SAFTE® FAST

The Science of Performance at Work

Sleep and Fatigue During COVID-19 Humanitarian Missions

Collaborators: Institutes for Behavior Resources, Inc., and Azul.

Analysis conducted by: Jaime K Devine, PhD., Caio R Garcia, Audrey S Simoes, Marina R Guelere, Bruno de Godoy, Diego S Silva, Philipe Pacheco, Jake Choynowski, Steven R Hursh, PhD.

Contents

Introduction

In response to the COVID-19 pandemic, Azul Airlines organized and conducted five separate humanitarian missions to China to bring respirators, COVID rapid tests, and medical supplies to Brazil. These missions were conducted between May and July of 2020. This was the first time in the history Azul Airlines had flown to China. Successful completion of these operations required extensive logistical preparation. Not only were the pilots unfamiliar with their destination airports in China, but the direct round-trip flights required 30 hours of in-flight travel. Moreover, to prevent exposure or potential transmission of the novel virus SARS-CoV-2, Azul's 8-pilot crew remained on the plane during time on-ground in China. To mitigate the fatigue risk factor associated with the missions' long flight duty periods (FDP) and poor-quality rest facilities, Azul used SAFTE-FAST fatigue modeling software to forecast an estimate of pilot fatigue and effectiveness using the Auto-sleep feature. Auto-sleep is an algorithm within SAFTE-FAST that constructs sleep events based on the likelihood of sleep occurring and the parameters of the scenario.

During the COVID-19 support flights, the 8-pilot crew remained on the plane during flight and layover periods. Crew slept either in crew rest facilities or in the business class section, per their preference, and did not disembark during their time in China. Pilots wore a sleep-tracking actigraphy device (Zulu Watch, Institutes for Behavior Resources), and reported the timing, duration, and quality of their in-flight rest periods using a sleep diary. This white paper describes the comparison of SAFTE-FAST's fatigue prediction using Auto-sleep against pilots' self-report sleep diary and objective sleep patterns as measured by the Zulu watch from Azul's five COVID-19 humanitarian missions.

Methods

Auto-sleep Parameters for COVID-19 Humanitarian Missions and Planned Work Sleep

A separate SAFTE-FAST project file was created for each of the humanitarian missions (Mission 1-5). Trip pairings were modelled in SAFTE-FAST using Azul's default setup package. Planned Work Sleep Event Rules (for example, inflight sleep) were manually entered for each trip to reflect mission specifics. The flights were designed to be carried out with 2 relay crews consisting of 8 pilots. The crews were organized so that all pilots would be available to work during any flight leg and that no one pilot would need to fly extra time. In-flight rest periods were freely chosen by the crew during the mission. Each flight leg was approximately 12 hours and the available rest time for each crew member per stage was approximately 9 hours. Auto-Sleep predicted planned work sleep using an augmentation rule that prohibited work sleep events from occurring with 30 minutes of beginning an FDP or within 90 minutes of ending an FDP. One work sleep event per crewmember was assumed to occur during any given FDP. The length of the sleep event was computed by subtracting the time when sleep was prohibited (120 minutes) from the total flight duration and then dividing that time by two (two crews) and multiplying by 0.75 (75% credit for sleep during that opportunity). Azul used no Event Rules, meaning that they manually chose to add the Sleep Rule calculation to the specific flights. Planned work sleep quality was set to "Good", assuming two interruptions per hour that each cost 5 min of sleep time, or 50 min of restorative sleep per hour.

Collecting Sleep Data During COVID-19 Humanitarian Missions

Pilots were assigned the Zulu watch (Institutes for Behavior Resources, Inc.) in May 2020 prior to COVID-19 support missions and wore the watches continuously until the completion of their mission (between May and July, 2020). The Zulu watch is a validated actigraphy device that has off-wrist detection to help differentiate between sleep periods and off-wrist periods, can detect multiple sleep episodes per day, and can detect naps as short as 20-minutes. The Zulu watch continuously collects sleep data and stores the most recent sleep events (up to 80 sleep events) on on-wrist. The Zulu watch also records sleep efficiency (SE), a measure of sleep quality which represents the amount of time spent actually sleeping over the total amount of time when sleep was attempted. Crews returned the watch to airline researchers directly upon returning to Brazil from their mission. Data were downloaded by airline researchers using the Zulu Data Extraction application (Institutes for Behavior Resources, Version 2.0) and saved as .CSV files.

