Fatiguing Effect of Multiple Take-Offs and Landings in Regional Airline Operations

Kimberly A. Honn^{a,c}, Briann C. Satterfield^{a,c}, Peter McCauley^a,
J. Lynn Caldwell^b, Hans P.A. Van Dongen^{a,c}

^aSleep and Performance Research Center, Washington State University, PO Box 1495,
Spokane, WA 99210, USA (kimberly.honn@wsu.edu, satterfield@vetmed.wsu.edu,
mccauleypete@gmail.com, hvd@wsu.edu)

^bNaval Medical Research Unit Dayton, Wright-Patterson Air Force Base, OH 45433, USA
(jo.caldwell@us.af.mil)

^cElson S. Floyd College of Medicine, Washington State University, PO Box 1495, Spokane, WA
99210, USA

Corresponding author:

Kimberly A. Honn
Sleep and Performance Research Center
PO Box 1495, Spokane, WA 99210-1495, USA
Tel: +1-509-358-7756
Fax: +1-509-358-7810
E-mail: kimberly.honn@wsu.edu

Abstract

Fatigue is a risk factor for flight performance and safety in commercial aviation. In US commercial aviation, to help to curb fatigue, the maximum duration of flight duty periods is regulated based on the scheduled start time and the number of flight segments to be flown. There is scientific support for regulating maximum duty duration based on scheduled start time; fatigue is well established to be modulated by circadian rhythms. However, it has not been established scientifically whether the number of flight segments, per se, affects fatigue.

To address this science gap, we conducted a randomized, counterbalanced, cross-over study with 24 active-duty regional airline pilots. Objective and subjective fatigue was compared between a 9-hour duty day with multiple take-offs and landings versus a duty day of equal duration with a single take-off and landing. To standardize experimental conditions and isolate the fatiguing effect of the number of segments flown, the entire duty schedules were carried out in a high-fidelity, moving-base, full-flight, regional jet flight simulator. Steps were taken to maintain operational realism, including simulated airplane inspections and acceptance checks, use of realistic dispatch releases and airport charts, real-world air traffic control interactions, etc. During each of the two duty days, 10 fatigue test bouts were administered, which included a 10-minute Psychomotor Vigilance Test (PVT) assessment of objective fatigue and Samn-Perelli (SP) and Karolinska Sleepiness Scale (KSS) assessments of subjective sleepiness/fatigue.

Results showed a greater build-up of objective and subjective fatigue in the multi-segment duty day than in the single-segment duty day. With duty start time and duration and other variables that could impact fatigue levels held constant, the greater build-up of fatigue in the multi-segment duty day was attributable specifically to the difference in the number of flight segments flown. Compared to findings in previously published laboratory studies of simulated night shifts and nighttime sleep deprivation, the magnitude of the fatiguing effect of the multiple take-offs and landings was modest. Ratings of flight performance were not significantly reduced for the simulated multi-segment duty day.

The US duty and flight time regulations for commercial aviation shorten the maximum duty duration in multi-segment operations by up to 25% depending on the duty start time. The present results represent an important first step in understanding fatigue in multi-segment operations, and provide support for the number of flight segments as a relevant factor in regulating maximum duty duration. Nonetheless, based on our fatigue results, a more moderate reduction in maximum duty duration as a function of the number of flight segments might be considered. However, further research is needed to include investigation of flight safety, and to extend our findings to nighttime operations.

Keywords: alertness; psychomotor vigilance; task load; high-fidelity flight simulator; commercial aviation; duty time regulations

1. Introduction

Passenger-carrying domestic flights in US commercial aviation were governed, until recently, by duty and flight time regulations stipulated in Title 14 Code of Federal Regulations (CFR) Part 121. In unaugmented flight operations (comprising only the minimum number of pilots required to operate the airplane), the regulations restricted pilot duty time to a maximum of 16 hours and flight time to a maximum of 8 hours per duty period. These requirements were fixed regardless of the number of flight segments (i.e., number of take-offs and landings) within the duty period. New duty and flight time regulations for US domestic operations, based in part on the extant science of fatigue and safety, took effect on January 4, 2014. In these new regulations, included in Title 14 CFR Part 117, duty duration is limited as a function of duty start time and as a function of the number of flight segments in the duty period.

In the new regulations, a duty period starting between 05:00 and 06:00, for example, is limited to 12 hours for between 1 and 4 scheduled flight segments, dropping stepwise to 10.5 hours (a 12.5% reduction) for 7 flight segments or more. Similarly, a duty period starting between 07:00 and 12:00 is limited to 14 hours for 1 or 2 scheduled flight segments, dropping stepwise to 11.5 hours (a 17.9% reduction) for 7 flight segments or more. The maximum duration of the duty period in unaugmented operations varies between 9 hours and 14 hours depending on the duty start time. The reduction in the maximum duration of the duty period as a function of the number of scheduled flight segments ranges from 0% to 25% depending on the duty start time.

The relevance of duty start time for limiting duty duration is corroborated by established fatigue science (Roach et al., 2012; Vejdova et al., 2014). Endogenous circadian rhythmicity produces daytime and nighttime peaks and troughs in subjective alertness and cognitive performance (Kerkhof and Van Dongen, 1996; Monk et al., 1997). Additionally, an endogenous sleep/wake-homeostatic process modulates alertness and performance as a function of time

87 awake (Daan et al., 1984; Dijk et al., 1992). The circadian and homeostatic processes interact
88 to generate predictable diel patterns of fatigue (Dijk and Czeisler, 1994; Gabelhart and Van
89 Dongen, in press), where the term fatigue refers to sleepiness and performance impairment (as
90 is common practice in operational settings; see Satterfield and Van Dongen, 2013). In this
91 context, regulating maximum duty durations as a function of start time to help mitigate fatigue
92 and improve safety makes sense (Van Dongen and Hursh, 2010), especially in daytime-oriented
93 operations (additional complexities limit the effectiveness of such regulations in nighttime
94 operations; e.g., see Rangan et al., 2013).

95 The fatigue science pertaining to number of segments flown is not as well developed. Most of
96 the literature on aviation and fatigue has focused on long-range and ultra-long-range flights
97 (Gander et al., 2014; Holmes et al., 2012) rather than short-range, multi-segment operations.
98 Limited evidence pertaining to short-range flight operations comes from field studies, in which
99 pilots' subjective ratings of fatigue were found to increase with the number of segments flown in
100 the duty period (Powell et al. 2007; Powell et al. 2008). In these studies, duty schedules and
101 operational conditions were not standardized, leaving too many potential confounds to be able
102 to draw conclusions regarding the fatiguing effect of flying multiple segments per se. No
103 objective data pertaining to fatigue in multi-segment, short-range operations have been
104 published. Still, it stands to reason that there may be a fatiguing effect of flying multiple
105 segments in a duty day – in particular, a fatiguing effect of multiple take-offs and landings.
106 These critical phases of flight are arguably the most safety-sensitive, and they typically have the
107 highest task load (Hoermann, Gontar, & Haslbeck, 2015).

108 A systems biology view of the brain mechanisms underlying cognitive impairment due to
109 fatigue (Van Dongen et al., 2011a) posits that the effect of fatigue on alertness and performance
110 is neuronal circuit use-dependent, and is thus predicted to be a function of task load. This
111 prediction has been confirmed in laboratory studies, which have shown that increased task load
112 accelerates the degradation of subjective alertness and cognitive performance due to circadian

rhythm and sleep loss (Goel et al., 2014; Van Dongen and Dinges, 2007). Yet, the effect of increased task load on alertness and performance in the laboratory appears to be modest in magnitude, and it is a priori unclear to what extent it translates to a significant effect of multiple take-offs and landings in real-world operations.

