U.S. Department of Transportation

Federal Railroad Administration

Validation and Calibration of a Fatigue Assessment Tool for Railroad Work Schedules, Final Report

Office of Research and Development
Office of Safety
Washington, DC 20590

DOT/FRA/ORD-08/04

Final Report
November 2008

This document is available to the U.S. public through the National Technical Information Service, Springfield, VA 22161.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 11, 2007	Final Report 7/2005 to 9/2006

4. TITLE AND SUBTITLE
Validation and Calibration of a Fatigue Assessment Tool for Railroad Work Schedules, Final Report

5. FUNDING NUMBERS

6. AUTHOR(S) Steven R. Hursh, Ph.D.,[1] Thomas G. Raslear, Ph.D.,[2] A. Scott Kaye,[3] and Joseph F. Fanzone, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
[1]Institutes for Behavior Resources [2]Office of Research & Development and [3]Office of Safety
2104 Maryland Avenue Federal Railroad Administration
Baltimore, MD 21218 Washington, DC 20590

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Department of Transportation
Federal Railroad Administration
Office of Research and Development
Washington, DC 20590

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
DOT/FRA/ORD-08/04

11. SUPPLEMENTARY NOTES
COTR: Thomas Raslear, Federal Railroad Administration

12a. DISTRIBUTION/AVAILABILITY STATEMENT
This document is available to the public through the National Technical Information Service, Springfield, VA 22161 and at www.fra.dot.gov.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
This report summarizes the results of a project to demonstrate a method to validate and calibrate a fatigue model. The project examined 30-day work histories of locomotive crews prior to 400 human factors accidents and 1000 nonhuman factors accidents. A biomathematical fatigue model estimated crew effectiveness (the inverse of fatigue) based entirely on work schedule information and opportunities to obtain sleep. A reliable linear relationship existed between crew effectiveness and the risk of human factors accidents ($r = -0.93$), but not for nonhuman factors accidents. A reliable time of day variation occurred in human factors accidents ($r = 0.71$). The risk of human factors accidents was elevated at any effectiveness scores below 90 and increased progressively with reduced effectiveness. At an effectiveness score ≤ 50, human factors accidents were 65 percent more likely than chance. Human factors accident risk increases reliably when effectiveness goes below 70, a value that is the rough equivalent of a 0.08 blood alcohol level or 21 hour wakefulness following 8-hour sleep. Below an effectiveness score of 70, accident cause codes indicated the kinds of operator errors consistent with fatigue, confirming that the relationship between accident risk and effectiveness was meaningful. Consistent night workers were estimated to spend a majority of time working between effectiveness 60 to 80; a significantly larger proportion of accidents had human factors causes (37%) compared to day workers (27%).

14. SUBJECT TERMS
Fatigue, biomathematical fatigue model, work schedules, accidents, train operators, performance, effectiveness, risk, safety, freight rail operations, sleep, circadian rhythm, model validation.

15. NUMBER OF PAGES
65

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	

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

Inches	0			1		2			3		4			5
Centimeters	0	1	2	3	4	5	6	7	8	9	10	11	12	13

QUICK FAHRENHEIT - CELSIUS TEMPERATURE CONVERSION

°F	-40°	-22°	-4°	14°	32°	50°	68°	86°	104°	122°	140°	158°	176°	194°	212°
°C	-40°	-30°	-20°	-10°	0°	10°	20°	30°	40°	50°	60°	70°	80°	90°	100°

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

Acknowledgements

The authors of this report wish to thank the following individuals for their invaluable contributions to the analysis, results, and documentation of the procedures:

- The Federal Railroad Administration for providing funding for this important study.
- Dr. Timothy Elsmore for the development of the batch processing software that made it possible to analyze the enormous amount of work schedule data used in this study.
- Jeff Moller and Peter French of the Association of American Railroads for encouraging this study, interfacing with industry, and assisting in the identification of a suitable sample of reportable accidents that formed the basis for the study.
- The managers and points of contacts of the five participating railroads for cooperation in conducting the study and contribution of time and effort to assemble the data needed to conduct the study.
- The Brotherhood of Locomotive Engineers and Trainmen and the United Transportation Union labor organizations for their support of the study objectives.
- Foster-Miller, Inc., the prime contractor, and Judith Gertler, the project manager, for the opportunity and generous support to conduct this study.

Preface

This the final report on a major study sponsored by the Federal Railroad Administration (FRA) to demonstrate a method to validate and calibrate fatigue models for use in predicting and managing fatigue in railroad workers. This report provides the details of methods and findings which a previous summary report (Hursh, Raslear, Kaye and Fanzone, 2006) described in a more abbreviated form. Furthermore, the researchers will make available the database used for the study for future research and analysis, with the contents coded to protect the privacy of the participants.

Executive Summary

Biomathematical fatigue models allow the objective assessment of fatigue, so that employees and employers can schedule work and rest to minimize the degradation of operator performance by fatigue. To be useful, a fatigue model must be validated. Validation means that the model must be a predictor of fatigue-related performance errors. Moreover, a model should be calibrated. Calibration means that the predictions from a model can be related to the level of risk of failures of human performance. One method of validating and calibrating a biomathematical fatigue model is to demonstrate that the model can predict an increased likelihood of human factors accidents relative to nonhuman factors accidents under conditions of fatigue. 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 a human factors accident exists. By comparison, fatigue levels should have a weaker or no relationship to the likelihood of nonhuman factors accidents. The Federal Railroad Administration (FRA) Office of Research and Development and the Office of Safety have partnered with the railroad industry to demonstrate a method to validate and calibrate fatigue models. This study collected 30-day work histories of locomotive crews prior to 400 human factors and 1,000 nonhuman factors accidents to demonstrate this validation method. More than 1 million 30-minute work intervals before the accidents, covering over 57,000 work starts, were evaluated for effectiveness (the inverse of fatigue) predicted by the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model using the Fatigue Avoidance Scheduling Tool (FAST). This served as the baseline of exposure to various levels of effectiveness. In addition, the effectiveness at the time of each accident was calculated.

The analysis used two criteria to evaluate if the SAFTE biomathematical fatigue model was a valid predictor of fatigue-related accidents:

- The proportion of human factors accidents that occur at low levels of effectiveness should be greater than the proportion of time working at those levels of effectiveness (exposure level). If the proportion of human factors accidents is reliably greater than the exposure level, and a statistically reliable relationship exists between decreases in effectiveness and an increase in human factors accident risk, then low levels of effectiveness (increased fatigue) predict increased human factors accident risk.
- By comparison, a statistically reliable relationship should not exist between effectiveness and nonhuman factors accident risk and at low levels of effectiveness (increased fatigue) the risk of human factors accidents should be greater than the risk of nonhuman factors accidents. If the human factors accident risk is reliably greater than the nonhuman factors accident risk, then low effectiveness is associated with the kinds of accidents that would be expected to be related to fatigue.

The statistical reliability of relationships was based on finding significant correlation coefficients (r). The chi square (χ^2) statistic was used to compare the distributions of human factors and nonhuman factors accidents to demonstrate significantly different distributions of risk. In each case, we rejected the null hypothesis (the hypothesis that there is no correlation, i.e., $r = 0$, or that no difference occurs in the distributions of human factors and nonhuman factors accidents) when the chance probability of a finding was less than 5 percent ($p < 0.05$).

The results of the study indicated that the biomathematical fatigue model met both criteria for validation and the results were statistically reliable:

- The biomathematical fatigue model estimates of effectiveness were significantly correlated with human factors accident risk ($r = -0.93$, $p < 0.01$, Figure 7).
- The biomathematical fatigue model distinguished between human factors and nonhuman factors accidents. Nonhuman factors accident risk was not correlated with estimated effectiveness ($r = -0.14$, $p > 0.05$, Figure 8), and at low estimated effectiveness the relative risk of human factors and nonhuman factors accidents were significantly different ($\chi^2 = 7.201$, $p < 0.01$).

This study found that an elevated risk of human factors accidents occurs at any effectiveness score below 90, and accident relative risk increased as effectiveness decreased. A reliable increase in human factors accident risk occurred when effectiveness scores were below 70 but nonhuman factors accident risk was not consistently elevated. Although an effectiveness level of 70 represents a point at which a detectible elevation of human factors accident risk occurs, various operational activities may occur below this level and still be conducted in a safe manner. Based on other research, effectiveness scores below 70 are the rough equivalent of a 0.08 blood alcohol level or being awake for 21 hours (hr) following an 8-hour sleep period the previous night.

An analysis of the cause codes associated with accidents that occurred at or below an effectiveness score of 70 showed an overrepresentation of the type of human factors accidents that might be expected of a fatigued crew (e.g., failure to comply with a stop signal). This confirms that the relationship between reduced effectiveness and elevated human factors accident risk is meaningful and not a mere correlation.

A small fraction of the accident cases involved night-shift workers. These cases were treated separately because these workers would be expected to be subject to more fatigue than the remainder of the sample. Consistent with this assumption, human factors accidents accounted for 36 percent of the accidents in this sample as compared to only 27 percent of the incidents for the variable schedule or daytime workers. This difference in distribution of human factors and nonhuman factors accidents is statistically significant (Chi square $p = 0.010$; Fisher's exact test $p = 0.003$, $p < 0.01$). Also, a nonsignificant trend between reduced effectiveness and elevated accident risk was smaller than the relationship for nonhuman factors accidents.

The study also demonstrated that human factors accidents follow a circadian pattern that is significantly correlated with the circadian rhythm of a fatigue model ($r = 0.71$, $p < 0.05$). The same model rhythm is not correlated with nonhuman factors accidents. The maximum human factors accident risk due to time of day alone was increased by less than 20 percent, while the maximum accident risk due to reduced effectiveness (fatigue) was increased by 65 percent, reflecting the combined effects of time of day and sleep opportunities.

This study was designed to demonstrate a method to test the validity of fatigue models. The data used accidents and the 30-day work histories that preceded them were not a random sample of all workers in freight rail operation. Hence, the levels of effectiveness calculated by the SAFTE model should not be interpreted as representative of the freight railroad work force. The study was not designed to determine the extent of fatigue in the freight rail industry. Furthermore, given the well-known variations in individual sleep requirements and absence of specific information on individual sleep habits, health, and circumstances, it was not the intent of this study to validate fatigue models based entirely on work schedule data as tools for determining the fatigue of particular individuals.

This study provides the first evidence that a biomathematical fatigue model can relate work schedules to an elevated risk of railroad accidents. This provides a strong scientific basis for evaluating work schedules with valid mathematical models to reduce worker fatigue. A mathematical model for detecting elevated fatigue risk could be part of a nonprescriptive, performance-based fatigue management plan that would supplement current regulations. Although fatigue models do not identify all sources of fatigue and will require a cooperative partnership among management, labor, and government regulators, they are an important tool in the identification of one of the causes of fatigue in the railroad industry.

1. Introduction

The Federal Railroad Administration (FRA) has sponsored a research program to demonstrate a method to validate and calibrate fatigue models for use in predicting and managing fatigue in railroad workers. Fatigue models allow the objective assessment and forecasting of fatigue so that employees and employers can schedule work and rest to minimize degradation of operator performance by fatigue. To be useful, a fatigue model must be validated. Validation means that a fatigue model predicts changes in job performance and/or job-related errors, such as incidents and accidents, caused by fatigue. A useful fatigue model must be calibrated. Calibration means that the predictions from the model can be related to the risks of meaningful failures of human performance. This report describes the results of the research program to demonstrate a method to validate and calibrate fatigue models.

1.1 Background

Human factors accidents have increased as a proportion of FRA reportable accidents over the past 5 years. No question exists that fatigue may be a factor in many human factors accidents. Without a detailed history of work and rest before an accident, however, it is difficult to determine the role of fatigue in that accident. The analysis of work/rest histories to rule out fatigue can be accomplished with many software models associated with fatigue, but until now, none of these models has been validated in commercial transportation operations. In the interest of developing validated fatigue models, the Office of Research and Development and the Office of Safety sponsored a study of work histories of locomotive crews associated with accidents to provide the necessary data to validate fatigue models. This report describes the details of the data collection, the methods used to analyze the data for signs of fatigue, and the results of the analysis.

1.2 Objectives

One method of validating and calibrating a fatigue model is to demonstrate that the model can predict an increased likelihood of human factors accidents relative to nonhuman factors accidents under conditions of fatigue. A valid fatigue model should predict higher levels of fatigue (based on opportunities to sleep and the time of day of an accident) when a greater likelihood of a human factors accident exists. By comparison, fatigue levels should have a weaker or no relationship to the likelihood of nonhuman factors accidents. The present study determined cognitive effectiveness (a predictor of speed of reactions and vigilance in laboratory tests that is inversely related to fatigue) from 30-day work histories of locomotive crews prior to 400 human factors and 1,000 nonhuman factors accidents. The objective was to determine if a statistically reliable relationship exists between reductions in effectiveness and the risk of human factors accidents. Further, the study sought to determine if the relationship of effectiveness to human factors accidents was larger and more consistent than that for nonhuman factors accidents, which would be expected to be much less sensitive to the effects of fatigue. The second objective was to determine the level of effectiveness at which an elevated likelihood of human factor accidents occurs relative to chance. That result served to calibrate the fatigue model for aggregate work schedule analyses of fatigue. The goal was to determine the nature of the relationship between effectiveness scores and statistically elevated accident risk. Given the well-known variations in individual sleep requirements and absence of specific information on individual sleep habits, health, and circumstances, it was not the intent of this study to validate fatigue models based entirely on work schedule data as tools for determining the fatigue of particular individuals.

The biomathematical fatigue model used for these studies, the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model, was originally developed for the U.S. Army and U.S. Air Force to predict potential fatigue caused by work schedules (Hursh, Redmond, Johnson, Thorne, Belenky, Balkin, Storm, Miller, and Eddy, 2004). The model has been incorporated into a fatigue assessment tool (Hursh, Balkin, Miller, and Eddy, 2004) called the Fatigue Avoidance Scheduling Tool (FAST). FAST uses work schedule information to estimate the amount of sleep and cognitive effectiveness. Cognitive effectiveness is a metric that tracks speed of performance on a simple reaction time test and is strongly related to overall response speed, vigilance, and the probability of lapses (Hursh, et al., 2004; Van Dongen, 2004). Cognitive effectiveness can be interpreted as the inverse of fatigue.

1.3 Overall Approach

FRA and the Association of American Railroads (AAR) determined which human factors accident cause codes would be appropriate to use. This was important because work histories for locomotive crews would be analyzed to determine effectiveness, and therefore, only cause codes that related to locomotive crew errors would be relevant. FRA and AAR also determined which nonhuman factors accident cause codes would be appropriate to use as a control group. Finally, the kinds of track and equipment most likely to be associated with irregularly scheduled line-haul (between cities or terminals) freight service was determined. For each accident, the researchers requested from that railroad the 30-day work history of the train crew just before the accident. The work histories were used to estimate the effectiveness of the crew at the time of the event and to determine the overall percentage of time typically spent at work at various levels of effectiveness. This information was used to determine the ability of the fatigue model to predict increases in human factors accident risk relative to chance. It was also used to determine if the relationship between effectiveness and accident risk was stronger for human factors than nonhuman factors accidents. Little or no relationship should exist between effectiveness and nonhuman factors accident risk. The reduction in effectiveness associated with statistically reliable cumulative increases in accident risk was determined in order to calibrate the effectiveness scores. Additionally, human factors accidents that were overrepresented at reduced levels of effectiveness were examined to determine if these accidents were consistent with the expected effects of crew fatigue (e.g., lapses increase with fatigue and are consistent with accidents in which signals are passed at danger). Descriptive information regarding work patterns and accident patterns that were independent of the model predictions were also developed.

1.4 Scope

The focus of this analysis was limited to FRA reportable incidents involving freight trains on mainline, siding, and industry track. The researchers considered only a subset of human factors incidents that would have involved an error by the operating crew and compared them to nonhuman factors incidents involving equipment or track related cause codes. The intent was to focus on line haul operations with crews (engineers and conductors) working irregular schedules. Nevertheless, because some engineers and conductors are qualified to work both yard and line-haul jobs, there was no way to positively exclude individuals who may have worked a yard assignment for a portion of the 30-day work history prior to the reportable incident. Those individuals who worked a majority of the time at a regular night job were, however, segregated for a separate analysis. As will be seen, those workers would be expected to have generally lower levels of effectiveness while at work because of the circadian loss of alertness at night between midnight and 0600 hrs.

2. Method of Analysis

2.1 General Method

The general method of the study was to use a biomathematical fatigue model (SAFTE) to estimate the cognitive effectiveness of locomotive crews at the time of accidents to determine if low levels of effectiveness are associated with higher than expected rates of accidents. The model computes cognitive effectiveness (the inverse of fatigue) from opportunities to sleep and the time of day during the 30 days before the accident and at the time of the accident. To control for extraneous factors, the analysis compared human factors and nonhuman factors accidents using the same methods. The following sections describe the validation method, the sample determination, the data requested from the railroads that describe the accidents, and the 30-day work histories of the associated train crews.

2.2 Validation Method

Thirty-day work histories were collected for train crews who were involved in FRA reportable accidents. FAST processed these histories with the SAFTE model to determine the crew's cognitive effectiveness at the time of the accidents. The histories were also used to estimate the proportion of time spent working at various levels of effectiveness during the preceding 30 days. This is the exposure of the locomotive crews to fatigue. If fatigue is not associated with accidents, then the proportion of accidents at each effectiveness level should be similar to fatigue exposure. Fatigue exposure provides an expected proportion of accidents at each effectiveness level. To determine if the obtained proportion of accidents matches the expected proportion, the obtained proportion is divided by the expected proportion. If accidents are independent of fatigue, the ratio should be close to one. This ratio is the accident risk. Statistically reliable deviations in the ratio indicate increased or decreased risk, depending on whether the ratio is greater or less than one. This analysis yields a function that relates accident risk to levels of effectiveness and estimates fatigue-associated risk.

It was expected that if fatigue is a factor in some accidents, then accident risk should be higher at lower levels of effectiveness (increased levels of fatigue), and the risk should be greater for human factors accidents than for nonhuman factors accidents. Testing these hypotheses takes several steps and makes several assumptions.

2.2.1 Assumptions

First, it is assumed that loss of alertness, confusion, or slowed reaction time causes some railroad accidents. These are all performance changes that are associated with fatigue. Since these accidents are the result of human error, it was expected that fatigue-associated accident risk would be greater for human factors than for nonhuman factors accidents.

Second, it is assumed that when the model predicts a loss in cognitive effectiveness, an increase occurs in the probability of a human error that could result in a railroad accident. For example, when the model predicts reduced effectiveness, the train crewmen are more likely to miss a signal or run through a switch. This connection is not deterministic; reduced effectiveness increases the probability of an error but does not determine that an error will occur. Furthermore, a similar error could occur in the absence of unusual fatigue, such as one caused by a distraction, so that not all occurrences of a particular sort of error can be attributable to fatigue.

Hence, a fatigue model can predict an increase in fatigue-associated risk, but not the specific occurrence of an accident.

Third, the SAFTE model incorporated in FAST makes a number of assumptions about the amount of sleep that can be achieved under specific work schedules, the amount of sleep the average person requires, and the susceptibility of the average person to time of day variations in alertness. Wide variations exist in individual sensitivity to the factors that cause fatigue, so, again, fatigue models can only predict an increased risk of fatigue, not a specific individual person's level of fatigue or performance.

2.2.2 Validation Criteria

If the fatigue model is valid, it will show that human factors-related accidents are more likely to occur at decreased levels of crew effectiveness, whereas the likelihood of nonhuman factors-related accidents are the same regardless of crew effectiveness.