Pilots completed the sleep diary during FDP. Pilots were not asked to complete the sleep diary during layovers in Europe or ground time in China. All times were reported in Brazilian time. Aircrew were instructed to remain on home base Brazilian time during layovers. The pilots reported the flight leg, duty start time, flight time, the timing, duration, and quality of their pre-flight and in-flight sleep, and provided self-assessment of fatigue using the Karolinska Sleepiness Scale (KSS) and the Samn-Perelli scale. Subjective sleep quality was rated on a 4-point scale as either Poor, Fair, Good, or Excellent by pilots. Diary sleep information was manually compiled into a .CSV file which could be imported into SAFTE-FAST. Sleep, fatigue, and performance metrics are summarized in Table 1.

Table 1. Sleep, Fatigue and Performance Metrics

Metric Name	Definition
Sleep Opportunity Time	The sleep opportunity time is the amount of non-crewing time, layover time, or other “down time” during which crew members have the opportunity to sleep.
Time in Bed (TIB)	TIB is a measure of the amount of time that crew attempted to sleep/dedicated to sleep. TIB is expressed as the total time, in minutes, between the beginning and the end of a sleep event.
Total Sleep Time (TST)	TST is a measure of actual sleep. TST is expressed as the total time, in minutes, during a sleep event when sleep was actually occurring.
Sleep Efficiency (SE)	SE is a measure of sleep quality which represents the amount of time spent actually sleeping over the total amount of time when sleep was attempted. It is commonly calculated as a percentage of TST/ TIB.
Sleep Quality	Sleep quality is a measure of the crewmember’s satisfaction with their sleep. Crew select from 4 sleep quality options based on their subjective experience following the sleep event. The

Metric Name	Definition
	options are: Excellent, Good, Fair, and Poor.
Sleep Environment	Sleep environment quality refers to the potential for interruptions to sleep due to the quality of the environment. There are 4 sleep environment quality options which can be selected based on knowledge of the sleep environment. Sleep environment quality uses the same labels as sleep quality but may not necessarily represent the crewmember's subjective sleep experience.
Karolinska Sleepiness Scale (KSS)	The KSS is a 9-point Likert scale used to report subjective fatigue. Scores range from 1-9, with 1 being extremely alert and 9 being very sleepy/great effort to keep awake.
Samn-Perelli Scale	The Samn-Perelli Scale is a 7-point Likert scale used to report subjective fatigue. Scores range from 1-7, with 1 being fully alert, wide awake and 7 being completely exhausted/unable to function effectively.
Crewing Effectiveness	Effectiveness represents speed of performance on the Psychomotor Vigilance Test (PVT), scaled as a percent of a fully rested person's normal best performance. Effectiveness corresponds to the speed of cognitive performance, is highly sensitive to fatigue, and correlated with many other cognitive performance metrics. The higher the score, the lower the fatigue risk. Crewing effectiveness refers to effectiveness during crewing events. Minimal effectiveness during critical phases of flight is considered to be 77%.

The original SAFTE-FAST models of COVID-19 humanitarian mission flights using Auto-sleep were duplicated to produce two comparison scenarios, one with explicit sleep scenarios based on Zulu watch objective sleep and one based on the subjective sleep diaries. Both diary and Zulu watch sleep scenarios used sleep start time (indicating sleep onset) and sleep end time (indicating the time of final awakening) to model explicit sleep duration.

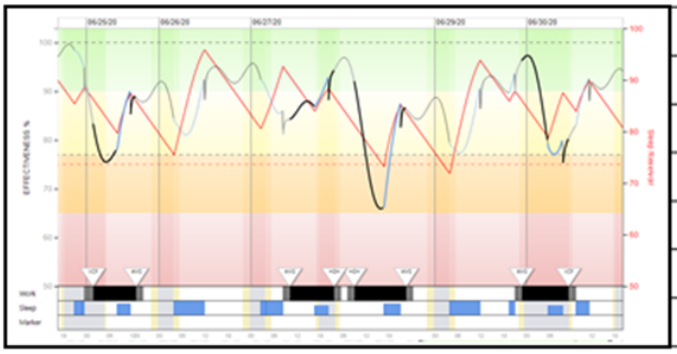
The subjective sleep diary scenario adjusted Environment settings based on subjective sleep quality ratings, such that Excellent sleep assumed no interruptions, or 60 min of restorative sleep per hour, Good assumed 2 interruptions per hour or 50 min of restorative sleep per hour, Fair assumed 4 interruptions per hour or 40 min of restorative sleep per hour and Poor assumed 6 interruptions per hour or 30 min of restorative sleep per hour.

