Advance Publication

INDUSTRIAL HEALTH

Received: May 3, 2025

Accepted: June 18, 2025

J-STAGE Advance Published Date: June 26, 2025

Field Report

Workload predictions from a biomathematical model compared to top-of-descent NASA Task Load Index scores in commercial pilots

Jaime K DEVINE, PhD[1]; Kae YOSHIDA[2], Takeshi TANAKA[2], Kohei IKUTA[2], Wataru TANAKA[2], Jake CHOYNOWSKI[1], Steven R HURSH[1,3], PhD

[1] Institutes for Behavior Resources, Inc., Baltimore, MD, 21218, USA.

[2] Japan Airlines, Flight Safety Management Department, Shinagawa, Tokyo, 140-8637, Japan

[3] Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Corresponding author: Jaime K. Devine, PhD

jdevine@ibrinc.org

Institutes for Behavior Resources, Inc.

2104 Maryland Ave

Baltimore, MD 21218

Suggested short title: Pilot Workload Modeling

Keywords: Workload; Fatigue; Aviation; Biomathematical modeling; NASA TLX

Abstract

Biomathematical models of fatigue (BMMFs) are commonly used to predict cognitive alertness in commercial aviation. Accounting for workload in association with routine job tasks may help BMMFs to more accurately predict fatigue in real world operations. This study compared the accuracy of BMMF workload predictions (SF Workload) against pilot self-report of workload during normal flight operations. N=99 pilots from a major Asia-based airline completed the NASA Task Load Index (TLX) at top of descent (TOD) during a multiple-flight three-day roster that consisted of daytime flying. SF Workload predictions and TLX scores were normalized to a 100-point scale and compared using equivalence testing. SF Workload predictions were statistically non-different from pilot TLX scores at the same TOD (64±7 vs. 65±15; both t=1.56, p=0.06) using the two one-sided t-test (TOST) approach, indicating high workload and that BMMF predictions are non-inferior to pilot self-report as a means of estimating workload. Establishing the accuracy of workload predictions against real-world reports in a commercial pilot population is an important step towards risk management in situations where high workload may create a safety risk.

INTRODUCTION

Fatigue, defined as “a physiological state of reduced mental or physical performance capability”, is a known contributor to aviation accident risk[1)]. An estimated 23% of all major aviation accidents between 2001 and 2012 could be at least partially attributed to fatigue[2)]. The International Civil Aviation Organization (ICAO) identifies fatigue as resulting from sleep loss, circadian phase, or workload[1)]. Workload, which can be either physical or mental activity related to job tasks[1)], is a growing concern for fatigue risk management in aviation[3-6)]. High workload has been estimated to contribute to upwards of 80% of aviation incidents and accidents that result from human error[7)]. Accounting for the potential impact of workload on performance and safety is therefore important to appropriate fatigue risk management in aviation.

Fatigue Risk Management Systems (FRMS) are data-driven means of managing fatigue-related safety risks with the goal of ensuring adequate levels of operator alertness[1)]. Biomathematical models of fatigue (BMMF) are one of the tools used to

support FRMS in commercial aviation. Most BMMFs are fundamentally influenced by Borbely's Two-Process Model of sleep regulation and use information about time of day and prior sleep history to predict alertness in relation to work schedules[8]. The Two-Process Model reflects biological processes that can be observed at the neuronal level not only in humans but in other animals as well[9]. In fact, the genetic circuits regulating circadian pacemakers are conserved across lifeforms from bacteria to humans[10]. BMMFs estimate when fatigue will occur as a function of these highly-conserved and observable biological processes. Fatigue due to sleep loss or circadian misalignment can be conceptualized as an unavoidable physiological consequence.

By comparison, workload is a modern concept and a uniquely human psychological experience[11]. The preferred approach for measuring workload is through individual self-assessment using surveys like the National Aeronautics and Space Administration Task Load Index (NASA TLX), Subjective Workload Assessment Technique (SWAT), Workload Profile (WP), or Air Force Workload Estimate Scale (ARWES)[4, 12-14]. Self-assessment of workload can capture the mental and emotional effort required to complete a task in a way that performance or physiological measurements cannot accomplish[12]. Mental workload can be conceptualized as the difference between the cognitive demands of a task and the operator's attentional resources[11, 12]. In other words, the level of difficulty associated with a work task depends on the mental state of the operator.

Many job tasks that contribute to fatigue are predictable enough that some BMMF applications have begun to incorporate workload into their calculations of fatigue estimation[3, 15-17]. For example, pilots often report fatigue in relation to the number of takeoffs and landings per duty period, the number of duties per day or consecutive duties, or difficulty landing at a specific airport[6, 18, 19]. These fatigue factors are easy to identify when looking at the pilot's work schedule. However, the question remains of how best to incorporate the impact of psychological fatigue related to workload into a model that is rooted in biological processes?

One method of modeling workload is to consider it as a demand on cognitive capacity rather than a variable that contributes to the calculation of physiological alertness. This is how workload is modelled by the BMMF software SAFTE-FAST.

SAFTE-FAST uses the Sleep Activity Fatigue Task Effectiveness (SAFTE) model to compute a metric called Effectiveness that reflects physiological fatigue due to sleep history and circadian factors[20]. There is a separate function in SAFTE-FAST called the Workload Calculator that is used to estimate job demands during operations. SAFTE-FAST Workload predictions will be referred to as “SF Workload” in this paper to differentiate BMMF predictions from the general concept of workload. Triggers can be set within SAFTE-FAST to indicate that SF Workload will increase with longer time on task independently of time of day or opportunities for sleep[17]. A crew member suffering from physiological fatigue is assumed to have diminished cognitive resources to meet workload demands relative to a well-rested crew member. Low Effectiveness in combination with high SF Workload is interpreted as a potential area of increased risk within the SAFTE-FAST system.

An important step towards accurate modeling of fatigue due to workload is to test the accuracy of SF Workload against a real-world measure, like the NASA TLX scale. An issue to consider in testing model predictions against self-report in operations, however, is that pilots will perceive workload as higher under conditions of physiological fatigue because of sleep deprivation or circadian effects rather than the effects of the work itself[21]. Previous studies have found that perception of workload fluctuates as a function of sleep loss and time of day[22-24]. It is therefore important to account for time of day and prior sleep opportunities when assessing workload during normal working conditions.

Comparison of means using t-tests or analysis of variance (ANOVA) statistical tests are inappropriate for analyzes that are attempting to show that two measures are the same since failure to show a difference between the means of two measures is not the same as showing equivalence between measures[25]. Equivalence testing is a statistical method that establishes that one measure is not unacceptably worse than another[25, 26]. Equivalence testing provides a clear demonstration that one measure is not inferior to another with regards to safety concerns in the context of FRMS[25].

