Human Reliability under Sleep Deprivation: Derivation of Performance Shaping Factor Multipliers from Empirical Data

Candice D. Griffith, Sankaran Mahadevan*
Vanderbilt University, Nashville, TN USA

Abstract

This paper develops a probabilistic approach that could use empirical data to derive values of performance shaping factor (PSF) multipliers for use in quantitative human reliability analysis (HRA). The proposed approach is illustrated with data on sleep deprivation effects on performance. A review of existing HRA methods reveals that sleep deprivation is not explicitly included at present, and expert opinion is frequently used to inform HRA model multipliers. In this paper, quantitative data from empirical studies regarding the effect of continuous hours of wakefulness on performance measures (reaction time, accuracy, and number of lapses) are used to develop a method to derive PSF multiplier values for sleep deprivation, in the context of the SPAR-H model. Data is extracted from the identified studies according to the meta-analysis research synthesis method and used to investigate performance trends and error probabilities. The error probabilities in test and control conditions are compared, and the resulting probability ratios are suggested for use in informing the selection of PSF multipliers in HRA methods. Although illustrated for sleep deprivation, the proposed methodology is general, and can be applied to other performance shaping factors.

Keywords: Human reliability analysis, Sleep deprivation, Performance shaping factors, Human error probability

* Corresponding author, Phone: 615-322-3040, Email: sankaran.mahadevan@vanderbilt.edu

1.0 Introduction

Human error has been associated with significant losses in many industries. Despite years of research, difficulties still exist in quantifying the contribution of human error to accidents that result in disaster and/or losses. Human reliability analysis (HRA) attempts to quantify human error probability under various situations. This information is useful in probabilistic risk analysis (PRA) of large complex systems, such as nuclear power plants. To date, HRA has mostly used expert opinion to quantify error probabilities; the subjectivity of expert opinion can result in discrepancies in risk assessments from one analyst to another (Boring, 2007). One way to improve the technical basis of current HRA methods is to quantify the effects of various performance shaping factors using empirical data. This paper investigates a quantitative methodology for including the effect of sleep deprivation in HRA based on empirical data.

Fatigue is considered a root cause of many large scale incidents; two examples are Three Mile Island in 1979 (Monk, 2007) and *Exxon-Valdez* shipwreck in 1989 (NTSB, 1990). While the effects of fatigue are pervasive in modern life, fatigue is not easy to define or measure and must be inferred from changes in performance. One particular cause of fatigue is sleep deprivation; its effect on performance is not explicitly included in current HRA methods (Griffith and Mahadevan, 2011). In the United States, it has been estimated that accidents caused by sleep deprivation have direct economic costs of more than \$50 billion annually; this cost rises to \$150 billion when productivity losses are considered (Ledger, 1994).

Three examples of large scale accidents involving sleep deprivation are the Chernobyl 1985 incident, 1986 VIA rail incident, and 1989 *Exxon-Valdez* incident. At approximately 01:30 am at the Chernobyl nuclear power facility in the Ukraine, workers turned off critical automatic safety systems and the reactor overheated. The sleep-deprived shift workers turned off the cooling system instead of turning on the automatic safety systems. This led to an explosion that

released 13 million curies of radioactive gases (Reason, 1990). In 1986, a VIA Rail passenger train collided with a Canadian National Railway freight train near Hinton (in Alberta, Canada), west of Edmonton. The accident was attributed to the crew of the freight train falling asleep (Mittler et al., 1988). In 1989, the *Exxon Valdez* oil tanker struck a reef in Prince Island Sound, Alaska, and spilled 11 to 32 million gallons of crude oil. The accident was caused by the third mate improperly maneuvering the vessel. The third mate had only had 6 hours of sleep during the previous 48 hours, while the first mate had been working for 30 hours continuously. The estimated clean-up cost was \$2 billion (Dement, 1992).

The factors that influence human performance in complex systems are referred to as performance shaping factors (PSF). These are environmental, personal, or task-oriented factors that influence the probability of human error. Some examples of psychological and physiological PSFs are task speed, distractions, monotonous work, emergency situations, fatigue, discomfort, and high temperature (Kirwan, 1994). Other examples of PSFs include: written procedures, training, personal skills, motivations, expectations, environment, and equipment design (Gertman and Blackman, 1994). Operator scheduling programs such as SAFTE (Sleep Activity, Fatigue, and Task Effectiveness) (Hursh, 1998), and FAST (Fatigue Avoidance Scheduling Tool) (Eddy and Hursh, 2001), consider performance moderating functions (PMF) based on physical and cognitive effects associated with circadian rhythms and hours of wakefulness, and evaluate the performance through a percentage effectiveness measure. However, this is primarily for use in scheduling, not to quantify human error probability.

In quantitative HRA models, multipliers are used corresponding to different levels of the individual PSFs to modify the base or nominal human error probability. Currently HRA methods mostly rely on expert opinion to select PSF multipliers. One way to improve the technical basis

of such HRA methods is to utilize available empirical data on human performance. Current research efforts are being made to use empirical data to move towards more data-informed HRA methods (Groth and Mosleh, 2012). Quantitative approaches that are often used in PRA could benefit HRA (Oxstrand, 2010). Several quantitative studies have been reported regarding the use of Bayesian methods to update the HEP and PSF multipliers based on empirical data (Mahadevan, 2007; Groth and Swiler, 2013).