The fatigue model estimates effectiveness during each work shift of each worker in the database up to the time of an accident and records the mean effectiveness level across each 30-minute work interval. From all work histories combined, the approximate proportion of work time spent by workers at given effectiveness levels can be computed. This is the exposure to fatigue and, as noted above, determines the expected proportion of accidents at any given effectiveness level. The model also estimates the crew effectiveness at the time of the accident. From these data the proportion of accidents occurring at any given crew effectiveness level can be computed. Combining the two proportions defines the risk ratio:

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

Accident risk, as used in this report, is defined entirely in terms of the risk ratio specified above. If operator effectiveness is unrelated to accident occurrence, then the proportion of accidents occurring at any given level of effectiveness should be approximately equal to the proportion of work time spent at that effectiveness level (i.e., the accident risk should be close to one). If lower effectiveness is related to a greater likelihood of an accident, then accident risk should increase with decreasing effectiveness (i.e., be greater than one at lower effectiveness levels and less than one at higher effectiveness levels).

Therefore, model-derived estimates of effectiveness that show the following would constitute validation that the model provides a measure of fatigue-associated accident risk:

1. A statistically reliable relationship between decreases in effectiveness and an increase in human factors accident risk
2. The absence of a statistically reliable relationship between effectiveness and <u>nonhuman</u> factors accident risk, and a greater risk of human factors versus nonhuman factors accidents at low levels of effectiveness

The statistical reliability of relationships was based on finding significant correlation coefficients (r). The chi square (χ^2) statistic was used to compare the distributions of human factors and nonhuman factors accidents to demonstrate significantly different distributions of risk. In each case, we rejected the null hypothesis (the hypothesis that there is no correlation, i.e., $r = 0$, or that no difference exists in the distributions of human factors and nonhuman factors accidents) when the chance probability of a finding was less than 5 percent ($p < 0.05$).

2.2.3 Calibration of the Fatigue Model

Up to this point, the analysis method has referred to low levels of effectiveness and levels of fatigue without specifying what those levels might be. This is because it is not known in advance how much fatigue or reductions in cognitive effectiveness are sufficient to cause a detectable increase in accidents or a detectable difference between human and nonhuman factors accidents. The level of effectiveness for human factor accidents that yields a reliable increase in accidents relative to chance and the effectiveness level for human factor accidents that is reliably different from the level for nonhuman factors accidents were used to calibrate the model.

2.2.4 Variables Not Considered

Given the nature of this analysis of historical work times associated with reportable accidents, the biomathematical fatigue model was not able to consider some variables that might ordinarily be part of an analysis. The database included work start and stop times, typical call times for each location, and estimated average commute times for each location. The analysis could not consider each worker's usual sleep habits or actual sleep times, quality of the sleep environment, or schedule predictability. The analysis did not consider the use of fatigue countermeasures or additional naps. The analysis could not consider unusual events that might have interfered with usual sleep opportunities. The analysis could not consider schedule delays or misinformation, medical conditions and sleep disorders, medications and/or drug use, observations of operator performance and appearance, concurrent stress, family issues, work demands, or crew resource management issues, such as communication problems. Many of these factors could cause additional fatigue or performance disruptions. Some of these factors could reduce fatigue, such as the use of napping strategies. The analysis had to assume that each individual obtained as much sleep as was afforded by the schedule, call times, commute times, and personal obligations, up to a nominal level of 8 hrs of sleep per day. Given these limitations, the predicted levels of effectiveness reported in this study are not necessarily the same as might be obtained if an accident investigator performed a detailed fatigue analysis at the time of the accident and caution should be exercised in extending the calibration levels to that application.

2.3 Sample Size Determination

A statistical process called power analysis, based on the results of a pilot study to evaluate the odds of finding a statistically significant difference in fatigue between human factor and nonhuman factor accidents given various sample sizes and assumptions about the signal to noise ratio, was used to determine sample size. The power analysis indicated that sufficient accidents occur within a 2½-year period to reach an acceptable statistical power with a sample size of 400 human factors cases and 1000 nonhuman factors cases. Since usually two employees are on each locomotive, an engineer and conductor, this results in approximately 2800 work histories and accident cases.

2.4 Selection of Accidents

Accidents that were assigned a human factors cause code (H###) for either the primary or secondary cause were considered human factors accidents. Human factors accidents caused by actions of non-operational personnel, such as maintenance or signal workers, cannot logically be associated with the fatigue or effectiveness levels of the crew operating the train. Accidents with cause codes of H305, H402, H501, H993, H994, and H997 were, therefore, excluded from the list of human factors accidents considered. All track- and equipment-related accident cause codes (Txxx or Exxx) were included for the nonhuman factors accidents.

2.5 Accident Samples for Each Participating Railroad

To achieve the sample size specified in Section 2.3, it was estimated that data would be needed for all non-excluded human factors accidents occurring during calendar years 2003, 2004, and the first 6 months of 2005 on mainline, siding, and industry track involving freight trains. All track- and equipment-caused accidents occurring over the same period on mainline, siding, and industry track involving freight trains were used as a comparison set of nonhuman factors accidents. The 5 participating railroads—BNSF, CSX, Kansas City Southern (KCS), Norfolk Southern (NS), and Union Pacific (UP)[1]—provided data on a total of 405 non-excluded human factors accidents and 1015 nonhuman factors accidents reported in this interval, which constituted the study.

2.6 Data Requested

Each participating railroad provided work schedule data for each worker for the 30 days before each of the identified accidents. Most critical to the analysis was a record of the reporting and release date/times of each work episode, including deadhead and limbo times, which for analytical purposes were combined as unavailable for sleep. (The duty periods reported are not all regulated by hours of service rules and may exceed the allowable limit of 12 hr on duty.) Data on terminals, standard call times, and estimated commute times were also requested and applied to further limit opportunities to sleep. Worker identification was used solely to associate each work history with the accident involved and was coded to insure confidentiality. The data requested were as follows:

30-day work histories prior to and including the event

- Operator identifiers, anonymous code
- Movement method: driving train, dead head, etc.
- Origin terminal (to be coded)
- Destination terminal (to be coded)
- Movement type: home-away, away-home, home-home, away-away
- Date and time of start (report time prior to any deadheading)
- Date and time of end (final release time after any deadheading)

[1] A list of abbreviations is included in Appendix A.

Terminal Information

- Terminal identifier (to be coded)
- Usual call time (amount of time prior to reporting when operator is called) for home and away terminals, if different
- Average estimated commute time

Accident/Event Data

- Operator identifiers, anonymous code
- Event date and time
- Primary event cause code (FRA classification code)
- Secondary event cause code (FRA classification code)
- Home terminal of operators
- Location of event

2.7 The SAFTE Model and the Work Schedule Fatigue Assessment Tool

The biomathematical fatigue model used for this specific analysis was the SAFTE model. The SAFTE model has received a broad scientific review and the US Department of Defense considers it the most complete, accurate, and operationally practical model currently available to aid operator scheduling. At the Fatigue and Performance Modeling Workshop held in 2002, six fatigue models from around the world were evaluated and the most recent version of SAFTE had the lowest error of all models evaluated (Van Dongen, 2004). During the past six years, the U.S. Air Force and FRA have sponsored the development a scheduling tool based upon the model that can be used to assess and manage fatigue in aircrews and railroad workers. The software was developed by Science Applications International Corporation (SAIC) and NTI and is called the Fatigue Avoidance Scheduling Tool (FAST). For the pilot project to develop a method for evaluating fatigue effects in railroad accidents, FRA sponsored the development of an extension of FAST that can process hundreds of work histories compiled in a database and provide aggregate estimates of fatigue across a spectrum of workers and for specific times associated with accidents. That batch processing version of FAST was used in this project.

2.7.1 The Work Schedule Fatigue Assessment Process

In general, the batch processing involves the following steps. First, the data are entered into a set of databases. Second, the program processed those work histories to determine when sleep could have occurred, given that workers normally cannot sleep at work, generally do not sleep after they are called to work, and must commute to and from their place of work. The tool uses an algorithm called AutoSleep to estimate how much sleep the average railroad worker would get given the sleep opportunities afforded by the work schedule and the call/commute times. AutoSleep is based on a study of railroad engineer work/rest diaries (Pollard, 1996). The AutoSleep procedure emulates how a person is most likely to arrange sleep under a particular work schedule. In general, it considers a standard bedtime, in this case set to 2200 hrs, and a typical maximum duration of sleep if time permits, set to 8 hr per day. In addition, it recognizes that the afternoon is a poor time to attempt sleep and generally does not schedule sleep between noon and 2000 hrs. On the basis of these constraints, AutoSleep schedules the maximum amount of sleep, given the sleep opportunities afforded by the work schedule, the estimated commute to and from work, and the standard call-time for each location.

The third step was to estimate the cognitive effectiveness of each worker for every minute that the worker was awake, using the SAFTE model. A detailed description of the SAFTE model has been previously published (Hursh, et al., 2004) and was applied to these data using the default parameter settings of the model. Simply stated, crew effectiveness[2] is an estimate of the speed of reaction time and alertness. Estimated effectiveness varies with the combined effect of the time of day and the pattern of sleep. Crew effectiveness follows a daily (circadian) rhythm that is much lower between 0000 and 0600 hrs rather than between 1200 and 1800 hrs. In addition, the model keeps track of the amounts of sleep opportunities a person gets as it contributes to performance and progressively degrades performance if the person experiences a loss of sleep from the nominal requirement of 8 hr per day. The batch processor summarized these detailed effectiveness estimates into 30-minute averages during every work period in the database. The program also summarized effectiveness in 30-minute intervals around the clock to provide a time of day estimate of work effectiveness. The analysis summarized effectiveness for each person's work shift by describing the distribution of the amount of time spent at work as a function of effectiveness. Finally, and most importantly, the program estimated the effectiveness of each worker at the time he/she was reported to have been involved in an accident.

2.7.2 The Dimensions of the Analysis

The batch processor created a schedule file for each worker that could be viewed in the standard version of FAST. These schedule files were used to troubleshoot any unusual values and in several cases led to corrections to the database. Of the 2962 workers involved in accidents (approximately 2 crewmembers per accident) that were reported, valid work histories were available for 2,843 workers, a 96-percent success rate. Of those, 790 were involved in human factors accidents, and 2,053 were involved in nonhuman factors accidents. Each work history was composed of a series of records that constituted a work shift. Across the five railroads, the analysis processed 57,537 work shifts. In total, the results reported are based on over 1 million 30-minute work intervals. Hence, the effectiveness exposure estimates and accident effectiveness values were based on over 2,800 work histories, and the results of all crewmembers contributed equally to the findings. The data were not coded in such a way to permit estimates of effectiveness by craft (engineer or conductor).

[2] Crew effectiveness can have values in the range of 0 to 100. A person who consistently obtains 8 hr of good quality sleep would have a peak effectiveness score of 100 during the following waking period.

3. Results of Analysis

3.1 Descriptive Analysis

The first results are simple descriptions of various features of the work schedule and accident data without any reference to the fatigue model predictions of effectiveness. The three basic descriptive charts are:

- Work durations
- Clock time of work
- Accidents by time of day

For this and all other sections of the report, the results are shown as an aggregate of the findings based on the data from all railroads.

3.1.1 Job-Related Non-Sleep Durations

The railroads reported all times crewmembers were doing work related activities, such as performing as a crewman, deadheading, or in limbo time before release. Together, these constituted job-related times unavailable for sleep. Figure 1 shows the number of these intervals sorted by duration in hourly intervals. Relatively few intervals were shorter than 5 hr. Work periods were about equally distributed from lengths of greater than 6 hr to less than 11 hr.

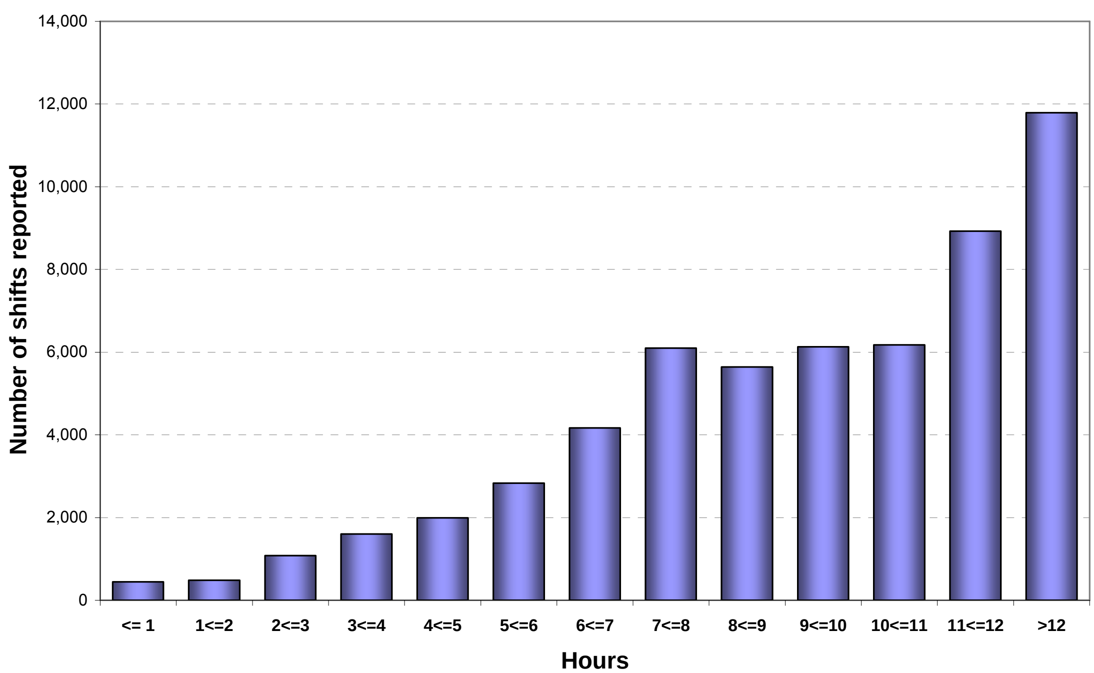

Figure 1. Durations of Reported Job-Related Non-Sleep Intervals

The most frequent job-related intervals were those between 11 and 12 hr and longer than 12 hr. This may not be surprising considering that these intervals would include deadhead time and limbo time. For example, if a typical deadhead time is at least 3 hr, then duty shifts of 9 hr or longer that were combined with deadhead time would all fall in the longer than 12 hr category.

3.1.2 Clock Time of Work

Figure 2 shows the number of work hours that occurred at hourly intervals around the clock. This analysis indicates the frequency of work at varying times of day and shows that, in general, work was fairly evenly distributed around the clock. For the population at large (the full height of the bars in Figure 2), work is a bit more concentrated between 0800 and 1600 hr and diminishes between 1600 and 0500 hr. A detailed analysis revealed that in the overall population involved in accidents, a subgroup of workers started a majority of their work periods between 2200 and 0400 hr. These workers are termed consistent night workers. Their data were separated from the larger population and evaluated as a distinctly different group, reported below. Their distribution of work is understandably more concentrated between 0000 and 1200 hr, shown as the dark blue bars in the graph.

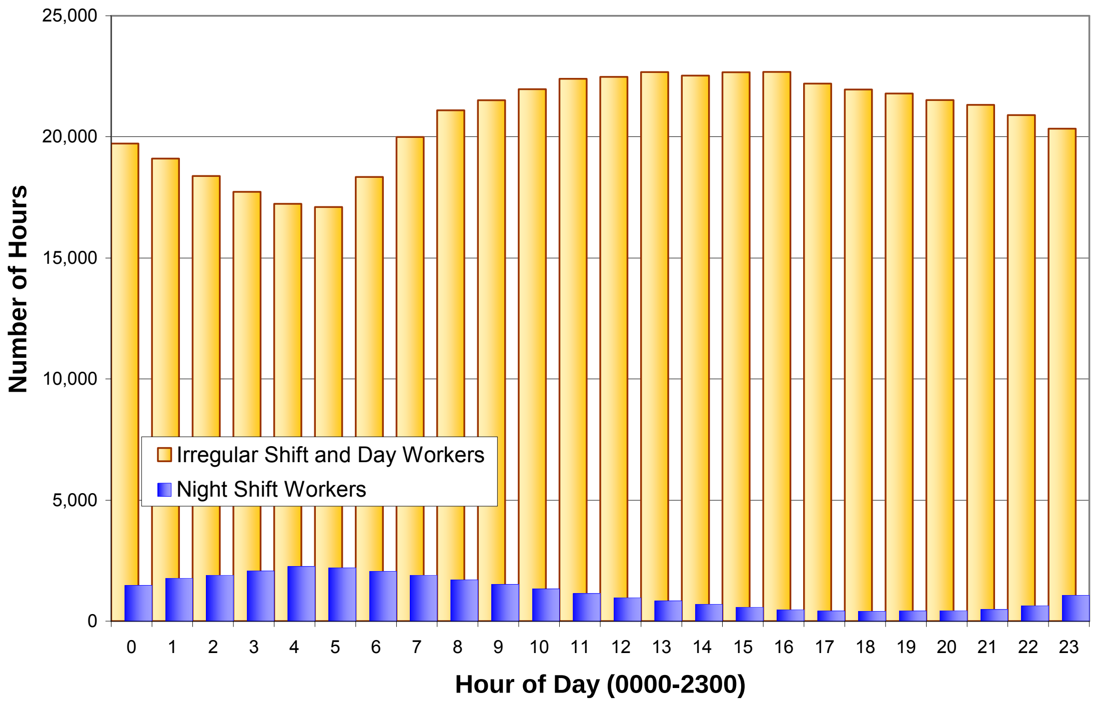

Figure 2. Frequency of Work Hours by Time of Day for Variable Schedule Workers (yellow bars) and for Night Workers (dark blue bars)

3.1.3 Accidents by Time of Day (Variable Schedule Workers)

Figure 3 shows the risk of human factors and nonhuman factors accidents in 3-hr intervals around the clock for the workers who had irregular shifts or day shifts. A similar analysis is not possible for consistent night workers because there are too few cases. For each clock interval the analysis calculated the proportion of accidents that occurred in that interval and divided that proportion by the proportion of work times that occurred at that clock time based on the work interval values in Figure 2. If accidents were randomly distributed around the clock, then the points would have a value of 1.0 and fall on the dashed line. The blue triangles in Figure 3 are for human factors accidents, and the green squares are for nonhuman factors accidents. A clear circadian pattern of human factors accident risk exists. Taking into consideration the distribution of work around the clock, accidents are relatively more likely in the early morning hours from

0000 to 0300 hr (the circadian nadir) and in the early afternoon from 1200 to 1500 hr (the postprandial dip). Accidents are much less likely in the late morning (0900 to 1200 hr) and in the early evening (1500 to 1800 hr). These patterns are predicted by a circadian pattern generated by the SAFTE model, shown as a red line. Because effectiveness is thought to decline with increased fatigue, the inverse of effectiveness is plotted against the right-hand axis. The maximum value of inverse effectiveness corresponds to the early morning peak in accident risk. The data are plotted twice along the x-axis to illustrate the rhythmic pattern of the results. Human factors time of day accident risk is reliably correlated with the circadian pattern derived from the fatigue model ($r = 0.71$, $p<0.05$). Circadian rhythms account for 51 percent of the variance in the time of day at which human factors accidents occur. By comparison, circadian rhythms only account for 6 percent of the variance in the time of day at which nonhuman factor accidents occur. Appendix C-1 shows the circadian patterns of accidents for each participating railroad.