The Flight Safety Management Department of a major Asia-based airline recently conducted a workload survey among pilots operating a selection of domestic three-day rosters during the summer of 2024 using the NASA TLX. The current study makes

secondary use of the data collected by the airline's Flight Safety Management Department to compare SF Workload predictions against NASA TLX taken by pilots during normal flight operations. To avoid any undue influence of physiological fatigue on the pilots' perception of workload, this test concentrated on regional flight operations that did not cross multiple time zones, occurred predominantly during daytime hours, and did not encroach on the window of circadian low (WOCL), a time when pilots are known to have increased feelings of fatigue[27]. Equivalence between NASA TLX and SF Workload was evaluated using two methods, the two-one sided t-test (TOST) approach and graphical non-inferiority[25, 26]. This study constitutes the first field test to evaluate the accuracy of SF Workload against pilot self-report during normal operations.

SUBJECTS AND METHODS

Ethics Approvals

A major Asia-based commercial airline agreed to share data with the IBR study team for secondary analysis. Participants were originally recruited through the Flight Safety Management Department of the airline. All participants provided informed consent for their participation. No personal identifying information was shared with the study team for the purposes of this study. Secondary use of de-identified data for research purposes was deemed non-human subjects research by Salus IRB on October 22, 2024 (Study ID: 23452) and these analyses were conducted in accordance with the Declaration of Helsinki.

Subjects and Study Design

Eligible subjects were airline-employed pilots based in Tokyo, Japan who were scheduled to fly a three-day domestic regional roster between May-August 2024. Pilots were considered eligible for inclusion in this study regardless of gender, ethnicity, age (over 18), sleep habits, or health status. Subjects were asked to complete a NASA TLX to reflect workload associated with the top of descent (TOD) shortly after the final landing of the duty day on the first and third days of a three-day roster. Subjects completed a modified version of the Japanese language NASA TLX[28, 29] using the online survey platform Google Forms (Google Inc., Mountain View, California, USA).

Subjects were also given the option to provide additional contextual information about workload using a text box with a 20k character limit at the end of the survey.

The NASA TLX assesses workload across six domains: 1) mental demand; 2) physical demand; 3) temporal demand; 4) performance; 5) effort, and; 6) frustration. Subjects were asked to rate the workload they experienced at TOD across each domain on a 100-point scale. Subscale ratings are then averaged to create a global TLX score with the following workload distinctions: Very Low=0-20; Low=21-40; Medium=41-60; High=61-80; Very High=81-100. Global NASA TLX scores were matched to pilots' schedule information using a numeric schedule identifier, translated to English, and compiled as a csv file by the airline's Flight Safety Management Department. De-identified files of NASA TLX scores and pilot schedules were provided to the study team after completion of the original data collection for comparison against SF Workload. Write-in comments about causes of workload were translated to English and provided to the study team for qualitative analysis[30].

Biomathematical Modeling

Pilots' schedules were modeled in SAFTE-FAST using custom-built settings that included airline-specific SF Workload triggers. SAFTE-FAST version 6.13 was used for all modeling in this project. SAFTE-FAST is the brand name of the FRMS software provided by the Institutes for Behavior Resources, Inc. (IBR, Baltimore, MD, USA) to a variety of operational organizations, including commercial aviation. SAFTE-FAST is a two-step, three-process model that estimates sleep patterns around work duties and then estimates performance levels. The three processes involved are circadian function, homeostatic sleep reservoir, and sleep inertia[20]. The SAFTE-FAST Workload Calculator (SF Workload) is comparable to NASA TLX global scores and is used to create weighted triggers for job tasks or operational factors that may impact cognitive demands[17, 31]. SF Workload is normalized to a 100-point scale, with a score of 0 representing no workload and 100 representing the highest level of workload.

Data Analysis

SF Workload predictions were exported as csv files. SF Workload predictions and global NASA TLX results were matched using the schedule identifier and compiled in Excel MP 15. NASA TLX scores were independently normalized to a 100-point scale

for comparison against SF Workload predictions. SF Workload predictions at TOD were compared against TLX scores for the corresponding flight. Between-group differences in TLX scores, time at TOD, and duty duration were explored using student's t tests and analysis of variance (ANOVA). Correlations between time at TOD, duty duration and TLX scores were explored using Pearson's r correlation. Within-subjects changes to global NASA TLX scores and SF Workload across the three-day roster were further evaluated using repeated measures ANOVA. Statistical significance was assumed at $p \leq 0.05$.

Equivalence testing compared SF Workload against global TLX using two separate methods: 1) the two-one sided t-test (TOST) approach and 2) graphical non-inferiority[25, 26]. Using the TOST approach, confidence intervals (CI) are set at 90%, and a conclusion of statistical equivalence is warranted when the larger of the two p values is smaller than alpha (α)=0.10[26]. In graphical non-inferiority, the mean difference between two measures is graphed on a two-dimensional grid framed by vertical lines that represent the lower equivalence bound (LEB) and upper equivalence bound (UEB). CIs were tested at 95% in accordance with previously-established non-inferiority testing procedures for fatigue risk management[25]. Equivalence is demonstrated when the mean difference and CIs fall between the LEB and UEB on the graph[25].

Key terms that summarized a workload factor were extracted from write-in comments using qualitative research methodology[30]. The Excel 13 Rank function was used to calculate the weighted mean rank order of workload factors extracted from the write-in comments. All statistical analyses were done in Excel MP 15 with the QI Macros Excel add-in (KnowWare International, Denver, CO) and BlueSky Statistics 10.3.4 for Windows (BlueSky Statistics, Chicago, IL, USA)[32].

RESULTS

Between June 2 to August 2, 2024, 102 pilots completed at least one NASA TLX over the course of a three-day roster. Responses from three pilots were excluded because NASA TLX scores could not be linked to work schedule data, leaving a final study population of N=99 pilots (51 First Officers; 48 Captains). Pilots completed two NASA TLX surveys on average during the three-month period (range: 1-6) for a total of

187 surveys. Pilot demographics, including age, gender, and flight hours, were not collected as part of this study.

Responses were distributed across three separate rosters that provided regional service in Japan. Table 1 depicts the target three-day rosters for this study. Roster legs were identical for Days 1 and 2 but went to different destinations on Day 3. Ninety-one (n=91) pilots completed NASA TLX for both Days 1 and 3 within a given roster period. Eight (n=8) pilots completed only one NASA TLX during the target three-day roster.

[Insert Table 1 here]

Average pilot NASA TLX scores across all six domains are depicted by roster and day in Fig. 1A and by rank in Fig. 1B. NASA TLX scores were not significantly different between rosters or days (all $p>0.08$). Captains tended to rate workload as lower on the mental demand domain than did First Officers (CPT: 72±18 vs. FO: 77±14; $t=2.21$, $p=0.03$). No other NASA TLX domains or the global NASA TLX scores were significantly different by rank (all $p>0.32$).

