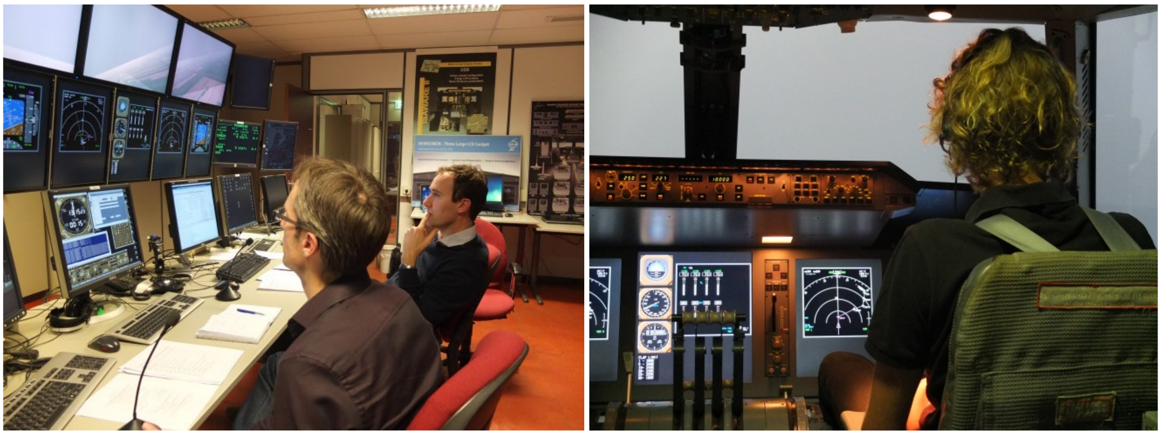


Figure 5.5: GRACE experiment set-up with control room (left) and cockpit (right)

The scenario used for this experiment represented the last phase of a flight from Los Angeles to runway 06 at Amsterdam Airport Schiphol. The experiment started with the subject pilot as the pilot not flying and the confederate pilot as the pilot flying.

The start position of the scenario was at Flight Level (FL) 150 about 45 NM north-east of Amsterdam Airport Schiphol approximately halfway along the MOLIX2A Standard Terminal Arrival Route (STAR). The flight continued toward point SUGOL, after which the aircraft was guided by radar vectors provided by ATC. The duration of the experiment was approximately 30 minutes.

The scenario contained eight measurements. An overview of the measurements is given in Figures 5.6 and 5.7. The triggers, corresponding pilot competencies, performance indicators and desired behaviour are stated in Table 5.2.