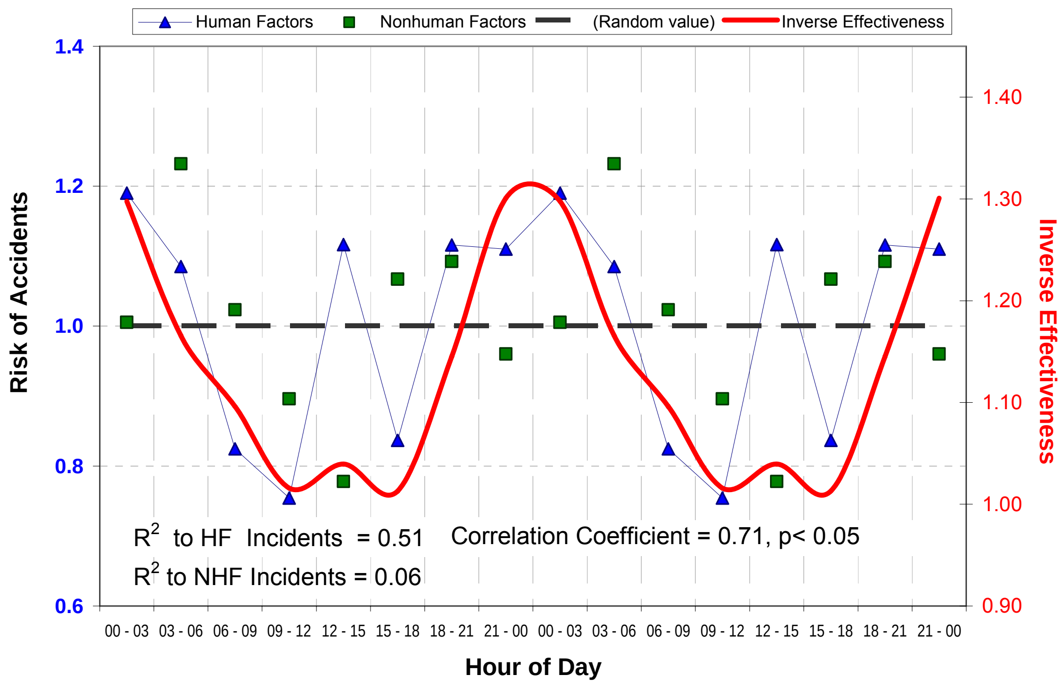

Figure 3. Accident Risk by Time of Day. Data Have Been Double-Plotted to Show the Repeating Circadian Pattern

3.2 Validation of a Fatigue Model

3.2.1 Work Time in Effectiveness Categories (Variable Schedule Workers)

Figure 4 shows the distribution of 30-minute work intervals for variable schedule workers as a function of effectiveness. The figure shows the proportion of time spent at work with an effectiveness score between 90 and 100, between 80 and 90, between 80 and 70, and so on down to the lowest bin with scores of 50 or less effectiveness. The right-hand axis shows that 42 percent of the time locomotive crews have effectiveness scores above 90. Another 23 percent of the time, they have effectiveness scores between 80 and 90. Hence, 65 percent of the time, locomotive crews have scores above 80. The percent of time declines consistently with decreasing effectiveness, and less than 5 percent of the time effectiveness is below a score of 50.

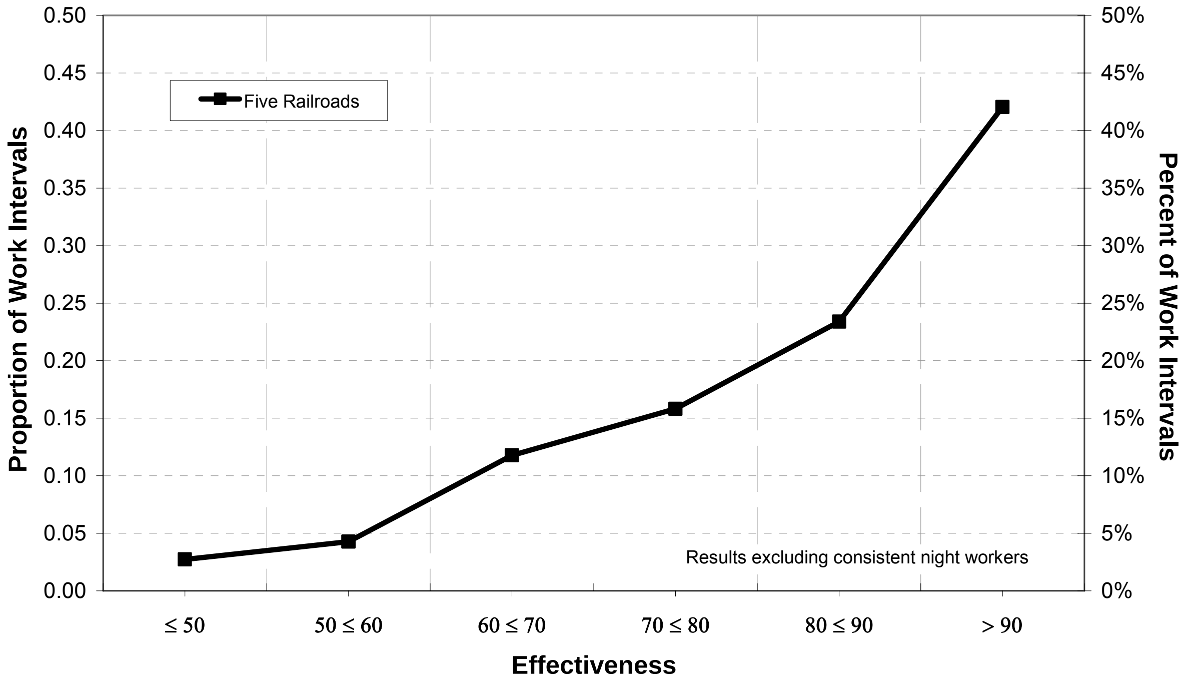

Figure 4. Work Interval Effectiveness Distribution for Five Railroads

These data provide an important reference for interpreting the estimated effectiveness values associated with accidents. If accidents are independent of effectiveness (or fatigue), then when accidents are sorted by crew effectiveness at their time of occurrence, the accidents should distribute exactly as work time effectiveness is distributed in Figure 4. However, if accidents are caused, in part, by low effectiveness (fatigue or lack of alertness), then one would expect that a greater proportion of human factors accidents would occur at low levels of effectiveness than the proportion of time or exposure to those levels of effectiveness. In other words, if the fatigue model is predictive of accidents, then low effectiveness should be associated with an elevated risk of an accident.

3.2.2 Effectiveness at the Time of the Accident (Variable Schedule Workers)

Figure 5 shows the distributions of human factors and nonhuman factors accidents for variable schedule workers by effectiveness scores of the locomotive crews at the time of the accidents. Figure 5 also shows the mean work interval effectiveness distribution from Figure 4 for comparison. The y-axis is logarithmic, so that the distance between the accident distribution lines and the time distribution line at any effectiveness level reflects the degree of risk at that level; the greater the distance between the lines, the larger the ratio that defines risk (see Section 2.2.2).

Figure 5 shows that the proportion of accidents that occur above an effectiveness score of 90 is less than expected by the distribution of time, indicating reduced risk. Below an effectiveness score of 90, the rate of human factors accidents is consistently above the work-time distribution (heavy line) and gradually separates from the line as effectiveness decreases. For nonhuman factors accidents, the proportion is sometimes above the line and sometimes below the line, and no consistent relationship exists between nonhuman factors proportions and decreases in effectiveness. Work time and accidents by effectiveness distributions for the individual railroads are included in Appendix C-2.

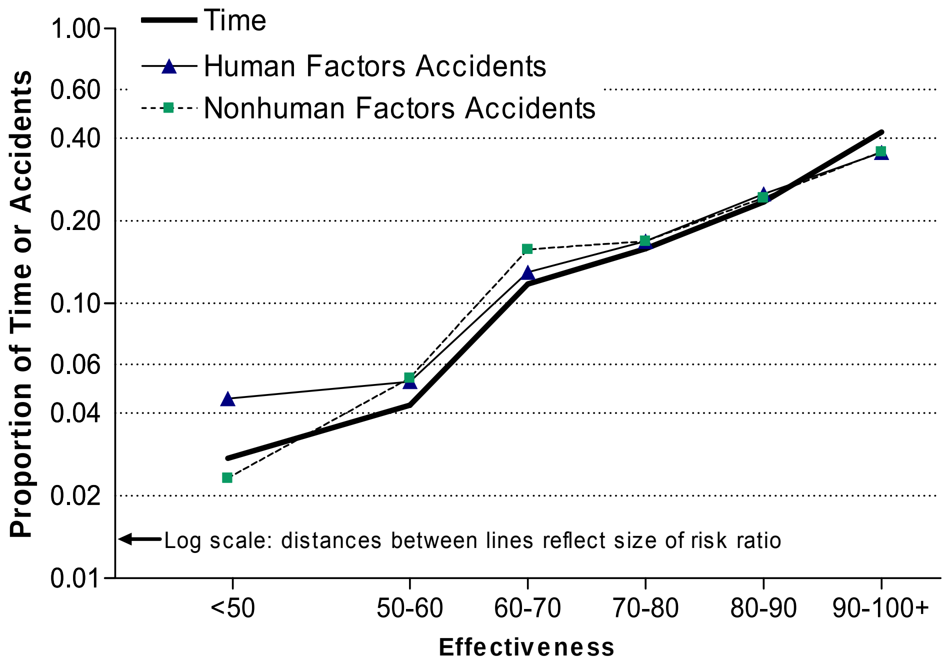

Figure 5. Proportion of Time or Accidents as a Function of Predicted Effectiveness Categories. Human Factors and Nonhuman Factors Accidents are Indicated as Blue Triangles and Green Squares, respectively

3.2.3 Consistent Night Workers

As was noted previously, certain workers had fairly consistent nighttime work shifts. About 5.7 percent or 161 out of 2,836 workers had this pattern that we have termed “consistent night workers.” The analysis defined consistent night workers as those individuals for whom 50 percent or more of their shifts started between 2200 and 0400 hrs. Since they spend much of their night time at work or commuting back and forth to work, the analysis used somewhat different AutoSleep assumptions to process their work histories; the sleep estimator reduced the afternoon “forbidden zone” to the 4 hr from 1400 to 1800 hrs, allowing them to obtain some sleep in the afternoon—a reasonable assumption for workers who lack the opportunity to sleep at night. The consistent night worker may be an engineer who regularly works a night job in the yard but was called for a line-haul job and had an incident. The work history would be that of a yard worker even though the incident was a mainline freight accident. Because the study is primarily concerned with the fatigue induced by irregular schedules of on-call engineers, the analysis separately analyzed all the night worker histories and incidents. In fact, the distribution of work times by effectiveness categories is distinctly different for this category of workers, as shown in Figure 6.

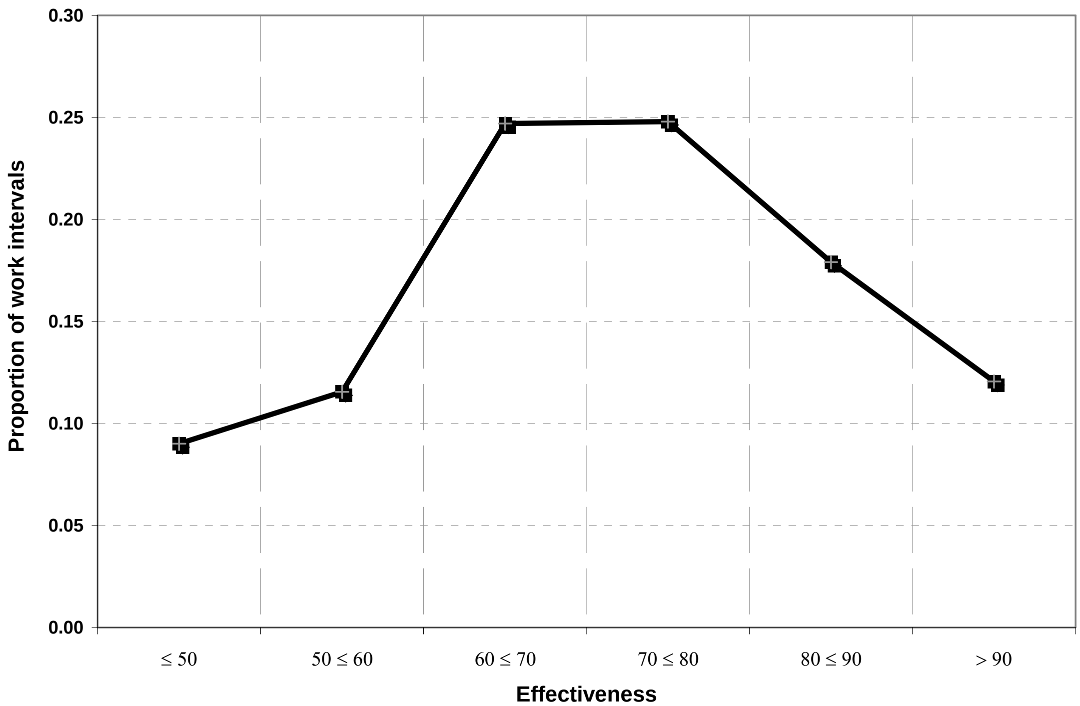

Figure 6. Work Intervals Sorted by Effectiveness Category for Consistent Night Workers, All Railroads

Figure 6 indicates that compared to variable schedule workers (Figure 4), the consistent night workers spend considerably more time at low levels of estimated effectiveness. Comparing Figs. 4 and 6, one can see that 35 percent of the time variable schedule workers have effectiveness scores less than 80, but this is the case for consistent night workers 70 percent of the time. It would not be appropriate to compare the incident effectiveness rates of the consistent night workers to the work time exposure profile of the variable schedule workers who spend much less time working at low levels of effectiveness; to do so would have inflated our estimate of incident risk. To avoid this, the incident effectiveness for consistent night workers was evaluated in comparison to their exposure levels in Figure 6.

3.2.4 Accident Risk as a Function of Effectiveness, Variable Schedule Workers

As discussed in Section 2.2.2, if effectiveness (fatigue) is not associated with accidents, then the ratio that defines relative risk will be equal to 1.0. In other words, if being exposed to a certain level of effectiveness (fatigue) does not alter chances of having an accident, then accident risk is approximately 1.0. On the other hand, a ratio greater than 1.0 indicates that accidents are more likely than chance at that level of effectiveness. A ratio of 1.5 means that a 50 percent increase in the risk of having an accident occurred at that level of effectiveness. To show the relationship between risk and effectiveness, the data points from Figure 5 were used to compute the risk ratio (using the expression from Section 2.2.2) for each effectiveness category. Figures 7 and 8 show the accident risk for the aggregated data from all five railroads for human factors and nonhuman factors accidents, respectively.

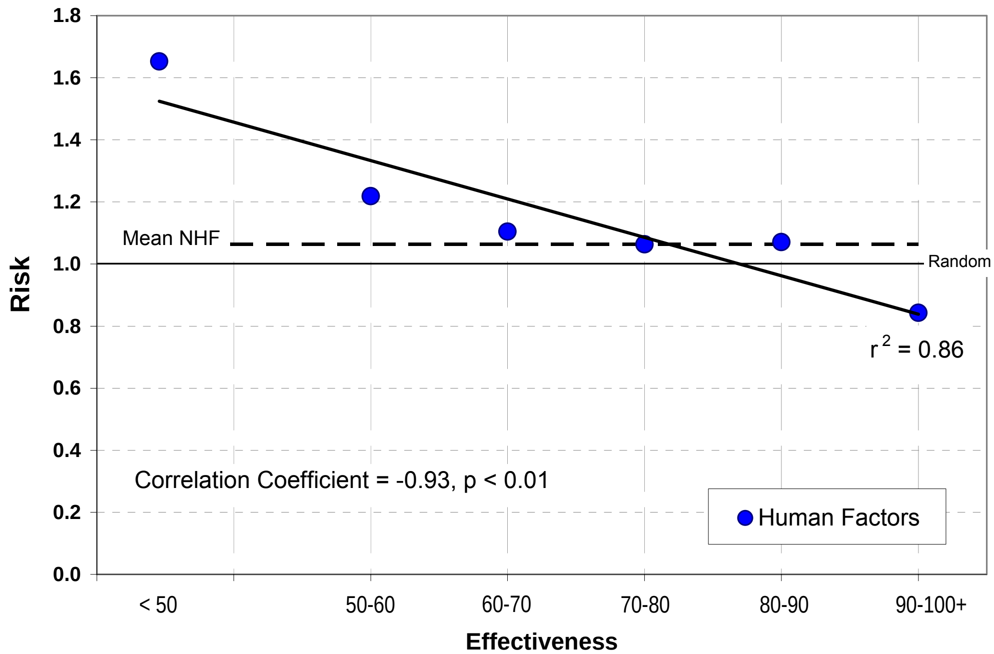

Figure 7. Human Factors Accident Risk at Each Level of Effectiveness Aggregated from Five Railroads

In Figure 7, the solid blue circles and the heavy solid line fit to them show the increase in risk for human factors accidents as a function of decreases in effectiveness (increasing fatigue) at the time of the accident. A risk value of 1.0 indicates no effect of fatigue on accident occurrence. The dashed line shows the mean accident risk of nonhuman factors accidents. As can be seen from Figure 7, human factors accident risk is well described by a linear function. The line that is fit to these data accounts for 86 percent of the variance in human factors accident risk and a significant inverse correlation exists between accident risk and effectiveness, the inverse of fatigue ($r = -0.93$, $p < 0.01$). The data show that for effectiveness scores between 90 and 100 (values associated with optimal prior sleep), a reduction of risk occurs. At effectiveness scores below 90, risk progressively increases. At the lowest level of effectiveness, a 65 percent increase in accident risk occurs (for effectiveness scores equal to or less than 50). The consistent relationship between reduced effectiveness (increased fatigue) and elevated risk indicates that the additional risk is associated with fatigue.

- This finding satisfies the first criterion for model validation: there was a significant correlation between model predicted reductions in effectiveness and an increase in human factors accident risk.

Importantly, the maximum increase in accident risk due to time of day alone was less than 20 percent (Figure 3), while the maximum increase in accident risk due to reduced effectiveness (fatigue) was 65 percent (Figure 7), reflecting the combined effects of time of day and sleep opportunities. Details of the statistical analysis are in Appendix B. Charts of human factors accident risk by effectiveness for individual railroads are included in Appendix C.

Figure 8 shows the risk for nonhuman factors accidents as a function of effectiveness. In Figure 8 no consistent, statistically reliable relationship exists between risk and effectiveness. Data points fall above and below the line that indicates a risk value of 1.0. The best fitting straight line has a slope that is not reliably different from zero, indicating that no consistent relationship exists between nonhuman factors accident risk and effectiveness ($r = -0.14$, $p > 0.05$). Only 2 percent of the variance in risk is accounted for by a linear function. Charts of nonhuman factors accident risk by effectiveness for individual railroads are included in Appendix C.

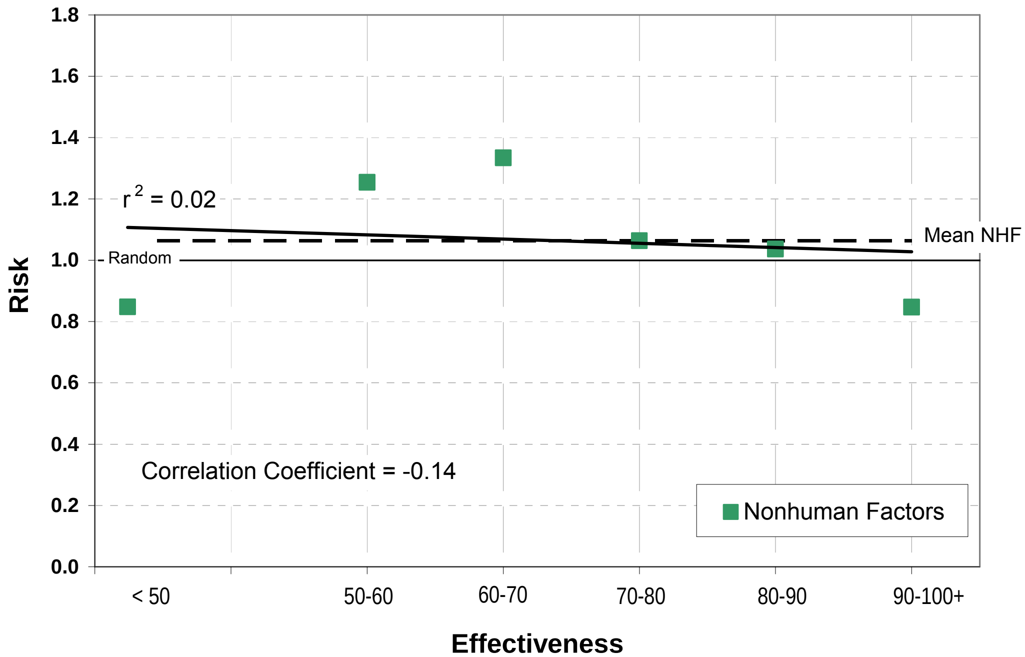

Figure 8. Nonhuman Factors Accident Risk at Each Level of Effectiveness Aggregated from Five Railroads

A large difference in relative risk occurred between human factors and nonhuman factors accidents in the lowest category of effectiveness and this difference was statistically reliable ($\chi^2 = 7.201$, $p < 0.01$). These findings satisfy the second criterion for model validation:

- No reliable relationship exists between effectiveness and nonhuman factors accident risk, and at low levels of effectiveness, human factors accident risk was elevated more than nonhuman factors accident risk.