[Insert Fig. 1 here]

The average NASA TLX score at TOD for all flights was 65±15 out of a possible 100 points, indicating high workload. On average, TOD occurred during the evening (21:17±2:04) for flights where pilots were asked to complete a NASA TLX; there were no significant differences in time of TOD between Days 1 and 3 or between rosters (all $p>0.05$). Pilots had been on duty between 7 to 9 hours prior to completing the NASA TLX at TOD for all time points (498±51 minutes). There were statistically significant differences in duty duration between Days 1 (486±62 minutes) and Day 3 (507±31 minutes, $t=2.76$, $p=0.006$) and for Day 3 duty duration between Rosters 1, 2, and 3 (Roster 1: 477±19 minutes; Roster 2: 520±20 minutes; Roster 3: 540±21 minutes; $F_{(3,1)}=10.10$, $p<0.001$). There was no correlation between global NASA TLX scores and time at TOD ($r=0.12$, $p=.21$) or duty duration ($r=0.06$, $p=0.45$).

SF Workload predictions at TOD for each flight leg in the roster and global NASA TLX scores at TOD for the final legs on Day 1 and 3 are depicted in Fig. 2. The average SF Workload score at TOD for all flights was 64±7, indicating high workload. SF

Workload estimates increased significantly across the roster ($F_{(2,16)}$=62.02, p<0.001). There were no significant differences in the means of SF Workload and NASA TLX controlling for flight number or roster day.

[Insert Fig. 2 here]

Equivalence between SF Workload and NASA TLX at TOD were tested using both the TOST and graphical non-inferiority methods. TOST results are depicted in Table 2 and indicated that the difference between the means of SF Workload and NASA TLX scores for the same TOD were less than α=0.10, indicating equivalence between the two measures (both p=0.06). Graphical non-inferiority results are depicted in Fig. 3. The observed difference between means was 1.86±1.19 standard deviation. The LEB was set at -3.23 and the UEB was set at 3.23. The difference between means fell within the LEB and UEB, but 95% CIs exceeded the UEB. Equivalence could neither be confirmed nor rejected using the graphical non-inferiority method.

[Insert Table 2 here]

[Insert Fig. 3 here]

Twenty-six (26) key terms were identified from pilots' write-in comments regarding causes of workload. Pilots were permitted to write freely in the text box, and so, comments could include multiple workload factors at one time. Pilots identified 3±2 workload factors per comment on average (range: 0-10 factors per comment). There were 164 write-in comments in total out of 187 opportunities to submit a comment. The most frequently reported workload factor was related to weather, with 48% of the pilot participants including phrases like "bad weather", "wind", "thunderstorms", "turbulence due to weather" or similar terms in their comments. Workload factors from pilots' write-in responses are included in Table 3 along with the frequency of reports from pilots by percentage. An annotated list of 40 English-translated pilot comments is available in the supplemental material (Supplemental Table 1). Comments have been edited for clarity and confidentiality. Specifically, the use of aviation jargon and references to the airline by name have been replaced with more general terminology without changing the intent of the original comment.

[Insert Table 3 here]

DISCUSSION

This study compares field assessments of workload using the NASA TLX against BMMF predictions over the course of three three-day rosters providing regional service in Japan. Pilots reported high workload on average for the final TOD on Days 1 and 3 across all rosters. NASA TLX scores were not largely influenced by rank, time of day, or duty duration.

SF Workload predictions were statistically non-different from pilot NASA TLX scores at TOD using the TOST method. However, graphical non-inferiority with 95% CI could neither confirm nor reject equivalence between SF Workload and TLX at TOD since the observed difference falls within the so-called “zone of indifference” (the graph area between the LEB and UEB) but the CIs extend beyond the UEB. Adjusting SF Workload triggers or weights based on pilot-identified factors could improve agreement between NASA TLX and SF Workload but runs the risk of overfitting as well. For example, an increase in SF Workload can be set to trigger at TOD. Triggers are configurable within the SAFTE-FAST system; it may be that different operation types require different configurations to yield the best possible workload predictive results. The weight of the SF Workload trigger can be adjusted to reflect the fact that landing at one airport (for example, an airport with a difficult approach) requires greater effort than at an easier-to-maneuver airport. Weights in SF Workload are adjustable and based on the NASA TLX scale ranging from very low to very high workload. Future investigations will be necessary to determine the best settings for SF Workload triggers that are accurate across multiple rosters and flight duty lengths.

Pilots identified several workload factors using the write-in comment box. Weather, delays, and schedule changes were the most commonly identified workload factors. Many of the pilot-identified workload factors were interrelated and it was infrequent that pilots could attribute their feelings of high workload to a single isolated factor. For example, pilots indicated that delays due to inclement weather or schedule changes caused a “ripple effect” that led them to feel stressed or pressured for time. These delays may also have cut into mealtimes or resulted in late finishes and a short rest period. Pilots also indicated feeling more fatigued if their workload was high on

previous days without sufficient recovery time in-between flights. A possible solution to the ripple effect that pilots reported may be to schedule the flight legs further apart to provide a buffer for delays due to weather or employ novel analysis techniques for predicting delays in relation to weather. An analysis published in the Journal of Big Data in 2019 created a predictive model of flight arrival times in relation to weather data for a Japanese-based low cost carrier airline[33].

Of course, weather can be unpredictable, and perhaps increasingly unpredictable in response to climate change[34]. Unpredictable weather, along with other pilot-identified factors that are not expected to occur regularly, like passenger issues, errors, or bird strikes, are not good targets for modelling in a BMMF. Some of these factors could be addressed by the airline without the involvement of modelling but those possible solutions are beyond the scope of this analysis to suggest. The pilot-identified workload factors that were related to scheduling or airport specifics could be included as SF Workload triggers weighted to reflect the pilot-reported workload associated with that factor (see Table 3).

Some of the pilot-identified factors could be considered a physiological fatigue factor rather than workload factor. “Insufficient rest” is clearly related to a lack of sleep. Reports of “fatigue” may be related to either physiological need for sleep or mental exhaustion related to workload. Factors like “early starts”, “late finishes”, or “night flying” may be related to circadian fatigue even though the rosters in this analysis included predominantly daytime operations that did not encroach on the WOCL. Circadian fatigue is modelled in SAFTE-FAST as part of Effectiveness metric, which computes cognitive alertness as a function of time of day and sleep history[20]. SF Workload is modelled separately from Effectiveness, but the two metrics can interact to create an area of compound risk. A previous analysis from the IBR science team indicated that SF Workload in combination with low Effectiveness appeared to be useful for predicting instances of fatigue risk compared to Effectiveness alone[35]. Including insufficient sleep, night flying, or circadian swaps as workload factors runs the risk of double counting the impact of these issues on overall fatigue risk since they already contribute to predictions of Effectiveness. However, it is also possible that early morning or late/overnight operations incur additional workload independently from circadian fatigue due to

logistics or visibility issues. For example, many comments indicated that night flying was related to runway closure (see Supplemental Table 1). Teasing apart the circadian versus logistic underlying reasons for increased workload will require further investigation.