Environment was kept at Excellent for all Zulu sleep data, since the watch considers sleep interruptions as awakenings (see Figure 1). SAFTE-FAST estimated performance metrics based on the crew member' flight schedules and sleep input. Crew schedules were identical across all scenarios within a mission project. Therefore, differences in performance metrics are assumed to be due to differences in the reporting of sleep (i.e., Auto-sleep vs. Diary vs. Zulu watch). An

example of the 3 sleep scenarios for an individual pilot is depicted in Figure 1.

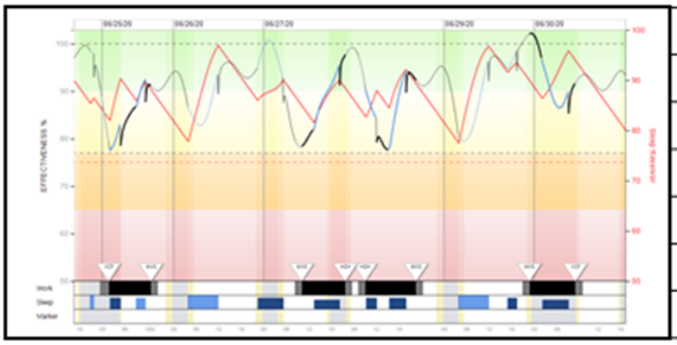
Figure 1. Example Comparison between Auto-sleep, Diary, and Zulu watch SAFTE-FAST Scenarios

A) AutoSleep



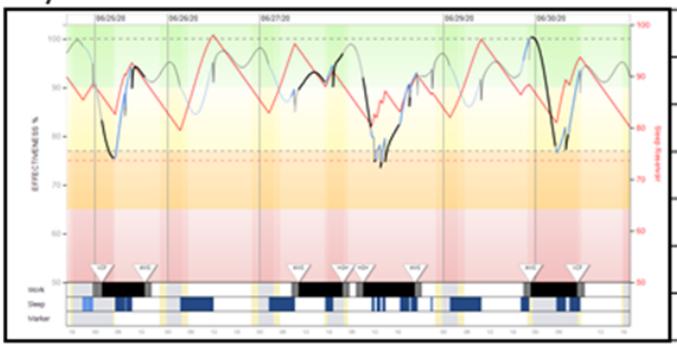
Sleep Opportunity	540 mins
Environment Setting	Good
# Work Sleep Events	1
TIB	229±23 mins
TST	188±19 mins
Effectiveness	83±6
Reservoir	83±4

B) Diary



Sleep Opportunity	540 mins
Sleep Quality	Fair/Good
# Work Sleep Events	1 to 2
TIB	360±128 mins
TST	264±91 mins
Effectiveness	88±6
Reservoir	88±2

C) Zulu watch



Sleep Opportunity	540 mins
Sleep Efficiency	87%±3%
# Work Sleep Events	1 to 5
TIB	270±115 mins
TST	240±106 mins
Effectiveness	87±5
Reservoir	87±3

Figure 1. SAFTE-FAST modelled performance using A) Auto-sleep, B) sleep diary or C) Zulu watch data for one pilot. Crewing information (flight leg dates, duration, destinations, etc.) and sleep opportunity time were the same across all 3 scenarios. Auto-sleep predicted shorter TIB and TST, lower effectiveness, and lower reservoir during crewing periods than either diary or Zulu watch data, but differences were not significant. Auto-sleep predicted 1 work sleep event per flight leg; the pilot reported taking 1-2 work sleep events per flight leg, with either fair or good sleep quality. Zulu watch data indicated that the pilot awoke multiple times while attempting sleep, which accounts for the higher reported number of work sleep events and sleep efficiency.

Sleep metrics were exported from SAFTE-FAST as .CSV files. Zulu watch SE percentages were converted into sleep quality categories as described in Table 3 to allow for a comparison to Auto-

sleep and diary sleep quality.

Table 3. Sleep Quality Category Equivalents to Sleep Efficiency Percentage Ranges

Sleep Quality Category	Sleep Efficiency Range
Excellent	>83%
Good	68%-83%
Fair	51%-67%
Poor	≤50%

