U.S. Department of Transportation

Federal Railroad Administration

Procedures for Validation and Calibration of Human Fatigue Models: The Fatigue Audit InterDyne Tool

Office of Railroad
Policy and Development
Washington, DC 20590

DOT/FRA/ORD-10/14

Final Report
November 2010

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

NOTICE

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 2010	Technical Report

4. TITLE AND SUBTITLE
Procedures for Validation and Calibration of Human Fatigue Models: The Fatigue Audit InterDyne Tool

5. FUNDING NUMBERS
DFRA.101152

6. AUTHOR(S)
Barbara Tabak and Thomas G. Raslear*

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
QinetiQ North America, Inc., Technology Solutions Group
350 Second Avenue
Waltham, MA 02451-1196

8. PERFORMING ORGANIZATION REPORT NUMBER
DFRA.101152

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Department of Transportation
Federal Railroad Administration
Office of Railroad Policy and Development
1200 New Jersey Avenue, SE
Washington, DC 20590

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
DOT/FRA/ORD-10/14

11. SUPPLEMENTARY NOTES
*Federal Railroad Administration

12a. DISTRIBUTION/AVAILABILITY STATEMENT
This document is available to the public through the FRA Web site at http://www.fra.dot.gov.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
This report presents the results of a study that illustrates a procedure for validating and calibrating a biomathematical fatigue prediction model for evaluating work schedules. The validation has two components: (1) establishing that the model is consistent with science in the area of human performance, sleep, and fatigue, and (2) determining that the model has a statistically reliable relationship with the risk of a human factors (HF) accident and lacks a relationship with the risk of other accidents. Calibration is achieved by showing a statistically increasing relationship between cumulative risk of an HF accident and fatigue level. A railroad accident database containing work intervals for individuals involved in 732 HF accidents and 1944 nonhuman factors accidents was used to apply this process to the Fatigue Audit InterDyne (FAID) tool. Validation of FAID was achieved, but an alternative method, comparing a previously validated and calibrated model, was necessary to calibrate FAID.

14. SUBJECT TERMS
Fatigue model, FAID, fatigue prediction

15. NUMBER OF PAGES
36

16. PRICE CODE

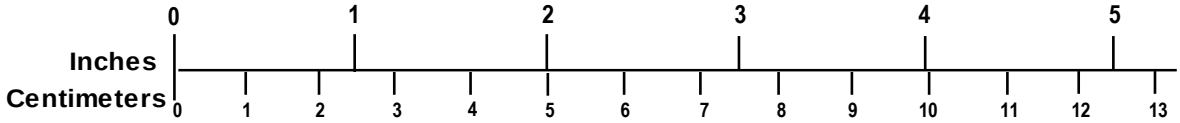
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	None

NSN 7540-01-280-5500
Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
298-102

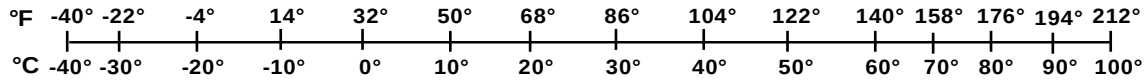
METRIC/ENGLISH CONVERSION FACTORS

ENGLISH TO METRIC	METRIC TO ENGLISH
LENGTH (APPROXIMATE)	**LENGTH (APPROXIMATE)**
1 inch (in) = 2.5 centimeters (cm)	1 millimeter (mm) = 0.04 inch (in)
1 foot (ft) = 30 centimeters (cm)	1 centimeter (cm) = 0.4 inch (in)
1 yard (yd) = 0.9 meter (m)	1 meter (m) = 3.3 feet (ft)
1 mile (mi) = 1.6 kilometers (km)	1 meter (m) = 1.1 yards (yd)
	1 kilometer (km) = 0.6 mile (mi)
AREA (APPROXIMATE)	**AREA (APPROXIMATE)**
1 square inch (sq in, in^2) = 6.5 square centimeters (cm^2)	1 square centimeter (cm^2) = 0.16 square inch (sq in, in^2)
1 square foot (sq ft, ft^2) = 0.09 square meter (m^2)	1 square meter (m^2) = 1.2 square yards (sq yd, yd^2)
1 square yard (sq yd, yd^2) = 0.8 square meter (m^2)	1 square kilometer (km^2) = 0.4 square mile (sq mi, mi^2)
1 square mile (sq mi, mi^2) = 2.6 square kilometers (km^2)	10,000 square meters (m^2) = 1 hectare (ha) = 2.5 acres
1 acre = 0.4 hectare (he) = 4,000 square meters (m^2)	
MASS - WEIGHT (APPROXIMATE)	**MASS - WEIGHT (APPROXIMATE)**
1 ounce (oz) = 28 grams (gm)	1 gram (gm) = 0.036 ounce (oz)
1 pound (lb) = 0.45 kilogram (kg)	1 kilogram (kg) = 2.2 pounds (lb)
1 short ton = 2,000 pounds (lb) = 0.9 tonne (t)	1 tonne (t) = 1,000 kilograms (kg)
	= 1.1 short tons
VOLUME (APPROXIMATE)	**VOLUME (APPROXIMATE)**
1 teaspoon (tsp) = 5 milliliters (ml)	1 milliliter (ml) = 0.03 fluid ounce (fl oz)
1 tablespoon (tbsp) = 15 milliliters (ml)	1 liter (l) = 2.1 pints (pt)
1 fluid ounce (fl oz) = 30 milliliters (ml)	1 liter (l) = 1.06 quarts (qt)
1 cup (c) = 0.24 liter (l)	1 liter (l) = 0.26 gallon (gal)
1 pint (pt) = 0.47 liter (l)	
1 quart (qt) = 0.96 liter (l)	
1 gallon (gal) = 3.8 liters (l)	
1 cubic foot (cu ft, ft^3) = 0.03 cubic meter (m^3)	1 cubic meter (m^3) = 36 cubic feet (cu ft, ft^3)
1 cubic yard (cu yd, yd^3) = 0.76 cubic meter (m^3)	1 cubic meter (m^3) = 1.3 cubic yards (cu yd, yd^3)
TEMPERATURE (EXACT)	**TEMPERATURE (EXACT)**
[(x-32)(5/9)] °F = y °C	[(9/5) y + 32] °C = x °F

QUICK INCH - CENTIMETER LENGTH CONVERSION



QUICK FAHRENHEIT - CELSIUS TEMPERATURE CONVERSION



For more exact and or other conversion factors, see NIST Miscellaneous Publication 286, Units of Weights and Measures.
Price $2.50 SD Catalog No. C13 10286

Updated 6/17/98

Contents

Illustrations

Tables

Executive Summary

The Federal Railroad Administration (FRA) now uses a validated and calibrated model of human performance and fatigue, the Fatigue Avoidance Scheduling Tool (FAST), to evaluate fatigue in accidents and work schedules. However, a large portion of the U.S. railroad industry uses other models that have not been similarly validated and calibrated. Because the Rail Safety Improvement Act of 2008 requires railroads to develop formal fatigue risk management plans, it is important that FRA understands the extent to which any fatigue model yields valid conclusions about accident risk and fatigue. This report presents a detailed methodology by which any fatigue model can be validated and calibrated. The Fatigue Audit InterDyne (FAID) served as an example for this study.

Validation

In the context of the present study, validation means determining that the output of a biomathematical model of human fatigue and performance actually measures human fatigue and performance. There are two dimensions to this validation. First, the model must be consistent with currently established science in the area of human performance, sleep, and fatigue. Second, the validation process involves determining that the model output has a statistically significant relationship with the risk of a human factors (HF) accident caused by fatigue, and the model output does not have such a relationship with nonhuman factors (NHF) accident risk.

The presence of a statistically significant relationship was evaluated using a correlation coefficient (r) with statistical significance requiring a p-value ≤ 0.05. By convention, a p-value > 0.05 indicates the absence of a statistically significant relationship. Analysis of the FAID scores produced correlation coefficients, and corresponding p-values, that met the validation requirements for both the HF ($p = 0.045$) and NHF ($p = 0.071$) risks.

Calibration

Calibration refers to the assignment of numerical values to represent aspects of empirical observations. In the case of human fatigue and performance, the calibration of a fatigue scale would start with the assignment of values to a well-rested or Not Fatigued state and to the most fatigued condition or Severely Fatigued. Given a scale for human fatigue and performance and a relationship between that scale and HF accident risk, a final calibration point would be identified: a point on the scale at which fatigue becomes unacceptable because the increase in accident risk compromises safety. This is known as the fatigue threshold. The fatigue threshold was defined statistically as that point at which the cumulative risk of a HF accident exceeds chance and the mean risk of a NHF accident with 95% confidence. Since a fatigue threshold for FAID could not be established this way, statistical regression of FAST and FAID scores was used as an alternate procedure.

When analyzed at the population level, the regression equation for FAID scores as a function of FAST scores, or FAST scores as a function of FAID scores has a correlation of 0.909.

Table 1 presents the approximate translation between FAID and FAST scores.