The objective of the present study was to help fill this science gap. In active-duty regional airline pilots, we compared the effect of fatigue on alertness and performance between a duty schedule containing multiple segments and a duty schedule containing only a single segment. Duty start time and duration were held constant, and the order of conditions was randomized and counterbalanced, with pilots serving as their own controls. The duty periods were carried out using a high-fidelity, moving-base, full-flight simulator. This enabled us to standardize and account for other fatigue factors commonly encountered in real-world commercial aviation (e.g., air traffic density, weather, schedule delays), while retaining a high degree of operational realism. Here we report the first objective data pertaining to the fatiguing effect of multiple take-offs and landings in regional airline operations.

2. Experimental Design and Data Analysis

2.1 Subjects

Twenty-four active-duty pilots – 12 captains (CAs) and 12 first officers (FOs) – of a US commercial regional airline were randomly recruited from among the airline's lineholder pilots certified to fly the Bombardier CRJ-200 regional airplane. One lineholder pilot scheduled for the study became unavailable and was replaced by a reserve. The pilots were 33.2 years old on average (range 24–49), and the sample included two women. The CAs had an average of 9,688 hours of total flight experience and 5,979 hours of CRJ-200 flight experience; the FOs had an average of 2,829 hours of total flight experience and 1,475 hours of CRJ-200 flight experience. The participating pilots were domiciled on the east coast of the US. On the day prior to the study, they were flown (deadheaded) to Charlotte, NC, where the study took place.

The study was approved by the Institutional Review Board (IRB) of Washington State University. All pilots gave written, informed consent. Data collected from the individual pilots were kept confidential and were not shared with the airline or any other parties besides the research team.

2.2 Experimental Design

2.2.1 Simulator Flights

The 24 pilots formed 12 flight crews, each comprised of one CA and one FO. Each crew flew a high-fidelity, moving-base, full-flight simulator (CAE Inc., Saint-Laurent, Québec) of the CRJ-200 regional jet airplane (Bombardier Inc., Montréal, Québec). Under the control of a simulator operator – a flight instructor who did not intervene with regard to the pilots' flight performance – two different flight schedules were simulated. One schedule was a 9-hour, multi-segment duty day in which pilots flew 5 short segments, from St. Louis, MO (STL) to Springfield, IL (SPI), to Dallas, TX (DFW), to Corpus Christi, TX (CRP), to Houston, TX (IAH), and lastly to Little Rock, AR (LIT). The other schedule was a 9-hour, single-segment duty day in which pilots flew from Miami, FL (MIA) to Seattle, WA (SEA). The airports in these schedules are associated with class B airspace representative of large airline hubs and smaller satellite airports commonly served by regional airlines. The specific airports selected, which in the simulator were modeled after the real airports, were not likely to be familiar to the pilots participating in the study.

In a randomized, cross-over study design, each flight crew participated in the study for two consecutive days. Six crews flew the multi-segment duty day first; the other six crews flew the single-segment duty day first. In the multi-segment duty day, the CA always flew segments 1, 3, and 5 while the FO monitored, and the FO flew segments 2 and 4 while the CA monitored. The single-segment duty day was randomized so that in half of the crews it was flown by the CA and in the other half it was flown by the FO.

For both the multi-segment and single-segment duty days, scheduled departure (“block out”) time for the first (or only) segment was 06:00, and scheduled arrival (“block in”) time for the last (or only) segment was 14:00. The single-segment flight from Miami to Seattle took an indirect route, flying west to California before turning north towards Seattle. This resulted in 8 hours of flight time, which by regulation was the maximum allowable flight time for unaugmented crews at the time of data collection. In the multi-segment duty day, the same period allowed for inclusion of five flight segments. On both duty days, the duty period began at 05:15 allowing for 45 minutes of preparation time, and ended at 14:15 after 15 minutes for post-flight duties, for a total of 9 hours on duty (05:15–14:15). See Figure 1.

At 17:00 on the evening before their two simulator flight days, researchers briefed each pilot crew on the study procedures. In a briefing session lasting up to an hour, the goals of the study and the schedule of the simulator days were discussed, including whether the multi-segment or the single-segment day would be flown first and who would fly each flight segment. Pilots were instructed to treat the simulated flights with the same professionalism as real passenger-carrying revenue flights. They were reminded that their performance in the study would not be shared with the airline. Pilots also received instruction on and practice with the procedures for fatigue testing (see sections 2.3.1 and 2.3.2). Finally, they were given a copy of the airport charts used in the study so that they could review them in advance.

Figure 1: Study timeline as the study was executed (“as flown”). The beginning (05:15) and end (14:15) of the 9-hour duty period for both the multi-segment and single-segment days are indicated by gray tick marks. The take-off (upward ramp), cruise (horizontal line) and landing (downward ramp) phases of each flight are indicated, to scale, in red (multi-segment duty period) and blue (single-segment duty period) with the respective airport cities identified. The yellow diamonds indicate the times of fatigue testing in both the multi-segment and single-segment duty days.

2.2.2 Realism

The experimental procedures were highly standardized, providing the same experience for all pilot crews. Yet, the duty periods were as realistic as possible, to resemble normal duty days in real-world operations. Pilots were told that the simulator flights were to be flown like real, passenger-carrying flights and under the prevailing regulatory provisions. They were asked to report fit for duty (they were not given any specific instructions on how to manage their sleep). They came to the simulator dressed in uniform and carrying their flight bags. Although the flight scenarios contained no unexpected procedures or emergencies, the pilots were instructed to be ready to respond to any emergencies that might occur, just like in the real world.

Realistic dispatch releases were reviewed and signed by the pilots prior to each flight segment. These contained the information that would be sent to Air Traffic Control (ATC), such

203 as the intended route of flight, speed, and altitude. Weather reports were provided, including the
204 (simulated) current and forecasted weather for the departure and destination airports and other
205 airports along the intended route. Also included in the paperwork provided prior to each flight
206 were the flight performance data for take-off, such as the speeds and aircraft weights, as well as
207 the required landing distance information at the destination airport, and the flight plan analysis
208 data, which gave detailed information about the planned route of flight with estimates of fuel
209 remaining at selected waypoints along the route. Navigation charts for each of the airports were
210 generated for the study and were available in a binder in the cockpit.

211 Pilots began each flight day taking a shuttle from a nearby crew hotel to the simulator facility,
212 where they began their duty period at 05:15. They met with a member of the research team in a
213 briefing room to review expectations for the day. They were reminded to keep their mobile
214 phones turned off in the simulator, and to give their best effort and refrain from talking during
215 fatigue testing. They also received the dispatch release and other paperwork for their first flight
216 (in the multi-segment duty day) or only flight (in the single-segment duty day). They proceeded
217 to enter the simulator and prepare as if they were readying a real airplane for their first flight of
218 the day. They settled into the cockpit, and conducted their first-flight-of-day checks. One of the
219 pilots conducted a simulated walk-around inspection of the airplane (using a marked path on the
220 ground below the flight simulator). Throughout the duty period, pilots interacted with ATC, where
221 the simulator operator served as the air traffic controller. Cued by ATC at the scheduled time,
222 they began taxiing and took off. While en route, they made normal use of the autopilot.
223 Airspace-specific ATC background chatter, recorded during real flights prior to the study, was
224 played back during flight to add to the realism. At the end of each flight segment, the pilots
225 conducted a routine descent and landed the airplane, again interacting with ATC, and taxied
226 from the runway to a designated gate. All take-offs and landings were conducted under
227 instrument flight rules (IFR) meteorological conditions.