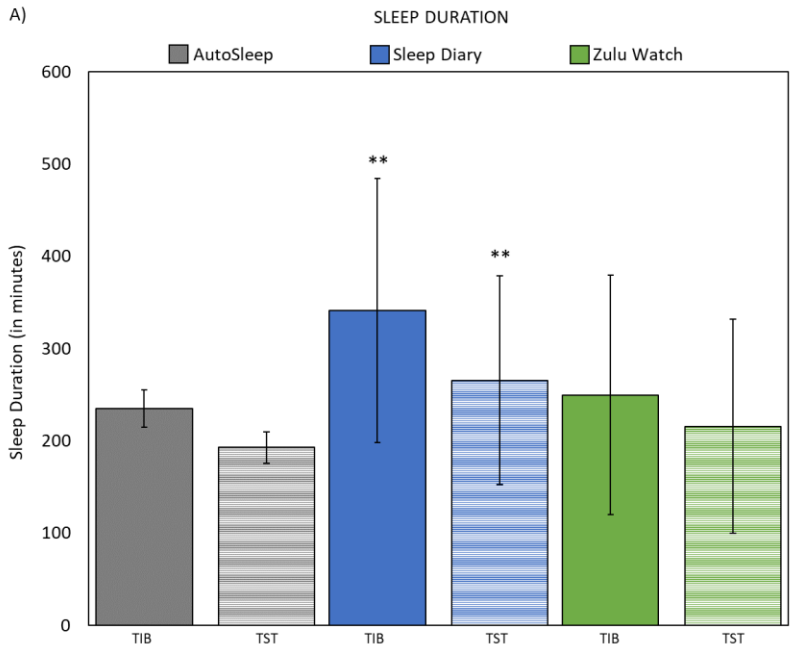
Thirty-two out of 40 (80%) pilots crewing a COVID-19 humanitarian mission completed the sleep diary and 22 out of 40 (55%) wore a Zulu watch between May and June 2021. Twenty (20; 50%) pilots completed both the sleep diary and the Zulu watch. Only pilots who both completed the sleep diary and provided Zulu watch data (N=20) have been included in these analyses.

Results

Sleep and Sleep Quality During COVID-19 Humanitarian Missions

Sleep metrics were compared between the 3 sleep scenarios (Auto-sleep, Zulu, and Diary) for each mission flight (Mission 1-5) using STATA MP 15.1 statistical analysis software and Excel 2013. Statistical significance was assumed at $p<0.05$. Differences in TIB, TST, and sleep quality between Auto-sleep, sleep diary, and Zulu watch measurements were explored using paired samples t-tests and one-way analysis of variance (ANOVA). Figure 2 shows A) in-flight sleep duration and B) the distribution of sleep quality across flight legs.

Figure 2. In-Flight Sleep Duration and Quality During COVID-19 Humanitarian Missions



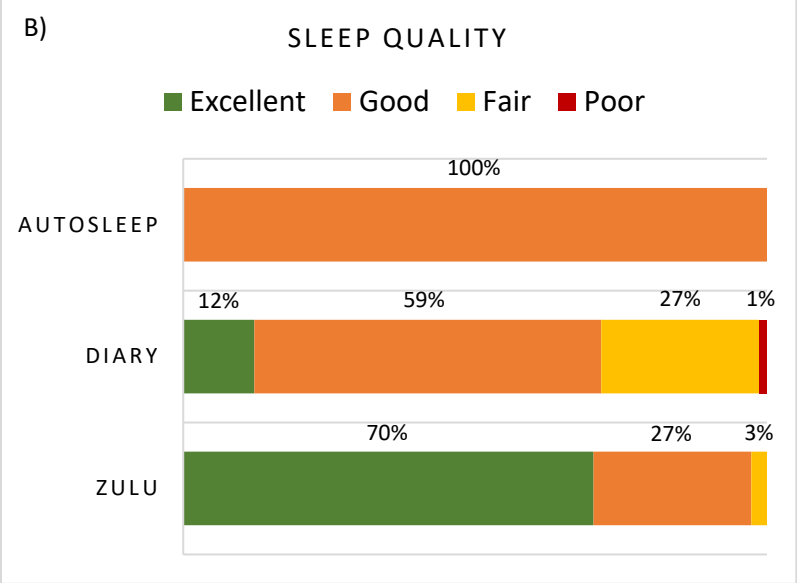


Figure 2. A) Diary reports of TIB (solid blue) and TST (blue stripes) were significantly higher than Auto-sleep predictions (in gray) of TIB (t=6.80, p≤0.001) and TST (t=5.60, p≤0.001) and Zulu watch (green) TIB (t=5.34, p≤0.001) and TST (t=3.31, p≤0.001). Auto-sleep and Zulu watch measures of TIB (t=6.80, p≤0.001) and TST (t=1.59, p=0.12) were not statistically different. Significance is indicated by an asterisk (*) at p≤0.05 and double asterisk (**) at p≤0.001. B) Auto-sleep assumed "Good" sleep quality for 100% of in-flight sleep. Pilots reported "Good" sleep quality for the majority (59%) of in-flight sleep, and Zulu watch sleep efficiency approximated "Excellent" quality sleep for 70% of in-flight work sleep events. See Table 3 for more information on sleep efficiency-to-quality approximations.