Table 1. Approximate Correspondence between FAID and FAST Scores

	Severely Fatigued	Extremely Fatigued	Very Fatigued	Moderately Fatigued	Fatigued	Not Fatigued
FAST	<50	<60	<70	<80	<90	<90
FAID	>80	>70	>60	>50	>40	>40

With reference to FAST, the fatigue threshold for FAID is approximately a score of 60.

Conclusion

This study successfully illustrated the application of procedures for validating and calibrating a fatigue model for use in assessing railroad worker schedules. The FAID model was validated with scores of 40 and 120, corresponding to Not Fatigued and Severely Fatigued. FAID scores showed a statistically reliable relationship with the risk of a HF accident but did not show such a relationship with other accident risks. The calibration of FAID indicated that FAID scores > 80 indicate a severe level of fatigue, and that FAID scores between 70 and 80 indicate extreme fatigue. A fatigue threshold (the fatigue level at which there is an unacceptable accident risk due to fatigue) of 60 was established for FAID. A recent Transport Safety Alert issued by the Independent Transport Safety Regulator in New South Wales, Australia, confirms that FAID scores of <80 do not necessarily indicate a lack of fatigue.

1. Introduction

1.1 Background

Over the past 15 years, numerous researchers have studied the effects of fatigue on human health, productivity, performance, and well-being. Until very recently, however, the railroad industry has had no statistically validated tools with which to measure and evaluate fatigue as a potential accident risk factor.

A recent study by Hursh, Raslear, Kaye, and Fanzone (2006, 2008) has demonstrated a method to validate and calibrate a biomathematical fatigue model, the Sleep, Activity, Fatigue and Task Effectiveness model (Hursh et al., 2004), operationalized as the FAST, to relate work schedules to the risk of HF-caused railroad accidents. FRA is now using FAST to rule out fatigue as a cause of accidents and to evaluate work schedules (Gertler & DiFiore, 2009; Gertler & Viale, 2006a, b, 2007). However, a large portion of the railroad industry uses other models, which have not been similarly validated and calibrated.

One such model, FAID, is currently in use by at least two Class I railroads to determine whether work schedules for train and engine crews are providing a safe work environment. FRA has encouraged the use of fatigue models to provide information about work schedules that railroads can use to better manage fatigue on their properties. FRA has also expressed the opinion that the several fatigue models now in existence all share a common concept of the underlying physiology of sleep and fatigue and are, therefore, likely to produce similar outcomes when analyzing a group of schedules.

However, the Rail Safety Improvement Act of 2008 (RSIA) now mandates that railroads develop and implement formal fatigue risk management plans under Section 103. Since FRA must review and approve (or disapprove) such plans, FRA must understand how a fatigue model yields valid conclusions about accident risk and fatigue.

RSIA also allows FRA to issue new regulations governing the hours of service of train employees engaged in commuter rail passenger transportation and intercity rail passenger transportation (Section 108). RSIA notes the promulgation of such regulations "...shall consider scientific and medical research related to fatigue and fatigue abatement, railroad scheduling, and operating practices...."

Biomathematical modeling of fatigue is a new scientific tool that FRA is using to support its passenger transportation hours of service rulemaking. The rules under consideration will require affected carriers to analyze the work schedules of train employees to determine whether the schedules will result in an unacceptable level of fatigue. Each railroad subject to the new rules would be required to perform an analysis of one cycle of the work schedules of its train employees engaged in commuter or intercity rail passenger transportation and identify those work schedules that, if worked by such a train employee, would put the train employee at risk for a level of fatigue at which safety may be compromised. A level of fatigue at which safety may be compromised, called "the fatigue threshold," would be determined by procedures that use a scientifically valid, biomathematical model of human performance and fatigue.

The rules under consideration do not specify a particular biomathematical model. Therefore, it is imperative that FRA understands the extent to which a fatigue model yields valid conclusions about accident risk and fatigue.

This study provides a methodology by which any fatigue model can be easily validated and calibrated. FAID served as an example for the purposes of the study.

1.2 What Is Meant by Validation?

"...the concept of validity concerns whether a measurement operation measures what it intends to measure" (Salvendy & Carayon, 1997). In the present context, validation means determining that the output of a biomathematical model of human fatigue and performance actually measures human fatigue and performance.

In general, it is expected that any model will be based on the scientific findings in that area of research. In the area of human fatigue and performance modeling, a valid model must demonstrate that it is consistent with currently established science regarding human performance, sleep, and fatigue. The scientific literature has documented that specific patterns of work and/or sleep (model inputs) have known patterns of effects on behavioral or performance-based indicators of fatigue (model outputs). Consequently, model inputs such as

- the amount of work and/or sleep over long and short time periods (chronic and acute sleep deprivation or restriction),
- the time of day that work and/or sleep occur (circadian rhythms), and
- abrupt changes in the time of day that work and/or sleep occur (phase changes),

should affect model outputs such as

- vigilance speed (e.g., time to switch between tasks),
- reaction time,
- lapses of attention,
- cognitive throughput (speed and accuracy of performing cognitive tasks),
- alertness, and
- tendency to fall asleep.

Specifically, any model should be able to demonstrate that appropriate model inputs result in acute and chronic sleep deprivation/restriction effects, circadian and phase adjustment effects, sleep recovery effects, and sleep inertia effects with regard to one or more of these model outputs. The consistency of model outputs with the pattern of time and magnitude of these effects, as documented in the scientific literature, is a basic requirement of valid measurement.

For reference, there are currently six scientific models that allow work schedules to be evaluated for the effects of fatigue on performance and alertness. They are:

- Two-Process model (Achermann, 2004)
- Sleep, Activity, Fatigue and Task Effectiveness model (Hursh et al., 2004)
- FAID model (Roach, Fletcher & Dawson, 2004)
- Three-Process model (Akerstedt, Folkard & Portin, 2004)
- System for Aircrew Fatigue Evaluation (Belyavin & Spencer, 2004)
- Circadian Alertness Simulator (Moore-Ede et al., 2004).

Each of these models has demonstrated sensitivity to circadian, sleep deprivation, sleep recovery, and sleep inertia effects on one or more well-known behavioral or performance-based indicators of fatigue, including reaction time, cognitive throughput, lapses, alertness, and tendency to fall asleep (Balkin, Braun & Wesensten, 2002; Balkin et al., 2000; Bonnet, 1997; Carskadon & Dement, 1977; Dinges, Orne & Orne, 1985; Dinges & Powell, 1985; Dinges & Powell, 1989; Folkard & Akerstedt, 1987; Froberg, 1977; Harrison & Horne, 1996; Jewett, 1997; Jewett & Kronauer, 1999; Lumley, Roehrs & Zorick, 1986; Mitler, Gujavarty, Sampson & Bowman, 1982; Monk & Embry, 1981; Richardson, Carskadon & Flagg, 1978; Thorne, Genser, Sing & Hegge, 1983; Wesensten, Balkin & Belenky, 1999). These models were recognized as adequate representations of the effects of fatigue on human performance by inclusion in a 2002 workshop on fatigue and performance modeling (Neri, 2004) and are recognized by FRA as satisfying basic validity as noted above.

FRA requires the use of railroad accident data or other railroad operational data to validate models of human fatigue and performance beyond basic validity. The use of accident or operational data is a more stringent method by which the validity of a model can be examined because assumptions concerning how fatigue affects human performance are tested in a nonlaboratory setting. FRA requires that any model of human fatigue and performance (including the six models referenced above) shall demonstrate sensitivity to rail operations HF accident risk to be qualified for use in evaluating work schedules.

In the study by Hursh et al. (2006, 2008), it was reasoned that a valid model of human fatigue and performance should show a statistically reliable relationship between model estimates of human fatigue and the risk of an accident caused by human error (i.e., an accident attributed to HF). In addition, a valid model of human fatigue and performance should NOT show a statistically reliable relationship between model estimates of human fatigue and the risk of an accident caused by mechanical, electrical, or equipment failures (i.e., an accident attributed to track, equipment, or signal failure causation). Stated differently, "A valid fatigue model should predict higher levels of fatigue (based on opportunities to sleep and an accident's time of day) when a greater likelihood of an HF accident exists. By comparison, fatigue levels should have a weaker or no relationship to the likelihood of NHF accidents" (Hursh et al., 2008).

1.3 What Is Meant by Calibration?

Webster's New Collegiate Dictionary defines calibration as "...a set of graduations to indicate values or positions...." For instance, if we wanted to construct a thermometer, we might first determine the thermometer readings (e.g., length of mercury in a capillary tube) at which water boiled and froze and assign the values of 0 and 100 to these readings. Values between 0 and 100 could then be marked off, equally spaced, to provide a continuous scale of temperature. Calibration is basically the assignment of numerical values to represent aspects of empirical observations.

In the case of human fatigue and performance, empirical observation indicates that most people can be considered well-rested or not fatigued if they consistently have 8 hours (h) of sleep at night and have not been awake for more than 16 h. Similarly, sleep deprivation induces fatigue and that fatigue increases with sleep deprivation. The calibration of a fatigue scale would logically start with the assignment of values to the least fatigued condition, or Not Fatigued, and then to the most fatigued condition that might be labeled as Severely Fatigued.

