Equivalence Testing as a Tool for Fatigue Risk Management in Aviation

Lora J. Wu; Philippa H. Gander; Margo van den Berg; T. Leigh Signal

BACKGROUND: Many civilian aviation regulators favor evidence-based strategies that go beyond hours-of-service approaches for managing fatigue risk. Several countries now allow operations to be flown outside of flight and duty hour limitations, provided airlines demonstrate an alternative method of compliance that yields safety levels "at least equivalent to" the prescriptive regulations. Here we discuss equivalence testing in occupational fatigue risk management. We present suggested ratios/margins of practical equivalence when comparing operations inside and outside of prescriptive regulations for two common aviation safety performance indicators: total in-flight sleep duration and psychomotor vigilance task reaction speed. Suggested levels of practical equivalence, based on expertise coupled with evidence from field and laboratory studies, are $\leq$ 30 min in-flight sleep and $\pm$ 15% of reference response speed.

METHODS: Equivalence testing is illustrated in analyses of a within-subjects field study during an out-and-back long-range trip. During both sectors of their trip, 41 pilots were monitored via actigraphy, sleep diary, and top of descent psychomotor vigilance task. Pilots were assigned to take rest breaks in a standard lie-flat bunk on one sector and in a bunk tapered 9" from hip to foot on the other sector.

RESULTS: Total in-flight sleep duration (134 $\pm$ 53 vs. 135 $\pm$ 55 min) and mean reaction speed at top of descent (3.94 $\pm$ 0.58 vs. 3.77 $\pm$ 0.58) were equivalent after rest in the full vs. tapered bunk.

DISCUSSION: Equivalence testing is a complimentary statistical approach to difference testing when comparing levels of fatigue and performance in occupational settings and can be applied in transportation policy decision making.

KEYWORDS: occupational sleep medicine, evidence-informed regulation, in-flight sleep and performance, pilot fatigue, flight safety.

Wu LJ, Gander PH, van den Berg M, Signal TL. *Equivalence testing as a tool for fatigue risk management in aviation*. Aerosp Med Hum Perform. 2018; 89(4):383–388.

The use of ongoing data collection and monitoring has great potential for advancing safety and promoting evidence-informed regulatory approaches in transportation. Data-driven fatigue risk management systems in commercial aviation are designed to use this potential.[10] The U.S. Federal Aviation Administration[6] and European Aviation Safety Agency[4] allow airlines to operate outside of prescriptive flight and duty time requirements if they demonstrate an alternative that provides a "level of safety that is at least equivalent to that" of the prescriptive regulations. For a fatigue risk management system to be approved as an "alternative means of compliance" in this context, an airline must estimate the level of fatigue-related risk associated with the operation(s), propose appropriate mitigations to manage that risk, and monitor fatigue and related risk on an ongoing basis.

Traditionally, difference testing is used to determine whether means or distributions of measurements differ between conditions by testing the null hypothesis that they are not different. If the probability that the conditions are not different is below a predetermined threshold (alpha level, usually < 0.05), the null hypothesis is rejected and it is inferred that the conditions are different. It is inappropriate to conclude that the conditions are the same when the null hypothesis is not rejected, although this is a common mistake in the interpretation of difference tests.[24] In this paper we discuss equivalence testing, an alternative statistical approach used to determine whether conditions are practically equivalent to one another by testing the

From the Sleep/Wake Research Centre, Massey University, Wellington, New Zealand.

This manuscript was received for review in November 2016. It was accepted for publication in December 2017.

Address correspondence to: Lora J. Wu, Ph.D., Sleep/Wake Research Centre, Massey University, P.O. Box 756, Wellington 6140, New Zealand; l.wu@massey.ac.nz.