Agreement between sleep metrics was furthermore evaluated using single rater, two-way random effects intraclass correlation coefficients (ICC) with absolute agreement. ICC values were classified as poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), or excellent (>0.90) based on established guidelines (Koo and Li 2016).

Table 5. Intraclass Correlation Coefficients between Auto-sleep, Sleep Diary, and Zulu Watch

Summary Statistic	ICC	95% CI	Probability That ICC=0 [1]	Inter-Rater Reliability
Time in Bed (TIB)	0.94	0.79–0.99	p≤0.001**	Excellent
Total Sleep Time (TST)	.91	0.65-0.99	p≤0.001**	Excellent
Sleep Quality	0.98	0.94-0.99	p≤0.001**	Excellent

[1] Significance for the probability that the coefficient is zero, implying no agreement; *p<0.05; **p≤0.001.

Fatigue During COVID- 19 Humanitarian Missions

Pilots reported their fatigue, as measured by KSS and Samn-Perelli, during each flight leg. Self-reported fatigue was compared against SAFTE-FAST predictions in an attempt to gauge the modeling of fatigue. Fatigue predictions in SAFTE-FAST are based on a transformation of predicted effectiveness (with 100% indicating an individual's best possible effectiveness) to provide additional context to the user (Figure 3). SAFTE-FAST predictions of subjective fatigue have not been previously calibrated to actual KSS or Samn-Perelli scores.

Figure 3. SAFTE FAST Effectiveness Percentage Comparison to KSS and Samn-Perelli Scale

A)

SAFTE-FAST Effectiveness	Karolinska Sleepiness Scale	Definition
103	1.0	
100	1.0	1 - Extremely alert
99	1.5	
97.7	2.0	2 - Very alert
96	2.7	
95.25	3.0	3 - Alert, Normal Level
95	3.1	
94	3.4	
93	3.7	
92	4.0	4 - Rather alert
91	4.3	
90	4.6	
88.5	5.0	5 - Neither alert nor sleepy
86	5.5	
83.5	6.0	6 - Some signs of sleepiness
82	6.3	
80	6.6	
77	7.0	7 - Sleepy, but no effort to keep awake
75	7.2	
72	7.5	
70	7.7	
67	7.9	
65	8.0	8 - Sleepy, some effort to keep awake
60	8.3	
40	8.8	
20	8.9	9 - Very sleepy, great effort to keep awake

B)

SAFTE-FAST Effectiveness	Samn-Perelli	Definition
103	1.0	
100	1.0	1 - fully alert, wide awake
99	1.3	
97.5	1.8	
96.8	2.0	2 - very lively, responsive, but not at peak
96	2.3	
95	2.6	
94	2.8	
93.2	3.0	3 - okay, somewhat fresh
92	3.3	
91	3.5	
90	3.7	
88.5	4.0	4 - a little tired, less than fresh
86	4.4	
83.5	4.8	
82	5.0	5 - moderately tired
80	5.2	
77	5.5	
75	5.7	
72	5.9	
70	6.0	6 - extremely tired, very difficult to concentrate
67	6.2	
65	6.3	
60	6.5	
40	6.8	
20	7.0	7 - completely exhausted, unable to function effectively

Figure 3. SAFTE-FAST effectiveness compared against A) KSS and B) Samn-Perelli ratings and definitions. The KSS ranges from 1-9. The Samn-Perelli scale ranges from 1-7. Higher scores indicate greater fatigue.

SAFTE-FAST predictions of fatigue from the Auto-sleep scenarios were compared to pilots' diary reports of KSS and Samn-Perelli for each mission flight leg using paired samples t-tests. The average crewing KSS and average crewing Samn-Perelli Scale prediction metrics were used for these comparisons. Statistical significance was assumed at $p<0.05$. SAFTE-FAST predicted higher scores on the KSS and the Samn-Perelli Scale than pilots' self-report for all flight legs (all $p \leq 0.05$). Average fatigue ratings for each flight leg across all missions are shown in Figure 4.