One available HRA method is SPAR-H (Standardized Plant Analysis Risk – Human reliability analysis) (Gertman et al., 2005); this method uses PSF multipliers to modify the nominal or base human error probability. In SPAR-H, the nominal HEP is 0.001 for action tasks and 0.01 for diagnosis tasks (Gertman et al., 2005). The overall HEP is the product of the nominal probability with values for various PSF multipliers. This paper proposes a methodology to develop numerical values of PSF multipliers from empirical data, in the context of the SPAR-H format. In order to include the effect of fatigue in the SPAR-H method, PSF multipliers representing the effect of fatigue on performance need to be derived. In this paper, error probabilities under sleep deprived condition (i.e., test condition) and non-sleep deprived condition (i.e., control condition) are calculated based on available empirical data. The comparison of error probabilities under the two conditions can be used to inform PSF multiplier selection.

The main objective of this paper is to investigate this idea of using empirical data to derive PSF multipliers, and to illustrate the proposed methodology with sleep deprivation data. The proposed methodology is general and can be applied to other PSFs, thus forming an improved technical basis for HRA models like SPAR-H by including quantitative analysis of empirical data to supplement subjective expert opinion-based selection of PSF multipliers.

Section 2 briefly reviews three existing HRA methods, and discusses their treatment of performance shaping factors, especially sleep deprivation. Section 3 develops the proposed methodology for deriving data-driven PSF multipliers. Section 4 provides details of implementation and results for sleep deprivation, and Section 5 provides concluding remarks.

2.0 Quantitative HRA and Performance Shaping Factors

Several different HRA methods have been proposed for use in probabilistic risk assessments. Some methods are for quantification of human error rates, some are for finding error-prone situations (i.e., error-forcing context), and some methods are complete HRA models that identify error-forcing contexts and quantify error probabilities.

This section discusses three HRA methods: (1) Technique for Human Error Rate Prediction (THERP) by Swain and Guttman (1983), (2) A Technique for Human Event Analysis (ATHEANA) developed in NUREG-1684 (USNRC, 2000), and (3) Standardized Plant Analysis Risk – Human reliability analysis (SPAR-H) by Gertman et al. (2005). This section also addresses some of the limitations of existing HRA methods in the treatment of sleep deprivation.

2.1 THERP

THERP is the most widely used first generation HRA method. This is due to its ease of application and readily available numerical results for inclusion in a PRA. Other first generation HRA methods include: Success Likelihood Index Method-Multi-Attribute Utility Decomposition (SLIM-MAUD) from NUREG/CR-3518 (Embrey et al., 1983), Human Cognitive Reliability model (HCR) (Hannaman et al., 1984), and Human Error Analysis and Reduction Technique (HEART) (Williams, 1986).

The foundation of the THERP method is event-tree modeling where each branch represents a combination of human activities, influences upon these activities, and results of

these activities. Conditional probabilities are assigned for each branch along the tree (with the possible exception of the first branch). The branches emit from binary decision points, where only correct or incorrect performances are the options. THERP uses tables to provide error probabilities. For example, consider Table 1, which is THERP *Table 20-14*. The table supplies HEPs for an operator failing to notice a stuck valve under different conditions. For an activity involving multiple tasks, HEPs in THERP are calculated using event-trees based on success or failure in performing the individual tasks, and the corresponding probabilities; multiple tables are needed for such calculation.

The probability data used in THERP are based on a mixture of empirical data and expert opinion (Forester et al., 2006). Note that THERP does not construct an HEP from a nominal value that is modified by context-specific PSF multipliers. It mentions a number of PSFs, but provides quantitative guidance only for a limited number of PSFs and even that for specific errors only. It does not provide guidance for error quantification in situations not addressed in the provided tables (Forester et al., 2006). Nor does it convincingly depict the context of the actual situation.

2.2 ATHEANA

Second generation HRA methods include task context; ATHEANA (A Technique for Human Event Analysis) developed in NUREG-1624 (USNRC, 2000) is an example. Second generation HRA methods normally involve the modeling of the cognitive roots of human error, analysis of the human-machine interface, quantification of error probability, and probabilistic simulation. Examples of other second generation HRA methods are: CREAM (Hollnagel, 1998), MERMOS (Bieder et al., 1998), IDAC (Mosleh and Change, 2007), and CAHR (Sträter and Bubb, 1998).

ATHEANA focuses on identifying error-forcing contexts (i.e., why errors occur) and their influence upon the task. Error-forcing contexts are a combination of PSFs and plant conditions, regularly unexpected plant conditions, and unfavorable PSFs (USNRC, 2000). ATHEANA identifies a systematic search scheme for the identification of error-forcing contexts and is focused on events that led directly to core damage and is not aimed at pre-initiator events. HEPs in ATHEANA are calculated by Equation 1:

$$HEP = \sum P(EFC_i)P(UA_{ij}|EFC_i)P(failure\ recovery|UA_{ij}, EFC_i) \quad (1)$$

where: $i = i^{th}$ scenario, $j = j^{th}$ unsafe act, EFC = error-forcing context, $P(EFC_i)$ = the probability of error-forcing context in the particular accident scenario of analysis, $P(UA_{ij}|EFC_i)$ = the conditional probability of the unsafe acts that can cause the human failure event, and $P(failure\ recovery | UA_{ij}, EFC_i)$ = the conditional probability that the unsafe action is not recovered prior to the catastrophic failure.

ATHEANA does not provide specific lists of error-forcing contexts; instead error-forcing contexts have to be developed from different scenarios. The ATHEANA methodology is developed at a conceptual level, but not in specific detail. Also, there is no specific quantification method associated with ATHEANA; any quantitative assessment to be used in PRA has to come from expert judgment informed by the ATHEANA analysis. Thus a team of HRA specialists are normally required to conduct an HRA when using ATHEANA.

