Fatigue Models as Practical Tools: Diagnostic Accuracy and Decision Thresholds

THOMAS G. RASLEAR AND MICHAEL COPLEN

RASLEAR TG, COPLEN M. *Fatigue models as practical tools: diagnostic accuracy and decision thresholds.* **Aviat Space Environ Med 2004; 75(3, Suppl.):A168–72.**

Human fatigue models are increasingly being used in a variety of industrial settings, both civilian and military. Current uses include education, awareness, and analysis of individual or group work schedules. Perhaps the ultimate and potentially most beneficial use of human fatigue models is to diagnose if an individual is sufficiently rested to perform a period of duty safely or effectively. When used in this way, two important questions should be asked: 1) What is the accuracy of the diagnosis for duty-specific performance in this application; and 2) What decision threshold is appropriate for this application (i.e., how "fatigued" does an individual have to be to be considered "not safe"). In the simplest situation, a diagnostic fatigue test must distinguish between two states: "fatigued" and "not fatigued," and the diagnostic decisions are "safe" (or "effective") and "not safe" (or "not effective"). The resulting four decision outcomes include diagnostic errors because diagnostic tests are not perfectly accurate. Moreover, since all outcomes have costs and benefits associated with them that differ between applications, the choice of a decision criterion is extremely important. Signal Detection Theory (SDT) has demonstrated usefulness in measuring the accuracy of diagnostic tests and optimizing diagnostic decisions. This paper describes how SDT can be applied to foster the development of fatigue models as practical diagnostic and decision-making tools. By clarifying the difference between accuracy (or sensitivity) and decision criterion (or bias) in the use of fatigue models as diagnostic and decision-making tools, the SDT framework focuses on such critical issues as duty-specific performance, variability (model and performance), and model sensitivity, efficacy, and utility. As fatigue models become increasingly used in a variety of different applications, it is important that end-users understand the interplay of these factors for their particular application.

Keywords: fatigue models, diagnostic accuracy, decision threshold, signal detection theory, decision theory, risk management.

THE FATIGUE AND Performance Modeling Workshop that was held in Seattle, WA in June 2002 identified three major goals for the workshop: 1) assess the state-of-the-art of biomathematical models of fatigue, sleepiness, and performance; 2) identify conceptual and technological barriers to these models; and 3) identify and communicate research needs in military and civilian applications. Although the Workshop was an overwhelming success in meeting goal 1, little was accomplished with regard to the remaining two goals. This paper addresses those goals through a risk management approach in commercial transportation. Any model can have numerous uses (and misuses) that depend on the needs of the end-user. The Workshop participants were primarily scientists who, as end-users of fatigue models, have very different uses of the models than commercial operators. Where scientists may see utility in a model that is a predictive research tool, commercial operators may see utility in a model that helps them manage operational risks. Unless such differences are clarified and understood, fatigue models are apt to be misused in civilian settings (a conceptual barrier to models under goal 2). Consequently, civilian research needs will not be identified and communicated (see goal 3).

Although fatigue models may be used for a variety of purposes (education, awareness, analysis of individual or group work schedules), perhaps the ultimate application of any fatigue model may be to diagnose if an individual is sufficiently rested to perform a period of duty safely or effectively. The Fatigue and Performance Modeling Workshop discussed the predictive ability and accuracy of the models relative to subjective reports of fatigue or sleepiness and/or neurobehavioral tests of performance. While this approach is familiar and useful to scientists, it begs the questions that end-users interested in practical risk management will ask: 1) What is the accuracy of the model for performance in my application (how well does it distinguish between performance degraded by fatigue from performance that is not degraded); and 2) What decision threshold is appropriate for my application (i.e., what fatigue "score" indicates that performance is degraded to an unsafe level). For instance, industry managers will want to know about model accuracy because they and their company can be sued if an employee has an accident. If the decision to allow that employee to work was based in part on a fatigue model, the accuracy of the model is important legal evidence that due care was taken (or not taken) by the manager and company. Managers, labor representatives, and employees will also want to know about model accuracy because it affects employees' ability to work and be safe. Accuracy will indicate the extent to which a fatigue model provides a fair, impartial, and objective assessment of fitness to work and enhances employee safety. Managers do not want to disrupt operations because a model with low accuracy falsely indicates that a majority of the night shift is unfit for work. Labor, similarly, does not want to lose income unnecessarily.

From the Office of Research and Development, Federal Railroad Administration, Washington, DC.