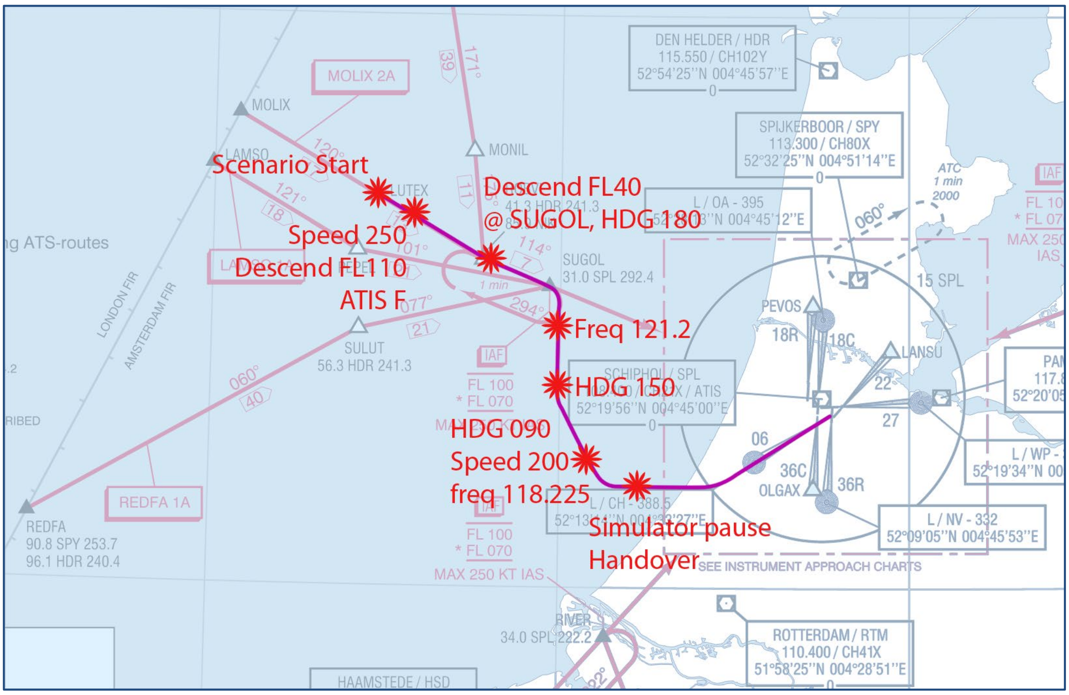
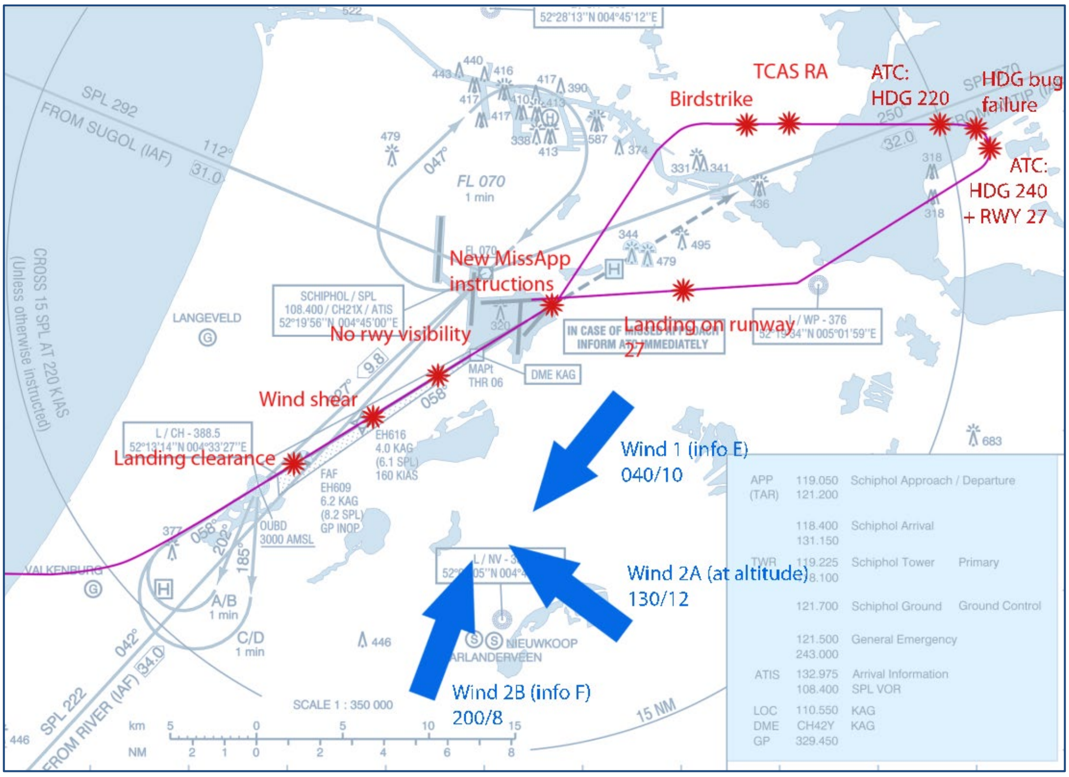


Figure 5.6: GRACE flight path and events pre-handover



APP (TAR)	119.050 121.200	Schiphol Approach / Departure	
	118.400 131.150	Schiphol Arrival	
TWR	119.225 118.100	Schiphol Tower	Primary
	121.700	Schiphol Ground	Ground Control
	121.500 243.000	General Emergency	
ATIS	132.975 108.400	Arrival Information SPL VOR	
LOC DME GP	110.550 CH42Y 329.450	KAG KAG	

Figure 5.7: GRACE flight path and events post-handover

Table 5.2: Events in GRACE

Event 1 - RT1	**Timing**	**Performance indicator**
ATC: "NLR 714, Information Foxtrot now current. Reduce speed to 250, then descend to FL110. Expect SUGOL at FL100 or below. Expect radar vectors"	IC + 1:00	Correct read-back of all items (clearance); listening to updated Automatic Terminal Information Service (ATIS; Information Foxtrot)
Desired behaviour		**Competency**
Participant accurately repeats the instructions in the correct order. ATIS updates are listened out immediately. Participant notices weather changes		Communication (clearance); SA (Information Foxtrot)