Given a scale for human fatigue and performance from which fatigue can be inferred, and a relationship between that scale and HF accident risk, a final calibration point would be the fatigue threshold or the point on the scale after which any increase in fatigue level compromises safety. The fatigue threshold would be a value somewhere between the categories Not Fatigued and Severely Fatigued. Other values that correspond to intermediate labels would then be assigned.

2. Fatigue Accident Validation Database

The accident database used for this analysis is the same one used by Hursh et al. (2006, 2008), which consists of 405 HF accidents and 1,015 NHF accidents involving engineers and conductors operating trains. Most accidents involved two crewmembers. The FRA cause code assigned by the railroad to the accident determined its categorization. Any FRA HF cause codes that were not related to the operation of a train were not included. The initial runs of the data through the FAID software identified some problems with the original database. These included:

- Duplicate work intervals
- Overlapping work periods for individual employees (start of one work interval occurred before the end of the prior one)
- Adjacent work intervals for individual employees (start of one work interval was the same as the end of the prior work interval)
- Lengthy work intervals

After correcting these problems, 51,034 work intervals were represented in the database. The distribution of the work intervals in the database is shown in Table 2. During these work intervals, a total of 1,336 accidents occurred; 366 were classified as HF accidents and 970 were classified as NHF accidents. Each accident involved two or more employees. In total, 528,782 hourly FAID scores were available for 2,673 employees. Three of the employees were involved in two accidents, both attributed to NHF for all three employees. It should be noted that Table 2 includes some extremely long work intervals. These work intervals were not eliminated from the database because they do not appear to be otherwise invalid.

Table 2. Distribution of Lengths of Work Intervals in Revised Database

No. of Hours	Count	Percent
0	0	0
>0–5	4,982	9.76
>5–10	21,935	42.98
>10–15	22,995	45.05
>15–20	1,086	2.13
>20–25	31	0.06
>25–30	1	0
>30–35	3	0.01
>35–40	0	0
>40–45	1	0
>45–50	0	0
>50	0	0

3. Validation

The validation process for a fatigue model involves determining that the model output has a statistically reliable relationship with the risk of an HF accident caused by fatigue, and that the model output does not have such a relationship with NHF accident risk.

3.1 Method

The validation process followed that of Hursh et al. (2006, 2008) and was based on the risk ratio defined as

$$\text{Risk Ratio} = \frac{\text{(Accidents at Fatigue Level)/(Total Number of Accidents)}}{\text{(Work Time at Fatigue Level)/(Total Work Time)}}$$

A risk ratio > 1 at any fatigue level indicates that a higher percentage of accidents occur at that fatigue level than the percentage of work-time spent at that level would indicate if fatigue level and risk were independent. In the case of FAID, a higher score indicates a higher level of fatigue. Therefore, for HF accidents, risk would be expected to increase with the FAID score.

In contrast, risk of NHF accidents would not be expected to relate to the FAID score. Validation depends on assessing the statistical significance of the correlation between the FAID score and risk for both HF accidents and NHF accidents.

The presence of a statistically significant relationship was evaluated using a correlation coefficient (r) with statistical significance requiring a *p*-value < 0.05. (The *p*-value represents the probability of concluding that there is a relationship between two measures when there is not.) The first step of the computation involves assigning each score to a bin. FRA plans to enact requirements that stipulate a total of six bins where the six bins are determined by a set of five evenly spaced edges or partitions, $\{e_1, e_2, e_3, e_4, e_5\}$ with $e_1 < e_2 < e_3 < e_4 < e_5$. The first bin consists of all FAID scores less than e_1, and the last bin consists of all FAID scores greater than or equal to e_5. The other four bins consist of FAID scores $\{x \mid e_i \leq x < e_{i+1}\}$ where i = 1, 2, 3, or 4. (Note that the side of the bin where the inequality includes "or equal to" is consistent with the direction of decreasing fatigue in FAID, or equivalently increasing effectiveness in FAST.) Statistical significance for six bins requires a correlation coefficient > 0.811 in absolute value.

The performance bin Not Fatigued is determined by the output of the model when sleep likely happens or can occur for ≥8 h, without abrupt interruptions, during the circadian trough between 2,200 and 1,000 h. This is similar to the level of fatigue produced by the standard 9 a.m. to 5 p.m., Monday through Friday workweek. The performance bin Severely Fatigued is determined by the output of the model in which there is total sleep deprivation for 42.5 h. This is similar to the amount of fatigue produced by a permanent night-shift schedule with six consecutive 12-hour work periods followed by 1 day (d) off. These two bins constitute the anchor bins for the validation procedure.

Roach, Fletcher, and Dawson (2004), provided the following definitions of fatigue levels for the FAID model:

> STANDARD (0–40): the upper limit of this range is similar to the maximum fatigue score produced by the standard 9 a.m. to 5 p.m., Monday through Friday workweek.
>
> MODERATE (40–80): the upper limit of this range is similar to the maximum fatigue score produced by a forward-rotating schedule (morning, afternoon, night) with five consecutive 8-hour work periods followed by 2 d off.
>
> HIGH (80–100): the upper limit of this range is similar to the maximum fatigue score produced by a forward-rotating schedule (morning, afternoon, night) with five consecutive 8-hour work periods followed by 1 d off.
>
> VERY HIGH (100–120): the upper limit of this range is similar to the maximum fatigue score produced by a schedule that rotates through two 12-hour day shifts, 2 d off, two 12-hour night shifts, and 2 d off.
>
> EXTREME (120+): fatigue scores of this magnitude are similar to those produced by a permanent night shift schedule with six consecutive 12-hour work periods followed by 1 d off.

These definitions suggested that the validation be performed with anchors of 40 and 120.

3.2 Results

As shown in Table 3, the resulting *p*-values meet the validation requirements both for the HF and NHF risks. Table 3 and Figure 1 through Figure 3 display the risk ratios and distribution of FAID scores for these anchors.

Table 3. Validation Statistics with Anchors 40 and 120

Type of Accident	Correlation Coefficient (r)	*p*-Value	Slope (m)	Intercept (b)
HF	0.82	0.045	0.0027	0.88
NHF	0.77	0.071	0.0037	0.82