In the multi-segment duty day, the pilot crews utilized the time spent at airports between flight segments, called “turns”, to use the restroom and to get food and drink. A few pilots went outside to smoke, as they would during real multi-segment operations. Before the next flight, one pilot of each crew picked up the new dispatch release and associated paperwork, while the other conducted a walk-around of the simulator. The pilots then completed their normal checks and prepared for the upcoming flight.

In the single-segment duty day, there were four opportunities to take the simulator off motion to allow for brief breaks in “the back of the plane.” These off-motion break opportunities were scheduled at the same times as when the turns occurred in the multi-segment duty day. If pilots opted to utilize a scheduled off-motion break opportunity, the simulator hydraulics were temporarily lowered and a gate was extended, allowing exit from and re-entry into the simulator. All normal flight processes and controls continued while off motion, and pilots could only leave the cockpit one at a time (while the other continued to fly the simulator). They could each quickly use the lavatory and get food and drink. However, they were not allowed to go outside to smoke, as they would not be able to do so in-flight during real-world operations.

Based on data provided by regional airlines prior to the study, which showed that five-segment duty periods include one airplane swap on average, an airplane swap was implemented in the multi-segment duty day after the second segment (in Dallas, TX). The pilots removed their belongings from the simulator, re-entered the simulator as a new airplane (with different aircraft number and airplane log), and conducted acceptance checks like they would normally do after an airplane swap.

Because pilots participating in this simulator study were exposed to time cues (zeitgebers) of the eastern US throughout the study, time zone changes in the simulated flight patterns were ignored. Also, the CRJ-200 does not normally have the capacity to carry enough fuel for an 8-hour flight. To circumvent this practical limitation, fuel was added artificially by the simulator operator during the single-segment day, so that pilots could fly the same type of airplane during

both duty days. This prevented the introduction of any airplane-specific confounds in comparisons of fatigue between the multi-segment and single-segment duty days.

2.3 Measurements

2.3.1 Performance Testing

Real-time, objective fatigue testing during the study was done with a 10-minute Psychomotor Vigilance Test (PVT; Lim and Dinges, 2008). The PVT is a simple reaction time task with stimuli occurring randomly at intervals of 2 to 10 seconds. The task is highly sensitive to fatigue (Balkin et al., 2004; Dorrian et al., 2005). It has been shown to track time of day (circadian) and time awake (homeostatic) factors underlying fatigue in real-world aviation operations (Gander et al., 2015).

The PVT was administered inside the simulator using a calibrated laptop (Pulsar Informatics Inc., Philadelphia, PA). The a priori selected, primary outcome variable was the total number of lapses of attention ("PVT lapses"), defined as the number of reaction times ≥ 500 milliseconds. Secondary outcome variables included the mean reaction time ("PVT mean RT") and the number of premature responses ("PVT false starts").

The PVT was administered 10 times on each simulator day (see Figure 1). During the multi-segment duty day, fatigue testing occurred twice during each segment – at top of climb right after pilots had reached cruise speed, while on autopilot, and immediately after landing, while parked on a taxiway before taxiing to the gate. During the single-segment duty day, pilots performed the PVT at the same times as during the multi-segment duty day, with nine tests en route and one after landing. During the extended climbing phase of the single-segment flight, at the designated testing times, pilots temporarily leveled off, used autopilot, completed the fatigue testing, and then resumed ascent.

Fatigue testing was done by each pilot individually, while seated in the cockpit, with the laptop positioned on the pilot's lap. A research assistant was present in the simulator to

administer the testing at the predetermined times. Pilots were trained on the PVT during the briefing on the evening before the two simulator flight days.

2.3.2 Subjective Fatigue

Each administration of the PVT was preceded by assessment of subjective fatigue, using a paper-and-pencil version of the Samn-Perelli (SP) fatigue checkcard (Samn and Perelli, 1982). On the SP, pilots rated whether they felt better than, the same as, or worse than 10 given statements related to fatigue. Scores on the SP scale can range from 0 (severe fatigue) to 20 (unusually wide awake).

Each administration of the PVT was followed by another assessment of subjective fatigue, using a paper-and-pencil version of the Karolinska Sleepiness Scale (KSS; Åkerstedt and Gillberg, 1990). The KSS is a Likert-type scale, on which pilots rated their sleepiness from 1 (extremely alert) to 9 (extremely sleepy).

Subjective fatigue assessments were done by each pilot individually, while seated in the cockpit, with the paper-and-pencil instruments provided to them on a clipboard at the predetermined times by a research assistant. Pilots were trained on the SP and KSS assessments during the briefing on the evening before the two simulator flight days. Nonetheless, one subject completed the SP fatigue checkcard incorrectly; his SP data are not included in the analyses.

2.3.3 Flight Evaluations

Each pilot's individual flight performance, whether they were flying or monitoring, was evaluated by the simulator operator after every flight segment using a standard flight evaluation form. The evaluation yielded scores for the pre-flight, gate departure, take-off, climb, cruise, descent, approach, landing, and gate arrival phases of flight, as well as an overall (i.e., calculated average) flight performance score. All scores were on a scale from 1 (unsatisfactory)

to 5 (no errors). The pilots were not provided with their evaluation scores or critiqued on their flight performance.

2.3.4 Sleep Measures

Sleep was measured during the study with an Actiwatch-2 wrist activity monitor (Philips Respironics, Bend, OR) worn on the non-dominant hand. Pilots wore the wrist activity monitor continuously from the briefing session on the evening prior to the first simulator flight day through to the end of the second simulator flight day, capturing both nights of sleep preceding the simulator flight days and any daytime naps. Sleep periods were identified and analyzed using Actiware 6.0 software (Philips Respironics, Bend, OR). For each sleep period, sleep duration was calculated as the interval between sleep onset and final awakening, subtracting any intermittent wakefulness.

At the briefing session, pilots were asked to record their sleep periods from the two days prior, using a sleep/wake/duty log. Four pilots reported taking a nap during the day before the first simulator flight day, prior to the briefing session. Wrist activity monitor-based estimates of sleep duration preceding the first simulator session do not include these naps.

3. Results

3.1 Objective and Subjective Fatigue

Statistical analyses involved head-to-head, within-pilot comparisons across test bouts between the multi-segment and single-segment duty days. Primary analyses involved mixed-effects analysis of variance (ANOVA) of condition (multi-segment versus single-segment) by test bout (1 through 10) with a random effect over subjects on the intercept (to account for inter-individual differences; see Van Dongen et al., 2004b). Secondary analyses included covariates for order of conditions, pilot age, total flight experience, CRJ-200 flight experience, prior sleep

duration, seat (CA or FO), pilot flying or monitoring, and time since the most recent rest break.

Results from secondary analyses are only reported if the covariate was statistically significant.

Figure 2 shows the results for objective and subjective fatigue in the study. The number of PVT lapses (the a priori selected primary outcome measure) was small and stable across time of day in the single-segment duty day, confirming that pilots gave their best effort to perform the task. There was a main effect of condition ($F_{1,437}=8.7$, $P=0.003$), indicating greater fatigue in the multi-segment duty day compared to the single-segment duty day (Figure 2, top left). Although the difference in PVT lapses between the multi-segment and single-segment duty days appeared to increase across the test bouts, the interaction of condition by test bout was not statistically significant ($F_{9,437}=0.9$, $P=0.52$), nor was the main effect of test bout ($F_{9,437}=1.1$, $P=0.38$).