Event 2 - Non-pertinent conversation	**Timing**	**Performance indicator**
Confederate pilot starts a non-pertinent discussion with the participant violating the sterile cockpit rule	Upon passing waypoint SUGOL	Stopping the conversation
Desired behaviour		**Competency**
Participant cuts the conversation short		Leadership

Event 3 - RT2	**Timing**	**Performance indicator**
ATC: "NLR 714, Descend and maintain FL40, after SUGOL fly heading 180"	Upon reaching FL110 + 0:15	Correct read-back of all items; noticing erroneous TL/FL
Desired behaviour		**Competency**
Participant accurately repeats the instructions in the correct order and notices the TL/FL error (FL40 should be 4,000ft)		Communication; SA

Event 4 - RT3	**Timing**	**Performance indicator**
ATC: "NLR 714, contact approach on 121.2"	Upon passing waypoint SUGOL	Correct read-back of all items
Desired behaviour		**Competency**
Participant accurately repeats the instructions		Communication

Event 5 - RT4	**Timing**	**Performance indicator**

ATC: “NLR 714, Fly heading 150”	IC + 8:00	Correct read-back of all items
Desired behaviour		**Competency**
Participant accurately repeats the instructions in the correct order		Communication

Event 6 - RT5	**Timing**	**Performance indicator**
ATC: “NLR 714, Reduce speed to 200 knots, fly heading 090. Contact Tower on 118.225”	IC + 10:20	Correct read-back of all items; noticing erroneous frequency
Desired behaviour		**Competency**
Participant accurately repeats the instructions in the correct order and notices the erroneous frequency (should be 119.225)		Communication; SA; leadership

Event 7 - Simulator pause	**Timing**	**Performance indicator**
Simulator freezes, displays black-out and participant is requested to reproduce the current fuel mass, distance to threshold, local QNH, cloud-base altitude, and wind speed When the simulator restarts, the participant is pilot flying and the confederate pilot is pilot not flying. The participant is instructed to perform a manual approach to runway 06	After last ATC instruction + 0:15 - 0:20 (≡ t(H))	Relative difference from actual values
Desired behaviour		**Competency**
Participant reproduces the parameter values within a reasonable margin		SA

Event 8 - ATC informs of nearby traffic	**Timing**	**Performance indicator**
ATC informs of nearby traffic (helicopter). The pilot asks the participant to repeat the message	t(H) + small margin	Correct memorisation of all items
Desired behaviour		**Competency**
Participant is vigilant and can correctly repeat all items from the warning (type, heading, altitude and direction)		Communication

Event 9 - Manual flight to 06	**Timing**	**Performance indicator**
Manual approach from SOKSI towards runway 06	Upon interception of the glideslope	Lateral and longitudinal deviation from FMS flight path; deviation from FMS target speed
Desired behaviour		**Competency**
Upon glideslope interception, the participant disables the autopilot and enables the approach mode. The participant follows the flight director instructions smoothly and accurately. Throttle is adjusted to match the FMS target speeds for the given section		Aircraft flight path management, manual control

Event 5 - Go-around decision	**Timing**	**Performance indicator**
When the aircraft descends below 400 ft, visibility decreases to values below minima	Upon dropping below 400 ft agl	Initiation of go-around; height of go-around decision
Desired behaviour		**Competency**
Participant initiates go-around before reaching the decision height at 190 ft agl		Problem solving and decision making; application of procedures
Event 6 - Birdstrike	**Timing**	**Performance indicator**
A sudden loud noise and is accompanied by a rapid raise of the temperature indication of engine 3, implying a birdstrike. While the crew executes the non-nominal checklist ENGINE LIMIT/SURGE/STALL, they are interrupted by a Traffic Collision Avoidance System (TCAS) alert that escalates into a Resolution Advisory (RA). The project pilot skips an item after resuming the checklist	Upon reaching 3,000 ft and heading 090 + 0:30	Execution of memory items; identification of correct engine; selection of correct non-normal procedure ENGINE LIMIT/SURGE/STALL; immediate and correct response to TCAS RA; identification of missing checklist step
Desired behaviour		**Competency**
Participant correctly quickly identifies the situation as an engine no 3 surge or stall and calls for the corresponding checklist. The participant executes the TCAS RA without		Application of procedures; problem solving and decision making; workload