3.2.5 Accident Risk as a Function of Effectiveness, Consistent Night Workers

To further assess the statistical significance of this possible risk factor, the analysis examined the distribution of human factors and nonhuman factors accidents between consistent night workers and the variable schedule workers. If, consistent night workers are more prone to fatigue-related incidents owing to the greater amounts of time spent at low levels of effectiveness, then one would expect that the consistent night workers would have proportionately more human factors accidents than nonhuman factors accidents compared to the variable schedule workers. Indeed, human factors accidents account for 36 percent of the incidents for consistent night workers and only 27 percent of the incidents for the variable schedule workers. This difference in distribution of human factors and nonhuman factors accidents is statistically significant ($\chi^2 = 6.56$ $p = 0.01$; Fisher's exact test $p = 0.003$, $p < 0.01$). A test of statistical reliability of the difference between the proportions of 27 percent and 36 percent is also statistically significant. Hence, one can conclude that the combined effect of fatigue and relatively large amounts of time working fatigued leads to an elevated risk of night workers having human factors accidents and this appears to occur at levels of estimated effectiveness below 70 percent. When effectiveness is above this level, it appears to be protective; fewer accidents occur than might be expected by chance (dashed line at 1.0), shown in Figure 9. Despite this significant relationship, the linear trend between reduced effectiveness and increased accident risk is not statistically significant nor is the correlation coefficient of -0.596. Results for night workers for individual railroads are not presented because the sample size is too small.

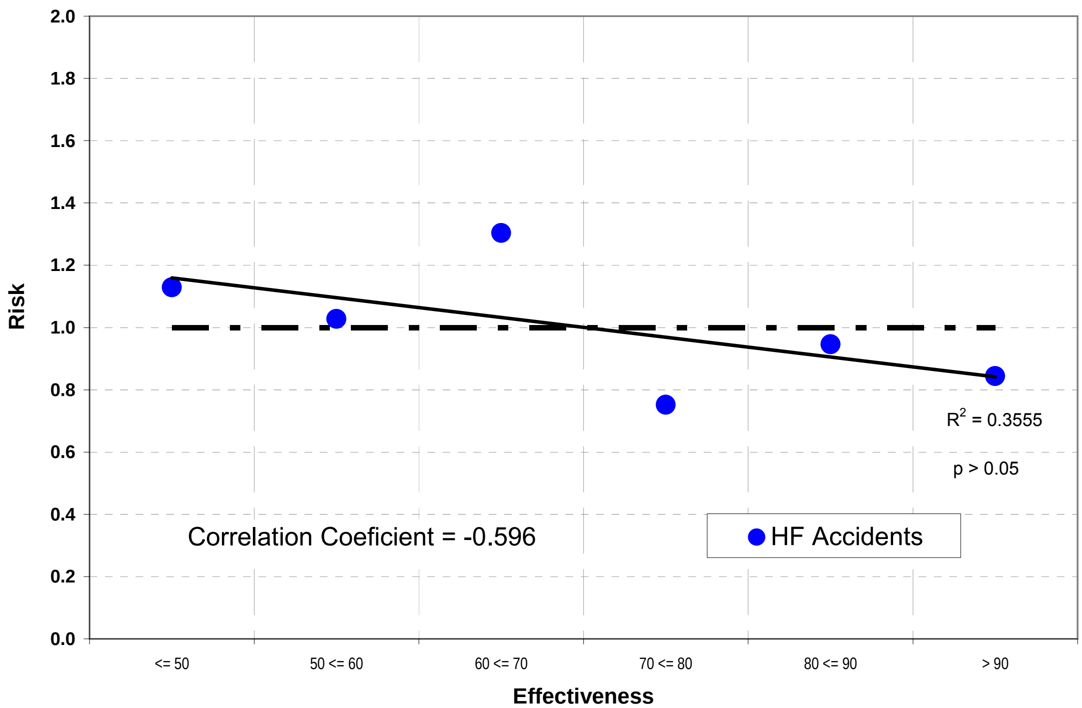

Figure 9: Human Factors Accident Risk at Each Level of Effectiveness Aggregated for Consistent Night Workers from Five Railroads

The relationship between nonhuman factors accident risk and effectiveness at the time of the accident for consistent night workers is shown in Figure 10. The correlation of -0.364 is lower than for human factors accidents and not statistically significant.

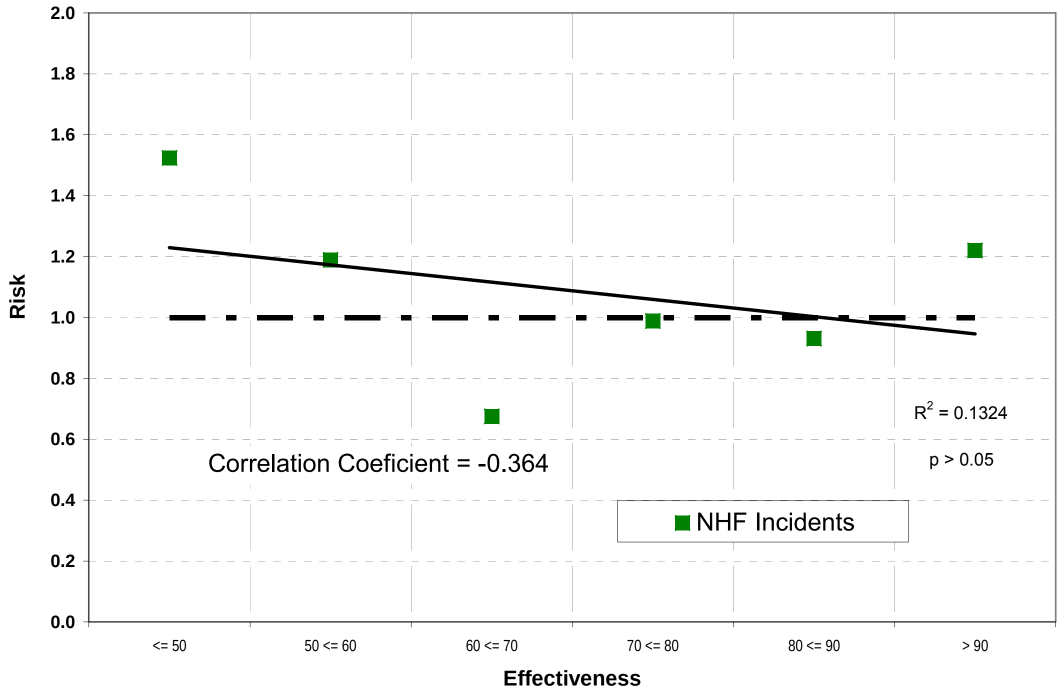

Figure 10: Nonhuman Factors Accident Risk at Each Level of Effectiveness Aggregated for Consistent Night Workers from Five Railroads

3.3 Calibration of a Fatigue Model

3.3.1 Accident Likelihood at Selected Criterion Levels (Variable Schedule Workers)

In addition to validating that a fatigue model can associate reductions in effectiveness with a reliable increase in human factors accidents, this study sought to calibrate the level of reduced effectiveness below in which a reliable increase occurs in accident risk (cumulative risk). To do this, a somewhat different analysis was conducted. In this case, a criterion level was set to include the proportion of all accidents that occurred with an effectiveness score at or below that level. The analysis then compared that proportion to the proportion of exposure time that occurred at or below that level. If the ratio of the two values is reliably greater than 1.0 at a particular criterion level, then that level could be considered the effectiveness level below which an increased cumulative risk of accidents would exist. As before, the analysis was conducted on the aggregated data, excluding the consistent night workers. Figure 11 shows the results, while Table 1 summarizes them. The blue symbols are the cumulative human factors fatigue-associated accident risk, the black line is the expected cumulative risk of accidents if they are distributed as work interval effectiveness exposure (see Figure 4), and the dashed line is the mean cumulative risk ratio for nonhuman factors accidents.

Each point shows the 95 percent confidence limits. If the lower confidence limit is above the black line (risk = 1), then the bar is reliably greater than 1.0 (chance). Interestingly, for any effectiveness criterion score below 90, human factors accidents have a cumulative risk reliably greater than expected by chance. Below an effectiveness score of 70, human factors cumulative risk is reliably greater than nonhuman factors cumulative risk and reliably greater than chance.

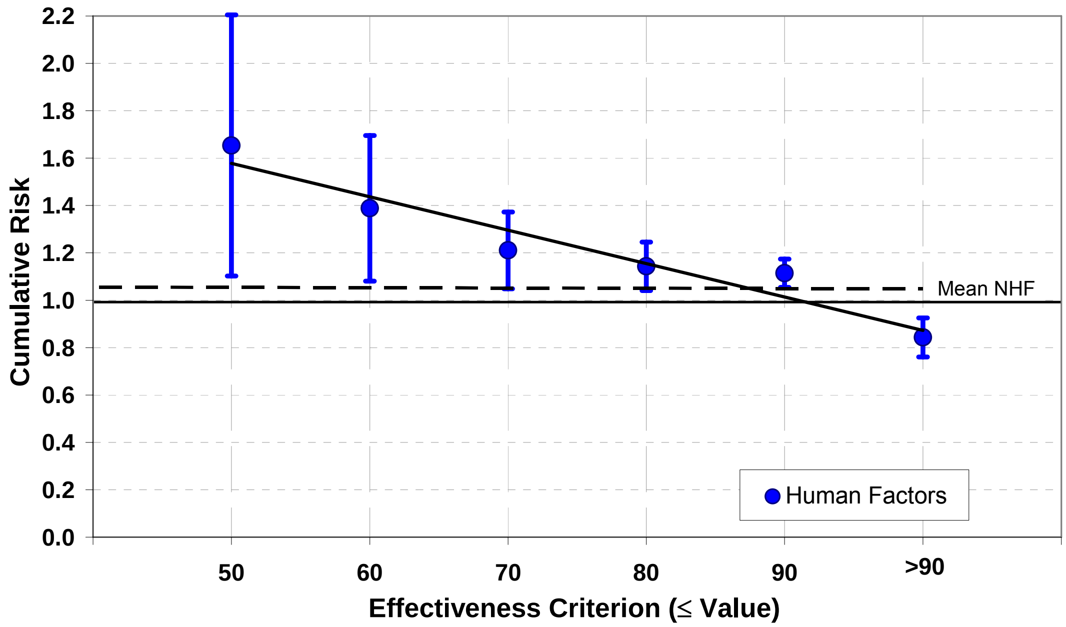

Figure 11. Human Factors Accident Risk by Criterion Levels of Effectiveness Aggregated for Five Railroads

Table 1 summarizes the results in Figure 11, along with the percent of work time and accidents at each criterion level. Several things are clear. First, no specific threshold defines the beginning of the relationship between estimated effectiveness and accident risk. At any criterion level of effectiveness below a score of 90, a significant, albeit a small, increase in risk occurs. Effectiveness above a score of 90 is actually protective; accident likelihood is significantly less than chance for both human factors and nonhuman factors accidents. An effectiveness level of 90 is meaningful; the average daytime worker who consistently gets 8 hr of sleep each night will never have an effectiveness score below 90 according to the model used here. If a person works at night or consistently misses some sleep, effectiveness drops below a score of 90, at least some of the time. Here the analysis indicates that whenever workers are below an effectiveness level of 90, they have a statistically significant increase in accident risk. However, an increase in risk, while detectable, may not be operationally meaningful. For example, Figure 2 indicates that a considerable amount of work occurs at night, which is an operational necessity. Undoubtedly, effectiveness is below a score of 90 for some or all of that time for those workers; however, that degradation is biologically determined. An 11 percent increase in accident risk may be an operational cost inherent to any industry that operates at night. What is more important to note is that fatigue risk gradually escalates with progressively lower levels of effectiveness.

Table 1. Human Factors Accident Cumulative Risk at Various Criterion Levels of Effectiveness

Criterion Effectiveness[#] Score	*Human Factors Accident Risk (%)*	*Percent of Work Time*	*Human Factors Cases Number (Percent)[+]*
> 90	**- 16 ***	42	259 (35%)
≤ 90	**+ 11 ***	58	472 (65%)
≤ 80	**+ 14 ***	35	289 (40%)
≤ 70	**+ 21 ***	19	166 (23%)
≤ 60	**+ 39 ***	7	71 (10%)
≤ 50	**+ 65 ***	2.7	33 (4.5%)

* Significantly different from chance ($p < 0.05$).

[#] Effectiveness at accident time based on 30-day work histories processed using the SAFTE biomathematical fatigue model.

[+] Human factors cases (two crewmembers per accident) in 2½ yrs, excluding accidents involving consistent night workers. The percentages above and below 90 sum to 100 percent. The percentages below 90 are cumulative and do not sum to 100 percent.

Choosing a criterion level for effectiveness, below which risk is intolerable, is an economic, operational, and safety decision beyond the scope of this report. From a purely statistical point of view, however, a criterion set at 70 and below indicates the point at which human factors fatigue-associated accident risk is reliably greater than chance and the risk for nonhuman factors accidents. Although an effectiveness level of 70 represents a point at which a detectible elevation of human factors accident risk occurs, various operational activities may occur below this level and still be conducted in a safe manner.

3.4 Other Evidence of Fatigue

The validation and calibration analyses documented in Figures 7 and 11 highlight human factors accidents that occur below an effectiveness score of 70 as the accidents most strongly related to the influence of fatigue or low effectiveness. The relationship between estimated effectiveness and accident risk is merely a correlation. This relationship could be due to fatigue, or it could be due to some other extraneous factor that is associated with this method for determining effectiveness. One way to more strongly implicate fatigue as an important part of this relationship is to identify evidence that the accidents were caused by errors that are typical of a fatigued state. To pursue this evidence, the analysis determined which cause codes were involved in the human factors accidents with effectiveness scores at or below 70 to see if they were the type of accident expected to be related to fatigue. If those accidents were caused by human errors that might reasonably be expected to result from fatigue, then that would constitute further validation of this method and increase confidence that the reliable statistical relationship is not a coincidental result of some extraneous factor(s).

Table 2. Human Factors Accident Cause Codes Related to Effectiveness ≤ 70

Rank	Cause Code	Frequency w/ Effectiveness ≤ 70	Frequency Total	Relative Likelihood	Category Description
1	H400	15	40	1.53	Main Track Authority
2	H700	42	149	1.15	Use of Switches
	All HF	215	880		All Human Factors Cause Codes, Primary and Secondary Total
Rank	**Cause Code**	**Frequency w/ Effectiveness ≤ 70**	**Frequency Total**	**Relative Likelihood**	**Individual Cause Code Description**
1	H701	4	4	4.09	Spring switch not cleared before reversing
2	H311	5	7	2.92	Moving cars while loading ramp, hose, or chute not in proper place
3	H019	8	15	2.18	Failure to release hand brake on cars
4	H222	4	8	2.05	Automatic block or interlocking signal displaying other than stop indication-failure to comply
5	H404	7	17	1.69	Train order, etc. failure to comply
6	H504	4	10	1.64	Buffing or slack action excessive, train makeup
7	H221	12	31	1.58	Automatic block or interlocking signal displaying stop indication-failure to comply
8	H605	18	53	1.39	Failure to comply with restricted speed in connection with the restrictive indication of a block or interlocking signal
9	H399	4	12	1.36	Other general switching rules
10	H503	20	60	1.36	Buffing or slack action excessive, train handling

Top 10 most overrepresented ≤ 70, excluding those with < 3 occurrences.
Note: A human factors accident could have two human factors cause codes; hence, more cases of human factors cause codes than human factors accidents occurred.

About 400 human factors accidents occurred, and usually two crew members were associated with each. Hence, about 800 cases occurred in which an effectiveness score was associated with a human factors accident, and 880 human factors cause codes were assigned because an accident could have more than one human factors cause code. Table 2 displays those categories of cause codes and individual cause codes that were over-represented in human factors accidents with an effectiveness score below 70: the proportion of these accidents at an effectiveness score below 70 was greater than their overall proportion for all human factors accidents. The top two rows of

Table 2 indicate the only two categories of causes that were overrepresented as a class: main track authority and use of switches. The lower part of the table lists the 10 most over-represented individual codes, excluding those that were rare (less than three occurrences in the entire sample). Most of these cause codes are associated with human errors that would be expected to increase with fatigue. The analysis cannot determine if these elevated risks are statistically reliable because the sample of events is too small, but this analysis suggests that the relationship between effectiveness and human factors accident risk is meaningful and not due to extraneous factors.

3.5 Interpretation of Effectiveness Scores

The results of this analysis of freight railroad accident risk indicate a reliable relationship between reduced effectiveness and an increased risk of human factors accidents. Below an effectiveness level of 70, the risk of human factors accidents is increased by about 20 percent; below an effectiveness level of 50, it is elevated by 65 percent. In this section, the effectiveness metric is related to various other metrics and sleep histories to provide a context for understanding and appreciating the kinds of circumstances that can lead to reduced levels of effectiveness of this magnitude.

The effectiveness values shown in Table 3 were derived from the SAFTE model and are based on an average person getting 8 hrs of sleep, awakening at 0700 hr, and remaining awake for the amount of time specified. A lapse is defined as an excessively long reaction time caused by loss of alertness or a micro-sleep. Lapse likelihood is the ratio of the expected frequency of lapses at an effectiveness level to the frequency of lapses of a well-rested person during a normal work day. A lapse likelihood score of 8 means that expected lapses are 8 times more frequent than for a well-rested person (Hursh, et al., 2004). The effects of wakefulness are based on studies of laboratory subjects who were kept awake after a full night of sleep and tested repeatedly on cognitive tests throughout the period of wakefulness (Angus and Heslegrave,1985; Belenky, et al., 1994). The blood alcohol concentration (BAC) equivalence is based on studies comparing the effects of alcohol and sleep deprivation on performance on a driving simulator (Arnedt, Wilde, Munt, and MacLean, 2001) and cognitive test performance (Dawson and Reid, 1997). As an example, Table 3 shows that at an effectiveness score of 70, lapses are 5 times more likely than for a well-rested person and that this score is the equivalent of being awake for 21 hr after awakening at 0700 h, or having a BAC of 0.08.

Table 3. The Relationship Among Various Effectiveness Scores and Other Meaningful Metrics: Likelihood of a Lapse, Continuous Hours Awake, and BAC

Effectiveness Score	Lapse Likelihood	Hours Awake (Hr:Min)	BAC Equivalent
98	0.2	14:00	
94	1.0	15:10	
90	1.5	16:00	
80	3	18:00	
77	**4**	**18:30**	**0.05**
70	**5**	**21:00**	**0.08**
69	5.4	22:00	
60	8	40:50	
50	12	42:30	
40	18	64:00	

Another way to understand effectiveness values is to consider the kinds of sleep patterns and times of day that can lead to different levels of effectiveness. The values in Table 4 were derived from the SAFTE model and are based on an average person awakening at 0700 hr after getting the hours of sleep shown in the first column or losing the amount of sleep (relative to 8 hr per day) in the second column. Performance at 1600 hr reflects near optimal performance for a person on that schedule; performance at 0400 hr reflects the combined effects of prior sleep, time of day, and 21 hr of wakefulness. The values at 0400 hr do not consider benefits of an evening nap; for example, a 2-hour nap at 2000 hr improves performance at 0400 hr by 4-6 percent.