DOI: https://doi.org/10.3357/AMHP.4790.2018

null hypothesis that they are not equivalent.[15] If the null hypothesis is rejected, equivalence is inferred; if it is not rejected, this does not imply the conditions are different. The two most common approaches for testing equivalence are: 1) the two one-sided *t*-test (TOST) procedure; and 2) the mathematically equivalent confidence interval approach.[17,24,25]

A critical component of equivalence testing is defining "practical equivalence," a tolerable difference that has no practical implication and thus is considered inconsequential. Defining the margin of equivalence requires expert knowledge of the field and the commonly used outcome measures, as well as the context of the operation. This paper examines approaches for defining practical equivalence in the field of aviation for two measures of pilot fatigue during long-haul flights and illustrates the use of equivalence testing through a case study.

Appropriate safety performance indicators[8,10,11] are used to compare alternative operation(s) with those that remain inside the prescriptive regulations. Two recommended safety performance indicators in long-haul flight operations are total in-flight sleep duration (measured with actigraphy or sleep diary) and psychomotor vigilance task (PVT) performance at critical phases of flight. While these indicators are commonly used, neither has an established margin of equivalence.

On long-haul flights operated by augmented crews (additional pilots), each crewmember has the opportunity for sleep in a bunk. The definition of practical equivalence for total in-flight sleep duration must represent a difference that is sufficiently small to yield no meaningful change in pilot functioning by the end of the flight. Available scientific evidence cannot define an absolute amount of sleep reduction that is associated with an operationally significant change in pilot performance capacity, for three main reasons. First, in most laboratory studies examining the effects of sleep restriction on waking function, participants slept at night in an ideal laboratory environment, which is not comparable to pilots' sleep during long haul flights. Second, there are stable, trait-like individual differences in sleep need and resilience to the effects of sleep loss on performance.[26] Third, the link between an individual pilot's sleep loss and changes in team performance of flight crew are complex and poorly understood.[12,14]

It is generally accepted that some sleep in-flight is preferable to none. In a study of pilots afforded a 40-min flight-deck nap or no nap, those who napped (average sleep duration 26 min) had better PVT performance toward the end of the flight.[16] It is unclear whether obtaining 26 min sleep in addition to the 200+ min of sleep typically obtained during long-haul flights[9,19] would make a difference in performance at the end of the flight.

A polysomnographic study conducted on ultra-long-range flights (mean duration 15.5 h) compared in-flight sleep duration of 14 command crew (performing takeoff and landing) and 16 relief crew on the outbound Singapore-Los Angeles sector (4-pilot crews; mean departure time 16:08 domicile time[21]). The relief crew took the first (short), third (long), and fifth (short) breaks, while the command crew took the second (long) and fourth (long) in-flight breaks. On average, the command crew had 47 min more break time available for sleep. At top of descent (TOD), the command crew obtained on average 52 min more in-flight sleep than the relief crew (201 vs. 149 min) and rated their fatigue and sleepiness lower than the relief crew. There was no statistically significant difference in PVT mean reaction time between command and relief pilots at TOD (command $= 265 \pm 56$ ms, relief $= 257 \pm 53$ ms). While 52 min additional sleep was associated with measurable changes in fatigue and sleepiness ratings, results suggest that a larger difference would be required to generate a statistically significant and operationally meaningful difference in PVT performance between the groups.

A laboratory sleep restriction study with healthy adult participants (range 24–62 yr, comparable to long-haul pilots) examined the effects of reducing time in bed from 8 h per night during 3 baseline nights (average sleep duration about 7.0 h, measured by polysomnography) to 5 h or 3 h per night for 1 wk.[2] Average sleep duration was 2.9 h for those allowed 3 h in bed and 4.7 h for those allowed 5 h in bed. The effects of sleep restriction on PVT performance averaged across four daily 10-min tests were cumulative and dose-dependent, but significant changes did not occur until the second night of sleep restriction in the 3-h group. Thus, significant decrements in PVT performance were not observed until after about 8.2 h cumulative sleep loss in the 3-h group and ~6.9 h in the 5-h group. This study considered the effects of sleep loss across multiple days with performance averaged across several tests each day. In contrast, in-flight sleep and performance at TOD occur in a relatively short timeframe, and it is possible that smaller differences in sleep loss have a greater impact in this scenario.