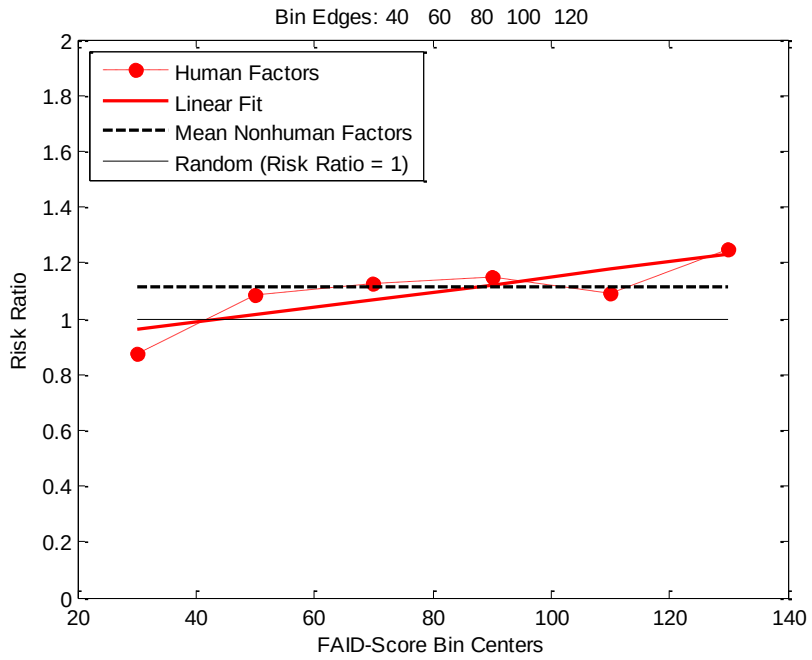


Figure 1. HF Risk Ratios for Anchors 40 and 120

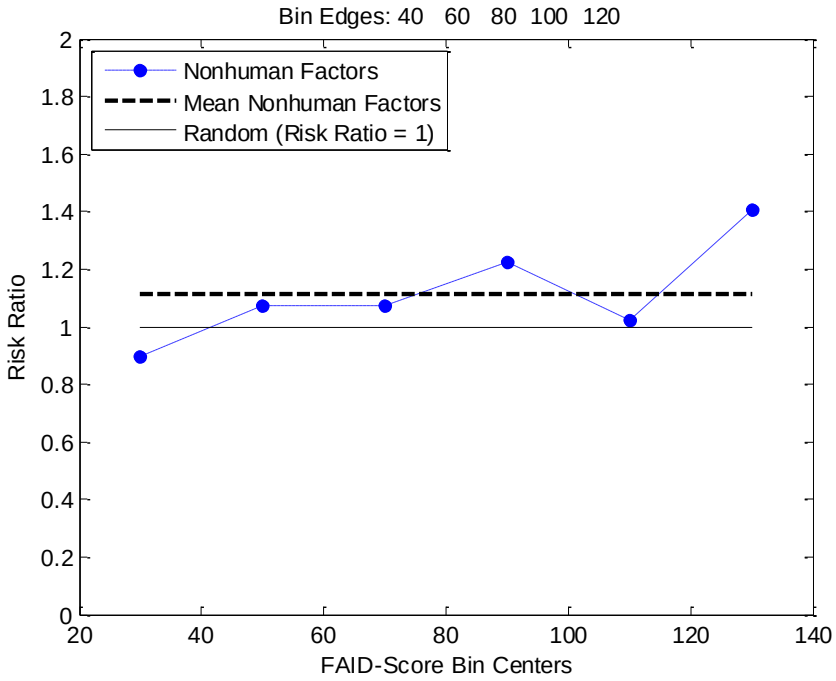


Figure 2. NHF Risk Ratios for Anchors 40 and 120

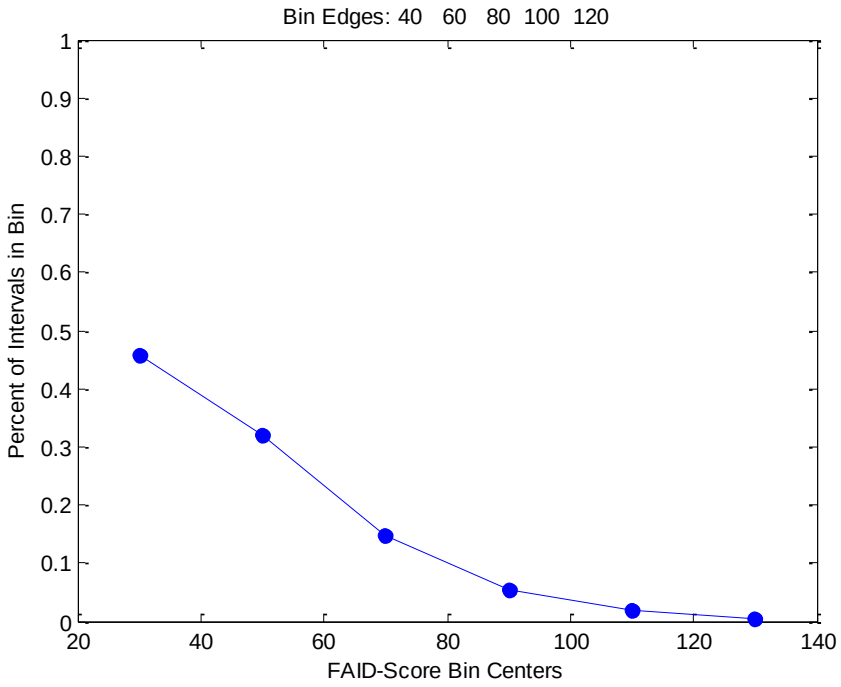


Figure 3. Distribution of FAID Scores with Anchors 40 and 120

4. Calibration

Calibration is the assignment of numerical values to represent aspects of empirical observations. Calibration starts during the validation process with the assignment of model output values to anchor bins for Not Fatigued and Severely Fatigued. The next step consists of determining the fatigue threshold, a procedure consisting of several computations. First, the cumulative risk for the six fatigue score bins is determined for HF and NHF accidents. Next, a 95 percent confidence interval is calculated for the cumulative risk in each bin. Finally, the fatigue score whereby HF cumulative risk exceeds both HF Accident Risk Ratio = 1 and the mean NHF risk is determined. This serves as the fatigue threshold for the model.

4.1 Method

Calibration followed the method used by Hursh et al. (2008). For each of the validated bin partitions of the previous section, the analysis included computation of the 95 percent confidence interval for cumulative risk, where cumulative risk is defined as:

$$\text{Cumulative Risk Ratio} = \frac{\text{(Accidents at or above Fatigue Level)/(Total Number of Accidents)}}{\text{(Work Time at or above Fatigue Level)/(Total Work Time)}}$$

The 95 percent confidence interval (CI) is

$$CI = \frac{T}{S},$$

where

$$T = P \pm 1.96 \sqrt{\frac{P(1-P)}{N}},$$

P is the proportion of HF accidents in each bin, *N* is the total number of accidents (based on formula 9.26.2; Hays, 1963, p. 291), and *S* is the cumulative proportion of employee time in each bin. If the 95 percent confidence interval for the cumulative risk at or above a given fatigue level was greater than the mean risk for the NHF accidents and neutral risk (risk = 1), then the increased risk of an accident at or above that fatigue level was considered to be statistically significant.

4.2 Results

4.2.1 Initial Calibration

As Figure 4 and Figure 5 indicate, none of the confidence intervals demonstrates a statistically significant increase in cumulative risk. This is true for both the HF accidents and the NHF accidents based on the bin anchors 40 and 120. Note that there is a significant decrease in risk < 40 for NHF accidents.

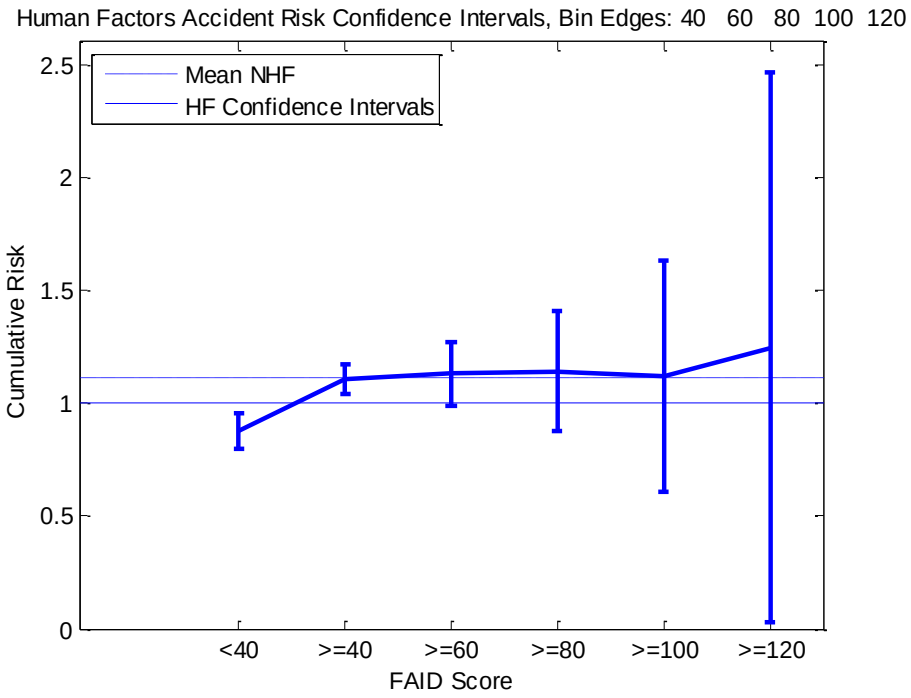


Figure 4. Confidence Intervals of Cumulative Risk for HF Accidents with Bin Anchors 40 and 120

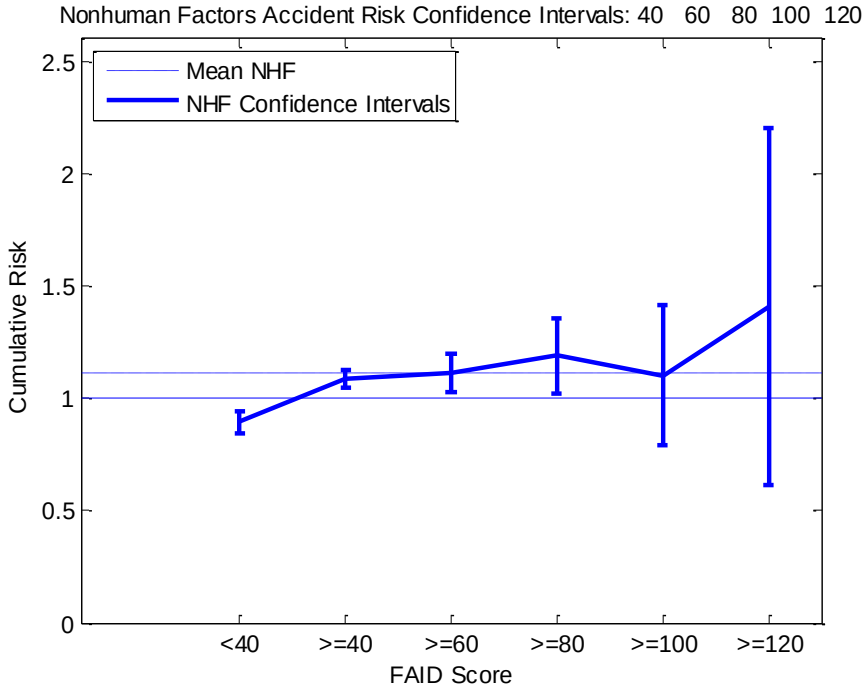


Figure 5. Confidence Intervals of Cumulative Risk for NHF Accidents with Bin Anchors 40 and 120

Because a fatigue threshold could not be determined by the preferred method, an alternative method, provided by FRA, was used in which calibration may be accomplished by demonstrating a statistically reliable correlation with a model that has been validated and calibrated as described above. The fatigue threshold for the model will be the value that corresponds with the fatigue threshold of the validated and calibrated model by use of a regression or other suitable mathematical equation.