delay. After completion of the TCAS manoeuvre, the participant identifies the missed checklist item		management
Event 7 - Heading bug failure	**Timing**	**Performance indicator**
After the crew is cleared to turn right from 090 to 220, the heading select mode of the autopilot fails when passing heading 120. The heading bug displays any heading dialled in, but the console will not relay the command to the flight computers. Quickly following the failure, the crew is cleared to extent their turn to 240	Upon RA solved + 0:30	Remarks system failure
Desired behaviour		**Competency**
The participant notices that aircraft levels out of the turn at heading 220 and continues the turn manually towards heading 240		Aircraft flight path management, automation
Event 8 - Manual flight to runway 27	**Timing**	**Performance indicator**
Manual approach towards runway 27	Upon passing waypoint EH639	Lateral and longitudinal deviation from FMS flight path; deviation from FMS target speed
Desired behaviour		**Competency**
The participant smoothly lands the aircraft on runway 27		Aircraft flight path management, manual control

Data recording plan

The performance of the participants in the flight tests was recorded by data logging of simulator parameters and observation sheets filled out by the researchers supervising the experiment (see Appendix D and Appendix E).

AIRSIM experiment pilot performance calculation

Auto-throttle 'off' detection time

Time from start of experiment until A/T engagement (in seconds).

Lower value indicated better performance.

Speed deviation during approach

Abs (mean) + 2 x standard deviation of the deviation in Kts from the first FMS reference speed. Measured with a frequency of 1 Hz from disengagement of the autopilot until selection of a new reference speed in the FMS.

Lower value indicated better performance.

Vertical deviation from glideslope during approach

Abs (mean) + 2 x standard deviation of the vertical deviation in ft measured with a frequency of 1 Hz from disengagement of the autopilot until activation of the wind-shear alert.

Approaches where the deviation exceeded 2500 ft were classified as invalid data.

Lower value indicated better performance.

Horizontal deviation from localizer during approach

Abs (mean) + 2 x standard deviation of the horizontal deviation in ft, measured with a frequency of 1 Hz from disengagement of the autopilot until activation of the wind-shear alert.

Approaches were the deviation exceeded 4000 ft were classified as invalid data.

Lower value indicated better performance.

Response to wind-shear warning

This was a score from 0 to 4 based on four criteria that defined appropriate response to a wind-shear warning [1]:

- Pitch increased to 12 degrees or higher after initiation of the wind-shear warning;
- Pitch increased to 12 degrees was obtained within 12.5 s after initiation of the wind-shear warning;
- Thrust advanced to 80% or higher after initiation of the wind-shear warning;
- Thrust level of 80% was obtained within 4.5 s after initiation of the wind-shear warning.

One point was awarded for each criterion that was met. A higher value indicated better performance.

Response to turn request from ATC during wind-shear manoeuvre

This was a score of 0 or 1. A higher value indicated better performance.

A value 0 was given if the aircraft heading changed more than 15 degrees between the moment of the ATC instruction until the moment the aircraft reaches 2000 ft (which was assumed to indicate the end of the wind-shear escape manoeuvre).

A value of 1 was given if the aircraft heading changed less than 15 degrees between the moment of the ATC instruction until the moment the aircraft reaches 2000 ft (which was assumed to indicate the end of the wind-shear escape manoeuvre).

Selection of alternate airport

This was a score from 0 to 5:

- 5 = Hannover was selected as a calculated decision;
- 4 = Hannover was selected as an educated guess;
- 2 = Norwich or Gatwick was selected;
- 0 = No alternate was selected.

A higher value indicated better performance.

AIRSIM experiment pilot performance results

Table 5.3: AIRSIM scores

Pilot	A/T off	Speed deviation	Y deviation	X deviation	Wind-shear response	Wind-shear turn	Alternate selection
AA	56.45	25.50	269.74	476.09	6.00	no data	5.00
AB	212.45	24.81	176.13	547.30	2.00	no data	4.00
AC	11.46	20.99	92.64	641.99	2.00	no data	4.00
AD	8.36	7.66	no data	no data	4.00	1.00	5.00
AE	221.13	35.39	224.56	547.46	3.00	0.00	5.00
AF	8.41	15.07	186.50	376.78	6.00	1.00	5.00
AG	219.22	no data	no data	no data	0.00	no data	no data
AH	162.23	27.65	117.42	731.51	3.00	0.00	4.00
AK	65.84	28.47	184.79	810.98	4.00	0.00	0.00
AL	206.65	8.75	117.34	506.96	6.00	1.00	2.00