For PVT mean RT, there was a main effect of condition ($F_{1,437}=15.4$, $P<0.001$), a main effect of test bout ($F_{9,437}=2.1$, $P=0.031$), and an interaction of condition by test bout ($F_{9,437}=2.6$, $P=0.007$). These results indicate greater fatigue in the multi-segment duty day compared to the single-segment duty day, with the difference increasing across test bouts (Figure 2, bottom left). Secondary analyses revealed that when pilots were monitoring, they had greater mean RT than when they were flying ($F_{1,436}=6.8$, $P=0.010$), but the difference was only 5 ± 2 ms (mean $\pm$ standard error).

For PVT false starts (not shown in Figure 2), there was a main effect of condition ($F_{1,437}=12.8$, $P<0.001$), reflecting more false starts in the multi-segment duty day (mean $\pm$ standard error: 1.8 ± 0.3) than in the single-segment duty day (1.3 ± 0.3). This is consistent with greater fatigue in the multi-segment duty day, as false starts tend to increase as a result of fatigue in tandem with PVT lapses (Doran et al., 2001). The interaction of condition by test bout was not statistically significant ($F_{9,437}=0.8$, $P=0.59$), nor was the main effect of test bout ($F_{9,437}=1.2$, $P=0.30$). Secondary analyses revealed a significant effect of prior sleep duration ($F_{1,436}=20.5$, $P<0.001$), where an hour of additional prior sleep led to a reduction of 0.37 ± 0.08

false starts (mean $\pm$ standard error). Controlling for prior sleep duration did not affect the significance of the condition and test bout main effects and interaction.

Figure 2: Performance on the 10-minute PVT (left) and scores on the subjective fatigue scales (right) over test bouts in the multi-segment (red) and single-segment (blue) duty days. The number of lapses on the PVT (top left) was selected a priori as the primary measure of fatigue for the study. Upwards corresponds to greater fatigue in all four panels (note the reversal of the axis for SP score). Whiskers indicate standard errors derived from mixed-effects ANOVA.

For SP subjective fatigue, there was a main effect of condition ($F_{1,418}=20.5$, $P<0.001$) and a main effect of test bout ($F_{9,418}=9.3$, $P<0.001$), but no significant interaction of condition by test

bout ($F_{9,418}=0.9$, $P=0.53$). These results indicate greater subjective fatigue in the multi-segment duty day compared to the single-segment duty day, and a build-up of subjective fatigue across the test bouts regardless of condition (Figure 2, top right).

Secondary analyses showed a significant effect of prior sleep duration on subjective fatigue measured by SP ($F_{1,417}=15.4$, $P<0.001$), where an hour of additional prior sleep led to a reduction of subjective fatigue of 0.5 ± 0.1 units (mean $\pm$ standard error). Controlling for prior sleep duration did not affect the significance of main effects and interaction for condition and test bout.

For KSS subjective sleepiness, there was a main effect of condition ($F_{1,437}=17.0$, $P<0.001$) and a main effect of test bout ($F_{9,437}=7.9$, $P<0.001$), but no significant interaction of condition by test bout ($F_{9,437}=1.3$, $P=0.21$). These results indicate greater subjective sleepiness in the multi-segment duty day compared to the single-segment duty day, and a build-up of subjective sleepiness across the test bouts regardless of condition (Figure 2, bottom right).

Secondary analyses revealed a significant effect of prior sleep duration on KSS subjective sleepiness ($F_{1,436}=28.5$, $P<0.001$), where an hour of additional prior sleep led to a reduction in subjective sleepiness of 0.4 ± 0.1 units (mean $\pm$ standard error). Controlling for prior sleep duration did not affect the significance of the main effects and interaction for condition and test bout.

The objective measures of fatigue, PVT lapses and mean RT, were highly correlated over subjects and across test bouts ($r=0.67$, $P<0.001$). Likewise, the subjective measures of fatigue, SP subjective fatigue and KSS subjective sleepiness, were highly correlated over subjects and across test bouts ($r=0.63$, $P<0.001$). The subjective measures were moderately correlated with the objective measures of fatigue (SP subjective fatigue versus PVT lapses: $r=0.26$, $P<0.001$; KSS subjective sleepiness versus PVT lapses: $r=0.28$, $P<0.001$).

3.2 Flight Evaluations

The simulator operator evaluations of pilots' flight performance ranged from 3 to 5, with a mean score of 4.34. Evaluation scores were analyzed with mixed-effects ANOVA over flight segments, with a random effect over subjects on the intercept. A planned contrast for condition compared the five segments in the multi-segment duty day to the one segment in the single-segment duty day.

There was a significant effect of condition ($t_{115}=2.6$, $P=0.010$), with pilots generally exhibiting better flight performance in the multi-segment duty day than in the single-segment duty day. The difference was 0.15 ± 0.6 units (mean $\pm$ standard error). For the distinct phases of flight, there were significant effects of condition for gate departure ($t_{115}=2.5$, $P=0.016$), climb ($t_{112}=2.9$, $P=0.004$) and cruise ($t_{112}=4.3$, $P<0.001$), with better flight performance in the multi-segment duty day than in the single-segment duty day in each case. Secondary analyses indicated that flight performance was considered better when pilots were monitoring than when they were flying during the approach ($F_{1,112}=2.9$, $P=0.002$) and landing ($F_{1,112}=20.9$, $P<0.001$).

3.3 Prior Sleep

The pilots varied in how much sleep they obtained, as assessed by wrist activity monitoring, before each of the two flight simulator sessions. During the night between the briefing session and the first simulator flight day, they slept 5.6 ± 1.1 hours (mean $\pm$ standard deviation). Before the second simulator flight day, they slept 6.4 ± 1.2 hours. The difference in sleep duration before the first and second simulator sessions was statistically significant ($F_{1,23}=9.4$, $P=0.005$). However, it is important to consider that the order of the multi-segment and single-segment duty days was randomized. Sleep duration prior to the multi-segment duty day was 6.0 ± 1.0 hours (mean $\pm$ standard deviation). Sleep duration prior to the single-segment duty day was 6.0 ± 1.4 hours. The difference in sleep duration before the multi-segment and single-segment duty days was not statistically significant ($F_{1,23}<0.1$, $P=0.98$).

On the first simulator flight day, the pilots woke up at $04:20 \pm 12$ minutes (mean $\pm$ standard deviation). On the second simulator flight day, they woke up at $04:12 \pm 15$ minutes. Wake-up time was significantly earlier on the second simulator flight day than on the first simulator flight day ($F_{1,23}=10.2$, $P=0.004$). However, wake-up time was not significantly different between the multi-segment and single-segment duty days ($F_{1,23}=0.3$, $P=0.58$).

For the day before the briefing (i.e., before the wrist activity monitors were worn), self-reported sleep duration was 7.7 ± 2.3 hours. Two days before the briefing, self-reported sleep duration was 8.3 ± 1.8 hours.

4. Discussion

4.1 Study Findings

In this randomized, counterbalanced, cross-over study, 24 active-duty regional airline pilots flew two duty days in a high-fidelity, moving-base, full-flight, regional jet flight simulator. One of the two duty days involved flying five segments, as is typical for regional airline operations; the other duty day involved flying a single (longer) segment, while the duty duration (9 hours) and start time (05:15) were the same (Figure 1). During each of the two duty days, fatigue test bouts were administered 10 times. Both objective measurements (Figure 2, left panels) and subjective reports (Figure 2, right panels) revealed greater build-up of fatigue in the five-segment duty day than in the single-segment duty day.

The experiment was designed such that the additional fatigue in the multi-segment duty day could be attributed specifically to the multiple flight segments, and the associated multiple take-offs and landings. Other factors that could systematically affect fatigue were purposely standardized and controlled for. As such, the study findings imply that the task load associated with multiple take-offs and landings in multi-segment operations results in increased fatigue over the duty day. This is consistent with data from laboratory-based fatigue studies in which task load was manipulated (Goel et al., 2014; Lim et al., 2010; Macdonald and Bendak, 2000; Van

Dongen and Dinges, 2007) and with the theoretical premise that fatigue is neuronal circuit use-dependent (Van Dongen et al., 2011a).