We propose a difference of ≤ 30 min in-flight sleep as a margin of practical equivalence by considering the available evidence. Results from the study of pilots sleeping during 15.5-h flights suggest that practical equivalence for in-flight sleep duration should likely be less than 52 min. The field and laboratory study findings taken together suggest that a reduction in total sleep time of up to 30 min in a 24-h period would not result in meaningful changes in PVT performance, sleep latency, or sleepiness ratings. A difference of ≤ 30 min is a conservative initial definition of practical equivalence. An important caveat is that flight duration determines the amount of time available for in-flight rest breaks, and is thus strongly associated with in-flight sleep duration. This definition can only be applied to flights of similar duration to those on which the arguments are based, and a change in sleep duration relative to available rest periods will be more appropriate in other operations.

The PVT is a widely used simple reaction time task[1,3] that can be administered on a portable device[23] and has no practice effect once a participant knows how to use the test device. The definition of practical equivalence in mean PVT response speed for pilots on long-haul flights should be sufficiently small, such that it can be expected to have no practical effect on flight safety. Reaction speed (mean of 1/reaction time*1000) is a relative measure of performance and, aside from using lapses (usually reaction times $>$ 500 ms), there is no laboratory benchmark

that defines PVT performance as too slow to be safe. The definition of "too slow" varies depending on the situation and changes within and between flight operations. For example, the reaction speed of one pilot operating as part of a two-pilot crew during the cruise portion of a long-haul flight would normally be less operationally significant than the same reaction speed during the critical landing phase. In addition, there is minimal information about the extent to which PVT performance of one crewmember influences the safety of a flight operated by two crewmembers, or indeed about the decision making of a two-pilot crew at all.[7]

A study of 67 fatigued and rested B747 2-pilot crews compared simulated flight performance during a 60–70 min flight with a critical decision event (whether or not to divert the flight from the scheduled destination).[13,14] Participants were categorized into low, moderate, or high fatigue groups based in part on 5-min PVT response speed prior to the simulated flight (specific details of categorization were not published). Preflight PVT performance was not a consistent predictor of changes in crew threat and error management. A greater proportion of moderate- and high-fatigue crews acquired information regarding the weather and trend forecasts for the destination airport and chose to divert compared to low-fatigue crews.[14] It is thus arguable that in this scenario, worse preflight PVT performance was associated with greater flight safety (although in the simulator, crews did not have to deal with the logistical and financial consequences of diverting a plane full of passengers). While PVT response speed may not be a useful predictor of complex performance, it remains sensitive to sleep loss.

Since the implications of differences in PVT response speed are unclear and likely vary according to the type and phase of flight operation, we considered the variability in PVT response speeds at TOD on long-range flights (compliant with flight and duty time limitations) and ultra-long-range flights (conditions approved by the regulator) flown with four-pilot crews. We argue that mean PVT response speed at TOD can be considered equivalent if the difference between two conditions is within the range of variability in response speeds observed on these flights. Analyses were based on data from previous studies which included 185 pilots from 3 airlines flying between 23 city pairs.[9,20,27] A 5-min PVT test (PalmPVT) with an interstimulus interval of 2–10 s was completed at TOD. Of the 424 performance tests (1–4 per participant), most (81%) were completed on B777 aircraft, with a minority completed on A340 aircraft. Mean PVT response speed at TOD was 3.97 ($\pm$ 0.55) and the coefficient of variation (standard deviation relative to the mean) was 14%. The distribution of TOD PVT response speeds is shown in **Fig. 1**. When TOD PVT response speed was compared between city pairs, the difference between the fastest and slowest mean speed was 0.58, or 15% of the mean of all TOD speeds.