2.3 SPAR-H

The SPAR-H method was developed to support plant-specific PRA and is discussed in NUREG/CR-6883 (Gertman et al., 2005). SPAR-H is a replacement and revision of the NRC's accident sequence precursor analysis (Swain, 1987). SPAR-H is primarily a quantification

method for human error probability. In the SPAR-H method, the nominal human error probabilities are multiplied by PSF multiplier values to obtain the overall error probability. SPAR-H uses a pre-defined set of PSF multipliers, which the analyst chooses according to situational context. A review of human behavioral science literature and HRA models was used to identify eight PSFs that are used in the SPAR-H method as listed in **Error! Reference source not found.** (Gertman et al., 2005). As mentioned earlier, second generation HRA methods consider context and cognitive roots of the error, such as plant conditions and situational factors. SPAR-H considers context but not the cognitive roots. Thus SPAR-H is between first and second generation methods.

The SPAR-H method separates human error into two separate categories: action (e.g., errors of commission -- active errors) and diagnosis (e.g., errors of omission -- latent errors). Separate SPAR-H worksheets are used for quantifying action and diagnosis error probabilities. SPAR-H blends the omissions and commissions broken down in THERP into a single HEP. Different PSF weights and nominal HEP values, which are pre-defined, are associated with action and diagnosis errors in the SPAR-H method. Action errors have a nominal probability of $1.0E-3$, while the nominal probability for diagnosis is $1.0E-2$. Therefore, in SPAR-H under normal operating conditions, an individual has a 1 in 100 chance of committing a cognitive error and 1 in 1000 of committing an action error. When the effect of PSFs is included, the overall error probability is expressed as:

$$HEP = P_0 \prod_{i=1}^8 PSF_i \quad (1)$$

where HEP = human error probability; P_0 = nominal probability value; and PSF_i = multiplier for performance shaping factor i in the situation investigated.

Appendix A shows the various PSF levels and the multiplier values in SPAR-H for both diagnosis and action errors; these worksheets produce a mean estimate of the HEP. The method uses standardized worksheets and is simple to implement; the worksheets ensure analysis consistency and documentation across all users. Although SPAR-H is relatively straightforward, one weakness is that interactions among the PSFs are not considered. SPAR-H is implicitly influenced by THERP for its base (i.e., nominal) failure rates. The main source of uncertainty in SPAR-H comes from selecting the multiplier values for the PSFs, which are largely based on expert opinion; different experts might assign different multipliers for the same situation.

2.4 Inclusion of Sleep Deprivation in HRA Methods

Sleep deprivation and human fatigue are not explicitly considered as a separate PSF in most commonly used HRA methods. Nor do current HRA methods adequately include explicit quantitative measures for the effects of sleep deprivation (Griffith and Mahadevan, 2011).

Sleep deprivation is not explicitly covered in THERP. The THERP tables deal with HEPs of actual events, e.g., *detecting stuck locally operated valves* (Swain and Guttman, 1983). In the case of ATHEANA, sleep deprivation could come under the error-forcing context (EFC) related to the effect of the initiator or accident sequence on the capacity of action of the control room before the incident. For retrospective analysis using ATHEANA (i.e., after the incident), sleep deprivation could be considered as a sub-factor under the PSF of human performance capabilities.

In SPAR-H, sleep deprivation could be considered a sub-factor of the fitness-for-duty PSF. However, it could be argued that sleep deprivation would have some effect on several of the eight PSFs given their generality. For example, SPAR-H PSFs such as stress, time available, and complexity would also be affected by sleep deprivation through its competition for limited

mental resource capacity. However, the main PSF that would be affected would be fitness-for-duty. The PSF fitness-for-duty is broken down into four levels in SPAR-H: *unfit* (which gives the probability of human error as 1.0), *degraded* (weight of 5), and *nominal* (weight of 1), with the additional option of selecting the category of insufficient information (weight of 1). These three levels of fitness-for-duty cover a wide range of influences on performance: fatigue, impairment due to drugs (prescription, over-the-counter, or illegal) or alcohol, distraction due to personal or family issues, whether a person is physically or mentally capable of performing task, and boredom (Whaley et al., 2011).

The selection of PSF multipliers in SPAR-H can be enhanced by using empirical data for quantification. This will lead to an improved technical basis of HRA, resulting in improved probabilistic risk assessment (PRA). An improved HRA will allow for an increased understanding of the safe operating bounds of human behavior in the twenty-four hour-economy world. In general, the quantitative and qualitative aspects of PSFs such as sleep deprivation needs to be better understood and represented in the HRA methods.

3.0 Proposed Methodology

In order to include quantitatively the effect of sleep deprivation on performance, empirical data on the effect of sleep deprivation must be collected and analyzed. The proposed methodology has three steps: (1) definition of performance metrics, (2) collection of relevant data, and (3) quantitative analysis to derive PSF multipliers.

3.1 Performance Metrics

There is no standard metric used to evaluate performance; studies may use single metric or multiple metrics. Typical performance evaluators could be subjective, physical, or psychomotor. Subjective measures, such as sleepiness scales, are more appropriate for sleepiness and

fatigue studies. Physical measures, such as temperature, electroencephalography (EEG) (Dahlgren, 1981), or critical flicker frequency threshold, are better measures of arousal or alertness and are very useful when investigating circadian influences (Czeisler, 1999). However, the focus in this paper is on how sleep deprivation affects performance; consequently, psychomotor tests are more appropriate. Psychomotor tests are based on mental processes (cognition) that require a motor reaction. Performance on working memory tasks is observed to be predictive of performance on a range of other cognitive tasks (Kushida et al., 2005).