The results are also congruent with predictions of a mathematical fatigue model (McCauley et al., 2009, 2013). This model was extended to account for use-dependent effects of task load on sleep/wake homeostasis, and then used to help evaluate the multi-segment aspect of the new US duty and flight time regulations before they were finalized (Federal Aviation Administration, 2011). The present data provided a first opportunity to validate the extended model; the results of this validation are presented in the Appendix.

4.2 Fatigue Outcomes

The primary measure of fatigue in this study, as selected in advance, was the number of lapses on the PVT. The PVT is a widely used, normative task for measuring the effect of fatigue on sustained attention (Lim and Dinges, 2008). Performance on the PVT is highly sensitive to fatigue (Balkin et al., 2004), exhibits no practice effect across repeated test bouts (Van Dongen et al., 2003), and displays minimal aptitude differences among healthy individuals (Dorrian et al., 2005). The task captures a hallmark effect of fatigue on brain functioning, state instability (Doran et al., 2001), which has been posited to underlie fatigue's contribution to risk of errors, incidents and accidents (Van Dongen and Hursh, 2010).

Other fatigue studies in which the number of PVT lapses has been assessed may be used as reference for evaluating the magnitude of fatigue in the present study. Figure 3 juxtaposes PVT performance during the single-segment and multi-segment duty days with PVT performance observed in laboratory studies with similar testing regimes during simulated day shifts, night shifts, and total sleep deprivation. As expected, the number of PVT lapses observed in the single-segment duty day was similar to that seen in a laboratory study with simulated day shifts. The number of PVT lapses in the multi-segment duty day was similar to that seen in the early night hours of a laboratory study with simulated night shifts (and 10-hour daytime sleep

periods), while staying below the number of PVT lapses in the late night / early morning hours of these simulated night shifts. Further, the number of PVT lapses in the multi-segment duty day was considerably smaller than that seen during a night of total sleep deprivation. Thus, as referenced against the effects of fatigue observed in laboratory studies of simulated night shifts and nighttime sleep deprivation, the magnitude of the fatiguing effect of multiple take-offs and landings in the present study was relatively modest.

Figure 3: Comparison of fatigue in the single-segment and multi-segment duty days with fatigue in previously conducted laboratory studies of simulated day and night shifts and total sleep deprivation. Means (circles) and standard errors (whiskers) are plotted against time of day. The single-segment (blue) and multi-segment (red) duty day curves are the same as shown in Figure 2 (top left). The shift schedule curves are from 14 healthy subjects in the day shift condition (yellow) and 13 healthy subjects in the night shift condition (gray) of a published laboratory study (Van Dongen et al., 2011b). Note that the curve for the night shift condition (gray) wraps around at midnight (24:00) on the right, continuing at midnight (00:00) on the left. The sleep deprivation curve (black) is from 13 healthy individuals during the first night (16–24

hours awake) of a 62-hour total sleep deprivation condition in a published laboratory study (Whitney et al., 2015).

Pilots' flight performance as evaluated by the simulator operator did not appear to be adversely affected by flying multiple segments. In fact, flight performance during the climb and cruise phases of flight was scored higher in the multi-segment duty day than in the single-segment duty day. However, both the climb and cruise phases were longer in the single-segment duty day than in any of the segments of the multi-segment duty day, so these differences in flight performance evaluations may be merely a consequence of differential flight duration. Furthermore, the study participants were regional airline pilots, who are accustomed to flying multiple, short flight segments and do not normally fly duty days with a single, long flight segment. No objective measures of flight performance were recorded during the study.

The group-average temporal profiles of subjective fatigue and subjective sleepiness during the multi-segment and single-segment duty days (Figure 2, right panels) resembled those of PVT performance outcomes (Figure 2, left panels). However, correlations between subjectively reported fatigue (SP and KSS) and objectively measured fatigue (PVT lapses) were low. This is largely due to inter-individual variability in vulnerability to fatigue (Van Dongen et al., 2004a), which has been demonstrated to be incongruent between objective and subjective outcome measures (Leproult et al., 2003; Van Dongen et al., 2004a, 2004b).

4.3 Experimental Control and Realism

In this study with randomized, counterbalanced design, pilot crews served as their own controls to compare fatigue between multi-segment and single-segment duty days. Experimental conditions other than the number of segments were standardized across the two duty days. This included fixed configuration of a high-fidelity flight simulator of the airplane the pilots were certified to fly; the same duty start and end times (less than 10 minutes standard

deviation); and administration of fatigue test bouts at equal times of day (less than 5 minutes standard deviation). A flight instructor (serving as the simulator operator) and a research assistant were present in the simulator throughout each duty period to ensure a high level of experimental control. Fatigue was measured while on autopilot (after take-off) or parked on a taxiway (after landing), under controlled circumstances with limited distractions. Pilots received practice on the fatigue testing procedures on the day before the simulator flight days, and were to give their best effort during performance testing.

The scheduled flights were flown in a high-fidelity, moving-base, full-flight, regional jet simulator. In this study, the simulator was not used to practice and evaluate specific maneuvers or procedures as is common in flight education and certification efforts; rather, entire duty periods were carried out in the simulator as if it involved a real plane in real-world operations. Significant steps were taken to incorporate a high level of operational realism into the study, as described in detail in section 2.2.2 above. With a high degree of experimental control as well as a high degree of operational realism, the study provided a solid foundation for head-to-head comparisons of fatigue between multi-segment and single-segment duty days.

4.4 Limitations

The study was simulator-based in order to achieve a high degree of operational integrity, standardization of conditions in the multi-segment and single-segment duty days, consistency in experiences among the participating pilot crews, and safe opportunities for uninterrupted fatigue testing during flight operations. This would not have been possible in a field study. The main strength of the study was that the additional fatigue in the multi-segment duty day could be attributed specifically to the task load from multiple take-offs and landings. This was also its chief limitation, as it required the exclusion of other factors that may be involved in differentiating fatigue in multi-segment flight operations from single-segment flight operations.

To quantify absolute levels of fatigue in multi-segment operations or to make direct comparisons with different types of commercial aviation operations (e.g., long-range flights), other operational differences potentially affecting fatigue besides multiple take-offs and landings need to be considered. These may include differences in aircraft models, airports, routes, traffic patterns, ATC interactions, avionics, weather, delays, airplane swaps, and maintenance and repair. The fatiguing effects of these factors, and their possible interactions, have not been systematically investigated. Their relevance notwithstanding, it would have been difficult to interpret the study findings and translate them to the airline operational environment if the study had been designed to attempt to measure the fatigue associated with these variables (if any) rather than the specific effect of multiple take-offs and landings. It would have required the disentanglement and quantification of too many variables for which fatigue science is limited or non-existent.

Although fatigue has the potential to erode the operational safety margin (Dawson and McCulloch, 2005; Van Dongen and Hursh, 2010), modestly greater fatigue, as observed in the multi-segment duty day, does not necessarily mean reduced flight safety. Fatigue is only one of many factors determining safety in commercial aviation; even if fatigue itself is well understood and quantified, its safety implications depend on the overall operational conditions. No abnormal circumstances, extreme weather, high-density air traffic, in-flight emergencies, or other complicating factors were included in the flight scenarios, so flight performance was not probed at the limits of safety. As the study was designed to measure fatigue, not flight safety, no conclusions regarding the safety risk associated with multi-segment flight schedules should be drawn from the study findings.