We argue that practical equivalence can be defined as PVT speeds within 15% of the reference value (i.e., the value on flights that are compliant with regulatory requirements), assuming that the data are normally distributed and have a similar standard deviation to that observed in these data. We now illustrate the use of equivalence testing in a study where safety performance indicators are compared between a standard (regulation-compliant) and nonstandard flight operation.

A number of regulations worldwide (e.g., United States, European Union, Singapore) specify that the maximum flight duration allowed for an aircraft with crew rest facilities depends on the quality of those facilities. The Federal Aviation Administration requires that Class 1 flight crew rest facilities have a minimum bunk size of 78″ x 30″.[5] The aim of this study was to compare total in-flight sleep duration and mean PVT response speed at TOD after pilots had slept in a full-size bunk vs. a bunk that had a taper on one side (due to the shape of the aircraft), from 30″ at hip-height to about 21″ at the foot.

METHODS

Subjects

Commercial pilots were recruited on a voluntary basis, provided written informed consent before participating, and were compensated for their participation in accordance with a union agreement. The study protocol was approved in advance by the Massey University Human Ethics Committee (Southern A).

Procedure

Pilots were observed across an out-and-back trip between Seattle, WA, United States, and Shanghai, China. Sleep was monitored via actigraphy and sleep diary throughout the trip and in-flight sleep was summed across rest breaks within each flight. Participants completed a 5-min PVT near TOD during each flight (PalmPVT). The design was within-subjects and participants were assigned to use either the full (reference) or tapered (experimental) bunk during outbound rest breaks and the other type during inbound breaks. Aside from bunk assignment, pilots were not instructed to change their typical rest break behavior.

Statistical Analysis

We hypothesized that total in-flight sleep in the tapered and full bunks would be equivalent within 30 min, and that the mean PVT response speed at TOD following rest in the tapered bunk would be within 15% of mean response speed following rest in the full bunk. The paired samples *t*-test was used to test for a significant difference between measurements following rest in the tapered or full bunk. Equivalence was tested with TOST in R (version 3.3.0, R Core Team, Vienna, Austria).

RESULTS

The 41 participants included 17 Captains and 24 First Officers operating as part of 4-pilot crews. Mean age was 56.2 $\pm$ 3.8 yr and pilots reported a median of 18,500 h flight experience. The outbound flight (Seattle-Shanghai) departed Seattle at 14:46 (median; range 14:02–15:37) local time and was 12.2 h