Performance tasks, such as Psychomotor Vigilance Task (PVT) used by Dinges and Powell (1985) are useful in tracking the ability to access the working memory of the prefrontal cortex of the brain. The metrics of reaction time, accuracy, and number of lapses are typically the reported results of tests with such working memory tasks. These three metrics are considered in this paper, since they are most frequently reported in the sleep deprivation literature and are common characteristics of working memory tasks. Although psychomotor tasks can be simple, they are cognitive performance tasks. Diagnosis tasks require cognitive engagement and determining the appropriate course of action (Gertman, et al., 2005). Action tasks relate to following procedures of actions. The psychomotor tasks evaluated in this study are aligned with diagnosis tasks.

3.2 Collection of Empirical Data

The identification of sleep deprivation as a significant source of risk has led to studies being conducted in a multitude of fields and settings. Most of the data used in this research comes from psychological experiments. These experiments, for the most part, have been conducted under the guidance of industries such as transportation, medical, and military. These industries rely on individuals to complete complicated tasks under extended hours on a regular

basis. A search for sleep deprivation studies evaluating performance was conducted. Only a brief summary is given here; refer to Griffith (2013) for details of data collection and meta-analysis.

Studies considered for inclusion in this study were published research studies on sleep deprivation and its effect on performance, in a pretest-posttest design. These studies use an experimental design comparing the performance before treatment (i.e., pretest) to the performance after treatment (i.e., posttest). The initial search strategy used computerized databases, (e.g., PubMed, Cochran, MEDLINE, PsycINFO, and PsychLit) to locate relevant publications. Search terms included fatigue, performance, sleep deprivation, shiftwork, and reaction time. The studies used (after the 1940's) were collected from multiple disciplines and journals in the English language.

The studies typically use simplified tasks in a laboratory setting to evaluate performance and study design variations instead of complicated tasks similar to those possibly experienced in working environments. Even though the studies considered different tasks they are of similar complexity and were collected together using the meta-analysis technique (Lipsey and Wilson, 2001).

Meta-analysis provides a structured format to extract data from multiple studies for the purpose of combining the findings through statistical procedure and coding. For example, if a study reported a change in a performance variable on a task over an increase in the hours of continuous wakefulness of the subject, the mean values and variance of the initial and ending performance and the hours of wakefulness could be extracted. This data could then be used to calculate an effect size, i.e., the degree of change in the variable under study.

Initially, 601 studies related to fatigue were found for review. Viable studies were evaluated to find if performance under sleep deprivation was reported; this step resulted in 108

studies. These studies were reviewed in more detail. The list of studies was further refined by including only studies that reported reaction time, number of lapses, and accuracy for both pre-sleep deprived (control) and post-sleep deprived (test) conditions. The 34 studies resulting from the search are listed in Table 3.

The 34 studies identified reported at least one performance measure (i.e., accuracy, number of lapses, and/or reaction time) under sleep deprived conditions. The studies listed in Table 3 were used to supply the data for the subsequent analysis and results in Tables 4 and 6. Table 3 uses R, L, and A to denote what type of data was extracted from each study (R–reaction time data, L–lapse data, and A–accuracy data). Several studies reported multiple types of data, e.g., Chee and Choo (2004) reported both reaction time and accuracy data. Only one study reported all three performance measures (Frey et al. 2004). The performance measure at the control condition and at each test condition was extracted from each study; typically multiple test conditions were reported. In cases where the standard deviation was reported or could be derived, these data were also recorded.

A subset of 18 studies (out of the 34 studies listed in Table 3) that used the same type of task (the psychomotor vigilance task by Dinges and Powell, 1985) gave an average effect size for the reaction time data as -1.173, for accuracy data as -0.920, and for number of lapses data as -2.388; see Griffith (2013) for details. A negative effect size shows a degraded (i.e., negative) effect of sleep deprivation on reaction time, accuracy, and lapses. However, it is not clear how to apply the effect size results to error probability calculation in HRA; therefore, a methodology to inform the selection of PSF multipliers similar to Equation 2 is developed below.

3.3 Proposed Method for PSF Multiplier Derivation

A quantitative measure of the effect of sleep deprivation on performance (i.e., estimation of human error probabilities) is derived in this subsection by comparing the probabilities of error in the control (pre-sleep deprivation) and test (post-sleep deprived) conditions to obtain a PSF multiplier value. This is approached in two ways: (1) by comparing both the control (C) and the test (T) performance measure to a threshold value, or (2) by simply comparing the control and the test performance measures (C or T) to each other. Both approaches are developed below.

Approach 1: Comparison of C and T to a threshold value

The performance measure (e.g., reaction time) under test condition and control condition can be compared to a threshold value k , in order to compute the error probability under each condition. The threshold value is selected by the analyst and is specific to the situation being considered. For example, the threshold values for reaction time could be selected based on the definition of a lapse for simple reaction time tasks, e.g., no response before 500 ms (Anderson et al., 2010).