AM	205.45	11.01	1662.00	1443.77	6.00	0.00	5.00
AN	207.49	18.28	175.84	493.28	6.00	1.00	4.00
AP	216.90	20.44	624.14	no data	1.00	0.00	4.00
AQ	204.10	9.47	1375.13	no data	6.00	0.00	2.00
AR	201.27	9.18	172.28	642.21	6.00	1.00	2.00
AU	211.25	8.83	2331.59	371.04	0.00	no data	no data
AV	205.35	24.60	780.52	798.06	6.00	1.00	4.00
AX	249.63	23.35	1269.46	869.49	0.00	no data	no data
AY	205.30	14.64	no data	2635.02	6.00	1.00	2.00
AZ	209.57	no data	no data	no data	0.00	no data	no data
E	216.00	55.53	no data	754.26	0.00	no data	0.00
F	103.55	17.63	242.07	462.05	6.00	1.00	4.00
H	209.28	22.86	184.67	743.12	4.00	0.00	5.00
J	218.05	25.19	347.33	622.22	0.00	no data	0.00
K	205.09	4.57	114.41	463.30	6.00	0.00	2.00
L	89.86	23.33	1207.21	2094.59	6.00	1.00	5.00
O	254.76	9.52	117.67	470.60	6.00	no data	4.00
R	284.20	16.03	1424.77	no data	1.00	0.00	2.00
S	166.94	36.07	no data	2421.54	3.00	1.00	5.00
V	241.99	5.80	269.32	476.69	6.00	1.00	2.00
X	224.90	116.45	914.91	1324.26	3.00	1.00	4.00
Y	73.15	7.45	149.30	322.78	0.00	no data	2.00

GRACE experiment performance calculation

Communication as pilot monitoring

Correct read-back of ATC messages. One point for each correct read-back.

Maximum score was 11. Higher value indicated better performance.

Situational awareness during handover

Accuracy of pilot estimated aircraft state values (recorded on the GRACE observation sheets) compared to the actual state values logged in the GRACE-log. The following parameters were included: (1) fuel mass, (2) distance to EHAM, (3) EHAM QNH, (4) cloud base altitude, (5) wind speed on navigation display. Depending on the accuracy of the estimate 0; 0.25; 0.5; or 1 point was awarded for each estimate.

Maximum total score was 5. Higher value indicated better performance.

Communication as pilot flying

Correct read-back of ATC message about traffic. One point was awarded for correct read-back of each information item in the message (location, direction, altitude and type of aircraft).

Maximum score was 4. Higher value indicated better performance.

Vertical deviation during approach to runway 06

Abs (mean) + 2 x standard deviation of the vertical deviation in ft, measured with a frequency of 1 Hz from disengagement of the autopilot until activation of the wind-shear alert.

Lower value indicated better performance.

Horizontal deviation during approach to runway 06

Abs (mean) + 2 x standard deviation of the horizontal deviation in ft, measured with a frequency of 1 Hz from disengagement of the autopilot until activation of the wind-shear alert.

Lower value indicated better performance.

Speed deviation during approach to runway 06

Abs (mean) + 2 x standard deviation of the deviation in Kts from the first FMS reference speed. Measured with a frequency of 1 Hz from disengagement of the autopilot until selection of a new reference speed in the FMS.

Lower value indicated better performance.

Go-around decision

One point was awarded if correct decision was made (i.e. decision to go-around).

Maximum score was 1. Higher value indicated better performance.

Decision making after birdstrike

One point was awarded for each correct decision after the birdstrike (diagnose engine #3, recall memory items, call for ENGINE LIMIT/SURGE/STALL checklist, correct and immediate response to TCAS RA, identification of missing step in execution of ENGINE LIMIT/SURGE/STALL procedure).

Maximum score was 5. Higher value indicated better performance.

Detection of autopilot heading failure

One point was awarded if the autopilot heading bug failure was detected by the pilot.

Maximum score was 1. Higher value indicated better performance.

Vertical deviation during approach to runway 27

Abs (mean) + 2 x standard deviation of the vertical deviation in ft, measured with a frequency of 1Hz from disengagement of the autopilot until activation of the wind-shear alert.

Lower value indicated better performance.

Horizontal deviation during approach to runway 27

Abs (mean) + 2 x standard deviation of the horizontal deviation in ft, measured with a frequency of 1 Hz from disengagement of the autopilot until activation of the wind-shear alert.