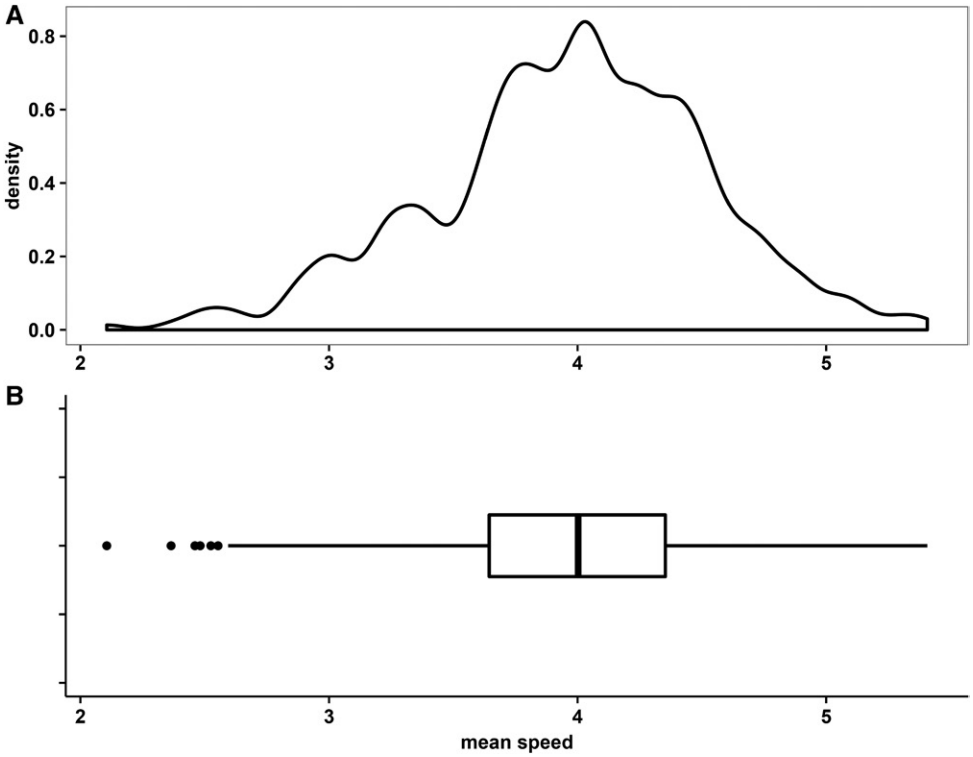

Fig. 1. A) Density estimate and B) boxplot of mean PVT response speed at top of descent in 424 long-range flights compliant with applicable flight and duty time regulations.

(median; range 11.5–12.5 h) in duration. The layover in Shanghai was 40.3 h (median; range 38.2–41.9 h; excluding four planned 64.4-h layovers). The inbound flight departed Shanghai at 21:12 (median; range 21:01–22:01) domicile (Seattle) time and was 11.3 h (median; range 10.4–14.3 h) in duration.

Assignment of bunks was balanced between outbound and inbound flights (21 pilots used the tapered bunk on the outbound and the full bunk on the inbound flight; 20 had the reverse pattern). All pilots attempted and obtained in-flight sleep in the bunks (range: full 28–241 min; tapered 22–229 min). Total in-flight sleep duration and the within-subjects difference in duration between bunks were normally distributed, and the variance between bunks was equal. The within-subjects difference in total in-flight sleep duration (mean ± SD) was not different between bunks [full = 134 ± 53 min; tapered = 135 ± 55 min; $t(40) = -0.04$, $P = 0.97$].

The proposed equivalence value of 30 min constituted 22% of the total in-flight sleep obtained in the full bunk in this flight. Total in-flight sleep duration was equivalent within ± 30 min between the tapered and full bunks (TOST $P < 0.01$). The confidence interval around the difference in total in-flight sleep during rest periods taken in the tapered vs. full bunk are shown in **Fig. 2**.

PVT data at TOD were available for 33 pilots after sleep in both bunks and were included in the within-subjects analyses. On average, TOD PVT tests occurred at 01:41 (range 00:45–03:08 domicile) on the outbound and 07:49 (range 06:25–10:15 domicile) on the inbound sectors. The comparison of performance following rest in the full or tapered bunk was counterbalanced. Mean PVT response speeds at TOD and the within-subjects difference in speed between bunk types were normally distributed, and variance was equal between bunks. Mean PVT response speed at TOD after sleep in the full bunk (reference value) was 3.94 ± 0.58; practical equivalence defined as 15% of this value is ± 0.59. Mean PVT response speed at TOD after sleep in the tapered bunk was 3.77 ± 0.58. Difference in PVT response speed was statistically different from 0 between bunks [$t(32) = -2.16$, $P = 0.04$]. Mean PVT response speeds at TOD were equivalent within 15% of the reference value after sleep in the tapered vs. full bunk (TOST $P < 0.01$). The confidence interval around the difference is shown in **Fig. 3**. There is an apparent paradox between the finding that the difference in response speed between bunks at TOD is significantly different (from 0), while also finding that response speed at TOD was equivalent. This is because the difference (0.17 units) represents only 4% of the reference (full bunk) mean response speed, which is well within the defined equivalence range of ± 15% of the reference mean PVT response speed.

Fig. 2. The 95% (outer tick) and 90% (bold inner tick) confidence interval around the mean difference in duration of in-flight sleep obtained during rest periods taken in the tapered vs. the full bunk in 41 pilots.