Table 4. The Effects of Various Daily Sleep Patterns on Effectiveness Estimates at 1600 hr and 0400 hr. Three schedules: 1, 2, 7 Days at the Specified Sleep Level

Prior Daily Sleep (Hr)	Prior Daily Sleep Loss (Hr)	Effectiveness Score After:					
		One Day		Two Days		Seven Days	
		1600 hr	0400 hr	1600 hr	0400 hr	1600 hr	0400 hr
8	0	97	70	97	70	97	70
7	1	96	69	95	68	93	67
6	2	94	68	92	66	88	63
5	3	92	65	89	62	82	57
4	4	90	63	84	58	72	48
3	5	87	59	78	51	57	34
2	6	83	55	70	42	*	*
1	7	78	49	58	30	*	*
0	8	73	43	46	15	*	*

* No data available for these conditions

Relating Table 4 to Table 2, the analysis shows that if a person gets less than 8 hr sleep on a regular basis, then effectiveness at 0400 hr (the circadian minimum) will be below a score of 70 and accident risk will be elevated by at least 21 percent. If the person gets less than 4 hr sleep, then effectiveness at 0400 hr is below a score of 65 in one day, less than a score of 60 in two days, and less than a score of 50 in seven days, at which point accident risk is elevated by 65 percent. After seven days of 4 hr sleep per day, effectiveness at the circadian peak (about 1600 hr) is nearing a score of 70 or an elevated risk of 14-21 percent in the daytime.

3.6 Interpretations and Limitations

This study was designed to test the validity of biomathematical models of fatigue, here the SAFTE model and the FAST software, as tools for evaluating schedules for increased risk of fatigue-related accidents. The data considered—accidents and the 30-day work histories that preceded them—were not a random sample of all locomotive crews in freight rail operations. Hence, the levels of effectiveness calculated by the model should not be interpreted as representative of the railroad work force in general. The study was not designed to determine the extent of fatigue in the freight rail industry. The objective was to test the validity of the mathematical model and to calibrate the model so that one could relate the effectiveness estimates to an estimate of elevated accident risk.

While the biomathematical model used is typical of other sleep and performance models and is designed to simulate the effects of circadian rhythms and sleep patterns on fatigue, it is important to emphasize that the data that drove these predictions were entirely related to work schedule and opportunities to obtain sleep. The study did not measure sleep duration and quality directly, nor did the study directly measure cognitive performance. Operationally, this study related work schedule to accident risk using a sophisticated mathematical algorithm, but it did not directly measure fatigue or the performance manifestations of fatigue. The relationships observed are entirely correlational, not causal. It is theoretically conceivable that some other factor correlated with the same work schedule factors considered by the fatigue model was the real operational cause of the elevated accident risk. That possibility cannot be entirely eliminated with the sorts of data used in this study. In practice, however, that logical possibility may not matter if elimination of those work schedule risk factors results in improved safety. The encouraging results of this study provide justification for such an experiment.

The inferences of sleep opportunity are based on a sleep estimation algorithm called AutoSleep, and the calculated effectiveness values were partly dependent on the settings of AutoSleep. Several studies have been completed (Gertler and Viale, 2006a,b, 2007) or are currently underway to measure actual sleep under typical railroad schedules, and those findings may lead to improvements in the sleep estimation accuracy of AutoSleep, which may lead to greater accuracy in estimates of the true risk of fatigue.

The accident risk levels that were found must be carefully limited to the operational factors that were considered: work schedule, time of day, call times, and commute times. Accidents that were estimated to have occurred with relatively high levels of effectiveness could have been fatigue related if the fatigue were caused by factors not considered by the analysis, such as sleep disorders, poor sleep hygiene, unusual circumstances in the operator's life, such as illness or family demands, or unusual sleep needs. This sort of error might be called a miss, a case of fatigue that was misidentified as not fatigue.

Even in cases where the fatigue model judged that fatigue could have occurred, it is possible that the human factors accident was caused by some other human error, such as distraction or poor training. This sort of error might be termed a false alarm, a case in which the analysis suspected fatigue as the cause but was really some other cause. These two sorts of errors, misses and false alarms, are typical of discriminations involving noise, and unidentified factors that lead to error in judgment. When making such judgments, the user of the model must weigh the relative costs associated with the two kinds of errors and set a criterion value for using the model, based on Table 1 that is operationally, economically, and socially optimal (see Raslear and Coplen, 2004). Making that decision is beyond the scope of this study.

Furthermore, fatigue models based entirely on the sorts of data used in this study are not suitable for determining the fatigue of particular individuals because of the kinds of inaccuracies discussed above, especially the wide variation in individual sleep requirements and absence of specific information on individual sleep habits, health, and circumstances. On the other hand, by adding such additional information to a fatigue analysis, it is possible to reduce the error sufficiently to use fatigue models as tools in accident investigations, combined with other important information about the performance of the individuals. That sort of detailed analysis, however, was clearly beyond the scope of this study.

Finally, it is important to properly interpret the risk values in Table 1. Those values represent the percent change in human factors accidents relative to the expected distribution of accidents based on work time effectiveness. The results indicate that the lower the effectiveness level, the greater the elevation of human factors accident risk. The percent changes in risk (second column, Table 1) are not to be confused with the percentages of human factors accidents associated with that level of risk (fourth column, Table 1). The percentages in second column represent the increase (or decrease) in the risk of fatigue-associated human factors accidents compared to chance when at a particular level of effectiveness.

The overall probability of a particular railroad worker on a particular work shift having a human factors accident is extremely low. This study only considered work histories related to accidents, not all work histories. Consequently, an overall probability of fatigue-caused human factors accidents in freight service cannot be calculated, but that probability will be a number with four or five zeros after the decimal point.

4. Summary and Conclusions

4.1 Summary

This report summarized the results of a project sponsored by FRA's Office of Research and Development and Office of Safety to develop a methodology to validate and calibrate biomathematical fatigue models for use as fatigue management tools. This study assessed whether a fatigue model can predict an increased risk of human factors caused accidents. Using 2½ years of accident data from five Class I freight railroads, the fatigue model used 30-day work histories before a sample of human factors and nonhuman factors accidents to determine the relationship between accident risk and crew effectiveness (the inverse of fatigue). This report summarized the scope of that effort, the methods used to conduct the analysis, and the results of the analysis to date. The study considered approximately 2,800 crew member work histories associated with 1400 accidents, 400 of which involved human factors errors. A biomathematical fatigue model (SAFTE) evaluated a total of over 1 million 30-minute work intervals, covering over 57,000 work starts, for predicted effectiveness. The 2,800 work histories served as a basis for determining the exposure level to various levels of fatigue in these locomotive crews. In addition, the fatigue model calculated estimated effectiveness at the time of each accident, again based on 2,800 estimates of effectiveness at the time of the accidents. The following summarizes the results of the study.

Human factors accidents follow a circadian pattern that is significantly correlated with the circadian rhythm of a fatigue model ($r = 0.71$, $p < 0.05$). The same model rhythm is not correlated with nonhuman factors accidents. The maximum human factors accident risk due to time of day alone was increased less than 20 percent (Figure 3), while the maximum increase in accident risk due to reduced effectiveness (fatigue) was 65 percent (Figure 7), reflecting the combined effects of time of day and sleep opportunities.

- Validation. The fatigue model met two validation criteria.
    1. Accident risk was significantly correlated with effectiveness for human factors accidents ($r = -0.93$, $p < 0.01$, Figure 7).
    2. The model distinguished between human factors and nonhuman factors accidents. Nonhuman factors accident risk was not correlated with estimated effectiveness ($r = -0.14$, $p > 0.05$, Figure 8), and the relative risk of human factors and nonhuman factors accidents were significantly different at low effectiveness levels ($\chi^2 = 7.201$, $p < 0.01$).
- Calibration. Effectiveness has a reliable correspondence to increases in accident risk.
    1. Risk is a relatively smooth increasing function with decreases in effectiveness.
    2. Above an effectiveness score of 90, risk is significantly reduced relative to chance.
    3. Below an effectiveness score of 70, risk is significantly elevated relative to nonhuman factors risk.
    4. Overall, railroad workers in this study spent about 40 percent of work time above an effectiveness score of 90 and about 20 percent of work time below an effectiveness score of 70.

- Human factors accidents accounted for 36 percent of the incidents for night workers and 27 percent of the incidents for the remaining population that do not work primarily at night (variable schedule workers). This difference in distribution of human factors and nonhuman factors accidents was statistically significant ($\chi^2 = 6.56$, $p = 0.01$; Fisher's exact test $p = 0.003$, $p < 0.01$).
- An effectiveness score of 90 is the minimum level for a person getting 8 hrs sleep per day and awake from about 0700 to 2300 hr (day shift).
- An effectiveness score of 70 is the minimum level for a rested person after being awake for 21 hr at 0400 hr.
- An effectiveness score of 70 is about equal to the effects of 0.08 BAC and lapse likelihood five times greater than a well-rested person during the daytime.

An analysis of the cause codes associated with the accidents that occurred at effectiveness at or below an effectiveness score of 70 indicated an overrepresentation of human factors errors associated with main track authority and use of switches. Most of the overrepresented individual cause codes reflect the kinds of operator errors one might expect of persons who are fatigued. This finding confirms that the relationship between effectiveness and human factors accident risk is meaningful and not a circumstantial coincidence.

4.2 Conclusion

This project established that a biomathematical fatigue model can be used to assess how much work schedule factors can contribute to increased fatigue and an elevated risk of railroad accidents. The virtue of having a validated fatigue model, especially if it is calibrated to accident likelihood, is that a carrier could use it to do a self-assessment of fatigue across its system. By evaluating work histories on a terminal by terminal basis and using the scores from the model as a metric, the carrier could determine which terminals are experiencing schedules that might be generating increased risk of fatigue in train operators. Perhaps none of the terminals have a problem, perhaps just a few. In any case, the carrier would be in a position to use this objective assessment as a way to focus its fatigue management efforts where the greatest payoff would be expected. This study provides evidence that such a strategy, using a validated fatigue model, can identify work schedule-related fatigue factors that contribute to an elevated risk of accidents.

5. References

Abelson, R.P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum.

Angus, R., and Heslegrave, R. (1985). Effects of sleep loss on sustained cognitive performance during a command and control simulation. *Behavior Research Methods, Instruments, and Computers, 17,* 1, 55-67.

Armitage, P., and Berry, G. (1994). *Statistical methods in medical research* (3rd edition). London: Blackwell Science.

Arnedt, J.T., Wilde, G.J., Munt, P.W., and MacLean, A.W. (2001). How do prolonged wakefulness and alcohol compare in the decrements they produce on a simulated driving task? *Accident Analysis and Prevention*, *33,* 3, 337-44.

Belenky, G., Penetar, D., Thorne, D., Popp, K., Leu, J., Thomas, M., Sing, H., Balkin, T., Wesensten, N., and Redmond, D. (1994). The effects of sleep deprivation on performance during continuous combat operations. In B.M. Marriott (Ed.), *Food components to enhance performance* (pp.127-135). Washington, DC: National Academy Press.

Belenky, G., Wesensten, N.J., Thorne, D.R., Thomas, M.L., Sing, H.C., Redmond, D.P., Russo, M.B., and Balkin, T.J. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of Sleep Research*, *12*, 1, 1-12.

Conover, W.J. (1999). *Practical nonparametric statistics* (3rd edition). New York: John Wiley and Sons.

Dawson, D., and Reid, K. (1997). “Fatigue, alcohol and performance impairment.” *Nature 388*, 23.

Fleiss, J.L., Levin, B., and Paik, M.C. (2003). *Statistical methods for rates and proportions* (3rd edition). Hoboken NJ: John Wiley & Sons.

Gertler, J., and Viale, A. (2006a). *Work schedules and sleep patterns of railroad signalmen* (Report No. DOT/FRA/ORD-06/19). Washington, DC: U.S. Department of Transportation.

Gertler, J., and Viale, A. (2006b). *Work schedules and sleep patterns of railroad maintenance of way workers* (Report No. DOT/FRA/ORD-06/25). Washington, DC: U.S. Department of Transportation.

Gertler, J., and Viale, A. (2007). *Work schedules and sleep patterns of railroad dispatchers* (Report No. DOT/FRA/ORD-07/11). Washington, DC: U.S. Department of Transportation.

Hursh S.R., Balkin, T.J., Miller, J.C., and Eddy, D.R. (2004). The fatigue avoidance scheduling tool: Modeling to minimize the effects of fatigue on cognitive performance. *SAE Transactions*, *113*, 1, 111-119.

Hursh, S.R., Raslear, T.G., Kaye, A.S., and 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.

Hursh S.R., Redmond, D.P., Johnson, M.L., Thorne, D.R., Belenky, G., Balkin, T.J, Storm, W.F., Miller, J.C., and Eddy, D.R. (2004). Fatigue models for applied research in warfighting. *Aviation, Space and Environmental Medicine*, *75*, 3, Suppl.: A44-53.

Lowry, R. (1998-2006). VassarStats: Web site for Statistical Computation. http://faculty.vassar.edu/lowry/VassarStats.html

Newcombe, R.G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, *17*, 873-890.

Pollard, J. K. (1996). *Locomotive engineer's activity diary* (Report No. DOT/FRA/RRP-96/02). Washington, DC: U.S. Department of Transportation.

Raslear, T. G., and Coplen, M. (2004). Fatigue models as practical tools: Diagnostic accuracy and decision thresholds. *Aviation, Space and Environmental Medicine*; *75,* 3, Suppl.: A168-172.

Van Dongen, H.P.A. (2004). Comparison of mathematical model predictions to experimental data of fatigue and performance. *Aviation, Space and Environmental Medicine*; *75,* 3, Suppl.: A15-36.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209-212.

Appendix A: Abbreviations/Acronyms

AAR	Association of American Railroads
BAC	blood alcohol concentration
FAST	Fatigue Avoidance Scheduling Tool
FRA	Federal Railroad Administration
SAFTE	Sleep, Activity, Fatigue, and Task Effectiveness Model

Appendix B.
Statistical Analysis of Incident Frequency Data

Analysis of data supplied by five railroads focused on the relationship between occurrence of human factors-related and nonhuman factors incidents and the amount of work hours recorded by employees at levels of effectiveness estimated by the fatigue model. The following analyses are based on data presented in Tables B-1 and B-2 below.

For each worker, the FAST model estimates an average effectiveness level over each half-hour interval of each work period recorded in the database. Table B-1 shows the distribution of these effectiveness levels over all work intervals in the data set. For example, it can be seen that of 1,052,978 half-hour intervals during which a worker in the database was on duty and that in 49,131 of these intervals the worker's estimated effectiveness level was between 50 and 60, a proportion of 0.047 or approximately 47 in 1,000. Of 790 human factors incidents recorded in the data set analyzed, 45 occurred when a worker's effectiveness level was between 50 and 60, a proportion of 0.057 or approximately 57 of 1,000. Of 2,046 nonhuman factors incidents recorded in the database, 118 occurred when a worker's effectiveness level was between 50 and 60, a proportion of 0.058 or approximately 58 of 1,000.

The concept of risk is discussed in Section 2.2.2 of the main report. Briefly, the risk of an incident occurring, when a worker's effectiveness level is in a specified range, is the proportion of all incidents (at any effectiveness level) in that range divided by the proportion of all work intervals in which worker effectiveness level in that range. A risk of 1 indicates that the proportion of incidents is exactly what would be expected if there is no correlation between incidents occur at random across all work intervals; a risk greater than 1 indicates that an incident is more likely than would be expected by chance, and a risk less than 1 indicates that an incident is less likely than would be expected by chance. In the case discussed above, the human factors incident risk associated with effectiveness levels between 50 and 60 is 0.057/0.047 or 1.221, while the risk of a nonhuman factors incident is 0.058/0.047 or 1.236, so both types of incidents are found in this range of effectiveness levels more often than would be expected by chance.

Table B-1 distinguishes between consistent night workers (those who regularly work at night and whose circadian rhythms and habitual sleeping patterns are likely to have been affected) and all other workers, including consistent day workers as well as those with variable work schedules (i.e., all those whose work schedules have presumably not affected their circadian rhythms or habitual sleep patterns). Of the 1,052,978 half-hour intervals, 57,730 of them were associated with a consistent night worker on duty, and in 6,665 of these the worker's effectiveness level was between 50 and 60, a proportion of 0.115. Out of a total of 59 human factors incidents occurring while consistent night workers were on duty, 7 occurred when the worker's effectiveness level was between 50 and 60, a proportion of 0.119; out of a total of 102 non human factors incidents that occurred while a consistent night worker was on duty, 14 occurred when that worker's effectiveness level was between 50 and 60, a proportion of 0.137. The corresponding HF risk is 0.119/0.115 = 1.028 and the NHF risk is 0.137/0.115 = 1.189.

Table B-2 presents the same measures as Table B-1 for cumulative effectiveness levels—i.e., all work intervals and incidents occurring when a worker's effectiveness level is at or below a given value.

Table B-1. Work Interval and Incident Data for Five Railroads by Estimated Effectiveness Levels

EFFECTIVENESS BREAKDOWNS	Modeled Worker Effectiveness in Interval						
	<=50	50<=60	60<=70	70<=80	80<=90	>90	TOTAL
HALF-HOUR TIME INTERVALS							
ALL SHIFTS	32,379	49,131	131,349	171,749	243,105	425,265	1,052,978
VARIABLE-SHIFT & DAY WORKERS	27,181	42,466	117,091	157,435	232,767	418,308	995,248
NIGHT WORKERS	5,198	6,665	14,258	14,314	10,338	6,957	57,730
FRACTION OF TIME INTERVALS							
ALL SHIFTS	0.031	0.047	0.125	0.163	0.231	0.404	1.000
VARIABLE-SHIFT & DAY WORKERS	0.027	0.043	0.118	0.158	0.234	0.420	1.000
NIGHT WORKERS	0.090	0.115	0.247	0.248	0.179	0.121	1.000
HUMAN FACTORS INCIDENTS							
ALL SHIFTS	39	45	114	134	193	265	790
VARIABLE-SHIFT & DAY WORKERS	33	38	95	123	183	259	731
NIGHT WORKERS	6	7	19	11	10	6	59
NON-HUMAN FACTORS INCIDENTS							
ALL SHIFTS	59	118	322	352	488	707	2,046
VARIABLE-SHIFT & DAY WORKERS	45	104	305	327	471	692	1,944
NIGHT WORKERS	14	14	17	25	17	15	102
PROPORTION OF HUMAN FACTORS INCIDENTS							
ALL SHIFTS	0.049	0.057	0.144	0.170	0.244	0.335	1.000
VARIABLE-SHIFT & DAY WORKERS	0.045	0.052	0.130	0.168	0.250	0.354	1.000
NIGHT WORKERS	0.102	0.119	0.322	0.186	0.169	0.102	1.000
PROPORTION OF NON-HUMAN FACTORS INCIDENTS							
ALL SHIFTS	0.029	0.058	0.157	0.172	0.239	0.346	1.000
VARIABLE-SHIFT & DAY WORKERS	0.023	0.053	0.157	0.168	0.242	0.356	1.000
NIGHT WORKERS	0.137	0.137	0.167	0.245	0.167	0.147	1.000
HUMAN FACTORS RISK							
ALL SHIFTS	1.605	1.221	1.157	1.040	1.058	0.831	1.000
VARIABLE-SHIFT & DAY WORKERS	1.653	1.218	1.105	1.064	1.070	0.843	1.000
NIGHT WORKERS	1.129	1.028	1.304	0.752	0.946	0.844	1.000
NON-HUMAN FACTORS RISK							
ALL SHIFTS	0.938	1.236	1.262	1.055	1.033	0.856	1.000
VARIABLE-SHIFT & DAY WORKERS	0.848	1.254	1.334	1.063	1.036	0.847	1.000
NIGHT WORKERS	1.524	1.189	0.675	0.989	0.931	1.220	1.000

One question is whether incidents, either human factors related or not, tend to occur more frequently at lower effectiveness levels. A related question is whether human factors-related incidents tend to occur more frequently relative to nonhuman factors incidents at lower effectiveness levels.