Address reprint requests to: Thomas G. Raslear, who is a Senior Human Factors Program Manager at the Federal Railroad Administration, Mail Stop 20, 1120 Vermont Ave NW, Washington, DC 20590; Thomas.Raslear@fra.dot.gov.

TABLE I. OUTCOMES FOR DIAGNOSES GIVEN THE FATIGUE STATE.

STATE	DIAGNOSIS	
	UNSAFE	*SAFE*
FATIGUED	True Positive (TP)	False Negative (FN)
NOT FATIGUED	False Positive (FP)	True Negative (TN)

Management and labor will also want to know what fatigue score indicates that an employee is too fatigued to work (i.e., question 2 above), given the accuracy of the model. If the criterion is too stringent, many employees will be considered "fatigued" when they are actually fit to work. But if the criterion is too lenient, the opposite will happen. Again, it is important for practical reasons (I may get sued, labor may strike, I may unjustly lose income, etc.) to decide how much fatigue is unsafe on a rational basis. For instance, it may be desirable to use different criteria for different operations. Employees who work night shifts may be known to be more fatigued than day shift employees, so a more stringent criterion may seem reasonable for the night shift. An operation involving the transportation of hazardous materials through a highly populated area might also seem to warrant a more stringent criterion because of the catastrophic consequences of an accident. But how does one set the decision threshold under these varying circumstances in a systematic, consistent, and rational way that will support risk management? Decision theory (6) provides a variety of methods to achieve this goal. Among these methods, Signal Detection Theory (SDT) is used in this paper to illustrate how end-users who are concerned with risk management can use fatigue models to diagnose fatigue in a flexible, defensible, and rational way.

Risk management, by definition, considers the probability of various decision outcomes and their associated benefits and costs (3). It will be shown below that, in the case of fatigue models, the probabilities of decision outcomes are jointly determined by model accuracy and the decision threshold. The decision threshold is dependent on benefits, costs, and the probability of various states of the world (e.g., "fatigued" vs. "not fatigued"). Benefits from the use of fatigue models may include reducing accidents, increasing operational efficiency, improving employee morale, and improved scheduling. Costs may include labor disputes, increased labor costs, and disruptions in service. Different groups within an industry (e.g., management, labor representatives, employees) will have different uses for fatigue models (e.g., risk assessment, collective bargaining, policy) and different benefits and costs associated with those uses. Unless end-users understand the utility of fatigue models for them, it is unlikely that they will use the models. Worse, they may inappropriately use the models to set important policy or other related decision outcomes. For example, it may be quite appropriate for a particular fatigue model to be used as a decision tool for ranking the relative risk of various work schedules, but quite inappropriate to use the output of that same model to support an absolute fitness for duty decision criteria; that is, whether or not an individual should be allowed to work, given their current state of fatigue. The following discussion of model accuracy and decision threshold is intended to illustrate the potential utility of any fatigue model for a variety of end-users. The methods described can be applied to any fatigue model that produces a quantitative output.

In the simplest situation, a diagnostic fatigue test must distinguish only between two states: fatigued and not fatigued, and the diagnostic outcomes are safe (effective) and not safe (not effective). **Table I** illustrates the situation.

The models are not perfectly accurate, so diagnostic errors (false positives, FPs; and false negatives, FNs) are expected. Moreover, diagnostic fatigue values vary between and within individuals under the same circumstances so that distributions of diagnostic fatigue values for the "fatigued" and "not fatigued" states overlap (**Fig. 1**). Consequently, true positives (TPs) and FPs covary with varying diagnostic "thresholds" or criteria (i.e., how "fatigued" do you have to be to be considered "unsafe"). The outcomes all have costs and/or benefits associated with them, so the choice of a criterion is extremely important. This is the type of situation in which SDT has been extremely useful (2,5), and it is suggested that an SDT analysis can provide the practitioner with an analytic framework to determine the accuracy of a model for a specific application and for determining the most appropriate decision criteria. It should be noted that although this discussion will assume normal distributions with equal variance for ease of exposition, this assumption is not critical to the use of SDT (5).