Figure 4. Predicted and Reported Fatigue During COVID-19 Humanitarian Mission Flight Legs

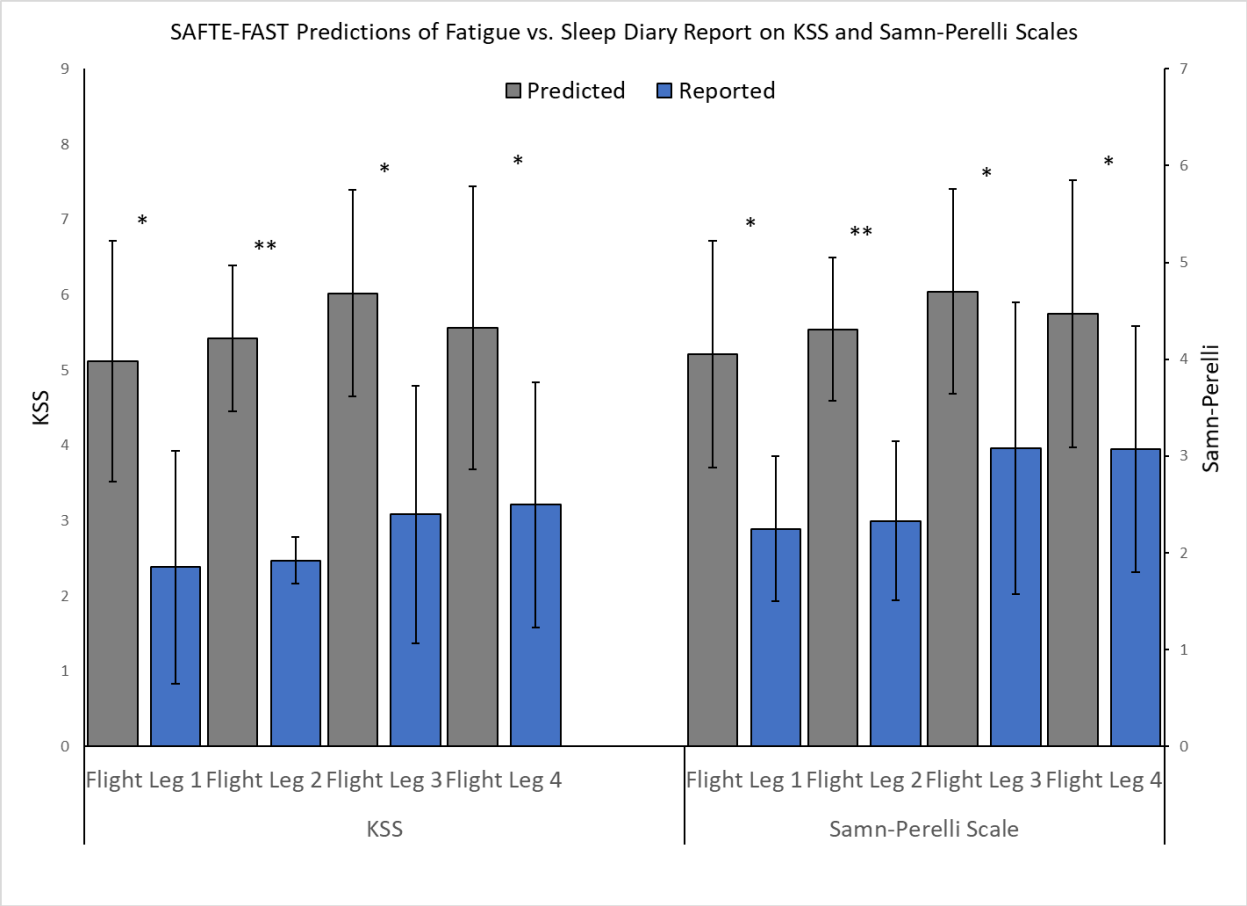


Figure 4. Auto-sleep predictions (in gray) of KSS (left) and Samn-Perelli Scale (right) were statistically higher from pilots' sleep diary reports (in blue). Significance is indicated by an asterisk (*) at p≤0.05 and double asterisk (**) at p≤0.001.

Effectiveness Distribution During COVID-19 Humanitarian Missions

The distribution of effectiveness across all crew minutes was compared between the 3 sleep scenarios (Auto-sleep, Zulu, and Diary) for each mission flight (Mission 1-5) using the SAFTE-FAST reporting tool. Results were exported to Excel 2013 and combined to examine effectiveness distribution across all missions, as shown in Figure 5A. Linear regressions examined the difference between the Auto-sleep predicted distribution of effectiveness compared to either sleep diary or Zulu watches across all missions.