To address the first question, the rates of occurrence of human factors-related incidents and of nonhuman factors-related incidents within a specified range of estimated effectiveness were compared with the fraction of time spent on the job within that range. A first analysis compared the proportion of human factors and nonhuman factors incidents for effectiveness levels of 50 or less with the proportion of half-hour time intervals worked at estimated effectiveness levels of 50 or less.

Table B-2. Cumulative Work Interval and Incident Data for Five Railroads by Estimated Effectiveness Levels

EFFECTIVENESS BREAKDOWNS	Modeled Worker Effectiveness in Interval						
	<=50	50<=60	60<=70	70<=80	80<=90	>90	TOTAL
HALF-HOUR TIME INTERVALS							
ALL SHIFTS	32,379	81,510	212,859	384,608	627,713	425,265	1,052,978
VARIABLE-SHIFT & DAY WORKERS	27,181	69,647	186,738	344,173	576,940	418,308	995,248
NIGHT WORKERS	5,198	11,863	26,121	40,435	50,773	6,957	57,730
FRACTION OF TIME INTERVALS							
ALL SHIFTS	0.031	0.077	0.202	0.365	0.596	0.404	1.000
VARIABLE-SHIFT & DAY WORKERS	0.027	0.070	0.188	0.346	0.580	0.420	1.000
NIGHT WORKERS	0.090	0.205	0.452	0.700	0.879	0.121	1.000
HUMAN FACTORS INCIDENTS							
ALL SHIFTS	39	84	198	332	525	265	790
VARIABLE-SHIFT & DAY WORKERS	33	71	166	289	472	259	731
NIGHT WORKERS	6	13	32	43	53	6	59
NON-HUMAN FACTORS INCIDENTS							
ALL SHIFTS	59	177	499	851	1,339	707	2,046
VARIABLE-SHIFT & DAY WORKERS	45	149	454	781	1,252	692	1,944
NIGHT WORKERS	14	28	45	70	87	15	102
PROPORTION OF HUMAN FACTORS INCIDENTS							
ALL SHIFTS	0.049	0.106	0.251	0.420	0.665	0.335	1.000
VARIABLE-SHIFT & DAY WORKERS	0.045	0.097	0.227	0.395	0.646	0.354	1.000
NIGHT WORKERS	0.102	0.220	0.542	0.729	0.898	0.102	1.000
PROPORTION OF NON-HUMAN FACTORS INCIDENTS							
ALL SHIFTS	0.029	0.087	0.244	0.416	0.654	0.346	1.000
VARIABLE-SHIFT & DAY WORKERS	0.023	0.077	0.234	0.402	0.644	0.356	1.000
NIGHT WORKERS	0.137	0.275	0.441	0.686	0.853	0.147	1.000
HUMAN FACTORS RISK							
ALL SHIFTS	1.605	1.374	1.240	1.151	1.115	0.831	1.000
VARIABLE-SHIFT & DAY WORKERS	1.653	1.388	1.210	1.143	1.114	0.843	1.000
NIGHT WORKERS	1.129	1.072	1.199	1.041	1.021	0.844	1.000
NON-HUMAN FACTORS RISK							
ALL SHIFTS	0.938	1.118	1.206	1.139	1.098	0.856	1.000
VARIABLE-SHIFT & DAY WORKERS	0.848	1.095	1.245	1.162	1.111	0.847	1.000
NIGHT WORKERS	1.524	1.336	0.975	0.980	0.970	1.220	1.000

For 32,379 half-hour work intervals out of 1,052,978 total work intervals (all shifts), estimated effectiveness was 50 or less (i.e., 3.07 percent of the intervals were spent at an effectiveness level of 50 or less). This was assumed to represent the proportion of work time spent by the population of all employees at such effectiveness levels. Of 790 work intervals encompassing an human factors incident, 39 (4.94 percent) found the employee at an estimated effectiveness level of 50 or less. This proportion is 1.605 as large as, or over 60 percent greater than, the proportion of work at these effectiveness levels in the population and represents a more than 60 percent increase in risk of a human factors accident when a worker's effectiveness level is 50 or less.

The data used in this study was a subset of all work records from the 5 railroads studied. To better estimate the range of risk values likely to be found overall, a statistical confidence interval was computed in standard fashion (Fleiss, 2003) for the proportion of human factors incidents at effectiveness levels of 50 or less. A 95 percent confidence interval for this proportion is given by $0.0494 \pm 1.96 \times \sqrt{[0.0494 \times (1-0.0494)/790]}$, so that one can be 95 percent certain that the proportion of human factors incidents occurring when a worker on duty is at an effectiveness level of 50 or less is greater than 0.03426 and less than 0.06447. Recalling that the proportion of half-hour intervals when a worker is at an effectiveness level of 50 or less was 0.307, then one can be 95 percent certain that the risk of a human factors incident with a worker in this range of effectiveness levels is greater than 0.03426/0.0307 = 1.102 and less than 0.06447/0.0307 = 2.097. In other words, the risk of an HF incident with a worker at an effectiveness level of 50 or less is at least 0.102 (or 10.2 percent) higher than would be expected if no relationship existed between HF incident risk and worker effectiveness level, and may be over twice as high (1.097 or 109.7 percent higher). Therefore it can be said with 95 percent confidence that the HF risk when a worker on duty is at an effectiveness level of 50 or lower is significantly higher than would be expected if risk and effectiveness were not related.

Similarly, in 59 of 2,046 work intervals (0.0288 of all work intervals) that included a nonhuman factors incident, a worker on duty's estimated effectiveness was ≤ 50. The corresponding 95 percent confidence interval for this proportion is $0.0288 \pm 1.96 * \sqrt{[0.0288*(1-0.0288)/2046]}$ or [0.02159, 0.03609]. At the lower limit of the confidence interval the risk of a nonhuman factors incident is 0.0288/0.0307 =0.702, and at the higher limit of the confidence interval, the risk is 0.03609/0.0307 =1.174. Thus, the nonhuman factors risk may be as much as 17.4 percent higher than would be expected by chance, but it may also be as little as nearly 30 percent lower than expected by chance—or it might be the same as that expected by chance. Therefore, it cannot be said with 95 percent confidence that the human factors risk—when a worker on duty is at an effectiveness level of 50 or lower—is significantly higher than would be expected if risk and effectiveness were not related.

Next the proportions of human factors and nonhuman factors incidents below and above an effectiveness level of 50 were compared by analyzing the 2 x 2 contingency Table B-3:

Table B-3. Human Factors and Nonhuman Factors Incidents Distribution Above and Below an Effectiveness Level of 50

	≤ 50 effectiveness	> 50 effectiveness	Total
HF incidents	39 (4.94%)	751 (95.06%)	790
NHF incidents	59 (2.88%)	1,987 (97.12%)	2,046
Total	98 (3.46%)	2,738 (96.54%)	2,836

Note that the proportion of human factors incidents at effectiveness levels of 50 or less is nearly 50 percent higher than that for all incidents combined. The chi-square test for a difference in proportions (Conover (1999), pp. 180-187) was applied to this table. The computed test statistic is 7.201 with 1 degree of freedom, corresponding to a significance level $p = 0.0073$. This indicates that that the type of incident (human factors vs. nonhuman factors) is reliably associated or correlated with the employee effectiveness level at the time of the incident with the rate of occurrence of human factors incidents at lower effectiveness levels significantly[3] higher than would be expected if this rate is unrelated to the effectiveness level (i.e., is the same as the percentage of work intervals where the employee's estimated effectiveness is 50 percent or lower).

[3] Significant in the statistical sense at a confidence level of 95 percent: If random samples of this size are repeatedly drawn from a population wherein the rate of HF incidents that fall within the specified effectiveness range is the same as the fraction of work intervals within that range, then an estimated proportion this much higher would be found by "luck of the draw" in less than 5 percent of the samples (here, in 0.73 percent of the samples, as indicated by the *p*-level).

A more precise means for testing association in a 2 x 2 contingency table is Fisher's exact test (Conover, pp. 187-191), based on the hypergeometric distribution. This test computes the probability of observing the value in the upper left hand cell of the table as

$$p = \frac{\binom{790}{39}\binom{2,046}{59}}{\binom{2,836}{98}} = \frac{\frac{790!}{39!\cdot 751!}\cdot\frac{2,046!}{59!\cdot 1,987!}}{\frac{2,836!}{98!\cdot 2,738!}}$$

which evaluates[4] to $p = 0.003$, confirming the finding of significance from the chi-square test.

A statistically significant difference in and of itself is not necessarily meaningful, noteworthy, or otherwise demanding of attention or action[5]. To establish the extent of the probable difference in proportions between the two types of incidents, it is reasonable to compute a confidence interval of the difference between the proportions in the table. From numerous methods for developing a confidence interval[6] the researchers chose to compute an interval derived from a procedure outlined by Wilson (1927) and incorporating a continuity correction.[7] The observed rate for human factors incidents at effectiveness levels of 50 percent or less is 39/790 = 4.94 percent and that for nonhuman factors incidence is 59/2,046 = 2.88 percent and the 95 percent confidence interval for the difference between the two is [0.45 percent, 3.98 percent]. This suggests that the human factors rate is quite likely to be larger than the nonhuman factors rate by a factor of at least 15 percent (0.45/2.88).

Computations analogous to those shown above were performed for the occurrence rates of human factors and nonhuman factors incidents at effectiveness levels no greater than 60 percent, 70 percent, 80 percent, and 90 percent, and at effectiveness levels greater than 90 percent.

---

[4] Although difficult to evaluate directly for large sample sizes (e.g. $2,836! \approx 2.2 \times 10^{8652}$, a preposterous number), logarithms may be used to simplify the computation; since

$$\log N! = \sum_{i=1}^{N} \log i\,,$$

$$\log p = \sum_{a=1}^{790}\log(a) - \sum_{b=1}^{39}\log(b) - \sum_{c=1}^{751}\log(c) + \sum_{d=1}^{2,046}\log(d) - \sum_{f=1}^{59}\log(f) - \sum_{g=1}^{1,987}\log(g) - \left[\sum_{h=1}^{2,836}\log(h) - \sum_{i=1}^{98}\log(i) - \sum_{j=1}^{2,738}\log(j)\right]$$

and $p = 10^{\log p}$. In the current instance, $\log p = -2.536$ and $p \approx 10^{-2.536} \approx 0.003$.

[5] This is discussed entertainingly by Abelson (1995).

[6] No less than eleven such methods are described in Newcombe (1998).

[7] Conveniently implemented by Dr. Richard Lowry on the VassarStats website in an interactive Web-based calculator at http://faculty.vassar.edu/lwry/prop2_ind.html .

Appendix C.
Results for Individual Railroads

The five participating railroads are code names Aspen, Beech, Cottonwood, Dogwood, and Elm. The next three sections provided individual railroad results for the circadian variations in human factors and nonhuman factors accident risk, section C-1, the charts of distributions of work time by effectiveness categories, section C-2, human factors accident risk, section C-3, and nonhuman factors accident risk, C-4.

C-1: Circadian variations in human factors and nonhuman factors accident risk.

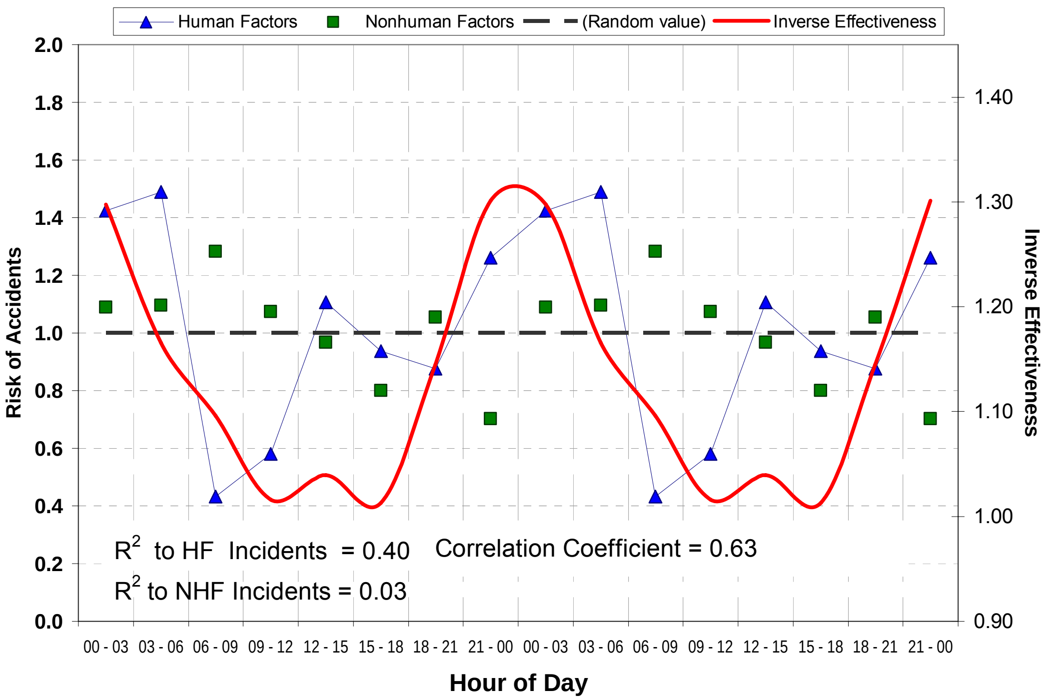

Figure 12: Accident Risk by Time of Day from Aspen Railroad. Data Have Been Double-Plotted to Show the Repeating Circadian Pattern

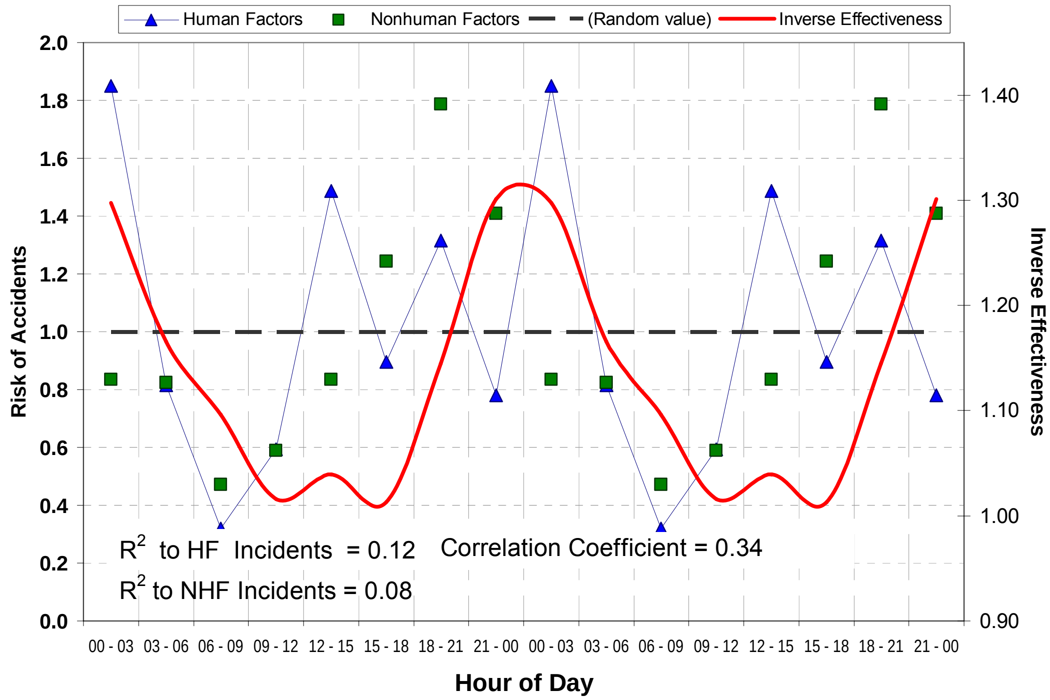

Figure 13: Accident Risk by Time of Day, Beech Railroad. Data Have Been Double-Plotted to Show the Repeating Circadian Pattern

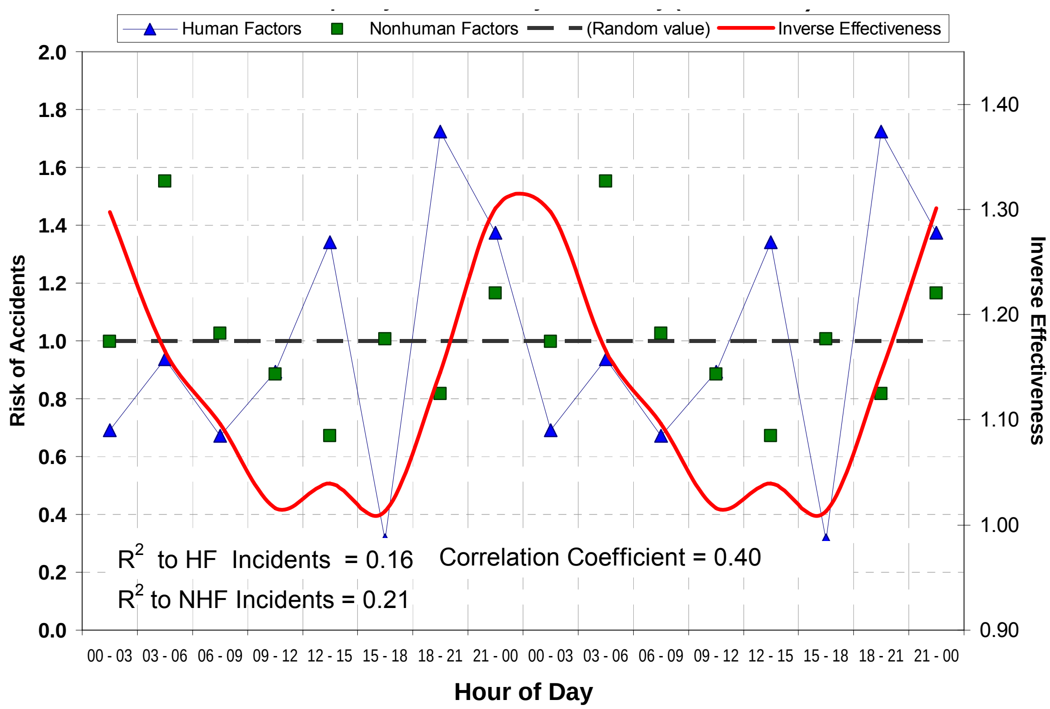

Figure 14: Accident Risk by Time of Day, Cottonwood Railroad. Data Have Been Double-Plotted to Show the Repeating Circadian Pattern

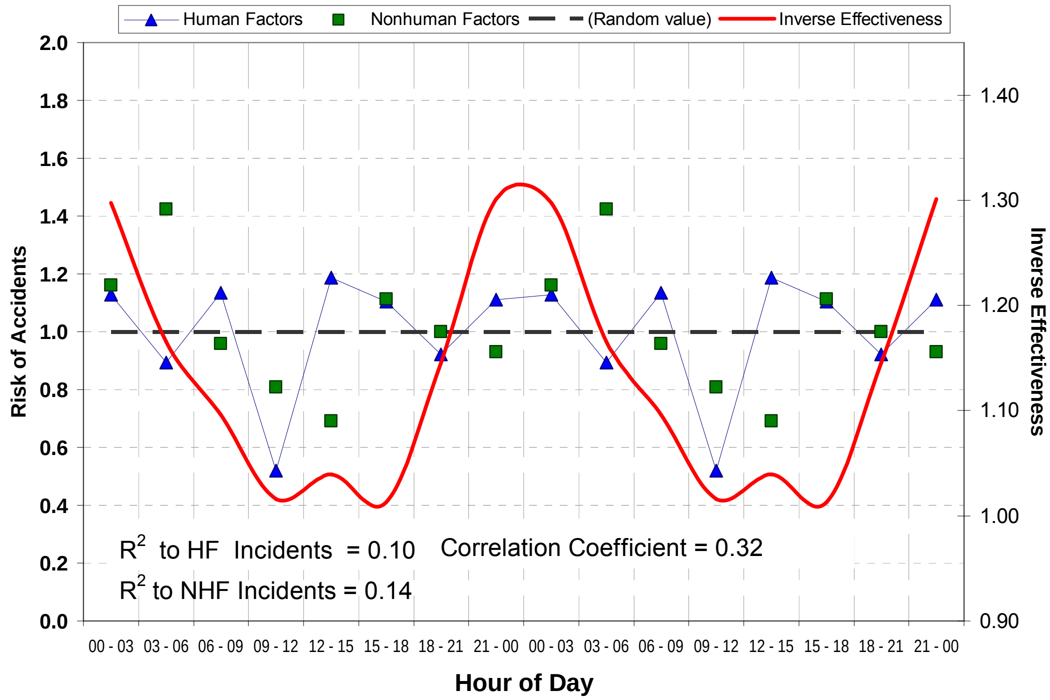

Figure 15: Accident Risk by Time of Day, Dogwood Railroad. Data Have Been Double-Plotted to Show the Repeating Circadian Pattern