Diagnostic Accuracy

As a practical matter, the managers of railroads, other transportation companies, and the military are con-

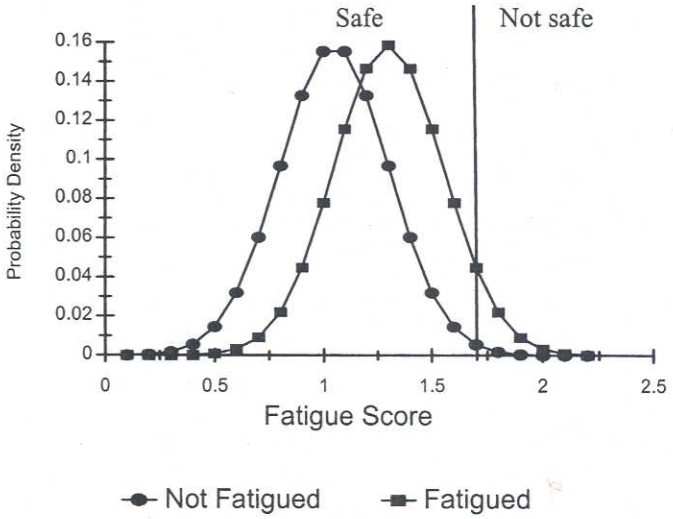

Fig. 1. Hypothetical distributions of fatigue scores for fatigued and non-fatigued individuals. The vertical line at a fatigue score of 1.7 is the criterion score above which individuals are considered to be unsafe due to fatigue. The criterion score divides the fatigued distribution into true positives (to the right of the criterion) and false negatives (to the left of the criterion), and divides the non-fatigued distribution into false positives (right of the criterion) and true negatives (left of the criterion).

cerned with the safe and efficient performance of specific tasks. Subjective feelings of fatigue or performance on neurobehavioral tasks do not directly indicate whether specific workplace tasks will be performed safely or efficiently. This leap from the laboratory to the workplace is not trivial. Many jobs are a mixture of cognitive and physical tasks, and workplaces vary considerably with regard to physical characteristics that may be conducive to alertness. Consequently, it is important for the users of fatigue models to know the accuracy of a model in detecting fatigue-induced changes in performance in a specific workplace and job. The ability of a model to detect fatigue may also depend on the subject population (old vs. young, male vs. female, etc.), work history (e.g., experienced vs. novice), and a host of other factors that may limit the generalization of results from one application to others.

In SDT, accuracy is indexed by sensitivity. In general terms, sensitivity is the difference between the means of the underlying distributions expressed in standard deviation units. For normal distributions, d' is the sensitivity index, but indices have been devised for other distributions (1) and for situations where the form of the distribution is not known (5). Sensitivity is independent of the decision criterion, and the outcome of a decision depends on both sensitivity and the criterion.

The determination of model sensitivity will depend on the goals of different users of the model. For example, a safety officer in a transportation company might want to reduce accidents caused by fatigue. An operations officer, on the other hand, might want to improve operational efficiency (e.g., delivery time, or fuel use). These different goals dictate the use of different performance measures and will result in different model sensitivities.

As an example, the safety officer who wants to reduce fatigue-caused accidents might sort accidents into two categories: 1) those for which fatigue can be entirely ruled out because mechanical or equipment failures are the primary causes of the accident; and 2) those for which fatigue is highly suspect because the accident did not involve a mechanical or equipment failure, but a memory lapse or inattention were indicated. These two categories of accidents can be labeled, per Table I, as "not fatigued" and "fatigued," respectively. In the railroad industry, for example, a rail that breaks under a car in the middle of the train and causes a derailment is not likely due to human fatigue. On the other hand, an accident in which a speed restriction or signal has not been obeyed are likely due to attention or memory problems which can be caused by human fatigue.

Given a reasonable set of "fatigue" and "non-fatigue" accidents, the work-rest records of involved operating personnel can be collected and analyzed with the model. It is expected that a range of fatigue scores will be obtained for operators in both sets of accidents. However, the mean score for "non-fatigue" accidents should be lower than the mean score for "fatigue" accidents. With means and standard deviations for both sets of accidents, the value of d' can be directly determined from

TABLE II. COVARIATION OF P(TP), P(FP), AND β AS A FUNCTION OF THE CRITERION FATIGUE SCORE FOR D' = 1 (SEE FIG. 1).