Since the only validated and calibrated model is FAST, the output values of FAID and FAST were compared to provide a calibration of FAID.

4.2.2 Alternative Calibration

Comparison of FAID and FAST required that the scores for each case be aligned in time. For each case, every hour overlapping with a work interval was associated with a pair of scores, (FAID score, FAST score), provided that both scores were available.

4.2.3 Correlation of FAID and FAST Scores

To better understand the relationship between FAID and FAST scores, a bin-by-bin comparison was performed. The FAST bins were already established in the earlier study while the FAID bins remained to be determined. As a first step, each FAST bin was considered and the distribution of the corresponding FAID data examined. The process was then reversed and the FAST distribution corresponding to FAID bins was considered. The section below describes the selection of FAID bins.

FAID versus FAST, Bin by Bin

The established six FAST bins are defined as below or ≤50, 50 to ≤ 60, 60 to ≤70, 70 to ≤80, 80 to ≤90, and >90. Figure 6 considers the distribution of FAST and FAID scores over the FAST bin 70 to ≤80. The FAST scores appear as a near uniform block, which is not surprising as the data points were selected based on their FAST values. In some of the bins, there is more variation, typically an incline, in the height of the histogram bars. This is consistent with the overall distribution of FAST values.

The histogram of Figure 6 indicates there is a clear region where the FAID scores are most dense, suggesting that the FAST bin could be identified with the FAID mean or median. Rounding to integer scores, either the mean or the median would identify the bin with a FAID score of 50. However, the variation in the FAID score remains large and, in fact, the values range from low to high fatigue levels. Table 4 exhibits the associated statistics for each of the FAST bins. The similarity between means and medians demonstrates that outliers are not of any significance so the remaining discussion will focus on means and standard deviations.

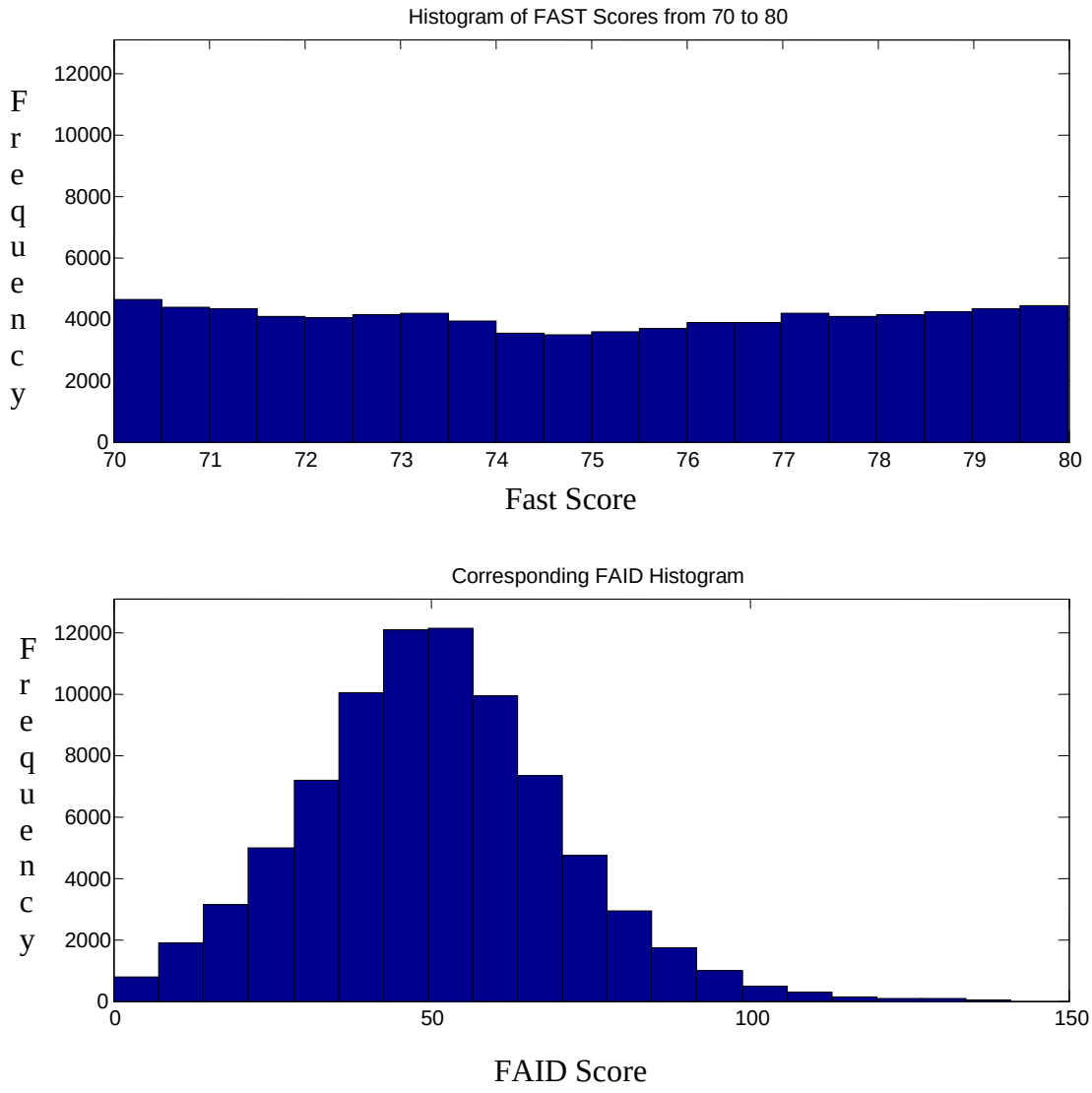


Figure 6. Distribution of FAST and FAID Scores Corresponding to FAST Bin of 70 to ≤80

Table 4. FAID Distribution Statistics by FAST Bin

FAST Bin	FAST Mean	FAID Mean	FAID Median	FAID Standard Deviation	FAID Interquartile Range (IQR)
≤50	38	91	90	21	30
50 to ≤60	56	77	76	19	26
60 to ≤70	66	62	62	20	27
70 to ≤80	75	50	50	20	25
80 to ≤90	86	43	43	17	21
>90	95	32	32	16	21

Figure 7 displays the plot of FAID means as a function of FAST means. The result is quite linear, as demonstrated by the correlation coefficient of -0.99 in Table 5. Parameters for this line are also exhibited in the same table. Although the average FAST bin scores for FAID and FAST correlate well, Figure 8 demonstrates that the variation of individual FAID scores over each bin is too large for the correlation to be of any practical value in linking fatigue levels *at an individual level*.

Table 5. Correlation and Regression Statistics for FAID Means with FAST Means over FAST Bins

	Correlation Coefficient (r)	*p*-Value	Slope (m)	Intercept (b)
FAID vs. FAST	-0.99	0.00006	-1.05	132.42

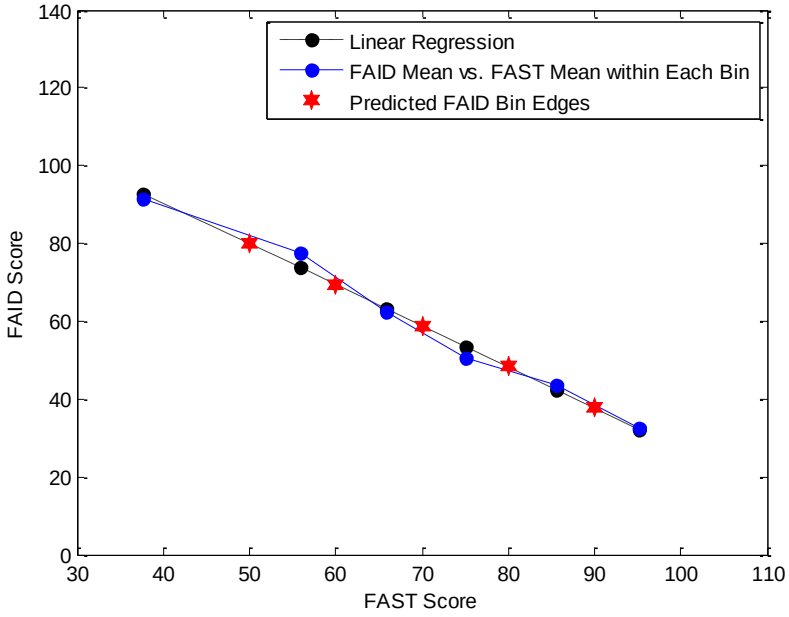


Figure 7. FAID Means vs. FAST Means and the Resulting Regression Line for FAID vs. FAST