Lastly, the study did not examine the interaction of fatigue from multiple take-offs and landings with earlier or later duty start times, longer duty days, flight delays, or more than two consecutive days on duty. The study findings must be considered in the context of these limitations.

4.5 Conclusion

With the new US flight and duty time regulations of Title 14 CFR Part 117, which took effect on January 4, 2014, duty duration in passenger-carrying domestic flights in commercial aviation is limited as a function of duty start time and the number of flight segments in the duty period. This study was the first to provide systematic evidence on the fatiguing effect of flying multiple segments as a factor relevant to duty duration. The data set comprised a total of 432 duty hours, 72 take-offs and landings, and 480 fatigue test bouts (including 4,800 minutes of PVT performance). In this regard, and in the degree of realism and operational detail, the experiment stood out among aviation fatigue studies in simulated operational settings (cf. Caldwell et al., 2000, 2004; Elmenhorst et al., 2009; Thomas et al., 2006).

Pilots' performance on the PVT indicated greater fatigue in the five-segment duty day than in the single-segment duty day. The magnitude of this effect appeared to be relatively modest. Simulated flight performance as evaluated by the simulator operator did not appear to be adversely affected, although whether this reflected a true absence of adverse effects on flight performance or insufficient sensitivity of the flight performance evaluations remains to be determined. As in earlier studies in simulated and real-world operational settings (Van Dongen et al., 2011b, 2011c), self-reported fatigue was not a reliable indicator of objectively measured performance.

While recognizing that, by practical necessity, a number of study limitations constrain the generalizability of the study findings to real-world multi-segment operations, an important strength of the study was that the increased fatigue in the multi-segment duty day could be attributed specifically to the additional take-offs and landings. As such, the study findings represent a robust first step into the scientific area of fatigue in multi-segment operations.

5. Acknowledgments

Scientists, airline management, labor representation, trade organization, and pilots worked together in this study to begin to fill a gap in fatigue science. The trade organization, i.e., the Regional Airline Association, provided research funding and logistical support. A US regional airline made substantial resources and personnel available for conducting the study. The manufacturer of the flight simulator provided access to the simulator and technical support. The pilot union expressed its essential support for the study. A scientific steering committee – consisting of scientists, airline representatives, pilots representing the pilot union, and a representative of the trade organization – oversaw the design of the study and reviewed the methodology for data analysis. We gratefully acknowledge these parties, as well as the simulator operators and the many other individuals whose concerted efforts made it possible to conduct this study.

A consensus report on the study from the scientific steering committee (Van Dongen et al., 2015) provided a basis for this paper. However, the scientific steering committee, the trade organization, the airline, and the pilot union were not involved in writing the paper, and it does not necessarily reflect their views.

The model development described in the appendix was funded in part by National Institutes of Health grant R01HL070154, and by the Office of Naval Research under contract N00014-11-C-0592 with Pulsar Informatics, Inc.

6. References

- Åkerstedt T, Gillberg M (1990). Subjective and objective sleepiness in the active individual. *Int J Neurosci.* 52: 29–37.
- Balkin TJ, Bliese PD, Belenky G, Sing H, Thorne DR, Thomas M, Redmond DP, Russo M, Wesensten NJ (2004). Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. *J Sleep Res.* 13: 219–227.
- Brooks S, Gelman A, Jones GL, Meng XL, Eds (2011). *Handbook on Markov Chain Monte Carlo*. CRC Press, London, UK.
- Caldwell JA, Caldwell JL, Smith JK, Brown DL (2004). Modafinil's effects on simulator performance and mood in pilots during 37 h without sleep. *Aviat Space Environ Med.* 75: 777–784.
- Caldwell JA, Smythe NK, Leduc PA, Caldwell JL (2000). Efficacy of Dexedrine for maintaining aviator performance during 64 hours of sustained wakefulness: a simulator study. *Aviat Space Environ Med.* 71: 7–18.
- Caldwell JL, Chandler JF, Hartzler BM (2012). Battling fatigue in aviation: recent advancements in research and practice. *J Med Sci.* 32: 47–56.
- Daan S, Beersma DGM, Borbély AA (1984). Timing of human sleep: recovery process gated by a circadian pacemaker. *Am J Physiol.* 246: R161–R178.
- Dawson D, McCulloch K (2005). Managing fatigue: It's about sleep. *Sleep Med Rev.* 9: 365–380.
- Dijk DJ, Duffy JF, Czeisler CA (1992). Circadian and sleep/wake dependent aspects of subjective alertness and cognitive performance. *J Sleep Res.* 1: 112–117.
- Dijk DJ, Czeisler CA (1994). Paradoxical timing of the circadian rhythm of sleep propensity serves to consolidate sleep and wakefulness in humans. *Neurosci Lett.* 166: 63–68.

636 Doran SM, Van Dongen HPA, Dinges DF (2001). Sustained attention performance during sleep
637 deprivation: evidence of state instability. *Arch Ital Biol.* 139: 253–267.

638 Dorrian J, Rogers NL, Dinges DF (2005). Psychomotor vigilance performance: neurocognitive
639 assay sensitive to sleep loss. In: Kushida CA, ed. *Sleep deprivation: clinical issues,*
640 *pharmacology, and sleep loss effects.* Marcel Dekker, New York, pp. 39–70.

641 Elmenhorst EM, Vejvoda M, Maass H, Wenzel J, Plath G, Schubert E, Basner M (2009). Pilot
642 workload during approaches: comparison of simulated standard and noise-abatement
643 profiles. *Aviat Space Environ Med.* 80: 364–370.

644 Federal Aviation Administration (2011). 14 CFR, Parts 117, 119, and 121: flightcrew member
645 duty and rest requirements. Docket no. FAA-2009-1093; Amendment nos. 117-1, 119-16,
646 121-357; RIN 2120-AJ58. Federal Aviation Administration, Washington, DC.

647 Gabehart RJ, Van Dongen HPA (2015). Circadian rhythms in alertness and performance. In:
648 Kryger MH, Roth T, Dement WC, eds. *Principles and practice of sleep medicine*, 5th ed.
649 Elsevier, St. Louis, in press.

650 Gander PH, Mulrine HM, van den Berg MJ, Smith AA, Signal TL, Wu LJ, Belenky G (2014). Pilot
651 fatigue: relationships with departure and arrival times, flight duration, and direction. *Aviat*
652 *Space Environ Med.* 85: 833–840.

653 Gander PH, Mulrine HM, van den Berg MJ, Smith AA, Signal TL, Wu LJ, Belenky G (2015).
654 Effects of sleep/wake history and circadian phase on proposed pilot fatigue safety
655 performance indicators. *J Sleep Res.* 24: 110–119.

656 Goel N, Abe T, Braun ME, Dinges DF (2014). Cognitive workload and sleep restriction interact
657 to influence sleep homeostatic responses. *Sleep.* 37: 1745–1756.

658 Hoermann HJ, Gontar P, Haslbeck A (2015). Effects of workload on measures of sustained
659 attention during a flight simulator night mission. In: *Proceedings of the 18th International*
660 *Symposium on Aviation Psychology.* Wright State University, Dayton, OH.

661 Holmes A, Al-Bayat S, Hilditch C, Bourgeois-Bougrine S (2012). Sleep and sleepiness during an
662 ultra long-range flight operation between the Middle East and United States. *Accid Anal*
663 *Prev.* 45 Suppl: 27–31.

664 Kerkhof GA, Van Dongen HPA (1996). Morning-type and evening-type individuals differ in the
665 phase position of their endogenous circadian oscillator. *Neurosci Lett.* 218: 153–156.