DISCUSSION

The definition of practical equivalence needs to be appropriate to the context and based primarily on expertise, scientific evidence, and operational knowledge. In bioequivalence studies, 20% of the reference value is often used

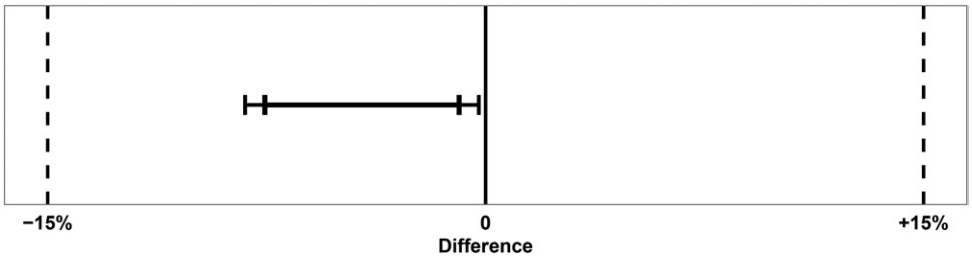

Fig. 3. The 95% (outer tick) and 90% (bold inner tick) confidence interval around the mean difference in PVT response speed at top of descent following rest in the tapered vs. the full bunk in 33 pilots.

as a benchmark of equivalence.[22] When developing equivalence thresholds in behavioral research, it is important to be aware of the operational relevance of the measure and the population to whom it is being applied. Furthermore, equivalence thresholds are likely to change depending on a number of factors and should be tailored to suit the application.[15]

The suggestion of using a difference of up to 30 min in total in-flight sleep as a definition of practical equivalence is based in part on a polysomnographic study of the in-flight sleep of pilots in four-pilot crews with access to gold-standard rest facilities (equipped with horizontal sleeping bunks).[21] The data from the laboratory study considered in this definition[2] suggests that 2 nights of sleep restriction (3–5 h in bed) is required before significant changes in PVT performance are observed. However, it is important to note the following caveats in the present operational context. First, the laboratory study involved a single sleep period during an appropriate time in the circadian cycle. In contrast, the in-flight sleep of pilots during long transmeridian flights is frequently taken during two rest breaks and at suboptimal times in the circadian cycle. Pilots' in-flight sleep is also lighter and more fragmented than their sleep on the ground,[18] and the impact of this on PVT performance is not known. In addition, layover sleep between flights is often split into multiple episodes. Second, the effects of the sleep restriction experienced by individual pilots on their functioning in a two-pilot flight deck crew remain poorly understood. Nevertheless, the findings of the laboratory study suggest that significantly more than a 30-min reduction in sleep in one 24-h period is needed to produce statistically significant changes in mean PVT response speed (if there is a sleep opportunity of 3–5 h at an appropriate time in the circadian cycle in ideal sleeping facilities).

The suggestion of using 15% of the reference value as a definition of practical equivalence for mean PVT response speed at TOD is based on current knowledge of PVT performance of long-haul pilots in four-pilot crews on routine flights operating within the prescriptive limits or with specific regulatory approval. Although the 15% recommendation applies only in this context, the approach on which it is established (examining the range of values at TOD on a large number of compliant flights) is generalizable. Performance within 15% of the reference value is likely a conservative estimate, as the data presented here were collected during flights in which pilots had access to Class 1 rest facilities, which provide the best environment for in-flight sleep. Both PVT tests occurred at an unfavorable circadian time (assuming no circadian adaptation on the layover) and after in-flight sleep. Notably, difference testing showed statistically significant differences in PVT speed at TOD between conditions, while equivalence testing showed that the speed was equivalent within a specified margin of variance. It is not expected that the difference in speed of 0.16 would be related to a meaningful difference in performance. This demonstrates the difference between statistical and practical significance, and highlights the utility of equivalence testing in applied fatigue risk management research. Future operational studies which include a greater variety of flights (different flight durations landing at different times of day) will further illuminate whether the equivalence ratios presented here are generalizable to other operations.