Criterion Fatigue Score	β	p(TP)	p(FP)	C_{FN}	p(F)
1.0	0.59	0.84	0.5	−$740,000	0.16
1.2	1.24	0.58	0.21	−$240,000	0.16
1.4	2.22	0.27	0.055	−$740,000	0.05
1.6	4.64	0.081	0.0082	−$240,000	0.05

$$d' = \frac{\mu_F - \mu_{NF}}{\sigma} \qquad \text{Eq. 1}$$

where μ_F is the mean fatigue score of the "fatigue" accident operators, μ_{NF} is the mean fatigue score of the "non-fatigue" accidents, and σ is the common standard deviation. For example, in Fig. 1, $\mu_F = 1.25$, $\mu_{NF} = 1.00$, and $\sigma = 0.25$, so that d' = 1.00. In general, the higher the d' value, the greater the separation of the distributions in Fig. 1, and the greater the diagnostic accuracy of a test.

Decision Criteria

A criterion for deciding what fatigue score indicates that performance is sufficiently degraded (to an unsafe state in the case of the safety officer example) can be set in a variety of ways depending on the decision goals of the decisionmaker. Several common decision goals include maximizing expected value, maximizing percent correct decisions, and satisfying the Neyman-Pearson objective.* Different model users cannot only have different model sensitivities based on different performance measures, but also different decision goals.

As an example, in Fig. 1, the criterion is the vertical line at a fatigue score of 1.7. Fatigue scores of 1.7 or higher would be considered "not safe", as indicated in the figure. Although sensitivity (d') does not change as the criterion is changed, the setting of the criterion does determine the probability of TPs and FPs [p(TP) and p(FP)]. **Table II** shows how p(TP) and p(FP) vary with the criterion for constant d'. Note that a change in d' (separation of the distributions) would, independently of the criterion, also change p(TP) and p(FP).

The criterion fatigue score is determined by the SDT parameter β. β is set according to decision goals and can be calculated from

$$\beta = \frac{p(NF)}{p(F)} \times \frac{(B_{TN} - C_{FP})}{(B_{TP} - C_{FN})}, \qquad \text{Eq. 2}$$

where p(NF) is the probability that individuals in the population under consideration are not fatigued, p(F) is the probability that individuals in the population under consideration are fatigued, B_{TN} and B_{TP} are the benefits of correct decisions, and C_{FP} and C_{FN} are the costs of incorrect decisions. If costs and benefits are all equal and the prior probabilities [p(F) and p(NF)] are also equal, then $\beta = 1$, which indicates an absence of diag-

* This should be familiar to those who have performed statistical tests. The objective is to hold p(FP) at some fixed level (e.g., 0.05) while maximizing p(TP).

TABLE III. PAYOFF MATRIX, $\beta = 0.59$.

STATE	DIAGNOSIS	
	UNSAFE	*SAFE*
FATIGUED	B_{TP} = \$220,000	C_{FN} = −\$740,000
NOT FATIGUED	C_{FP} = −\$100,000	B_{TN} = \$10,000

nostic bias. Values of $\beta < 1$ indicate a bias to diagnose unsafe levels of fatigue, while values of $\beta > 1$ indicate a bias to diagnose safe levels of fatigue. In Table II, one criterion score was set with a bias to diagnose unsafe levels of fatigue. A bias to diagnose unsafe levels of fatigue corresponds with higher values of p(TP) and p(FP). In Fig. 1, this means that the criterion score moves to the left.

According to Eq. 2, an optimal decision requires information about the probability of fatigue in the population under consideration and information about the benefits and costs associated with the four decision outcomes in Table I. In essence, this is an assessment of the relative risk of a binary decision (unsafe vs. safe), where risk (3) is defined as the product of the probability of an event (fatigued vs. not fatigued) and its outcome (e.g., an accident). Associating costs and benefits with the cells in Table I results in a "payoff matrix," and this device is often used to explicitly summarize the outcome of a cost-benefit analysis. As an example, **Table III** shows the complete payoff matrix for the first row in Table II, $\beta = 0.59$. It was assumed that the end user of the model was a railroad safety officer whose goal was to reduce accidents in a work population consisting of extra board (on-call) locomotive engineers.

In Table III, the value of B_{TP} is set at \$220,000. This includes a \$200,000 benefit associated with avoiding a fatigue-caused accident, a cost of \$20,000 for a labor dispute based on the decision to declare an employee "unsafe" due to fatigue, and \$40,000 in miscellaneous benefits such as reduced health costs and more efficient operations. The value of C_{FP} is the cost of a labor dispute caused by falsely declaring an employee "unsafe" due to fatigue when they are actually not fatigued. The value of C_{FN} includes a \$200,000 cost of an accident, \$40,000 in costs associated with less efficient operations and increased health costs, and a cost of \$500,000 due to damage to the company's business reputation as a result of an accident and negative press coverage. The value of B_{TN} is the benefit accruing to improvement in employee morale. It was also assumed that an individual in this population has a probability of 0.16 of being fatigued [p(F)], as estimated by the Pollard report (4).