These findings should be interpretated in light of several limitations. Data were collected during three-day rosters that had been selected specifically because of workload concerns in the absence of circadian disruption and may not generalize to other roster patterns within the airline, to operations in other airlines, or to different global regions. Pilot responses may have been biased in a manner that cannot be assessed with the given data since demographics were not collected, and participation was voluntary. Moreover, these data only reflect workload at two timepoints over a 3-day period. Limiting data collection to two timepoints was intentional for this study because the flight rosters were known to be high workload and data were collected during actual commercial flight duty periods. Operational concerns about adding to pilots' workload by asking them to take multiple surveys overrode scientific curiosity in this field study. Assessing TLX more frequently throughout the study would have allowed us to potentially understand how workload changed over the roster with greater granularity but may have also contributed to the pilots' workload and potentially inflated their perception of the difficulty of other tasks.

Interpretation of the write-in factors may have lost some context during translation from Japanese to English even though translation was done by native Japanese speakers with extensive knowledge of aviation English. No data was collected on pilots' sleep history, work-life balance, or other individual factors that may have influenced perception of workload. Finally, comparisons of SF Workload against NASA TLX cannot be generalized to other predictive models of workload outside the SAFTE-FAST system, even if those models utilize the SAFTE model but incorporate workload in a different way.

Despite these limitations, the data in this analysis provides a detailed glimpse into pilot perception of workload during regional operations and a comparison to BMMF predictions of workload across days. SF Workload did a fair job of predicting workload equivalently to NASA TLX. This study represents an important step in establishing the

accuracy of workload predictions, but follow-up analyses will be necessary to ensure continued accuracy in workload predictions. Our current focus is to understand how to model aviation workload accurately in a manner that takes into account the demands of that specific operation. Once there is a clear picture of workload associated within operations, future analyses can compare between operations, such as comparing workload by geographical region, between operational lengths (e.g., short-haul versus long-haul), or between rosters that have a reputation for being high vs. low workload. Pilots made ample use of the write-in comment option, yielding a robust set of qualitative data about workload during these rosters. Comments can be used to further hone BMMF predictions or identify workload issues occurring in situ. These findings build upon the growing field of research investigating the impact of workload on fatigue in aviation[3-6, 30, 36].

Acknowledgements

The authors would like to acknowledge and thank all the pilots who volunteered their time to complete this study.

Authors' Contributions

All authors collaborated on the study design and data analysis plan. Authors JKD, KY, and JC analyzed and interpreted the data. Author JKD drafted the article. Authors KY, TT, KI, WT, JC and SRH critically revised the article. All authors read and approved of the final manuscript.

Ethical Approval

This study analyzed secondary data provided by Japan Airlines. Secondary use of de-identified data for research purposes was deemed non-human subjects research by Salus IRB on October 22, 2024 (Study ID: 23452) and these analyses were conducted in accordance with the Declaration of Helsinki.

Data Availability

The data that support the findings of this study are not available due to the sensitive and proprietary nature of the original data collection.

Conflict of Interest

Authors KY, TT, KI, and WT are affiliated with Japan Airlines but do not benefit financially or non-financially from this study protocol or the presentation of the study results. Authors JKD, and JC are affiliated with the Institutes for Behavior Resources, which provides sales of SAFTE-FAST, but do not benefit financially or non-financially from sales of SAFTE-FAST. Author SRH is the inventor of the SAFTE-FAST biomathematical model, and a fraction of his compensation is based on sales of the software.

REFERENCES

1) Fatigue Management Guide for Airline Operations. 2nd ed: International Air Transport Association (IATA); 2015.
2) Masi G, Amprimo G, Ferraris C, Priano L (2024) Correction: Masi et al. Stress and Workload Assessment in Aviation-A Narrative Review. Sensors 2023, 23, 3556. Sensors (Basel) **24**.
3) Bartulović D, Steiner S, Fakleš D, Mavrin Jeličić M (2023) Correlations among Fatigue Indicators, Subjective Perception of Fatigue, and Workload Settings in Flight Operations. Aerospace **10**, 856.
4) Masi G, Amprimo G, Ferraris C, Priano L (2023) Stress and workload assessment in aviation—A narrative review. Sensors **23**, 3556.
5) Arsintescu L, Chachad R, Gregory KB, Mulligan JB, Flynn-Evans EE (2020) The relationship between workload, performance and fatigue in a short-haul airline. Chronobiology International **37**, 1492-4.
6) van den Berg MJ, Signal TL, Gander PH (2019) Perceived workload is associated with cabin crew fatigue on ultra-long range flights. The International Journal of Aerospace Psychology **29**, 74-85.

7) Human Factors Aspects in Incidents/Accidents. Flight Operations Briefing Notes2022.
8) Mallis MM, Mejdal S, Nguyen TT, Dinges DF (2004) Summary of the key features of seven biomathematical models of human fatigue and performance. Aviat Space Environ Med **75**, A4-14.
9) Borbély A (2022) The two‐process model of sleep regulation: Beginnings and outlook. Journal of sleep research **31**, e13598.
10) Bhadra U, Thakkar N, Das P, Bhadra MP (2017) Evolution of circadian rhythms: from bacteria to human. Sleep medicine **35**, 49-61.
11) Wickens CD, Helton WS, Hollands JG, Banbury S (2021) Engineering psychology and human performance, Routledge.
12) Rubio S, Díaz E, Martín J, Puente JM (2004) Evaluation of subjective mental workload: A comparison of SWAT, NASA‐TLX, and workload profile methods. Applied psychology **53**, 61-86.
13) Jensen R (2015) Field study confirms the belief that keeping busy helps control room operators sustain alertness during the night shift. Procedia Manufacturing **3**, 1297-304.
14) Charlton SG, O'Brien TG (2019) Handbook of human factors testing and evaluation, CRC Press.
15) Baldo L, Giardino M, Floriani A, Gajetti M, Maggiore P (2023) Development of a Bio-Mathematical Crew Fatigue Model for Business Aviation Oper-ators.
16) Li Y, He J, Cao S, Zheng J, Dou Y, Liu C, Liu X (2023) Assessing Flight Crew Fatigue under Extra Augmented Crew Schedule Using a Multimodality Approach. Aerospace **10**, 933.
17) SAFTE-FAST Workload Calculator. SAFTE-FAST White Papers. SAFTE-FAST.com2020.
18) Gander PH, Mulrine HM, van den Berg MJ, Smith AAT, Signal TL, Wu LJ, Belenky G (2014) Pilot fatigue: relationships with departure and arrival times, flight duration, and direction. Aviation, space, and environmental medicine **85**, 833-40.
19) Powell D, Spencer MB, Holland D, Broadbent E, Petrie KJ (2007) Pilot fatigue in short-haul operations: Effects of number of sectors, duty length, and time of day. Aviation, space, and environmental medicine **78**, 698-701.
20) Hursh SR, Redmond DP, Johnson ML, Thorne DR, Belenky G, Balkin TJ, Storm WF, Miller JC, Eddy DR (2004) Fatigue models for applied research in warfighting. Aviat Space Environ Med **75**, A44-53; discussion A4-60.
21) Bourgeois-Bougrine S, Gabon P, Mollard R, Coblentz A, Speyer J-J (2018) Fatigue in aircrew from shorthaul flights in civil aviation: The effects of work schedules. In: Human factors and aerospace safety, 177-87, Routledge.
22) Marando I, Matthews RW, Grosser L, Yates C, Banks S (2022) The effect of time on task, sleep deprivation, and time of day on simulated driving performance. Sleep **45**, zsac167.
23) Zeller R, Williamson A, Friswell R (2020) The effect of sleep-need and time-on-task on driver fatigue. Transportation research part F: traffic psychology and behaviour **74**, 15-29.
24) Feltman KA (2016) The effects of time of day and circadian rhythm on performance during variable levels of cognitive workload.
25) Lamp A, Chen JMC, McCullough D, Belenky G (2019) Equal to or better than: The application of statistical non-inferiority to fatigue risk management. Accid Anal Prev **126**, 184-90.
26) Lakens D, Scheel AM, Isager PM (2018) Equivalence testing for psychological research: A tutorial. Advances in methods and practices in psychological science **1**, 259-69.
27) Powell D, Spencer MB, Holland D, Petrie KJ (2008) Fatigue in two-pilot operations: implications for flight and duty time limitations. Aviation, space, and environmental medicine **79**, 1047-50.