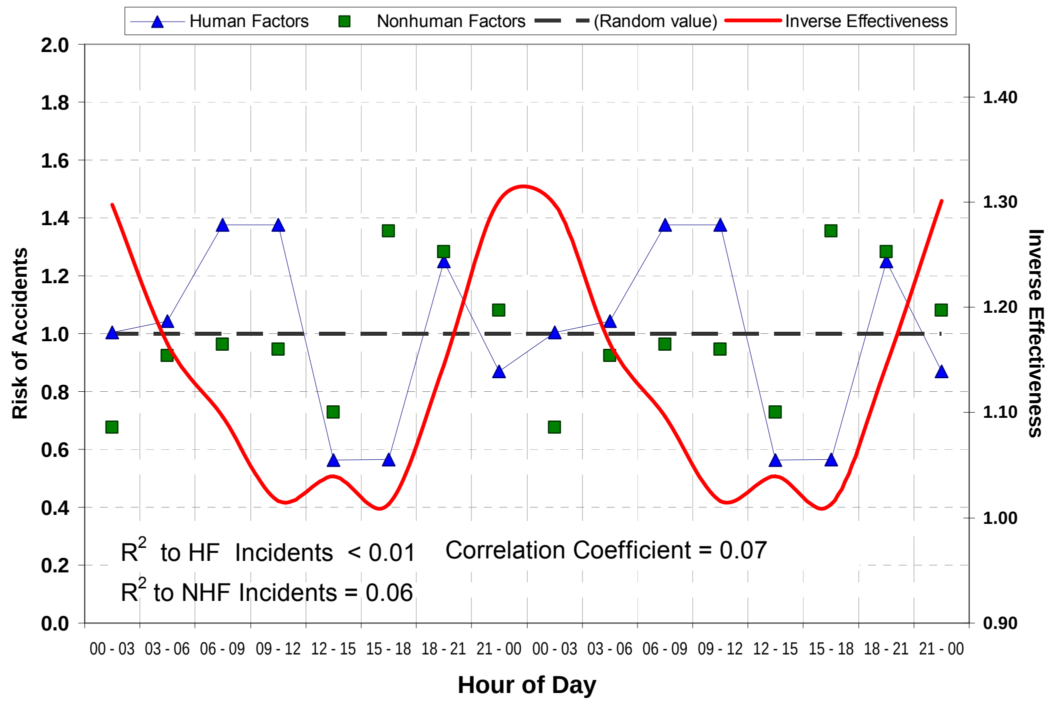

Figure 16: Accident Risk by Time of Day, Elm Railroad. Data Have Been Double-Plotted to Show the Repeating Circadian Pattern

C-2: Distributions of work time and accidents by effectiveness categories.

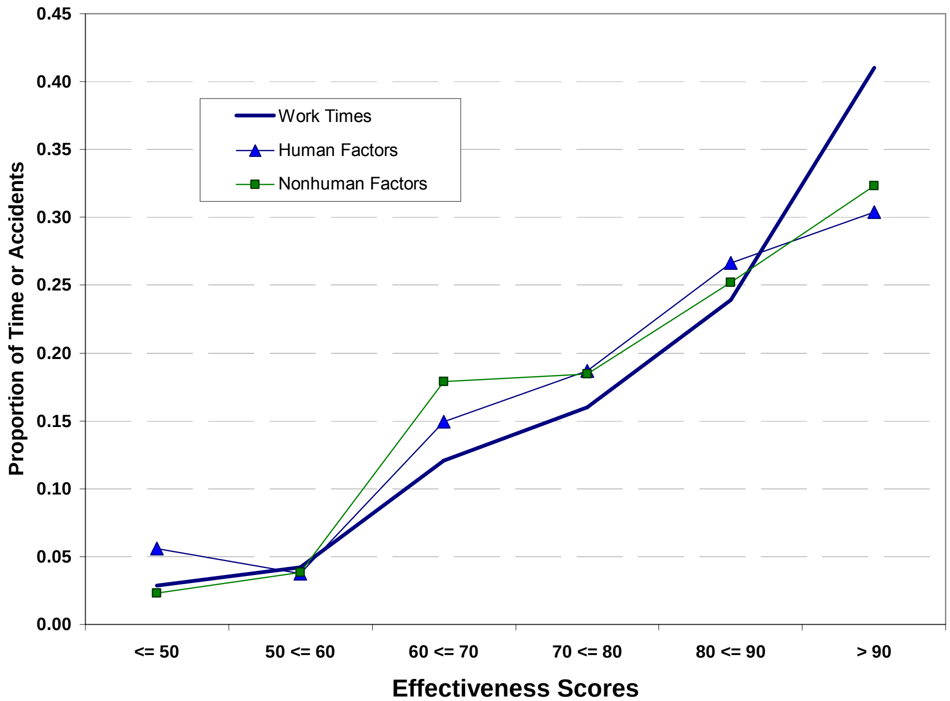

Figure 17: Proportion of Time or Accidents as a Function of Predicted Effectiveness Categories, Aspen Railroad. Human Factors and Nonhuman Factors Accidents are Indicated as Blue Triangles and Green Squares, respectively

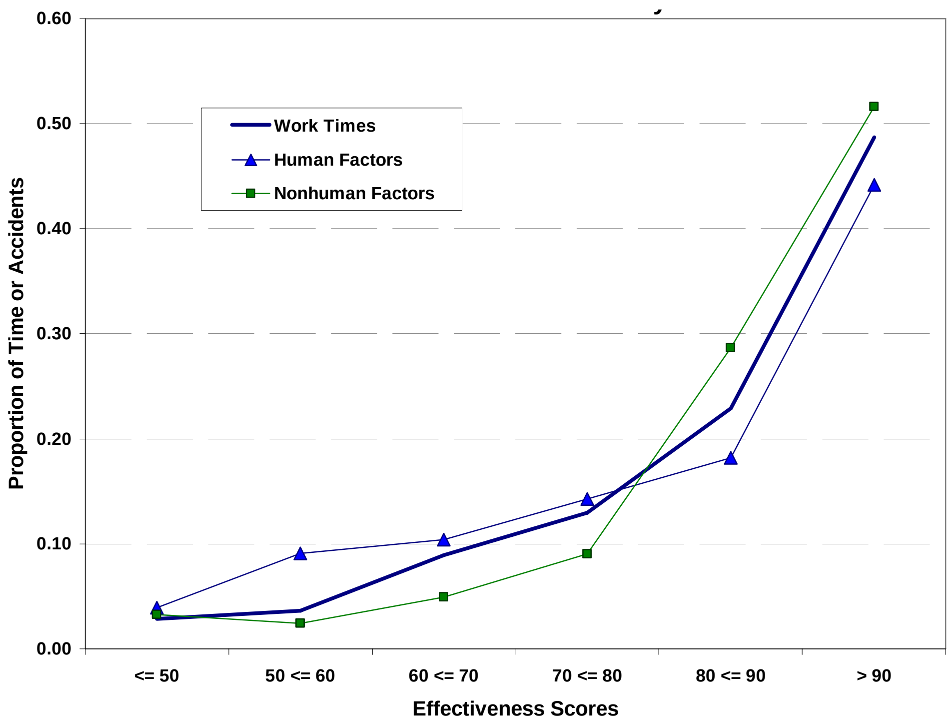

Figure 18: Proportion of Time or Accidents as a Function of Predicted Effectiveness Categories, Beech Railroad. Human Factors and Nonhuman Factors Accidents are Indicated as Blue Triangles and Green Squares, respectively

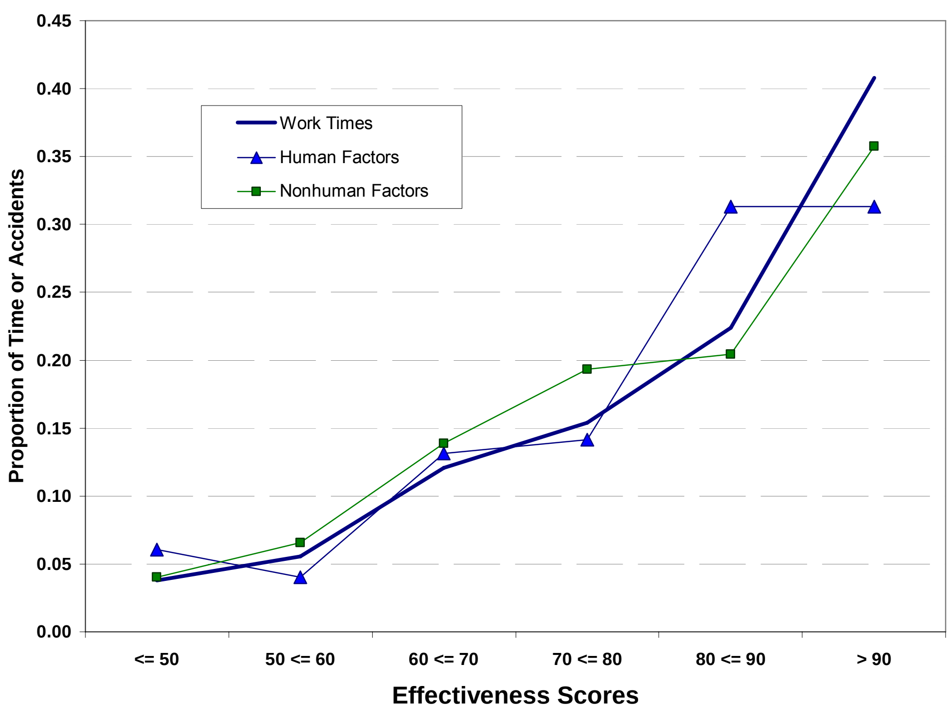

Figure 19: Proportion of Time or Accidents as a Function of Predicted Effectiveness Categories, Cottonwood Railroad. Human Factors and Nonhuman Factors Accidents are Indicated as Blue Triangles and Green Squares, respectively

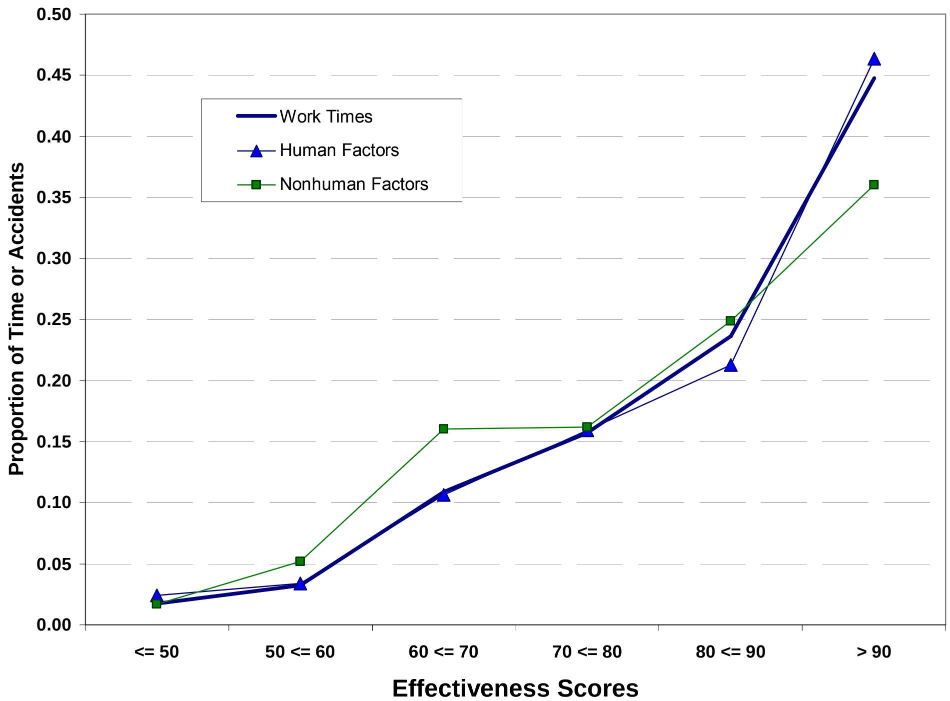

Figure 20: Proportion of Time or Accidents as a Function of Predicted Effectiveness Categories, Dogwood Railroad. Human Factors and Nonhuman Factors Accidents are Indicated as Blue Triangles and Green Squares, respectively

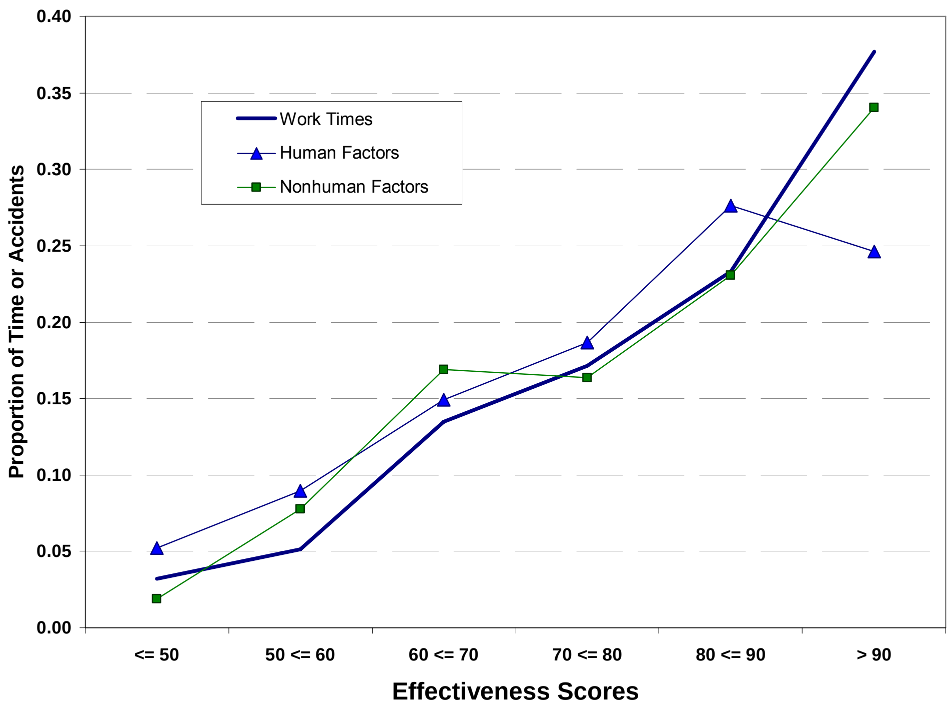

Figure 21: Proportion of Time or Accidents as a Function of Predicted Effectiveness Categories, Elm Railroad. Human Factors and Nonhuman Factors Accidents are Indicated as Blue Triangles and Green Squares, respectively

C-3: Human factors accident risk by effectiveness at the time of the accident.

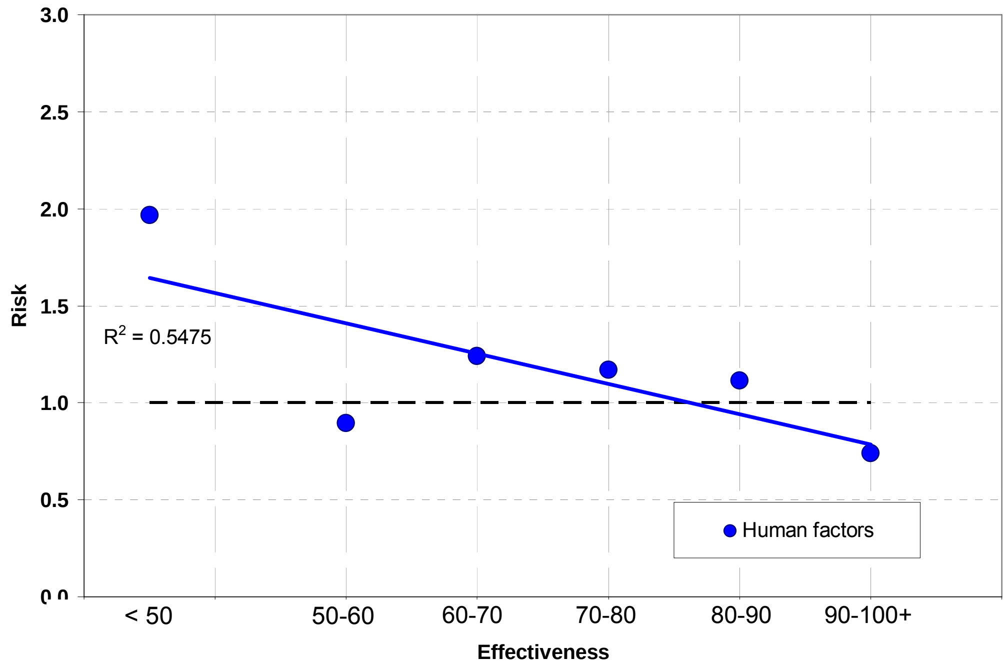

Figure 22: Human Factors Accident Risk by Criterion Levels of Effectiveness Aggregated for Aspen Railroad

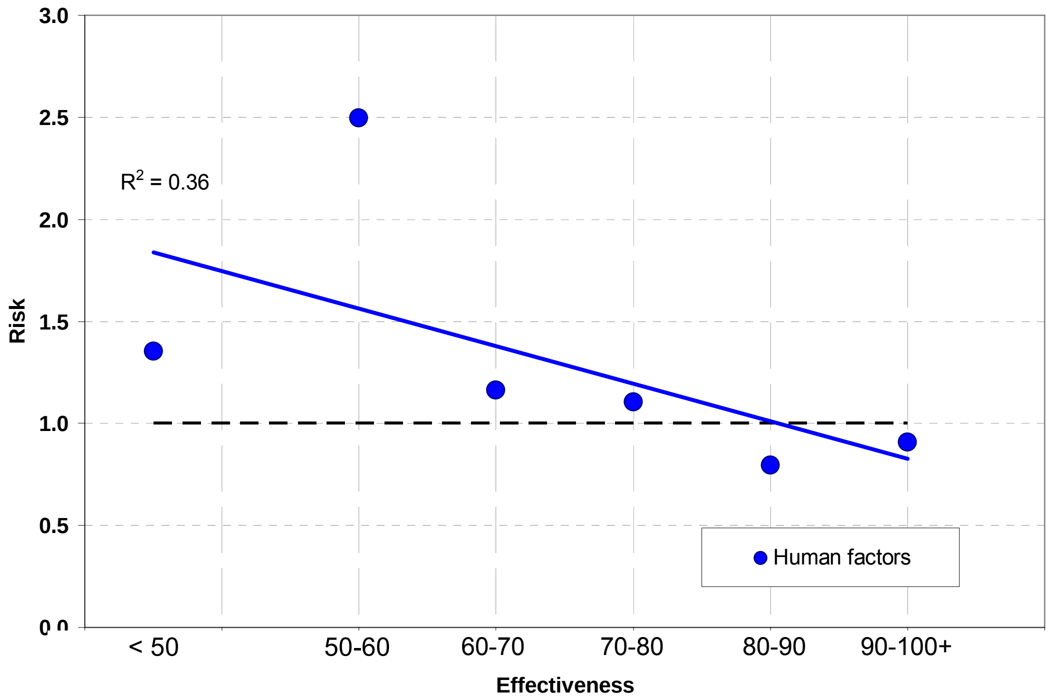

Figure 23: Human Factors Accident Risk by Criterion Levels of Effectiveness Aggregated for Beech Railroad