666 Leproult R, Colecchia EF, Berardi AM, Stickgold R, Kosslyn SM, Van Cauter E (2003).
667 Individual differences in subjective and objective alertness during sleep deprivation are
668 stable and unrelated. *Am J Physiol Regul Integr Comp Physiol.* 284: R280–R290.

669 Lim J, Dinges DF (2008). Sleep deprivation and vigilant attention. *Ann N Y Acad Sci.* 1129:
670 305–322.

671 Lim J, Wu WC, Wang J, Detre JA, Dinges DF, Rao H (2010). Imaging brain fatigue from
672 sustained mental workload: An ASL perfusion study of the time-on-task effect. *NeuroImage.*
673 49: 3426–3435.

674 Macdonald W, Bendak S (2000). Effects of workload level and 8-versus 12-h workday duration
675 on test battery performance. *Int J Ind Ergon.* 26: 399–416.

676 McCauley P, Kalachev LV, Mollicone DJ, Banks S, Dinges DF, Van Dongen HPA (2013).
677 Dynamic circadian modulation in a biomathematical model for the effects of sleep and
678 sleep loss on waking neurobehavioral performance. *Sleep.* 36: 1987–1997.

679 McCauley P, Kalachev LV, Smith AD, Belenky G, Dinges DF, Van Dongen HPA (2009). A new
680 mathematical model for the homeostatic effects of sleep loss on neurobehavioral
681 performance. *J Theor Biol.* 256: 227–239.

682 Monk TH, Buysse DJ, Reynolds CF 3rd, Berga SL, Jarrett DB, Begley AE, Kupfer DJ (1997).
683 Circadian rhythms in human performance and mood under constant conditions. *J Sleep*
684 *Res.* 6: 9–18.

685 Powell DMC, Spencer MB, Holland D, Broadbent E, Petrie KJ (2007). Pilot fatigue in short-haul
686 operations: effects of number of sectors, duty length, and time of day. *Aviat Space Environ*
687 *Med.* 78: 698–701.

688 Powell DMC, Spencer MB, Holland D, Petrie KJ (2008). Fatigue in two-pilot operations:
689 Implications for flight and duty time limitations. *Aviat Space Environ Med.* 79: 1047–1050.

690 Rangan S, Bowman JL, Hauser WJ, McDonald WW, Lewis RA, Van Dongen HPA (2013).
691 Integrated fatigue modeling in crew rostering and operations. *Can Aeronaut Space J.* 59: 1–
692 6.

693 Roach GD, Sargent C, Darwent D, Dawson D (2012). Duty periods with early start times restrict
694 the amount of sleep obtained by short-haul airline pilots. *Accid Anal Prev.* 45S: 22–26.

695 Samn WS, Perelli LP (1982). Estimating aircrew fatigue: a technique with application to airlift
696 operations. Brooks Air Force Base Report SAM-TR-82-21.

697 Satterfield BC, Van Dongen HPA (2013). Occupational fatigue, underlying sleep and circadian
698 mechanisms, and approaches to fatigue risk management. *Fatigue Biomed Health Behav.*
699 1: 118–136.

700 Tassi P, Muzet A (2000). Sleep inertia. *Sleep Med Rev.* 4: 341–353.

701 Thomas MJW, Petrilli RM, Lamond N, Dawson D, Roach GD (2006). Australian long haul
702 fatigue study. In: *Proceedings of the 59th Annual International Air Safety Seminar*. Flight
703 Safety Foundation.

704 Van Dongen HPA, Baynard MD, Maislin G, Dinges DF (2004a). Systematic interindividual
705 differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential
706 vulnerability. *Sleep.* 27: 423–433.

707 Van Dongen HPA, Belenky G, Krueger JM (2011a). Investigating the temporal dynamics and
708 underlying mechanisms of cognitive fatigue. In: Ackerman PL, ed. *Cognitive fatigue:*
709 *multidisciplinary perspectives on current research and future applications*. American
710 Psychological Association, Washington, DC, pp. 127–147.

711 Van Dongen HPA, Belenky G, Vila BJ (2011b). The efficacy of a restart break for recycling with
 712 optimal performance depends critically on circadian timing. *Sleep*. 34: 917–929.
 713 Van Dongen HPA, Caldwell JA, Caldwell JL (2011c). Individual differences in cognitive
 714 vulnerability to fatigue in the laboratory and in the workplace. *Prog Brain Res*. 190: 145–153.
 715 Van Dongen HPA, Caldwell JL, Frisch B, Gibson R, Preidecker P, Swindell R, Hintze M,
 716 Demitry PF, Foose S (2015). WSU–RAA fatigue study: simulator experiment on fatigue in
 717 multi-segment operations. Regional Airline Association, Washington, D.C.
 718 Van Dongen HPA, Dinges DF (2007). Individual differences in response to sleep deprivation.
 719 Final Technical Report, NIH R01 HL070154.
 720 Van Dongen HPA, Hursh SR (2010). Fatigue, performance, errors, and accidents. In: Kryger
 721 MH, Roth T, Dement WC, eds. *Principles and practice of sleep medicine*, 5th ed. Elsevier
 722 Saunders, St. Louis, pp. 753–759.
 723 Van Dongen HPA, Maislin G, Dinges DF (2004b). Dealing with inter-individual differences in the
 724 temporal dynamics of fatigue and performance: importance and techniques. *Aviat Space*
 725 *Environ Med*. 75: A147–A154.
 726 Van Dongen HPA, Maislin G, Mullington JM, Dinges DF (2003). The cumulative cost of
 727 additional wakefulness: dose-response effects on neurobehavioral functions and sleep
 728 physiology from chronic sleep restriction and total sleep deprivation. *Sleep*. 26: 117–126.
 729 Vejvoda M, Elmenhorst E, Pennig S, Plath G, Maass H, Tritschler K, Basner M, Aeschbach D
 730 (2014). Significance of time awake for predicting pilots' fatigue on short-haul flights:
 731 implications for flight duty time regulations. *J Sleep Res*. 23: 564–567.
 732 Whitney P, Hinson JM, Jackson ML, Van Dongen HPA (2015). Feedback blunting: total sleep
 733 deprivation impairs decision making that requires updating based on feedback. *Sleep*. 38:
 734 745–754.
 735

7. Appendix

7.1 Background

Refinement of biomathematical models of fatigue and performance is among the most immediately suggested improvements for battling fatigue in the aviation environment (Caldwell et al., 2012). One such biomathematical model, first introduced by McCauley et al. (2009) and recently expanded by McCauley et al. (2013), predicts performance impairment from fatigue based on the homeostatic and circadian processes of sleep/wake regulation (Daan et al., 1984) as well as a slow allostatic process modulating homeostatic equilibrium. The model has been validated for a number of different sleep/wake schedules, including total sleep deprivation, sustained sleep restriction, split sleep schedules, night shifts, and recovery sleep schedules (McCauley et al., 2013).

The model builds on a theoretical framework for the underlying neurobiology (Van Dongen et al., 2011a), which posits that task load interacts, in a neuronal circuit use-dependent manner, with the homeostatic process of sleep/wake regulation – such that both task load and time awake drive fatigue, independently as well as synergistically. We implemented this concept in the model of McCauley et al. (2013), and then used it to help evaluate the multi-segment aspect of the new US duty and flight time regulations before they were finalized (Federal Aviation Administration, 2011). Here we show that the predictions of the model are congruent with the findings of the present flight simulator experiment.