Lower value indicated better performance.

Speed deviation during approach to runway 27

Abs (mean) + 2 x standard deviation of the deviation in Kts from the first FMS reference speed. Measured with a frequency of 1 Hz from disengagement of the autopilot until selection of a new reference speed in the FMS.

Lower value indicated better performance.

Grace experiment pilot performance results

Table 5.4: GRACE scores

	Pilot Y	AA	AB	AC	AD	AE	AF	AH
Communication pilot not flying	8.00	10.00	8.00	8.00	11.00	9.00	9.00	6.00
SA	2.50	1.75	3.00	4.00	3.50	2.50	3.75	2.25
Communication pilot flying	3.00	2.00	3.00	3.00	2.00	1.00	2.00	1.00
Y deviation 06	24.83	22.83	23.44	23.54	40.87	28.08	35.91	22.51

X deviation 06	41.11	79.95	54.59	76.92	70.57	59.36	34.90	66.02
Speed deviation 06	5.08	6.57	6.40	9.94	15.81	65.31	8.34	6.67
Go-around decision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Birdstrike	4.00	4.00	5.00	4.00	3.00	4.00	3.00	4.00
Failure detection	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
Y deviation 27	13.59	17.83	35.74	20.34	23.18	23.01	26.89	28.29
X deviation 27	43.28	148.97	71.16	97.62	83.27	73.82	129.23	92.92
Speed deviation 27	6.57	7.02	11.67	9.12	13.28	8.24	10.86	15.81

Results and conclusions of Task 4

Comparison of performance of active and non-active pilots

In order to determine whether the AIRSIM test was a correct way to estimate flight crew performance, the average scores of the active pilots are compared with the average scores of the non-active pilots (i.e. pilots with a frozen ATPL). The hypothesis was that active pilots have better performance and therefore score higher on the AIRSIM experiment.

Table 5.5: Comparison of AIRSIM scores of active and non- active pilots

	Active pilots	Non-active pilots	
A/T off	94.20	200.94	Lower value was better performance
Speed deviation	20.57	23.18	Lower value was better performance
Y deviation	173.75	711.34	Lower value was better performance
X deviation	520.56	968.60	Lower value was better performance
Wind-shear response	3.25	3.67	Higher value was better performance
Wind-shear turn	0.50	0.59	Higher value was better performance
Alternate selection	4.25	2.90	Higher value was better performance

Table 5.5 shows that active pilots scored significantly better in:

- Detection that the autopilot is 'off' at the start of the experiment;

- Vertical deviation from the approach path during the (manually flown) approach to runway 06;
- Horizontal deviation from the approach path during the (manually flown) approach to runway 06;
- Selection of the alternate airport.

For the other parameters, the difference between scores of active and non-active plots was not significant.

It can be concluded that AIRSIM indeed was suitable to measure flight crew performance.

Effect of fatigue in flight crew performance

To determine if fatigue had an effect on flight crew performance during the AIRSIM and GRACE experiments, the PVT and KSS scores were compared with the performance scores for the various tasks during the AIRSIM and GRACE tests.

Table 5.6: Correlation (Spearman's rho) between PVT/KSS and AIRSIM scores

	PVT reaction time	PVT lapses	KSS
A/T off	-0.203	0.453	-0.114
Speed deviation	-0.257	0.253	0.192
Y deviation	0.238	-0.072	-0.159
X deviation	-0.232	0.381	-0.0755
Wind-shear response	0.378	-0.372	-0.266
Wind-shear turn	0.320	-0.279	0.080

Table 5.7: Correlation (Spearman's rho) between PVT/KSS and GRACE scores

	PVT reaction time	PVT lapses	KSS
Communication pilot not flying	0.244	-0.171	0.404
SA	-0.107	0.480	0.482
Communication pilot flying	0.037	0.468	-0.069
Y deviation 06	0.179	0.300	0.826

X deviation 06	-0.607	0.060	-0.165
Speed deviation 06	-0.393	0.450	0.699
Go-around decision	0.154	0.324	-0.169
Birdstrike	0.286	-0.150	0.089
Failure detection	0.036	-0.570	-0.368
Y deviation 27	-0.214	-0.180	0.089

The results from Tables 5.6 and 5.7 showed weak correlations between the fatigue measures (PVT and KSS) and performance on the flight simulator experiments. An explanation for this could be that there was a relatively high variation in pilot performance that was not related to fatigue, both between different pilots and within the same pilot. This can possibly be identified by monitoring pilot performance over a longer period of time under varying conditions such as during a routine operational setting rather than a laboratory experiment. It is therefore recommended to monitor flight crew performance and fatigue over a longer period of time during routine flight operations.