Fatigue remains a major human factors-related risk to safe transportation systems. We introduce the use of equivalence testing as a methodology for comparing safety performance indicators between flight operations, develop suggested margins of practical equivalence for commonly used fatigue metrics in transportation (sleep during breaks, reaction speed) and apply the margins to a study of fatigue in long-haul pilots. The methodology and findings can impact safety policy decision making and provide guidance to transportation regulators, operators, and researchers. We conclude that equivalence testing is a fundamental tool where performance-based regulatory approaches are available and operators are required to demonstrate that an alternative way to conduct operations can provide a level of safety that is "at least equivalent" to that afforded by remaining within the prescriptive regulations. However, care is needed when defining "practical equivalence" in different contexts, and multiple safety performance indicators need to be evaluated.[8,10] Monitoring in-flight sleep duration and PVT performance and the use of equivalence testing in fatigue risk management strategies are only a part of monitoring and mitigating fatigue in aviation operations, and these methods are not to be used in isolation of other important aspects of systemic fatigue risk management.

ACKNOWLEDGMENTS

Thanks to Professor Dennis G. Dyck for introducing the first author to equivalence testing as an adjunct to difference testing. We are grateful for statistical support from Dr. Alexander A. T. Smith and Dr. Jonathan Godfrey of Massey University; to Dr. Jarnail Singh for helpful comments and Singapore Airlines and The Civil Aviation Authority of Singapore for permission to report on previous analyses; and the following for permission to include data in these analyses: Captain Wynand Serfontein and South African Airways; Professor Gregory Belenky and United Airlines; and Captain Jim Mangie and the Delta Air Lines Fatigue Risk Management Team. Funding for the original studies was provided by the respective airlines. In two studies, airlines had representation on a Scientific Steering Committee which developed study methodology and the airlines assisted in data collection. Funders did not independently influence data

analysis, interpretation, or reporting. The analyses in the introduction were completed without external funding; the case study was partially based on analyses completed for a project funded by Delta Air Lines.

Authors and affiliation: Lora J. Wu, Ph.D., M.S., Philippa H. Gander, Ph.D., M.Sc. (1st Class), Margo van den Berg, B.S., and T. Leigh Signal, Ph.D., M.A. (Hons.), Sleep/Wake Research Centre, Massey University, Wellington, New Zealand.