Goodness of fit was evaluated using the R2 statistic. An R2 value of 0.5 means that half of the variance in the outcome variable is explained by the model. The linear regressions for effectiveness distribution across all missions are shown in Figure 5. R2 values were greater than 0.9 when compared against sleep diary (Fig. 5B) or Zulu watch (Fig. 5C). These findings indicate that the distribution of effectiveness across crewing work minutes may be equally estimated in SAFTE-FAST using either Auto-sleep, sleep diary, or Zulu watch data.

Figure 5. Auto-sleep Effectiveness Predictions During COVID-19 Humanitarian Missions Compared to Sleep Diary and Zulu Watch

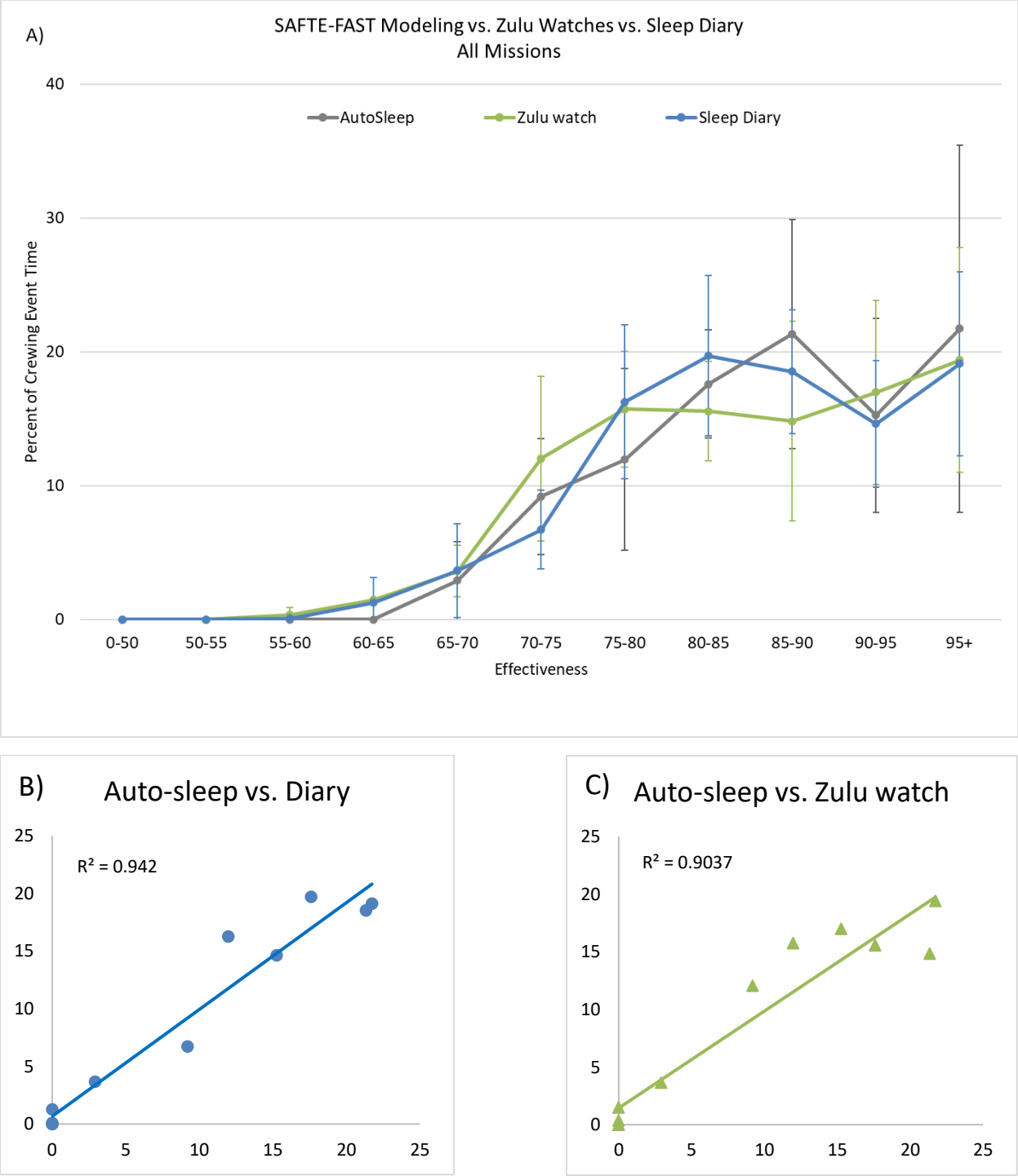


Figure 5. A) Auto-sleep predictions (in gray) of effectiveness distribution were not significantly different from distribution as determined by diary (in blue) or Zulu watch (in green) scenarios. Significance is indicated by an asterisk (*) at p≤0.05 and double asterisk (**) at p≤0.001. B &C) Linear regression and R2 value indicating the goodness of fit for effectiveness distribution between B) Auto-sleep and sleep diary and C) Auto-sleep and Zulu watch. Higher R2 values represent smaller differences between the model (in this case, Auto-sleep) and the observed data (i.e., diary or Zulu watch).