28) Hart S (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Human mental workload/Elsevier.
29) Haga S, Mizukami N (1996) Japanese version of nasa task load index sensitivity of its workload score to difficulty of three different laboratory tasks. The Japanese journal of ergonomics **32**, 71-9.
30) Hilditch CJ, Gregory KB, Arsintescu L, Bathurst NG, Nesthus TE, Baumgartner HM, Lamp AC, Barger LK, Flynn-Evans EE (2023) Perspectives on fatigue in short-haul flight operations from US pilots: A focus group study. Transport policy **136**, 11-20.
31) Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in psychology, 139-83, Elsevier.
32) Ashour L (2024) A review of user-friendly freely-available statistical analysis software for medical researchers and biostatisticians. Research in Statistics **2**, 2322630.
33) Etani N (2019) Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data. Journal of big data **6**, 85.
34) Ryley T, Baumeister S, Coulter L (2020) Climate change influences on aviation: A literature review. Transport Policy **92**, 55-64.
35) Devine JK, Choynowski J, Hursh SR (2025) Evaluation of a Biomathematical Modeling Software Tool for the Prediction of Risk in Flight Schedules Compared Against Incidence of Fatigue Reports. Safety **11**, 4.
36) Zhao D, Shenyang ZL, Hu L, Xin C, Zhou Y. Research on the evaluation model of pilot workload under day night alternation conditions. 2024 12th International Conference on Information Systems and Computing Technology (ISCTech): IEEE; 2024. p. 1-6.

Table 1. Roster Information

		Day 1	Day 2	Day 3
Roster 1	Leg 1	Tokyo- Shirahama	Sapporo-Fukuoka	Aomori-Tokyo
	Leg 2	Shirahama-Tokyo	Fukuoka-Tokyo	Tokyo-Misawa
	Leg 3	**Tokyo- Sapporo**	Tokyo-Aomori	**Misawa-Tokyo**
Roster 2	Leg 1	Tokyo- Shirahama	Sapporo-Fukuoka	Aomori-Tokyo
	Leg 2	Shirahama-Tokyo	Fukuoka-Tokyo	Tokyo-Obihiro
	Leg 3	**Tokyo- Sapporo**	Tokyo-Aomori	**Obihiro-Tokyo**
Roster 3	Leg 1	Tokyo- Shirahama	Sapporo-Fukuoka	Aomori-Tokyo
	Leg 2	Shirahama-Tokyo	Fukuoka-Tokyo	Tokyo-Asahikawa
	Leg 3	**Tokyo- Sapporo**	Tokyo-Aomori	**Asahikawa-Tokyo**

Table 2. TOST Equivalence Test: SF Workload Compared to NASA TLX at TOD

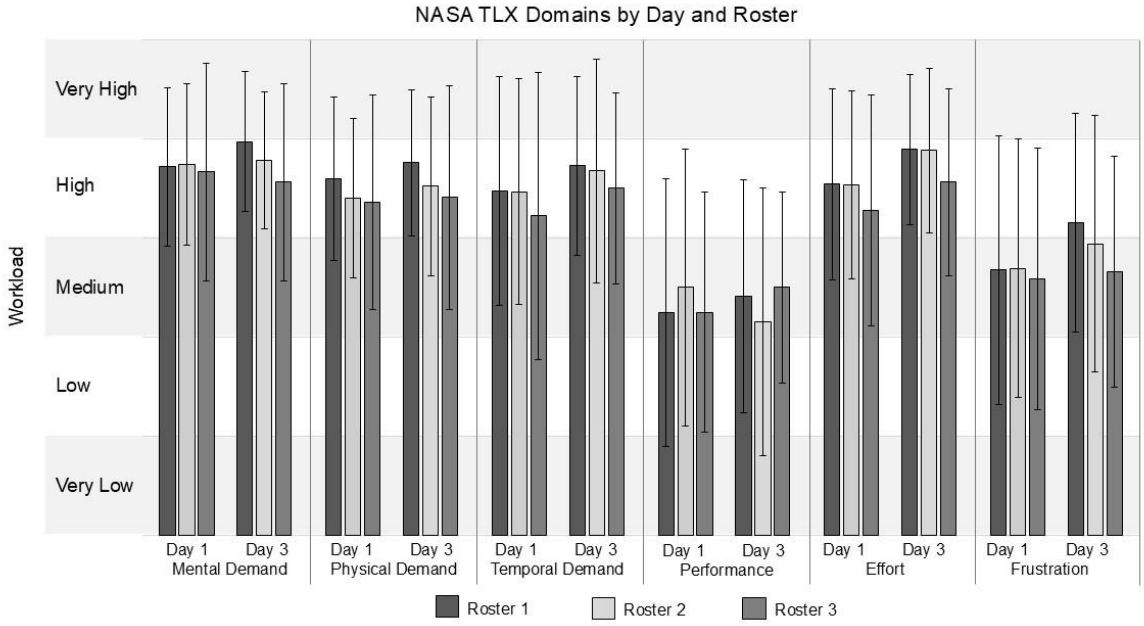
	SF Workload	NASA TLX
Mean	63.59	65.45
Variance	53.91	238.01
Observations	187	187
Equivalence Limits	0	0
Degrees of freedom	186	186
t statistics	1.56	-1.56
p values	0.06	0.06