In the derivation below, error is defined as the performance measure being larger than the threshold value; the derivation can be easily modified if the error is defined in the opposite manner. For example, suppose errors with respect to the control condition reaction time (C) and the test condition reaction time (T) are mathematically defined as $C > k$ and $T > k$. The reaction times under control condition and test condition are both assumed to be normal random variables with a mean value μ_C and variance σ_C^2 for the control condition and a mean value μ_T and variance σ_T^2 for the test condition. Since error is defined as $C > k$ or $T > k$, the probability of error is calculated as:

$$P_C = P(C > k) = 1 - \Phi((k - \mu_C) / \sigma_C) \quad (2)$$

$$P_T = P(T > k) = 1 - \Phi((k - \mu_T) / \sigma_T) \quad (3)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-x^2/2} dx \quad (4)$$

where k = threshold value, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal variable.

The next step is to compare the error probabilities from Equations 3 and 4 for the control and test conditions. The probability of error in the test condition can be divided by the control condition probability of error, to show the amount of change in failure probability due to sleep deprivation, which may be used as the PSF multiplier value in Equation 2. Thus,

$$\text{PSF multiplier value} = P_T / P_C. \quad (5)$$

Approach 2: Comparison of C and T to each other

Another way to derive PSF multiplier values is to not use a threshold value, but to compare directly the error probabilities in control and test conditions to each other. In this method, deterioration of performance is indicated by the test performance measure T being greater than the control performance measure C (e.g., number of lapses), i.e., $T > C$; or by the test performance measure being smaller than the control performance measure (e.g., number of correct responses), i.e., $T < C$. The PSF multiplier is derived by comparing the probability that the test condition value (T) is greater or smaller than the control condition value (C) against the probability of T being equal to C . If no deterioration is expected, the probability that C and T are equal is 0.5. The probability ratio, with no threshold value, is then defined as:

$$\text{for } T > C, \text{ Probability Ratio} = P(T > C) / P(T = C) = P(T > C) / 0.5 \quad (6)$$

$$\text{for } C > T, \text{ Probability Ratio} = P(C > T) / P(C = T) = P(C > T) / 0.5. \quad (7)$$

Note that in this approach, the results are bounded by 0 as the lower limit and 2 as the upper limit. This may or may not be appropriate for certain situations, especially when used with other PSFs that are given much larger values by expert opinion. In general, different situations might call for either Approach 1 or Approach 2.

4.0 Implementation for Sleep Deprivation

This section implements the methodology developed in Section 3 using available empirical data on the effect of sleep deprivation on performance. The data used is from the studies listed in Table 3. The implementation of this method for sleep deprivation is separated into three sub-sections, one for each of the performance metrics (reaction time, number of lapses, and accuracy).

4.1 Reaction Time

Approach 1: Comparison of C and T to a threshold value

The threshold values were selected based on the definition of a lapse in simple reaction time psychomotor tasks; a lapse is defined as failure to respond within 500 milliseconds (Anderson et al., 2010). Different threshold values were selected: $k = 250, 500, 750$, and 1000 . The number of data points available, and therefore the number of probability ratio values that could be calculated decreases as the threshold value increases, due to increasing number of outliers as explained below. The number of probability ratios that could be calculated for each k value is listed in Table 4.

The data for the probability calculations were available in 30 out of the 34 studies listed in Table 3. For $k = 500$, the thirty studies reported reaction time data at 249 time recordings. For each of these 249 points, $P_T = P(T > k)$ and $P_C = P(C > k)$ were calculated. The points with $P_T = 0$ or $P_C = 0$ were unusable and discarded. This is a limitation of the proposed approach. The

denominator in the ratio P_T/P_C cannot have zero values, since the ratio will become infinite. Therefore data points with P_C close to zero were removed from the data set for analysis. Data points with P_T close to zero were also removed, since it is not reasonable to have the error probability drastically reduced to zero in the test condition. Thus, for $k = 500$, 150 points had $P_T = 0$ and/or $P_C = 0$. Among the remaining 99 points (row 3 in Table 4), outliers were also removed from the analysis in order to prevent the probability ratio plots from being skewed toward one or two studies that had markedly different observations from the rest of the data (Grubbs, 1969). In 15 data points, the probability ratio (P_T / P_C) approaches values greater than 10 when the error probability for the control condition (i.e., denominator of the ratio) is small; these 15 points were also removed as outliers, since a majority of ratios are observed to be between 1 and 2. This resulted in 84 probability ratio values that remained for $k = 500$. Note that as k increases, the probability P_C decreases, leading to increased numbers of outliers (i.e., $P_T / P_C > 10$).

Figure 1 plots the probability that the test reaction time T is larger than the threshold value $k = 500$, for all 249 reaction time data points. The probability of error in the test condition clusters around two values (0 and 0.35) until 20 hours of wakefulness, and three values after 40 hours of wakefulness (0, 0.6 and 1). It should be noted that different studies used different numbers of hours of wakefulness for the control condition. That is, the test began immediately after waking up in some studies, and after a certain number of hours of wakefulness in some other studies.

The probability ratios P_T/P_C calculated using Equation (6) are plotted in Figure 2. The majority of the probability ratio values before 20 hours of wakefulness are approximately one, i.e., no change in performance. From 20 to 40 hours of wakefulness, the probability ratio value fluctuates between one and two, then approaches a steady value of about 2.0 after 40 hours of

wakefulness. That is, the error probability (with respect to reaction time here) appears to approach a constant impairment value and does not continue to increase linearly as hours of wakefulness increase. The probability ratio values show a trend similar to the raw data, i.e., an increase in simple reaction time after twenty-four hours of wakefulness.