Discussion

The data presented in this white paper overwhelmingly suggest that the SAFTE-FAST Auto-sleep function can be used as a reasonable prediction of pilot sleep patterns during ultra-long flights. Sleep duration and crewing effectiveness were similar whether sleep was estimated by Auto-sleep, reported via sleep diary, or collected objectively using the Zulu watch. There were minor differences between scenarios, but none so great as to create a safety oversight. For example, pilots overreported the amount of sleep that they received during flights compared to Auto-sleep or Zulu watch (Fig 2A). Conversely, Zulu watch-measured sleep efficiency was higher than sleep quality from the diary, or Auto-sleep assumptions of sleep environment (Fig 2B). However, it should be noted that the Zulu watch does not take into account sleep onset latency, or the time it takes to first fall asleep, and treats periods of sustained wakefulness as the termination of a sleep event. As shown in Fig 1C, there were frequently multiple Zulu sleep events occurring during one sleep diary reported sleep period. Those periods of wakefulness were not considered when computing sleep efficiency. However, neither the overestimation of TIB by diary nor the overestimation of sleep efficiency by the Zulu watch resulted in downstream effects with regards to effectiveness or goodness-of-fit. Variability in actual sleep behavior is to be expected, and no model can perfectly account for this unpredictability. Notwithstanding that limitation, SAFTE-FAST did an excellent job of forecasting sleep during ultra-long humanitarian flights.

SAFTE-FAST predictions of KSS or Samn-Perelli fatigue were higher than actual reported fatigue (see Fig. 4). There are several caveats to consider when interpreting these results. Firstly, fatigue predictions in SAFTE-FAST are only intended to provide consistent range ordering of fatigue hazard when evaluating potential rosters and are not intended as a prediction of actual fatigue. Secondly, SAFTE-FAST computes KSS and Samn-Perelli estimates continuously as a transformation of predicted effectiveness while pilots reported their fatigue at a single moment for each flight leg. For these analyses, average crewing KSS and Samn-Perelli estimates were used for the comparisons, but SAFTE-FAST can provide an estimate for any point during crewing or non-crewing periods. Without knowing the exact minute during FDP that pilots completed the diary, it is impossible to select the appropriate time-matched momentary estimate from SAFTE-FAST. Thirdly, SAFTE-FAST fatigue estimates from the Auto-sleep scenarios were used to compare against diary fatigue. Auto-sleep TIB and TST were shorter than diary or Zulu watch TIB or TST (see Fig. 2). Fatigue estimates for the sleep diary scenario may have been closer to pilot-reported fatigue, but these data are not included since the purpose of these analyses was to examine forecasted fatigue against reported fatigue rather than to calibrate SAFTE-FAST estimations of KSS or Samn-Perelli.

Low pilot fatigue during the COVID-19 humanitarian missions could be attributed to Azul's well-executed planning of rest periods, such as providing each pilot with a 9-hour sleep opportunity per flight leg or having pilots continue to operate in base time (i.e., Brasilia Standard Time) throughout the mission. Subjective fatigue may also have been influenced by the pilots' motivation with respect to the humanitarian goal of the missions.

Conclusion

The COVID-19 pandemic crisis has disrupted almost every facet of modern society, but particularly, the aviation industry is now facing unprecedented changes to daily operations. The purpose of SAFTE-FAST fatigue modeling software is to provide a science-based process for the analysis of on-the-job fatigue risk and to identify fatigue factors that inform mitigation decisions. For aviation, this means the assessment of suitable pairings and planned rosters to ensure safety during flights. Azul Airlines tested the limits of their pilots' capabilities, and the capabilities of SAFTE-FAST, to conduct five unprecedented humanitarian missions during an international crisis. The ability of the SAFTE-FAST Auto-sleep function to accurately predict actual pilot rest patterns during the humanitarian missions is impressive, but not nearly as impressive as the efforts put forth by Azul's human factors team or the dedication of Azul's pilots to plan and execute these missions.

References

Koo, T. K. and M. Y. Li (2016). "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." J Chiropr Med **15**(2): 155-163.