Table 3. Write-in workload factors ranked by frequency of report by pilots

Rank	Workload Factor	Number of Reports	Percent of Pilots Reporting	Average TLX Workload Rating Associated with Report
1	Weather	90	48%	High
2	Delays	64	37%	Medium
3	Schedule Changes	37	23%	High
4	Time Pressure	30	15%	High
5	Airport Uniqueness	30	15%	High
6	Short Turnaround	25	13%	Medium
7	Stress or Frustration	22	13%	High
8	Inexperience	17	9%	High
9	Airport Congestion	17	8%	High
10	Overscheduled	13	5%	Medium
11	Mechanical Issue	13	6%	High
12	Late Finish	12	5%	Medium
13	Insufficient Rest	10	6%	Medium
14	Duty Length	9	5%	Medium
15	Number of Legs	8	4%	Medium
16	Night Flying	8	5%	Medium
17	Hunger	7	4%	Medium
18	Fatigue	6	3%	Medium
19	Air Crew Issues	5	3%	High
20	Passenger Issues	5	2%	High
21	Error	4	1%	High
22	Hotel Issue	3	1%	High
23	Airport Crew Issues	3	1%	High
24	Bird Strike	3	2%	High
25	Early Start	2	1%	High
26	Sits	1	0.4%	Low

Figure 1. Average Workload Ratings by NASA TLX Domain

A.



B.

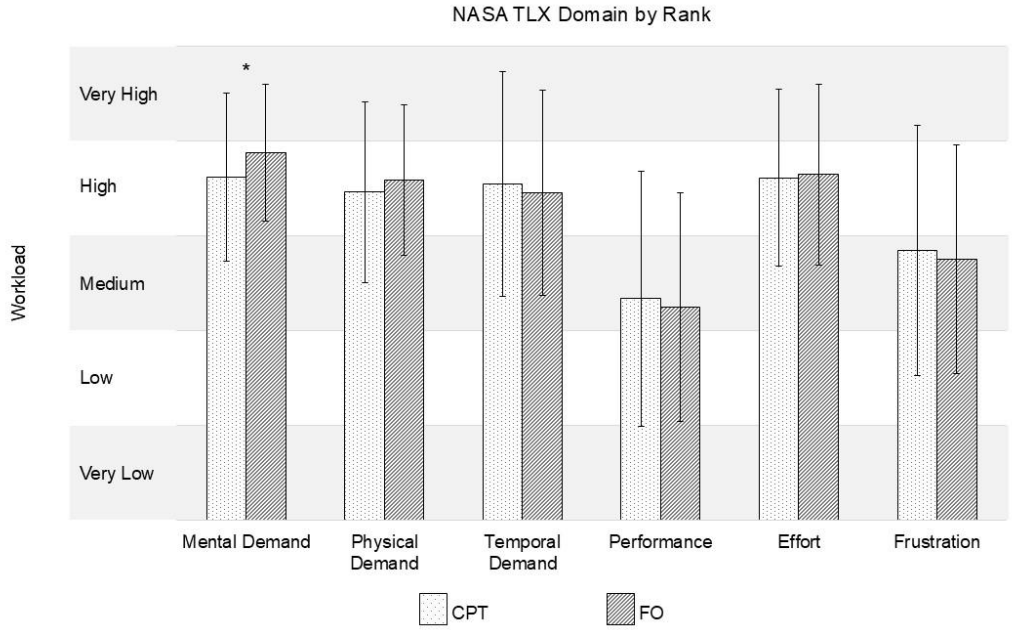


*Figure 1. Average Workload Ratings by NASA TLX Domain by A) Rosters and Days and B) Pilot Rank. Workload ratings are shown on a scale of Very Low to Very High on the y-axis. A) Rosters are shown as clustered bars organized by Day 1 and Day 3 across NASA TLX (mean and standard deviation) sub-domains on the x-axis. B) Rank is shown as clustered bars across NASA TLX (mean and standard deviation) sub-domains on the x-axis. * represents significance at p≤0.05.*

Figure 2. Predicted SAFTE-FAST Workload Compared to NASA TLX Global Scores

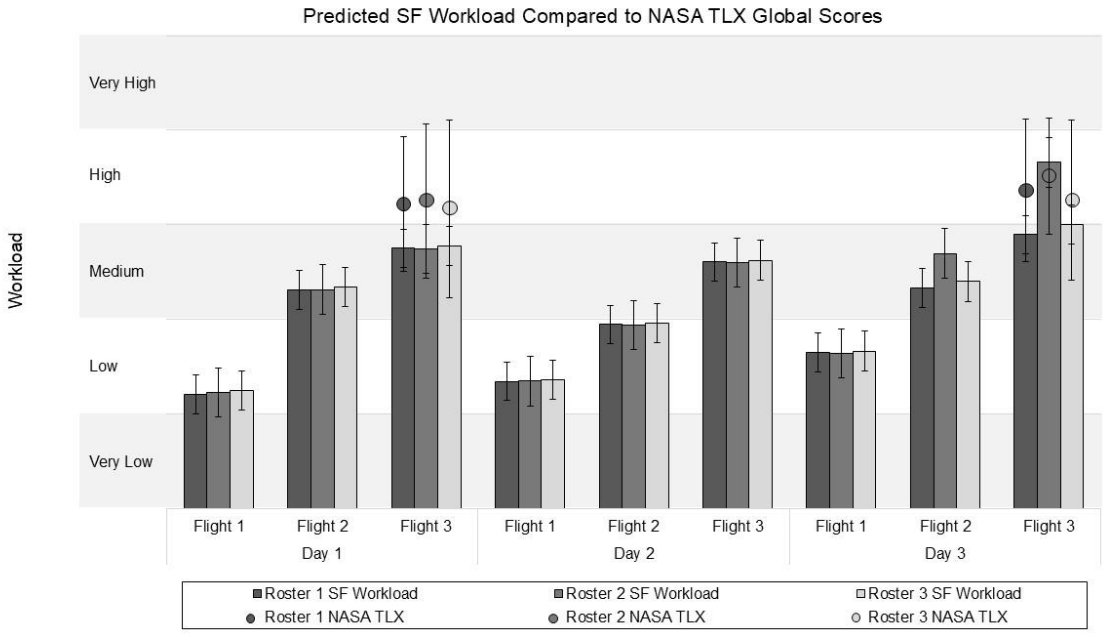


Figure 2. Predicted SF Workload (mean and standard deviation) are depicted as clustered bars across roster days and flights on the x-axis. NASA TLX global scores (means and standard deviations) are shown as markers above Flight 3 for Days 1 and 3 by roster. Workload ratings are shown on a scale of Very Low to Very High on the y-axis.

Figure 3. Equivalence Plot of Mean Difference in SFC Workload and NASA TLX at TOD ± 95% Confidence Interval

Figure 3. Graphical non-inferiority equivalence plot depicting the difference between SF Workload predictions and NASA TLX at TOD with 95% CIs. LEB (-3.23) and UEB (3.23) are depicted by the dashed vertical lines, and the equivalence limit (0) is depicted by a solid vertical line. The solid horizontal line depicts 95% CIs surrounding the mean difference between SF Workload and NASA TLX global scores. The mean difference between means (1.86) is indicated by the black circular marker at the center of the horizontal line.