In Table II, for simplicity, the remaining values of β were obtained by either changing the value of p(F) and/or C_{FN}. In practice, all of the costs and benefits could have been changed. Thus, for $\beta = 1.2$, C_{FN} has a reduced cost because the \$500,000 damage to the company's business reputation has been removed. All other cells of the payoff matrix and p(F) remain the same. For $\beta = 2.22$, the payoff matrix of Table III remains unchanged, but p(F) = 0.05 because of changes in company policy to allow engineers to refuse a work assignment when they feel fatigued. Finally, for $\beta = 4.64$, both p(F) and C_{FN} have been changed as indicated in Table II. Clearly, changes in the costs and benefits associated with decision outcomes and p(F) affect bias, the setting of the criterion score, and the probabilities of correct and erroneous diagnoses.

Discussion

The framework of SDT supports the ability of any quantitative fatigue model to address the goals of various model users (reduce accidents, improve efficiency, etc.) and to make optimal decisions based on those goals. The SDT framework makes explicit the connection between diagnostic accuracy and the goals of the model user. This is important because it encourages the careful and appropriate application of models to specific situations. The sensitivity of a model applied to detect unsafe levels of fatigue in combat pilots may not be the same as the sensitivity of the same model to detect unsafe levels of fatigue in bus drivers. Even if the sensitivities are the same, the optimal criterion score may not be the same, because the outcomes of the "safe" vs. "unsafe" decision may be vastly different. A general might risk the loss of an aircraft due to a fatigue-caused accident if the cost of not allowing fatigued pilots to fly was defeat in battle. At the same fatigue score, however, the manager of a bus company might consider the crash of a bus a higher cost than the loss of revenue from a cancelled bus route.

Decisionmakers often have unformulated assumptions concerning the costs and benefits of decision outcomes. SDT encourages the critical examination of those assumptions to determine if they best serve the goal of the decisionmaker. For instance, one decision goal is to maximize percent correct decisions, and many decisionmakers automatically opt for this outcome without realizing that this sets the criterion at the intersection of the two distributions in Fig. 1 (i.e., $\beta = 1$). This then entails the assumption that p(F) = p(FN) = 0.5, and that costs and benefits are all equal. These assumptions are often not true, and the resulting decisions are not optimal.

Tables II and III demonstrate how changes in estimates of the probability that an employee will be fatigued and changes in estimated costs and benefits associated with indicating that an employee is unsafe because of fatigue combine in Eq. 2 to affect the criterion fatigue score. The considerations described here are, however, not exhaustive. The setting of the criterion fatigue score can be affected by many factors that influence the elements of Eq. 2. An optimal setting for the criterion can always be obtained in this way, and this ensures the usefulness of the decision. For some work settings, estimates of fatigue probability already exist. Such estimates need to be improved and tailored to the specific population under consideration. In the absence of good estimates, Swets (5) suggests that a range of values be examined to determine the effect on TP and FP. Costs and benefits, likewise, can be directly estimated, or a range of cost-benefit ratios can be examined.

The use of any fatigue model can, and should, be as varied as the specific needs and characteristics of the users. The value of the SDT framework is that it explicitly addresses the issues of sensitivity and bias for dif-

ferent users. By doing so, SDT forces model users to define their goals in applying the model and identifies assumptions concerning the prevalence of fatigue in the workforce and the costs and benefits associated with decision outcomes. By making these goals and assumptions explicit, SDT can enable the appropriate use of fatigue models by fully informed decisionmakers.

REFERENCES

1. Egan JP. Signal detection theory and ROC analysis. New York: Academic; 1975.
2. Green DM, Swets JA. Signal detection theory and psychophysics. Huntington, NY: Krieger; 1974.
3. Kumamoto H, Henley EJ. Probabilistic risk assessment and management for engineers and scientists. New York: IEEE Press; 1996.
4. Pollard JK. Locomotive engineer's activity diary. Washington, DC: U.S. Department of Transportation; 1996. Report No: DOT/FRA/RRP-96/02.
5. Swets JA. Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers. Mahwah, NJ: Earlbaum; 1996.
6. von Winterfeldt D, Edward W. Decision analysis and behavioral research. Cambridge, UK: Cambridge University Press; 1986.