Although the studies show no significant change in performance before 20 hours of wakefulness, note that all these studies were conducted in a laboratory setting using simple tasks. The result would be different for different types of tasks, and more complex tasks would likely see significant changes in performance earlier. However, in SPAR-H the task complexity is addressed by a separate PSF; therefore the overall error probability would be enhanced by the corresponding multiplier for task complexity. The current study serves to isolate the effect of sleep deprivation on performance, by considering similar (simple) tasks across studies. Considering sleep deprivation and task complexity together would have confounded the effect. Further, the results appear to support the current regulations regarding a two-shift maximum (16 hours) in several industries. However, the 16-hour maximum is during normal operations; the maximum might be exceeded during abnormal conditions, and in that case the results for the extended hours might become more relevant.

The same probability ratio procedure was implemented for all four selected threshold values ($k = 250, 500, 750$, and 1000), and the results are plotted in Figure 3. Probability ratios larger than 10 were eliminated as outliers; as mentioned earlier, the number of such outliers increase as the threshold value increases, since the error probability in the control condition decreases as the threshold value increases. In Figure 3, the probability ratios for the threshold value of $k = 1000$ have a noticeably magnified trend than for the other k values.

Approach 2: Comparison of C and T to each other

The probability ratios resulting from Equation 6 and plotted in Figure 3 are affected by the choice the threshold value (k). The probability ratios indicate the enhancement of probability in the test vs. the control condition. Therefore the second approach directly compares the test condition value being greater than the control condition value (for reaction time), using Equation 7 (i.e., $P(T > C) / 0.5$, since the probability $P(T = C)$ is 0.5). The probability of error are calculated for all the 249 data points, and plotted in Figure 4.

In this approach, there are no outliers that need to be removed, since the upper bound of $P(T > C)/0.5$ is 2, as mentioned in Section 3. The probability ratio over all time periods has a mean value of 1.321 and standard deviation of 0.62. The results for Approach 2 are not dependent on the choice of a threshold value (k); however, the probability ratios are limited to a range of $[0, 2]$.

Next, probability ratio data from both approaches (1 and 2) were subdivided into twenty hour intervals (i.e., < 20 , 20-40, 40-60, 60-80, and > 80). Figure 5 shows the 95% bounds of the calculated mean probability ratios for each time block, for $k = 500$ (in Approach 1). The mean values and standard deviations of the probability ratios over each twenty hour time block are summarized in Table 5, for both Approach 1 and Approach 2. For Approach 1, the results are given for four different threshold values ($k = 250, 500, 750$, and 1000). For $k = 500$, all five time-blocks report mean probability ratios in the range $[1, 2]$.

Note that the variation of the probability ratio in relation to the number of hours of wakefulness is not linear. The raw reaction time data (under the test condition) collected directly from the studies shows that a linear fit for decline in performance vs. hours of wakefulness has a

low R^2 value (Figure 6). This is to be expected, especially due to the confounding effects of several other factors; such as circadian rhythms, time-of-day influence, and time-on-task.

4.2 Number of Lapses

Approach 1: Comparison of C and T to a threshold value

A lapse is defined as a failure to respond within 500 milliseconds (Anderson, et al., 2010). The data on number of lapses came from eight of the 34 studies in Table 3. The analysis of this data resulted in 68 data points. The studies reported mean number of lapses for the test and control conditions, over all the subjects. The average of the control condition mean values (number of lapses) was 4.31 and the standard deviation was 4.12. The average test condition mean value was 9.27 and the standard deviation was 6.51. The standard deviations are quite high due to the wide range of reported number of lapses. Four different threshold values were selected based on the mean values of the test and control conditions ($k = 1, 3, 6$, and 9).

The number of probability ratios that could be calculated varied for the different k values. As the threshold value increases, the number of lapses greater than the threshold value is reduced. The number of probability ratios that could be calculated for each k value is listed in Table 6. Probability ratio greater than 10 were removed as outliers. The number of outliers removed for $k = 9$ was significantly large. The combined results of the probability ratios for $k = 1, 3, 6$, and 9 , using Approach 1, are plotted in Figure 7.

Approach 2: Comparison of C and T to each other

The probability ratios for Approach 2 are calculated using Equation 7, and plotted in Figure 8. No outliers needed to be removed in this approach; thus all 68 ratios are plotted. The probability ratio over all time periods has a mean value of 1.42 and standard deviation 0.38. The

results show two dips occurring around forty hours and just after sixty hours of wakefulness, which may be due to the effect of diurnal or circadian rhythms on performance.

The probability ratios calculated show variability due to sparse data. Data are subdivided into twenty hour intervals (i.e., < 20 , 20-40, 40-60, 60-80, and > 80). Using the sample size, mean, and standard deviation of the data for each time block, the 95% confidence bounds for $k = 1$ are calculated using the t -distribution and the results are provided in Figure 9 and Table 7.

For $k = 1$, all five time-blocks report mean probability ratio values in the range [0.89, 1.12] as shown in Table 7. The largest mean probability ratio value occurs between 20 and 40 hours of wakefulness and the smallest occurs between 40 and 60 hours of wakefulness. The mean values and standard deviations of the probability ratios over each twenty hour time block are summarized in Table 7, for both approaches.

For Approach 1, the results are shown for four k values ($k = 1, 3, 6$, and 9). There was insufficient data to calculate a reliable mean for $k = 9$ beyond 40 hours. The mean values for $k = 3, 6$, and 9 are greater than 1, except for $k = 3$ under 20 hours of wakefulness. The mean probability ratios for $k = 1$ were less than one. The probability ratios less than 1 are not realistic, since the number of lapses is expected to increase or at least remain the same as hours of wakefulness increases; therefore it may be desirable to use a value of 1 in these cases. For practical applications, an appropriate value for the threshold k needs to be selected based on the task being studied.