Supplementary Table 1. Annotated List of Write-In Comments about Workload

Comment Number*	Comment
1.	I experienced time stress as the delay of the connecting flight caused a ripple effect, leading to delays in all subsequent flights. Additionally, the adverse weather added to the stress I was feeling.
2.	On the third day, we had marginal weather forecast and it ended up being an Autoland in borderline weather conditions. All three days involved long duty with three legs and short rest periods. Even without any issues, this pattern is tough. With Ground Turn Back and significant delays on the first day, and delays on the second day as well, it becomes mentally exhausting. Having marginal weather on the third day makes it extremely challenging. I think we should consider reducing it to two legs or extending the rest periods to allow for more recovery time.
3.	I was aware of the decreased performance, so I needed to proceed cautiously. The delay in the arrival of the assigned aircraft. Developed convective clouds. The unique procedure at RJSM airport. Weather avoidance. The need to switch control due to the First Officer's limitation.
4.	I didn't feel that the high workload was caused by flight patterns or aircraft issues. The poor weather increased the workload, especially since the first flight was delayed, causing a ripple effect for the next three legs. Constantly being concerned about delays and operating in bad weather made the operation feel overworked.
5.	I felt that the workload was higher than usual due to the infrequent opportunities to visit and the slightly more complex operations compared to other airports. Additionally, being the last day of the pattern, the accumulation of fatigue from the first two days also had some influence. On the third day, the weather was unstable compared to the earlier days, requiring closer attention to altitude selection and punctuality. As a result, the workload increased.

6.	On the second flight, the arrival was delayed due to congestion, and during the short flight interval, we discovered an equipment malfunction during the external inspection and took appropriate measures. However, there were deficiencies in verifying the procedures, which further contributed to the delay. It was unusual to have two copies of the same procedures on board. Since there was no notification about this situation, I checked the procedures provided by the maintenance personnel and processed accordingly. Eventually, I felt something was amiss with having two copies and realized that I had mistakenly used the non-effective procedure, prompting me to check again and process using the effective procedure. As a result, the departure of the final flight was delayed, and there was a time-pressure to make up for the delay. During the descent, we were slow to react to the changing winds, resulting in an increase in descent speed. After arriving at the hotel, I made the required documents. The next day, I had to show up early again, resulting in a shorter rest period.
7.	Adverse weather conditions and managing turbulence. Consideration towards service. During the final leg of the flights, there were reports of severe turbulence due to clear air turbulence and cloud effects. Flying at night with limited visibility of clouds requires extra attention to seat belt operations. Fatigue from the previous patterns may have resulted in reduced judgment and situational awareness.
8.	There was an overall delay due to bad weather at destination, resulting in delays for all flights. In the midst of this inclement weather and on short legs, there were challenges in avoiding clouds and selecting the appropriate approach type. Additionally, the airport has a short runway, making approaches challenging in strong winds. There was heavy rain to the extent that a warning was issued. There was a sudden air traffic controller instruction for a low-altitude level off at 2000ft shortly after takeoff due to the presence of a go around aircraft. There were active rain clouds along the departure route.
9.	The flight was handled may return due to heavy fog. We actually made a landing that was right at the minimums. The following day, I was scheduled for a long-awaited flight and had to prepare for it. Due to insufficient rest at home, I was feeling a bit fatigued.
10.	It was slightly taxing to deal with the multitude of local procedures, which is known for having the most among domestic airports. Although the weather wasn't poor, the strong turbulence at higher altitudes kept us on edge and contributed to a sense of fatigue.

11.	We experienced equipment malfunction. The cabin temperature control was clearly faulty. This meant frequent manual adjustments were necessary to prevent the cabin from becoming extremely hot or cold, including during critical phases like takeoff and landing. I probably operated the control switch at least 50 times across the three legs. On the third leg, we encountered moderate turbulence at a location that couldn't have been predicted beforehand. Our approach preparation fell behind as a result. While we can't control bad weather, the organization must prioritize aircraft maintenance. Neglecting it only increases workload and pressure on the crew.
12.	Regarding the schedule, we were unable to keep the on-time performance for even one leg over the three days. The reasons were: on the first day, there was a ground turn back; on the second day, there was an assignment of delayed expected departure clearance time; and on the third day, the initial flight was delayed due to boarding (it took time for 16 young passengers to disembark). On each of these days, the delays extended to the final leg, lengthening our flight time. This undeniably increased our fatigue. Additionally, the weather was extremely poor, may return to HND condition. We monitored the situation closely, focusing on the instruments more than usual due to the tense situation of whether we could land or not. We arrived at the stay hotel around 11 PM, and it cannot be said that we had sufficient rest. Furthermore, personally, it was a training pattern, and since it was an airport I had no experience with, I felt even more stimulated during the flight.
13.	During the final flight, there was a very large cluster of cumulonimbus clouds in the northern Kanto region, and it took considerable effort to decide on a course to avoid them. Although we did not get directly hit by lightning, we saw lightning flashes very close by, which significantly increased mental fatigue. While not related to workload, from the perspective of fatigue, I felt that the hotel arrival time after the flight was late, yet the pickup time the next day was early. I did not find the pattern on the first day to be particularly fatiguing.
14.	The flight started with an aircraft arrival delay and could not catch up with the delay throughout the three legs. There was no margin between flights, so the workload increased due to the inability to make up for the delay. Additionally, there was an impact from the rainy season front and strong winds.
15.	Airport curfews, competing flights with other airlines, and turbulence caused by jet streams.
16.	I felt that the workload was high on the second day, and I carried that fatigue over to the third day. Although the fatigue level was also high on the third day, strictly speaking, the fatigue was higher after the completion of the second day. As for the pattern, I would like to request Duty Off with 2 legs. On a side note, the Cabin Crew had higher workload