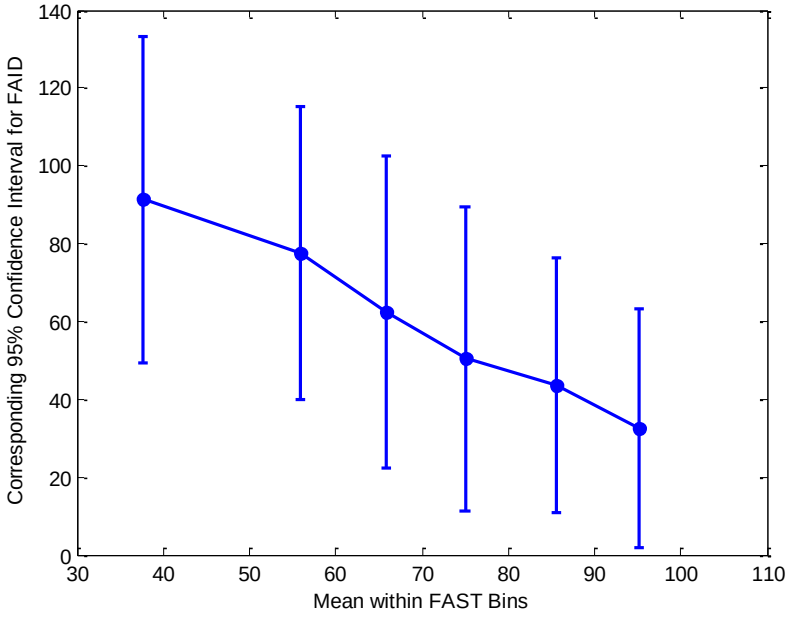


Figure 8. 95% Confidence Intervals for FAID Scores Associated to FAST Bins

FAST versus FAID, Bin by Bin

The process was then reversed with the distribution of FAST scores examined for FAID bins. The regression from FAST scores to FAID scores in the previous section suggested FAID bins as shown in Table 6. Table 6 and Figure 9 display the examination of FAST distributions over FAID bins. As in the previous section, the correlation is strong (see Table 7 and Figure 10), but now there is a high level of variation in the individual FAST scores within a FAID bin (see Figure 11). Again, linking fatigue scores on an individual level is not feasible.

Table 6. FAST Distribution Statistics by FAID Bins

FAID Bin	FAID Mean	FAST Mean	FAST Median	FAST Standard Deviation	FAST Interquartile Range (IQR)
<40	26	90	93	9	10
40 to <50	45	85	87	10	15
50 to <60	55	80	82	12	19
60 to <70	65	75	75	13	19
70 to <80	75	69	69	14	18
≥80	95	59	61	17	19

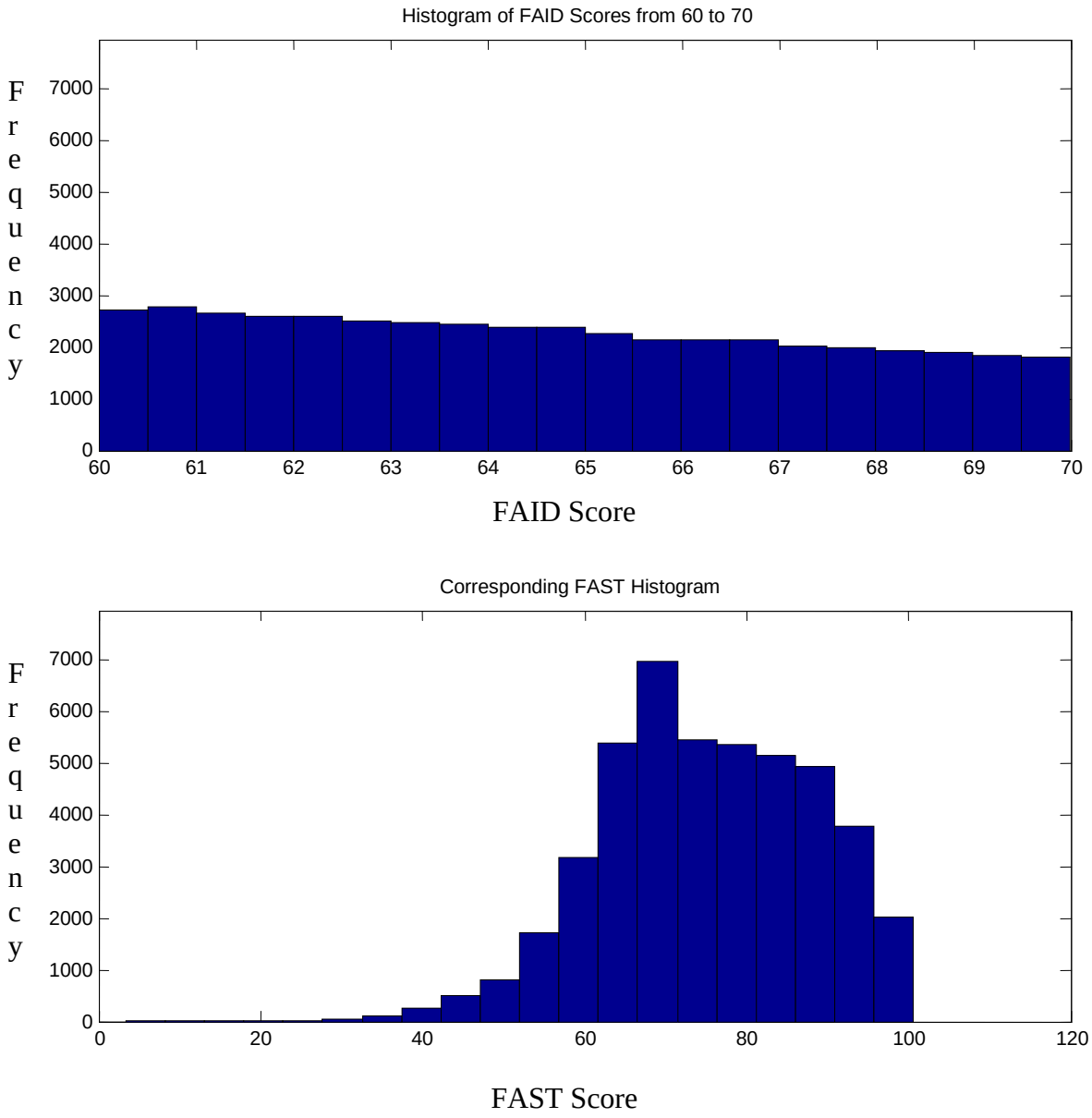


Figure 9. Distribution of FAST Scores Associated to FAID Bin of 60–70

Table 7. Correlation and Regression Statistics for FAST Means with FAID Means over FAID Bins

	Correlation Coefficient (r)	*p*-Value	Slope (m)	Intercept (b)
FAID vs. FAST	-0.99	0.0002	-0.46	104.38

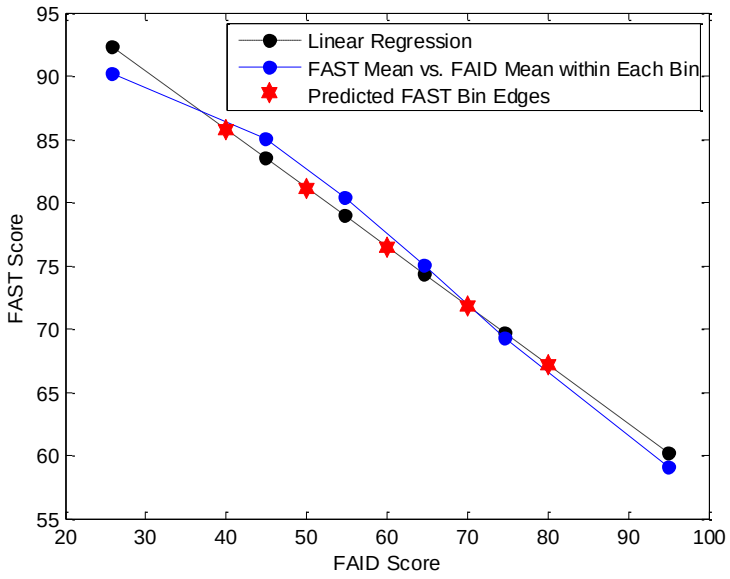


Figure 10. FAST Means vs. FAID Means and the Resulting Regression Line for FAST vs. FAID

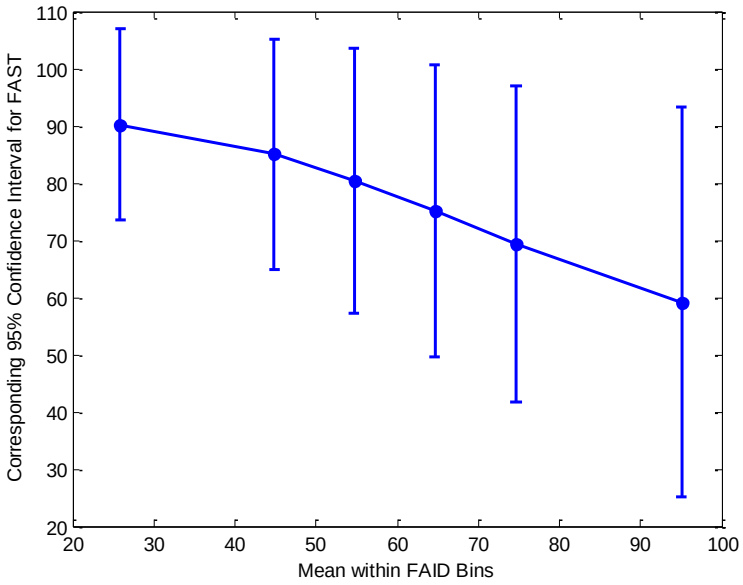


Figure 11. 95% Confidence Intervals for FAST Scores Associated to FAID Bins

Comparison of Mean Bin Scores

The analysis of "FAID versus FAST, Bin by Bin" and "FAST versus FAID, Bin by Bin" sections calculated confidence intervals for individual scores. An alternative method involves calculating confidence intervals for the population or mean score (Hays, 1963). Because biomathematical models are known to be more accurate at predicting population rather than individual behavior, the confidence intervals of the bin means were examined. From Figure 12 and Figure 13, it is apparent that accurate estimates of mean FAID and FAST scores can be made at the population level.