Figure 10 shows the increase in the mean number of lapses over hours of wakefulness. Note that the linear trend-line has a low R^2 value. Non-linearity is to be expected, especially due to the confounding effect of several other factors; such as circadian rhythms, time-of-day influence, and time-on-task.

4.3 Accuracy

Approach 1: Comparison of C and T to a threshold value

Twelve studies reported five different measures of accuracy: number correct, percentage correct, number of errors, percentage error, and number of false positives. The accuracy data were segregated and analyzed by the type reported in the original study. A reduction of performance occurs when the test value is larger than the control value, for three measures: number of errors, percentage error, and false positives. Degradation in performance occurs when the control value is larger than the test value for two measures: number correct and percentage correct. For practical applications, an appropriate value for the threshold k needs to be selected based on the situation being studied and the accuracy measure considered.

The analysis of Approach 1 was carried out with different threshold values for different accuracy measures. Detailed results are reported in Griffith (2013). Unfortunately when using Approach 1, after the data was segregated into five groups, the sample sizes of the data were not large enough to draw significant conclusions. Therefore the results of Approach 1 are not reported here.

Approach 2: Comparison of C and T to each other

Using the second approach to calculating probability ratios, Equation 7 was used for the number of errors, percentage error, and false positives, whereas Equation 8 for number correct and percentage correct measures. Collecting all types of accuracy measures together at the level of probability ratios, all the data points are plotted in Figure 11.

The probability ratio results are segregated into 20-hour time blocks. The mean and 95% confidence interval using a t -distribution are plotted in Figure 12. The mean and standard deviation are presented in Table 8. The mean varied from 0.85 to 1.69 over all five time blocks,

showing an increasing trend first and then a dip after 80 hours. However, only 2 data points were able to be used in the evaluation of the 60 to 80 hour time block; also, the largest probability ratio mean occurred at this interval 1.696 (0.01). The significance of the results is affected by sparse data.

The effect of sleep deprivation on accuracy is influenced by such factors as circadian rhythms, time-of-day influence, and time-on-task. The effect of the *speed-accuracy tradeoff* also confounds the accuracy data results (Fitts, 1954). The draw on cognitive capacity that results from sleep deprivation may help increase arousal levels (Yerkes and Dodson, 1908) and help to maintain, or even improve, performance on the simple tasks that were the primary tasks employed in the studies under investigation. Figure 13 shows the lack of trend in the test condition data for accuracy (i.e., low R^2 value for the linear trend-line). In general, accuracy does not appear to be affected as significantly as reaction time and number of lapses under sleep-deprived conditions.

5.0 Conclusion

This paper used available data from the literature to quantify the effect of sleep deprivation on performance, in order to improve the technical basis for the selection of PSF multipliers in SPAR-H. Sleep deprivation, defined as continuous hours of wakefulness, was considered to be an independent influence on performance and not correlated to other PSFs. In the context of HRA models, currently the PSFs are considered to be independent, and no interaction effects are considered. This limitation needs to be addressed in future work. Also,

future empirical studies should report time of day to help account for the effect of circadian rhythms.

Studies found in the literature with data on sleep deprivation effects on performance were extracted according to the meta-analysis research synthesis method (Griffith, 2013). The quantitative data extracted from the literature were used to analyze the effect of sleep deprivation on performance. This can be accomplished through various measures, e.g., percentage change, effect size, error probabilities, etc. The error probabilities are considered in this paper, and are derived using two approaches: (1) by contrasting the performance under test and control conditions to a threshold value; and (2) by contrasting the test and the control performance to each other.

There are mathematical limitations in both approaches. Approach 1 needs to exclude data points where the probability values are close to zero. When the denominator (i.e., error probability in the control condition) approaches zero, the ratio approaches infinity. In Approach 2, the probability ratio is bound mathematically between values of 0 and 2; however, this may not be suitable when compared with multiplier values used currently in SPAR-H, which may be as high as 50 (see Appendix A). In Approach 1, the upper bound used in this paper for the probability ratio is 10; this can be increased by changing the outlier definition. On the other hand, the advantage of Approach 2 is that there is no need to assume a threshold value for the performance measure, thus removing some of the subjectivity in the analysis; also, in this approach, no outliers need to be removed. Thus both approaches have advantages and disadvantages, and each approach might be preferable to the other depending on the application.

Note that the error probability ratios are different for different performance measures (e.g., reaction time, number of lapses, and accuracy), and for different hours of wakefulness. The

selection of PSF multipliers in HRA may be guided by the probability ratios for different measures in different ways, e.g., average over all three performance measures, or the use of probability ratios from the performance measure that is most relevant to the HRA situation of interest, varying with hours of sleep deprivation. In other words, the decision maker could fuse the three different multiplier values by giving different weights to each performance measure based on his/her risk preferences and the application problem, and arrive at the overall PSF multiplier values.

Note however, that the three performance measures are not independent of each other. In the case of speed vs. accuracy tradeoff, reaction time may be sacrificed to maintain or improve accuracy (Fitts, 1954). The statistical correlations between the performance measures may provide indication of interaction or tradeoff between them. However, it was not possible to quantify the correlations from the available data. Only one study (Frey et al., 2004) reported all three performance measures (R, L, and A). However, even in this study, the individual subject performance was not reported; only an average performance measure was reported for the entire group. In future efforts, the statistical correlations between the different performance measures could be investigated if they are reported for each individual subject.