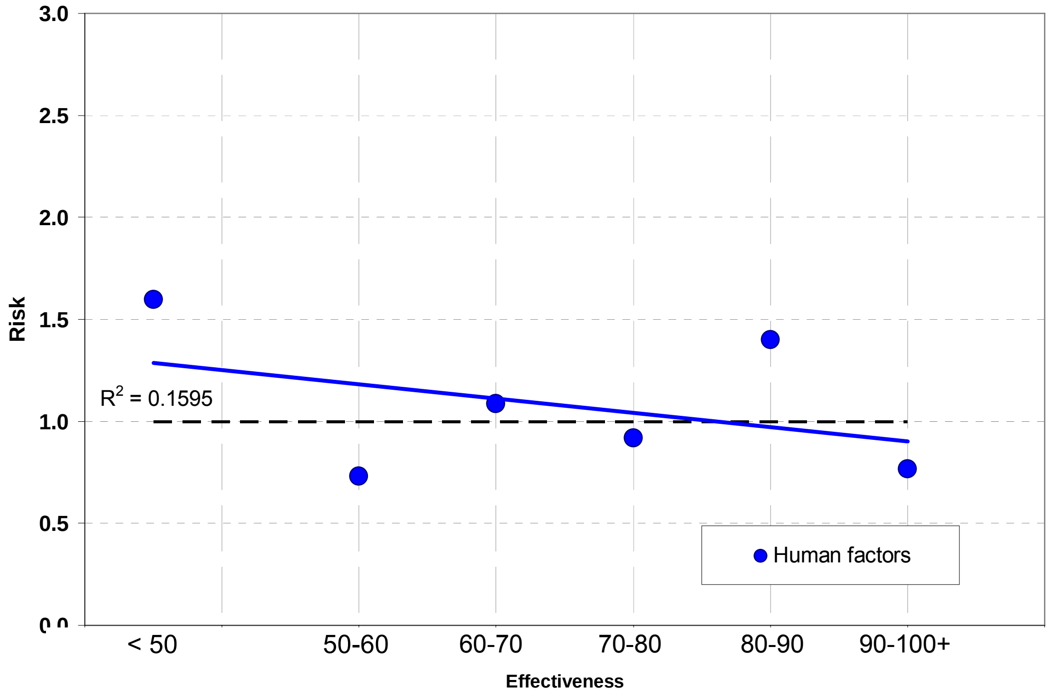

Figure 24: Human Factors Accident Risk by Criterion Levels of Effectiveness Aggregated for Cottonwood Railroad

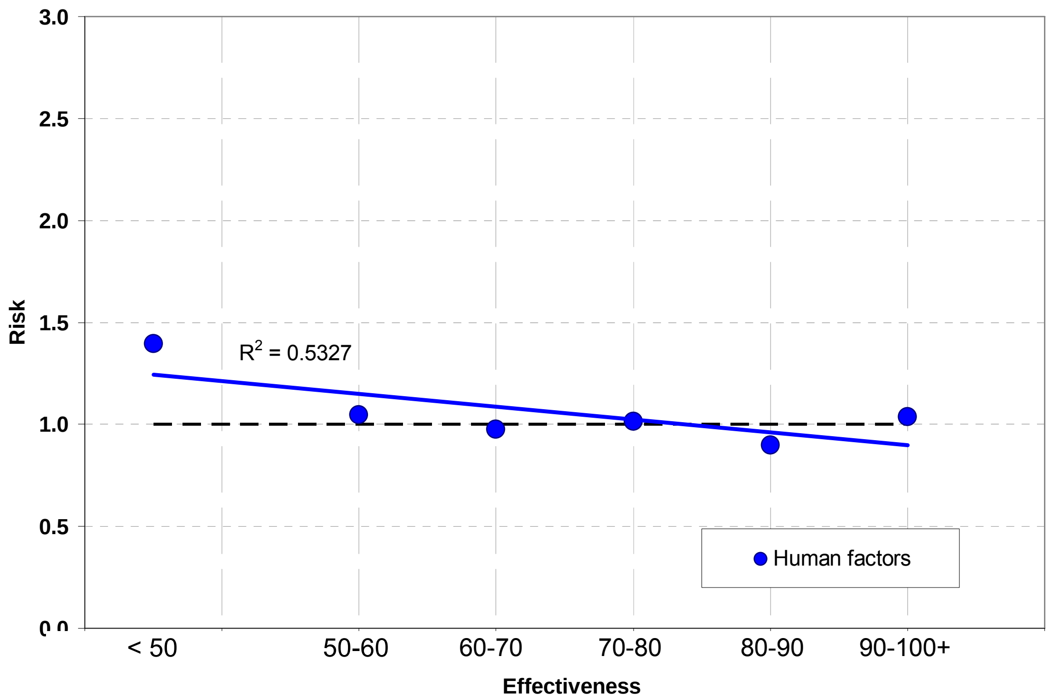

Figure 25: Human Factors Accident Risk by Criterion Levels of Effectiveness Aggregated for Dogwood Railroad

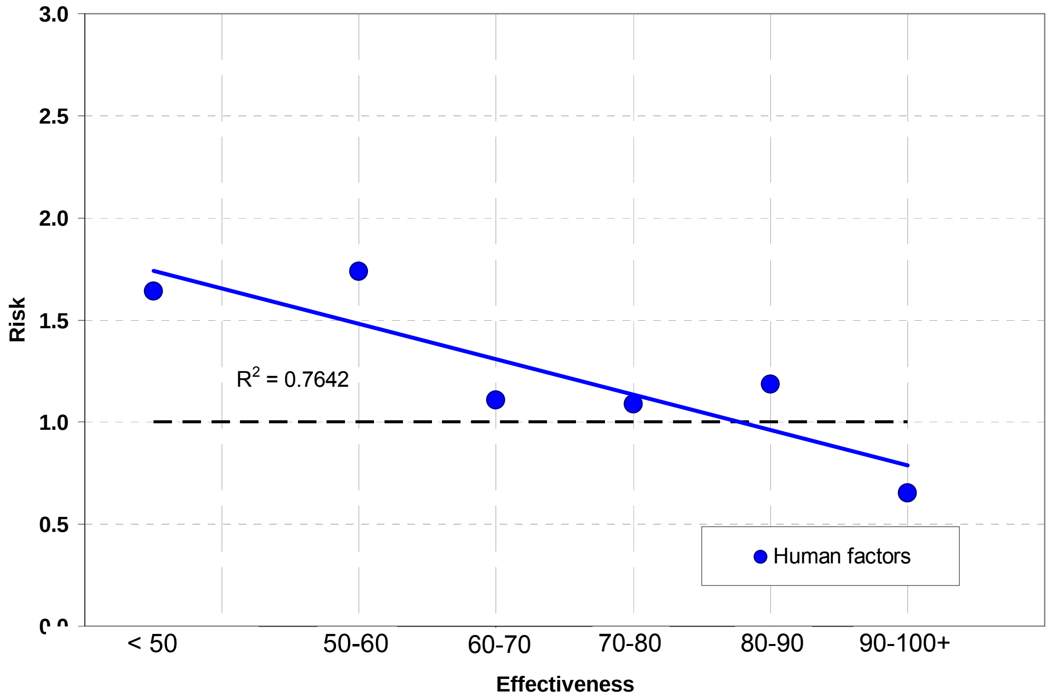

Figure 26: Human Factors Accident Risk by Criterion Levels of Effectiveness Aggregated for Elm Railroad

C-4: Nonhuman factors accident risk by effectiveness at the time of the accident.

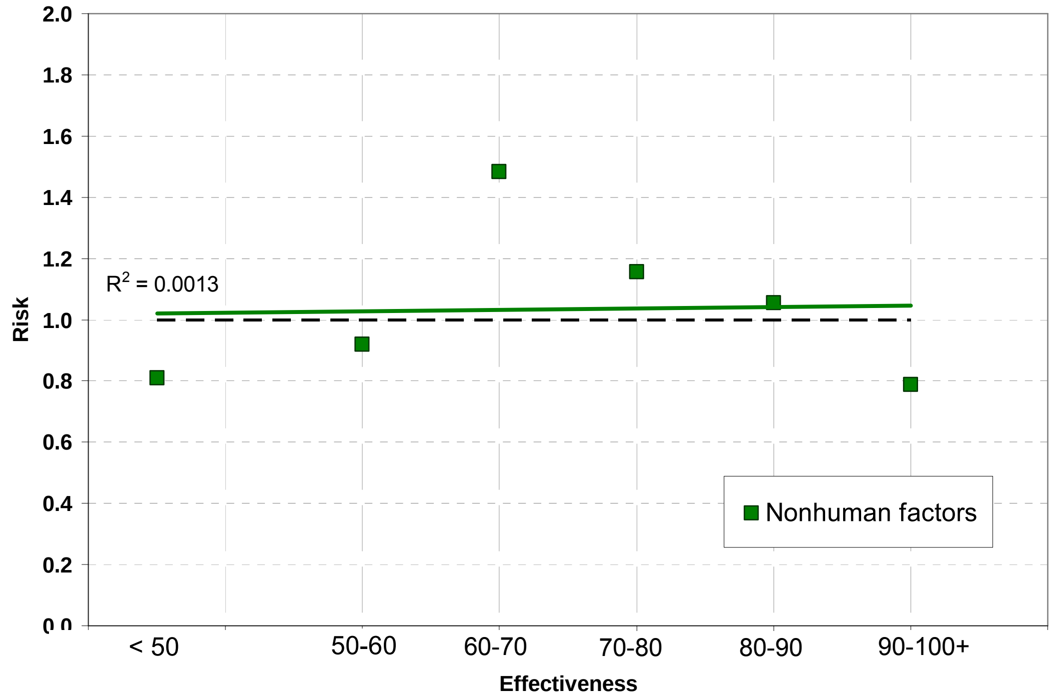

Figure 27: Nonhuman Factors Accident Risk at Each Level of Effectiveness Aggregated from Aspen Railroad

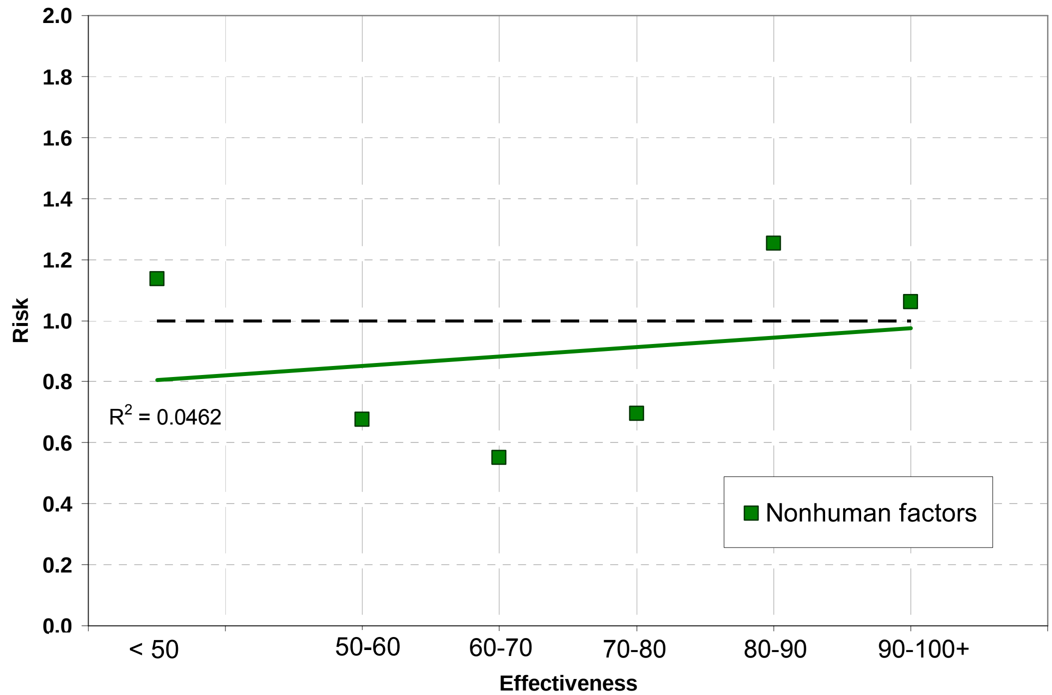

Figure 28: Nonhuman Factors Accident Risk at Each Level of Effectiveness Aggregated from Beech Railroad

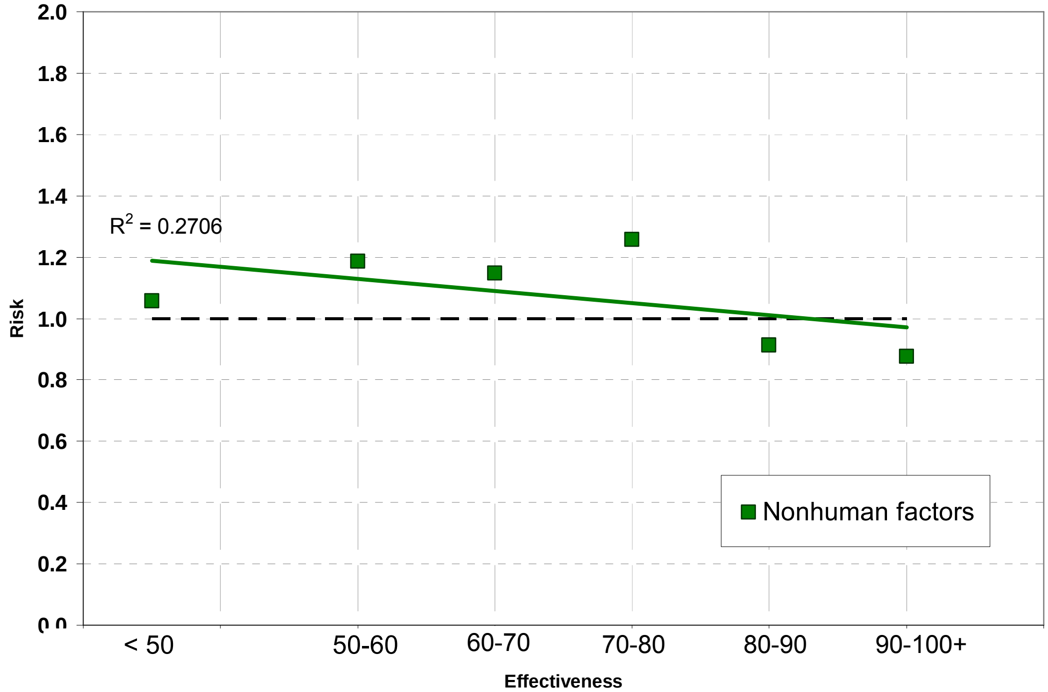

Figure 29: Nonhuman Factors Accident Risk at Each Level of Effectiveness Aggregated from Cottonwood Railroad

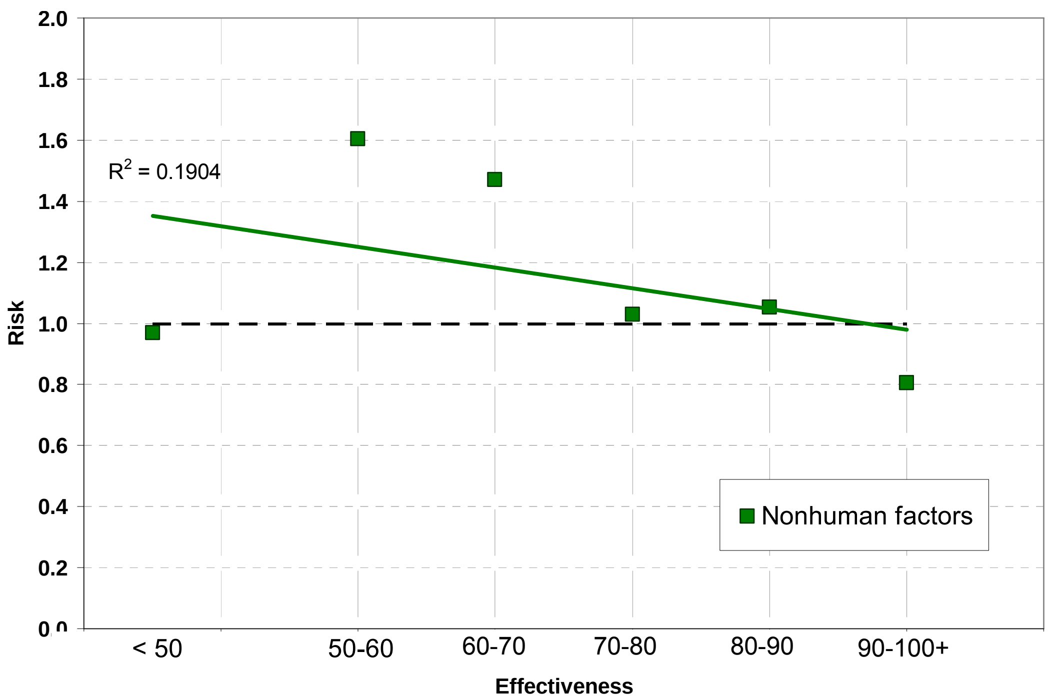

Figure 30: Nonhuman Factors Accident Risk at Each Level of Effectiveness Aggregated from Dogwood Railroad

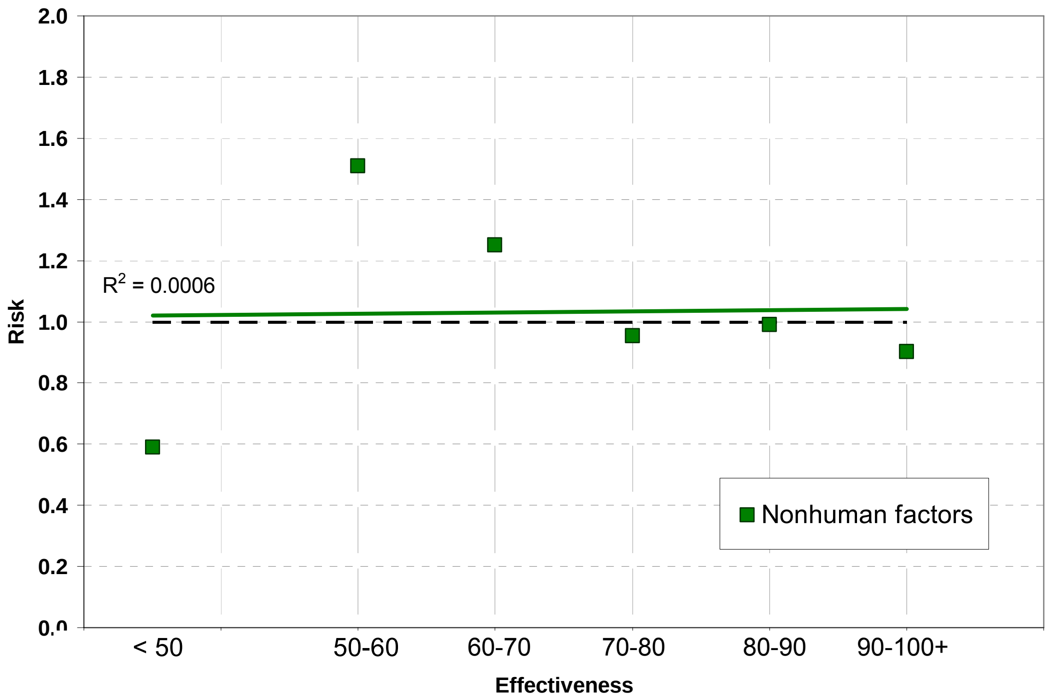

Figure 31: Nonhuman Factors Accident Risk at Each Level of Effectiveness Aggregated from Elm Railroad