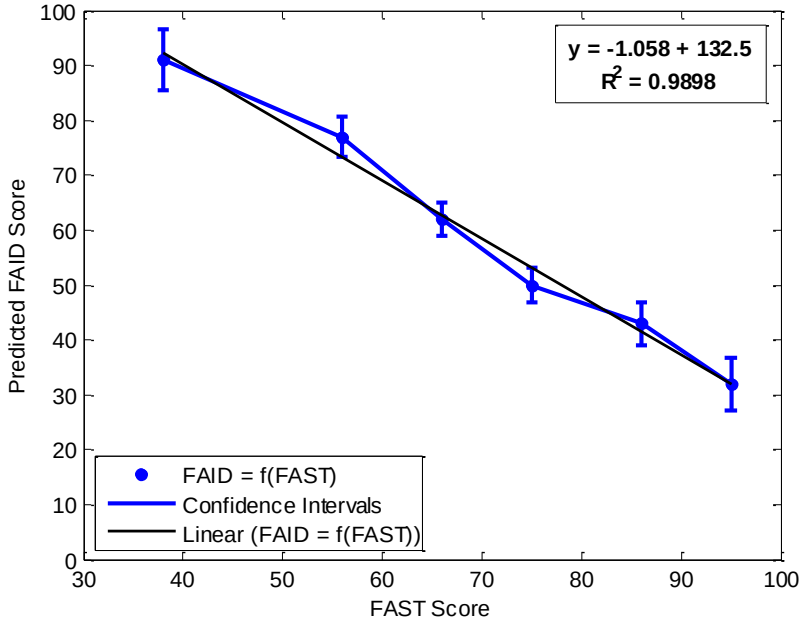


Figure 12. 95% Confidence Intervals for Mean FAID Scores Associated to FAST Bins

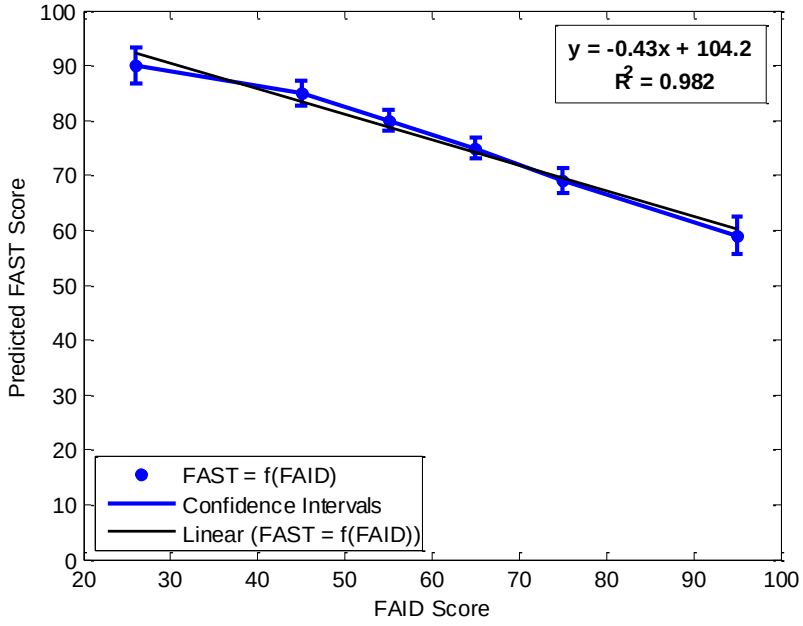


Figure 13. 95% Confidence Intervals for Mean FAST Scores Associated to FAID Bins

Because there is error in predicting FAID scores from FAST scores and vice versa, even at the population level, the regression equations for FAID = f(FAST) and FAST = f(FAID) will often provide conflicting predictions of scores. A partial solution to this problem is to derive regression equations from the combined analyses presented in Table 4 and Table 6. Figure 14 and Figure 15 show, respectively, the mean FAID and FAST scores as a function of mean FAST and FAID scores.

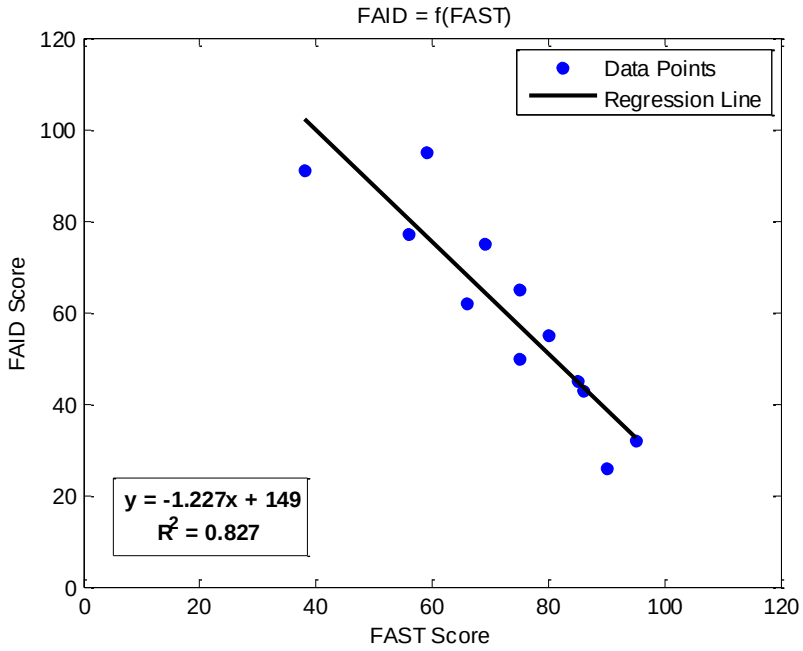


Figure 14. FAID Scores as a Function of FAST Scores

In this case, the correlation between FAST and FAID is reduced to 0.90948, but there is far less propagation of error. For example, if the regression equation in Table 5 is used to first predict FAID scores from FAST scores, and then those FAID scores are used with the regression equation in Table 7 to predict FAST scores, the root mean squared error in FAST scores is 13.01. The same process, using the regression equations in Figure 14 and Figure 15, produces a root mean squared error of 3.23.

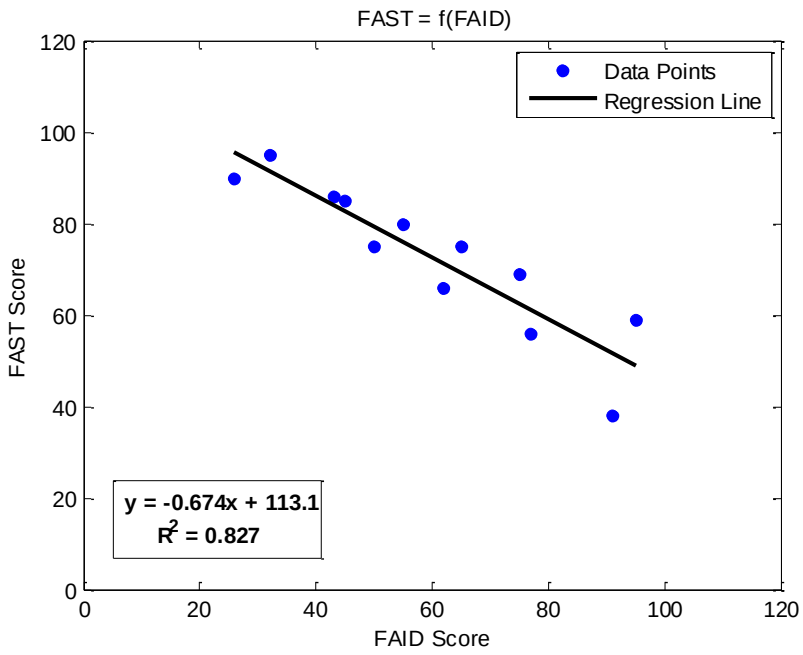


Figure 15. FAST Scores as a Function of FAID Scores

4.3 FAID Calibration

Based on the preceding discussion, Table 8 shows the approximate translation between FAID and FAST scores.