CHAPTER 6
Final conclusions

Research conclusions

Fatigue occurs frequently in aviation and is associated with long duty periods, irregular rest-wake cycles, circadian disruptions and sleep loss [1]. Sleep disruption-induced fatigue degrades most aspects of performance, like situational awareness, reasoning, multi-tasking and sustained attention [2], although greater performance decrements are observed on simple rather than on complex tasks [3]. Technological advancements have made flying a less demanding and more monotonous task, which may contribute to performance decrements.

This CAA-UK Pilot Fatigue Measurement study addressed key issues towards a goal of developing methodologies to better manage pilot fatigue. The main research question to be answered was: "Can fatigue in individuals be measured with sufficient reliability in order to make a relationship with changes (deterioration) in flying performance?"

Although fatigue has been extensively investigated over many years, there is no consensus on a golden standard for the measurement of the actual and predicted level of fatigue. Given this state of the art, it was timely to leave the univariate approach behind and pursue optimisation of fatigue assessment and prediction using multivariate assessment.

The research question of measuring fatigue came with the added requirement that the measurement should be operationally practicable. This provided the additional challenge of allowing for only a limited range of assessments. The research aimed to select, from a very complete multivariate assessment, the most discriminative minimal dataset that was feasible in practice yet provided a robust estimate of task-relevant present and projected pilot fatigue.

Some interesting findings that were found throughout the pilot fatigue measurement study:

- The most frequently mentioned competency for accidents and incidents involving fatigue is 'problem solving and decision making'. However, the analysis showed that all ICAO-defined core pilot competencies were mentioned in fatigue related accidents;

- Self-directedness, the individual ability to govern behaviour according to situational demands, helps people to be resilient to the effects of restricted sleep. This ability is definitely required in aviation and should warrant future investigations in a sample of active duty airline pilots;
- Fatigue questionnaires taken only once can be indicative of average fatigue levels throughout the week and can be considered as a simple means to estimate fatigue, although limited in operational use as the assessment may be manipulated;
- Subjective sleepiness is more sensitive to sleep(-disruption) than objective measures of sleepiness. However, in an operational context subjective reporting may be sensitive to the desired outcome. In such cases subjective reports should be complemented by objective assessments such as the PVT;
- Recording sleep, time awake, posture, light exposure and skin temperature all provide relevant information for estimating real-time levels of sleepiness. Of these variables, (unsurprisingly) time awake (since waking up that day) was the most potent predictor, followed by posture, light exposure, time spent in bed the previous night and skin temperature;
- Whereas previous laboratory studies have shown that each of these factors individually affects vigilance, the present study was the first to assess their combined effects. The findings indicated that combined measurements are desirable if not essential to optimize estimates of current and predicted fatigue in an operational setting;
- Fit-to-fly tests should consider or control, at a minimum, the time of day, skin temperature and posture when conducting the test. Supine posture seems to affect sleepiness most pronouncedly, followed by a small DPG and a later time of day. This may influence the practicality of the fit-to-fly tests taking into account the time to induce this effect on sleepiness;
- The desktop flight simulator AIRSIM turned out to be a valid way of testing flight crew performance;

- The results of both AIRSIM and GRACE flight simulator experiments indicated weak correlations between the fatigue measures and flight performance. It is expected that this correlation will be stronger in a more routine, operational situation.

In conclusion, the results from this study clearly showed that optimisation of fatigue assessment and prediction in an operational setting should consist of a multivariate assessment of physiology, behaviour, performance, environmental exposure, and sleep history. Fit-to-fly tests should consider standardization of the recent and current environment of testing to maximize the precision and sensitivity of the measured levels of fatigue.

Recommendations for further research

Regarding the primary research question of the Pilot Fatigue Measurement study, a relevant next step would be to feed the 'fit-to-fly' measurements with operational data.

In doing so, more insight would also be gained on research questions 2 and 3; i.e. can the implications of any change in flying performance be quantified in safety terms, and if so, what would the implications be for operational safety management, if such fatigue measurement techniques were to be employed? This could also explore informal techniques used by pilots to manage fatigue in an operational context.