REFERENCES

1. Basner M, Dinges DF. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. Sleep. 2011; 34(5):581–591.
2. Belenky G, Wesensten NJ, Thorne DR, Thomas ML, Sing HC, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. J Sleep Res. 2003; 12(1):1–12.
3. Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. Behav Res Methods Instrum Comput. 1985; 17(6):652–655.
4. European Aviation Safety Agency. European Aviation Safety Agency, Commission Regulation (EU) No 83/2014. Brussels, Belgium: Official Journal of the European Union; 2014:17–29.
5. Federal Aviation Administration. Advisory Circular 117-1: Flightcrew Member Rest Facilities. Washington (DC): Federal Aviation Administration; 2012.
6. Federal Aviation Administration. Flight crewmember duty and rest requirements. FAA-2009-1093. 14 CFR Parts 117, 119, and 121; Amendment Nos. 117-1, 119-16, and 121-357. Washington (DC): Federal Aviation Administration; 2012.
7. Foushee H, Lauber J, Baetge M, Acomb L. Crew Factors in Flight Operations, III: The Operational Significance to Short-Haul Air Transport Operation. NASA Technical Memorandum 88322. Moffett Field, CA: NASA Ames Research Center; 1986.
8. Gander PH, Mangie J, van den Berg MJ, Smith AAT, Mulrine HM, Signal TL. Crew Fatigue Safety Performance Indicators for Fatigue Risk Management Systems. Aviat Space Environ Med. 2014; 85(2):139–147.
9. Gander PH, Signal TL, van den Berg MJ, Mulrine HM, Jay SM, Mangie J. In-flight sleep, pilot fatigue and Psychomotor Vigilance Task performance on ultra-long range versus long range flights. J Sleep Res. 2013; 22(6): 697–706.
10. International Air Transport Association, International Civil Aviation Organization, International Federation of Airline Pilots' Associations. Fatigue Management Guide for Airline Operators, 2nd ed. Montreal (Canada): International Civil Aviation Organization; 2015.
11. International Civil Aviation Organization. Fatigue Risk Management Systems: Manual for Regulators. Doc 9966. Montreal: International Civil Aviation Organization; 2012.
12. Petrilli RM, Roach GD, Dawson D, Lamond N. The sleep, subjective fatigue, and sustained attention of commercial airline pilots during an international pattern. Chronobiol Int. 2006; 23(6):1357–1362.
13. Petrilli RM, Thomas MJW, Dawson D, Roach GD. The decision-making of commercial airline crews following an international pattern. Seventh International AAvPA Symposium; 9–12 November 2006; Manly, NSW, Australia. Victoria (Australia): Australian Aviation Psychology Association; 2006.
14. Roach GD, Petrilli RM, Dawson D, Thomas MJW. The effects of fatigue on the operational performance of flight crews in a B747-400 simulator. Seventh International AAvPA Symposium; 9–12 November 2006; Manly, NSW, Australia. Victoria (Australia): Australian Aviation Psychology Association; 2006.
15. Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimental groups. Psychol Bull. 1993; 113(3):553–565.
16. Rosekind MR, Gander PH, Miller DL, Gregory KB, Smith RM, et al. Fatigue in operational settings: examples from the aviation environment. Hum Factors. 1994; 36(2):327–338.
17. Seaman MA, Serlin RC. Equivalence confidence intervals for two-group comparisons of means. Psychol Methods. 1998; 3(4):403–411.
18. Signal TL, Gale J, Gander PH. Sleep measurement in flight crew: comparing actigraphic and subjective estimates to polysomnography. Aviat Space Environ Med. 2005; 76(11):1058–1063.
19. Signal TL, Gander PH, van den Berg MJ, Graeber RC. In-flight sleep of flight crew during a 7-hour rest break: implications for research and flight safety. Sleep. 2013; 36(1):109–115.
20. Signal TL, Mulrine HM, van den Berg MJ, Smith AA, Gander PH, Serfontein W. Mitigating and monitoring flight crew fatigue on a westward ultra-long-range flight. Aviat Space Environ Med. 2014; 85(12):1199–1208.
21. Signal TL, van den Berg MJ, Gander PH. Phase 3 ultra-long-range validation: polysomnographic sleep and psychomotor performance (final report). Wellington (New Zealand): Massey University; 2004.
22. Stegner BL, Bostrom AG, Greenfield TK. Equivalence testing for use in psychosocial and services research: an introduction with examples. Eval Program Plann. 1996; 19(3):193–198.
23. Thorne DR, Johnson DE, Redmond DP, Sing HC, Belenky G, Shapiro JM. The Walter Reed palm-held psychomotor vigilance test. Behav Res Methods. 2005; 37(1):111–118.
24. Tryon WW. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. Psychol Methods. 2001; 6(4):371–386.
25. Tryon WW, Lewis C. An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. Psychol Methods. 2008; 13(3):272–277.
26. Van Dongen HPA, Belenky G. Individual differences in vulnerability to sleep loss in the work environment. Ind Health. 2009; 47(5):518–526.
27. Wu LJ, Zaslona JL, Van Dongen H, Belenky G. Psychomotor vigilance performance at top of descent during ultra-long range flights as compared to long range flights. [Abstract.] Sleep. 2011; 34(Abstract supplement):A60.