	on the third and fourth days of the four-day pattern, as they were on the same flight as us.
17.	Regarding the second day, I believe it would be beneficial to end the day with the two-leg journey. On the second day, there is a high probability of flow control and delays during the evening flight, which results in longer duty hours. Additionally, there is a risk of thunderstorms, adding to the already high workload. Given the potential fatigue from the previous day, I feel that going to an airport which is known for its high Threat level, for the last two legs on the third day could potentially induce errors.
18.	The runway was closed due to a bird strike, causing fuel levels to drop below the planned reserve. Therefore, a decision was required to divert to the alternate airport.
19.	There was a sense of regret towards the customers due to the unrealistic schedule, which made it impossible to arrive on time during late hours. As a result, although the frustration was not high, there was a significant sense of resignation. This situation has been ongoing for many years, and to manage stress, I have been coping by not accumulating frustration over delays and accepting the situation. I believe this is not a good approach, but I hope that surveys like this will lead to some improvements. The flight schedule of this route falls into the late-night category, and there was limited standard terminal approach route information available upon arrival, which increased the workload during the preparation stage and when transferring. The final round trip had very few passengers on both flights, allowing for on-time operations with ease and unusually low workload.
20.	The first day's flight is likely to become special night arrival, so there is time pressure to avoid that, and the stress level is high due to the short rest time at the stay. Additionally, despite it being a three-day pattern, all three days are busy with flights.
21.	Due to night operations and the landing scheduled just 5 minutes after the runway closed, it was a challenging approach to anticipate. In fact, it was the first time an unfamiliar approach procedure had been assigned.
22.	The short stay time and the fact that the start-up times on the second and third days are earlier compared to the first day make it feel like I haven't fully recovered from fatigue at the stay location. I felt that the short rest time, rather than the specific airports in this pattern, contributed to the fatigue on the final day. Additionally, arriving late at night and having an early start-up time leaves little time to go out for lunch, which also adds to the stress.
23.	Despite our flight being the only one during that time slot, the security checkpoint was inexplicably crowded, causing delays in boarding.

24.	On the first flight, there was about a one-hour delay due ground turn back. Before the departure of the third flight, maintenance was required, making it very difficult to keep the schedule.
25.	There is nothing specific, but if I had to say, it would be the late-night arrival which is not common. Because it is late at night, instead of the usual standard terminal approach route, there are several possibilities which might be assigned, so there is some stress in preparing for that and knowing what will be specified. However, there are no radar vectors, so once it is determined, it is normal from there.
26.	I can imagine that there are a lot of things to consider, especially after dealing with a challenging airport followed by an unfamiliar late-night flight.
27.	Due to congestion caused by bad weather and the fact that the bus for deplaning passengers at the remote stand took more than 15 minutes to arrive, there was a delay of nearly 30 minutes. As for the third day's pattern, I feel that the physical burden on the third day is lower compared to the second day because there is sufficient rest time before the third day, unlike the second day which starts with insufficient rest.
28.	The tiredness from the second day of a long flight is significant. The route is known for its high number of delays caused by foreign passengers with excessive baggage and not adhering to boarding times. Additionally, there are unique approaches and a multitude of local procedures involved.
29.	We performed a round trip in situations where weather deviations were required. It was a constantly time-pressured situation on short legs, and after completing the round trip, the other pilot and I had a conversation saying, "We're tired." The turn around was also short, and we couldn't make time to have a proper meal. We managed to curb our hunger with our own energy gel substitutes.
30.	If there is a delay in the first leg (which I think was inevitable due to spot occupied), it will further delay the second leg during the time when the runway is in use, causing ground congestion. As a result, there was no buffer time between flights, and I felt time pressure in preparing during the short turn around. In the final leg, there was turbulence and a change in approach type, which made it a bit hectic.
31.	We had planned for a standard terminal approach route during the late night hours, but we received different clearance, so we had to modify the flight path and temporarily experienced an increased workload due to inputting it into the flight management system.
32.	As it was my first time experiencing the night operations at the airport, there were moments where my workload increased due to having to deal with situations that I couldn't understand even after reading company materials.
33.	I operated this route for the first time. There was a higher level of tension and fatigue than usual due to various first-time experiences, from preparations to actual setup. Additionally, the delayed arrival of the

	previous flight and shorter flight intervals, along with an inquiry about minimum equipment list application, made for a busy setup. There were high workloads due to altitude selection reconsideration and deceleration instructions caused by congestion, and there were also feelings of disappointment for not being able to keep to the scheduled time. There were preparations before the flight and handling during the flight based on the special runway use and approach late-night, resulting in a higher workload compared to regular flights.
34.	The time required for transferring passengers from canceled flight created a delay and increased workload due to the need for pilot in charge consultation for additional fuel considering the heavier zero fuel weight. Even a slight delay can lead to increased workload due to ground holding and possible runway changes. Therefore, we anticipated these scenarios and added more fuel on top of the original plan. Although there was no runway change, there was heavy congestion. The pilot in charge shared an experience of significant delay caused by a runway change after blocking out. We were able to take preemptive measures this time, but it is difficult to predict such situations when there is no prior experience or when there is busyness between flights. It is expected that the planned fuel will be insufficient, resulting in increased workload. I strongly recommend considering departure times and fuel plans based on the assumption of runway operations. Although there were no significant irregularities, having 9 legs in 3 days increased workload and fatigue with even minor issues. The workload on the second day, in particular, felt extremely high. Generally speaking, it seems that the measures taken to maintain punctuality during south wind operations are quite fragile. It is easy to imagine that delays would increase the workload for everyone involved. The 3 days were completed without any incidents, but I believe there were quite a few threats.
35.	The stress levels increased due to the congestion in the parking area, resulting in a higher workload for preparing for the next flight.
36.	The flight patterns were causing delays. In such circumstances, efforts are made to operate in a way that maintains the schedule, but it creates stress. Considering that the next day's pickup time is early despite being on a 3-day pattern, and the fact that it is a 3-leg pattern for the entire day, and the long flight duty period, it is tiring. I would like to have a working schedule that allows for longer rest periods, at least.
37.	There is a major issue with the pairing itself. Based on over 30 years of experience, it is believed that there are two solutions. One is to end the pairing on the second day second flight. The second is to end the pairing on the third day first flight. Completing this pairing, with only one day off, seems like a chaotic pairing. I would like to request those who created and approved this pairing to experience the difficulty themselves. I strongly urge you to taste the hardship firsthand.
38.	I felt that it would be physically challenging to have two consecutive days with a pairing consisting of three legs and exceeding a total block time of

	5 hours. In the final leg, I found myself considering altitude selection based on my own fatigue rather than operational efficiency. Specific signs of fatigue included having to ask air traffic controller for clarifications multiple times and needing to take a short break between flights before starting the next task.
39.	I felt the elevated workload due to the last-minute approach change, as the scheduled closure of runway was set after 22:00, while the standard was 22:05. Despite having an estimated time of arrival of 21:58, we were forced to change runways. Additionally, landing on the unfamiliar runway required considerable attention for energy management. With a small number of passengers, I didn't feel the significance of this charter flight.
40.	First, due to a significant delay caused by a aircraft change, I was assigned to another flight and then took charge of the flights with a delay of approximately 1.5 hours. In addition to the initial delay in the arrival of the aircraft, the duty flights were changed several times, requiring multiple re-tests of the alcohol check and briefings. Furthermore, taking charge of the flights, which were delayed by 2 hours, increased fatigue and time pressure. Originally, we had only a 35-minute turnaround time, and with few passengers, it was likely that they would be waiting for our cockpit preparation. Despite proceeding in an orderly manner, we were repeatedly contacted by the station officer regarding a 5-minute delay in acknowledging the weight and balance, which caused significant disturbances. This part was where the workload increased the most.

**Comments are not presented in any specific order*