7.2 Original Model Formulation

The model of McCauley et al. (2013) is written as a system of first-order ordinary differential equations (ODEs). For periods of scheduled wakefulness, the ODE system is given by:

$$\frac{dp(t)}{dt} = \alpha_w [p(t) + \beta_w u(t)] + g_w(t), \quad (1a)$$

$$\frac{du(t)}{dt} = \eta_w u(t), \quad (1b)$$

$$g_w(t) = \kappa(t)[c(t) + \mu_w], \quad (1c)$$

$$\frac{d\kappa(t)}{dt} = \lambda_w \kappa(t) \left(1 - \frac{\kappa(t)}{\xi}\right). \quad (1d)$$

For periods of scheduled sleep (i.e., time in bed), the ODE system is given by:

$$\frac{dp(t)}{dt} = \alpha_s [p(t) + \beta_s u(t)] + g_s(t), \quad (2a)$$

$$\frac{du(t)}{dt} = \eta_s u(t) + 1, \quad (2b)$$

$$g_s(t) = \kappa(t)[c(t) + \mu_s] + \frac{\alpha_s \beta_s}{\eta_s}, \quad (2c)$$

$$\frac{d\kappa(t)}{dt} = \lambda_s \kappa(t). \quad (2d)$$

Here, $p(t)$ is the performance prediction (observable during wakefulness, nominal during sleep), expressed in terms of lapses (reaction times ≥ 500 ms) during a 10-minute PVT. Further, $u(t)$ is the slow allostatic process modulating homeostatic equilibrium; the nonhomogeneities $g_w(t)$ and $g_s(t)$ capture circadian rhythm; and $\kappa(t)$ represents dynamic changes in circadian amplitude. The circadian process $c(t)$ is modeled as independent of sleep/wakefulness and is given by:

$$c(t) = \sin\left(2\pi \frac{t - \phi}{\tau}\right). \quad (3)$$

See McCauley et al. (2013) for detailed information regarding the model parameters.

7.3 Model Formulation Accounting for Task Load

Increased task load has been found to accelerate the homeostatic build-up of performance impairment (Goel et al., 2014; Van Dongen and Dinges, 2007). To track this dynamic, we add a

781 new, linear function $h_w(t)$ to the nonhomogeneity $g_w(t)$ in Eq. (1c):

$$782 \quad g_w(t) = \kappa(t)[c(t) + \mu_w + h_w(t)]. \quad (4)$$

783 Furthermore, assuming that task load also affects the slow process $u(t)$, we introduce a
784 nonhomogeneity in Eq. (1b):

$$785 \quad \frac{du(t)}{dt} = \eta_w u(t) + h_w(t). \quad (5)$$

786 In scenarios characterized by wakefulness with basal task load interspersed with brief
787 periods of high task load, $h_w(t)$ may be formulated as follows:

$$788 \quad h_w(t) = \psi \ell(t). \quad (6)$$

789 Here $\ell(t)$ represents the temporal profile of task load, where $\ell(t) = 0$ during periods with basal
790 task load and $\ell(t) = 1$ during periods with high task load. The parameter ψ determines the
791 magnitude of the effect of cognitive task load. It is the only free parameter considered here; all
792 other parameters are kept fixed at their previously estimated values (see McCauley et al.,
793 2013).

794

795 **7.4 Calibration Data Set**

796 To estimate ψ , we used data from 21 healthy subjects (age range 22–40), who completed a
797 12-day (11-night) laboratory study (Van Dongen and Dinges, 2007). The subjects were
798 exposed, on three consecutive occasions, to 36 hours of total sleep deprivation. Each sleep
799 deprivation period was preceded and followed by two 12-hour sleep opportunities (22:00–
800 10:00).

801 Throughout scheduled waking periods, cognitive performance was measured every 2 hours
802 by means of a standardized test battery. For each subject, two of the three sleep deprivation
803 periods involved basal (moderate) task load, involving a 30-minute performance test bout every
804 2 hours. The other sleep deprivation period, occurring in randomized, counterbalanced order,

involved high (doubled) task load, involving a 60-minute performance test bout every 2 hours. Each test bout included a PVT, which was 10 minutes in duration in the basal task load conditions and 20 minutes in duration in the high task load condition.

The number of lapses observed during each administration of the PVT during the sleep deprivation periods was used to calibrate the model, using only the first 10 minutes of the 20-minute PVT sessions in the high task load condition. Data from the test bouts occurring immediately after a sleep period were discarded to avoid confounds from sleep inertia (Tassi and Muzet, 2000).

7.5 Model Calibration

The 21 subjects were divided into three groups of 7 subjects, depending on whether they had been randomized to high task load during the first, second or third sleep deprivation period. The PVT lapse data were averaged over subjects in each group, yielding a total of 51 data points per group (153 data points in total). For each group, the timeline of scheduled sleep opportunities and scheduled wakefulness across the study duration – starting with the two sleep opportunities preceding the first sleep deprivation period, and ending with the third sleep deprivation period – was then used to select the wake and sleep equations of the ODE system. Furthermore, we set $\ell(t) = 1$ during the 1-hour periods of high task load during the sleep deprivation condition with high task load; else $\ell(t) = 0$.¹

Initial values were derived from steady state over days, assuming 8 hours time in bed for sleep (23:00–07:00) daily and no high task load during wakefulness. To estimate ψ , the model was then fitted to the data set using Metropolis-Hastings-type Markov chain Monte Carlo

¹ Alternatively, we could have set $\ell(t) = 1$ only during the second 30-minute interval of each 1-hour test bout in the high task load condition, or we could also have set $\ell(t) = 1$ during all performance testing in the study, including each 30-minute test bout in the basal task load condition. However, the latter approach would have required re-estimation of other, currently fixed parameters of the model.

(MCMC) fitting (Brooks et al., 2011) with a chain length of 2,000. The estimated parameter value was $\psi = 1.16 \pm 0.11$ (estimate $\pm$ standard error). The root mean square error (RMSE) was reduced from 5.8 (original model formulation) to 4.7 (new model formulation), representing a significant improvement (variance ratio test: $F_{1,10} = 5.21$, $P = 0.046$).

7.6 Model Predictions

We applied the calibrated model, with $\psi = 1.16$ fixed, to predict PVT lapses in the flight simulator experiment described in this paper, accounting for the pilots' prior sleep schedules individually. Take-offs and landings were considered high task load, represented with $\ell(t) = 1$; else $\ell(t) = 0$. Group-average model predictions for the difference in PVT lapses between the multi-segment and single-segment duty days in the flight simulator experiment showed a 79.9% reduction in mean squared prediction error, relative to the original model without task load, from 0.179 to 0.036 ($\chi^2=16.0$, $P<0.001$). Thus, the new model formulation showed improved predictive accuracy under conditions of increased fatigue due to high task load; see Figure 4.

This result shows that the predictions of the model are congruent with the results of the flight simulator experiment. As the model parameters were fixed, the finding provides evidence of the predictive validity of the model described by Eqs. (1a), (1d), (3)–(6) and (2a)–(2d). This novel, task load enabled, biomathematical model of fatigue and performance may be useful for comparing relative performance in daytime duty schedules in regional airline operations. Further model validation and refinement is needed to extend the scope of the model to night work and other operational conditions.

Figure 4: Comparison between predictions of the biomathematical model, with and without task load component, and the observed difference in performance between the multi-segment and single-segment duty days in the flight simulator experiment. The purple data points show the mean relative difference (centered on zero) in performance on the 10-minute PVT over test bouts between the multi-segment and single-segment duty days (cf. Figure 2, top left). Upward corresponds to greater performance impairment in the multi-segment duty day; whiskers indicate standard errors. The black curve shows the predicted difference between the two duty days based on the original model without task load (McCauley et al., 2013). Note that the marginal deviations of this curve from zero are caused by minor differences in pilots' prior sleep schedules between the two duty days. The green curve shows the predicted difference between the two duty days based on the new model formulation, accounting for task load.