Table 8. Approximate Translation between FAST and FAID Scores

	Severely Fatigued	Extremely Fatigued	Very Fatigued	Moderately Fatigued	Fatigued	Not Fatigued
FAST	<50	<60	<70	<80	<90	>90
FAID	>80	>70	>60	>50	>40	<40

The prediction of FAID scores from FAST scores and vice versa can be obtained from the following regression equations:

FAID score = 149–1.227 (FAST score)

FAST score = 113–0.674 (FAID score).

The fatigue threshold for FAID is approximately a score of 60. The exact fatigue threshold for FAID is a score of 63.18.

5. Summary and Conclusions

This report illustrates the application of procedures for validating and calibrating a fatigue model for use in assessing railroad worker schedules. The FAID model was validated, using a previously established method, with scores of 40 and 120 corresponding to Not Fatigued and Extremely Fatigued. FAID scores showed a statistically reliable relationship with the risk of an HF accident, but did not show such a relationship with other accident risks. This pair of anchor points did not lead to a successful calibration of the model, but an alternative method that allows for calibration by reference to an already-calibrated model did allow for calibration of FAID.

The results of the calibration by reference to FAST show that FAID scores > 80 indicate a severe level of fatigue and that FAID scores between 70 and 80 are associated with extreme fatigue. A fatigue threshold, the fatigue level at which there is an unacceptable accident risk, of 60 was established for FAID. This is consistent with a recent Transport Safety Alert from New South Wales, Australia (Independent Transport Safety Regulator, 2010): "A FAID *score* of less than 80 does not mean necessarily that a person is not impaired by fatigue, or that (sic) a work schedule is appropriate from a fatigue risk management perspective (p. 2)." The Safety Alert also states that, "It has not been established if FAID *scores* can predict risk of incidents or accidents." The Safety Alert, furthermore, notes a lack of calibration for FAID scores. The present report addresses these criticisms of FAID and establishes a fatigue threshold for those who use this model.

6. References

Achermann, P. (2004). The two-process model of sleep regulation revisited. *Aviation, Space, and Environmental Medicine, 75*, A75–A83.

Akerstedt, T., Folkard, S., & Portin, C. (2004). Predictions from the three-process model of alertness. *Aviation, Space, and Environmental Medicine, 75*, A44–A53.

Balkin, T.J., Thorne, D., Sing, H., et al. (2000). *Effects of sleep schedules on commercial driver performance.* Report No. DOT-MC-00-133. Washington, DC: U.S. Department of Transportation, Federal Motor Carrier Safety Administration.

Balkin, T.J., Braun, A.R., & Wesensten, N.J. (2002). The process of awakening: a PET study of regional brain activity patterns mediating the reestablishment of alertness and consciousness. *Brain, 125*, 2308–2319.

Belyavin, A.J., & Spencer, M.B. (2004). Modeling performance and alertness: The QinetiQ approach. *Aviation, Space, and Environmental Medicine, 75*, A93–A103.

Bonnet, M.H. (1997). Sleep fragmentation as the cause of daytime sleepiness and reduced performance. *Wien Med Wochenschr, 146*, 332–334.

Carskadon, M., & Dement, W. (1977). Sleep tendency: an objective measure of sleep loss. *Sleep Research, 6*, 200.

Dinges, D., & Powell, J.W. (1989). Sleepiness impairs optimum response capability. *Sleep Research*, 18, 366.

Dinges, D.F., & Powell, J.W. (1985). Microcomputer analysis of performance on a portable, simple visual RT task during sustained operations. *Behavioral Research Methods, Instrumentation, and Computing, 17*, 652–665.

Dinges, D.F., Orne, M., & Orne, E. (1985). Sleep depth and other factors associated with performance upon abrupt awakening. *Sleep Research, 14*, 92.

Folkard, S., & Akerstedt, T. (1987). Towards a model for the prediction of alertness and/or fatigue on different sleep/wake schedules. In A. Oginiski, J. Pokorski, & J. Rutenfranz (Eds.), *Contemporary advances in shiftwork research* (pp. 231–240). Krakow: Medical Academy.

Froberg, J. (1977). Twenty-four-hour patterns in human performance, subjective and physiological variables and differences between morning and evening active subjects. *Biological Psychology, 5*, 119–134.

Gertler, J., & DiFiore, A. (2009). *Work schedules and sleep patterns of railroad train and engine service workers* (Report Number DOT/FRA/ORD-09/22). Washington, DC: U.S. Department of Transportation.

Gertler, J., & Viale, A. (2006a). *Work schedules and sleep patterns of railroad signalmen* (Report Number DOT/FRA/ORD-06/19). Washington, DC: U.S. Department of Transportation.

Gertler, J., & Viale, A. (2006b). *Work schedules and sleep patterns of railroad maintenance of way workers* (Report Number DOT/FRA/ORD-06/25). Washington, DC: U.S. Department of Transportation.

Gertler, J., & Viale, A. (2007). *Work schedules and sleep patterns of railroad dispatchers* (Report Number DOT/FRA/ORD-07/11). Washington, DC: U.S. Department of Transportation.

Harrison, Y., & Horne, J.A. (1996). Long-term extension to sleep—are we really chronically sleep deprived? *Psychophysiology, 33*, 22–30.

Hays, W.L. (1963). *Statistics for Psychologists*. New York: Holt, Rinehart and Winston.

Hursh, S.R., Raslear, T.G., Kaye, A.S., & Fanzone, J.F. (2006). *Validation and calibration of a fatigue assessment tool for railroad work schedules, summary report* (Report No. DOT/FRA/ORD-06/21). Washington, DC: U.S. Department of Transportation. (http://www.fra.dot.gov/downloads/Research/ord0621.pdf)

Hursh, S.R., Raslear, T.G., Kaye, A.S., & Fanzone, J.F. (2008). *Validation and calibration of a fatigue assessment tool for railroad work schedules, final report* (Report No. DOT/FRA/ORD-08/04). Washington, DC: U.S. Department of Transportation. (http://www.fra.dot.gov/downloads/Research/ord0804.pdf)

Hursh, S.R., Redmond, D.P., Johnson, M.L. Thorne, D.R., Belenky, G., Balkin, T.J., Storm, W.F., Miller, J.C., & Eddy, D.R. (2004). Fatigue models for applied research in war fighting. *Aviation, Space, and Environmental Medicine, 75*, A44–A53.

Independent Transport Safety Regulator. (2010). *Use of bio-mathematical models in managing risks of human fatigue in the workplace* (TSA no. 34). New South Wales, Australia: Author.

Jewett, M. (1997). *Models of circadian and homeostatic regulation of human performance and alertness.* [Dissertation]. Cambridge, MA: Harvard University.

Jewett, M., & Kronauer, R. (1999). Interactive mathematical models of subjective alertness and cognitive throughput in humans. *Journal of Biological Rhythms, 4*, 588–597.

Lumley, M., Roehrs, T., & Zorick, F., et al. (1986). The alerting effects of naps in sleep-deprived subjects. *Psychophysiology, 23*, 403–408.

Mitler, M., Gujavarty, S., Sampson, G., & Bowman, C. (1983). Multiple daytime nap approaches to evaluating the sleepy patient. *Sleep, 5*, 119-127.

Monk, T., & Embry, D. (1981). A field study of circadian rhythms in actual and interpolated task performance. In: A. Reinberg, N. Vieux, & P. Andlauer (Eds.), *Night and shift work: biological and social aspects* (pp. 473–480). Oxford: Pergamon Press.

Moore-Ede, M., Heitmann, A., Guttkuhn, R., Trutschel, U., Aguirre, A., & Croke, D. (2004). Circadian alertness simulator for fatigue risk assessment in transportation: Application to reduce frequency and severity of truck accidents. *Aviation, Space, and Environmental Medicine, 75,* A107–A118.

Neri, D.F. (2004). Fatigue and performance modeling workshop, June 13–14, 2002. *Aviation, Space, and Environmental Medicine, 75,* A1–A3.

Richardson, D., Carskadon, M., & Flagg, W. (1978). Excessive daytime sleepiness in man: multiple sleep latency measurement in narcoleptic and control subjects. *Electroencephalography and Clinical Neurophysiology, 45,* 621–627.

Roach, G.D., Fletcher, A., & Dawson, D. A model to predict work-related fatigue based on hours of work. *Aviation, Space, and Environmental Medicine, 75,* A61–A69.

Salvendy, G., & Carayon, P. (1997). Data collection and evaluation of outcome measures. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics*. New York: Wiley. pp. 1451–1470.

Thorne, D.R., Genser, S., Sing, H., & Hegge, F. (1983). Plumbing human performance limits during 72 hours of high task load. In: *Proceedings of the 24th DRG seminar on the human as a limiting element in military systems* (pp. 17–40). Toronto: Defense and Civil Institute of Environmental Medicine

Wesensten, N.J., Balkin, T.J., & Belenky, G. (1999). Does sleep fragmentation impact recuperation? A review and reanalysis. *Journal of Sleep Research, 8,* 237–246.

Abbreviations and Acronyms

d	day(s)
FAID	Fatigue Audit InterDyne
FAST	Fatigue Avoidance Scheduling Tool
FRA	Federal Railroad Administration
h	hour(s)
HF	human factor(s)
NHF	nonhuman factor(s)
RSIA	Rail Safety Improvement Act