The 34 studies investigated in this paper typically used simplified laboratory tasks to evaluate performance. It is challenging to use the probability ratios based on this data to inform the selection of PSF multipliers for application to practical real-world tasks. Real-world tasks might have quite different complexity and time demands compared to laboratory tasks. However, in SPAR-H, task complexity is included as a separate PSF and the HEP is enhanced by the corresponding multiplier (see Tables 9 and 10 and SPAR-H worksheets in Gertman et al., 2005). Additionally, the use of simulator studies to gather data on performance under realistic situations

would be valuable (Lois et al, 2009). Future simulator studies that collect reaction time, accuracy, and/or lapse data could be used to produce an HEP to evaluate and compare both the proposed approaches.

This paper demonstrated an improved technical basis for selecting PSF multiplier values based on statistical analysis of available literature data in the context of the SPAR-H model, where PSF multiplier values are currently based mostly on expert opinion. In addition to such a statistical approach that improves first generation HRA methods, it is also important to pursue research in the context of second-generation HRA methods in order to advance the fundamental understanding of human cognition and response under various situational contexts.

Acknowledgments

This research was partially sponsored by a graduate fellowship funded by the NSF IGERT program in Reliability and Risk Engineering at Vanderbilt University, and by INL Human Factors, Instrumentation, and Control Systems Department through the INL Educational Programs (Mentors: Dr. Bruce Hallbert, Jeffrey Joe). The authors would also like to thank Dr. Harold Blackman at INL, and Dr. Erasmia Lois at USNRC, for support and guidance during this study.

References

1. Anderson, C., A. Wales, et al. (2010). "PVT Lapses Differ According to Eyes Open, Closed, or Looking Away." *Sleep* **33**(2): 197–204.
2. Bieder, C., P. Le-Bot, et al. (1998). "MERMOS: EDF's New Advanced HRA Method." Probabilistic Safety Assessment and Management. PSAM 4. New York, Springer-Verlag.
3. Boring, R., T. Tran, et al. (2006). "Users' Guide for the NASA Human Reliability Database." INL/EXT-06-11944. Idaho National Laboratory, Idaho Falls, Idaho.
4. Boring, R. (2007). "A Review of Expertise and Judgment Processes for Risk Estimation." INL/CON-07-12774. Idaho National Laboratory, Idaho Falls, Idaho.
5. Caldwell, J., B. Prazinko, et al. (2003). "Body Posture Affects Electroencephalographic Activity and Psychomotor Vigilance Task Performance in Sleep-Deprived Subjects." *Clin Neurophysiol* **114**(1): 23-31.

6. Chee, M. and W. Choo (2004). "Functional Imaging of Working Memory after 24-Hr of Total Sleep Deprivation." The Journal of Neuroscience **24**(19): 4560-4567.
7. Choo, W., W. Lee, et al. (2005). "Dissociation of Cortical Regions Modulated by Both Working Memory Load and Sleep Deprivation and by Sleep Deprivation Alone." Neuroimage **25**: 579-587.
8. Czeisler, C., J. Duffy, et al. (1999). "Stability, Precision and Near 24-hour Period of the Human Circadian Pacemaker." Science **284**(5423): 2177-2181.
9. Dahlgren, K. (1981). "Adjustment of Circadian Rhythms and EEG Sleep Functions to Day and Night Sleep among Permanent Nightworkers and Rotating Shiftworkers," Psychophysiology **18**: 381-391.
10. Dement, W. (1992). The Sleep Watchers. Stanford, CA, Stanford Alumni Association.
11. Dinges, D. and J. Powell (1985). "Microcomputer Analyses of Performance on a Portable, Simple Visual RT Task during Sustained Operations." Behavior Research Methods, Instruments, & Computers **17**: 652-655.
12. Dinges, D. and J. Powell (1988). "Sleepiness is More Than Lapsing." Sleep Research **17**: 84.
13. Doran, S., H. Van Dongen, et al. (2001). "Sustained Attention Performance during Sleep Deprivation: Evidence of State Instability." Archives Italiennes de Biologi **139**: 253-267.
14. Drake, C., T. Roehrs, et al. (2001). "Effects of Rapid versus Slow Accumulation of Eight Hours of Sleep Loss." Psychophysiology **38**: 979-987.
15. Drummond, S. and G. Brown (2006). "The Effects of Total Sleep Deprivation on Cerebral Responses to Cognitive Performance." Neuropsychopharmacology **25**(5, Supplement 1): S68-S73.
16. Eddy, D. and S. Hursh. (2001) "Fatigue Avoidance Scheduling Tool (FAST)." Technical Report No.AFRL-HE-BR-TR-2001-0140, Brooks AFB, TX.
17. Embrey, D., P. Humphreys, et al. (1983). "SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert." NUREG/CR-3518. U.S. Nuclear Regulatory Commission, Washington, DC.
18. Forester, J., A. Kolaczowski, et al. (2006). "Evaluation of Human Reliability Analysis Methods against Good Practices." NUREG-1842. U.S. Nuclear Regulatory Commission, Washington, DC.
19. Frey, D., P. Badia, et al. (2004). "Inter- and Intra-individual Variability in Performance near the Circadian Nadir during Sleep Deprivation." Journal of Sleep Research **13**(4): 305-315.
20. Gertman, D. and H. Blackman (1994). Human Reliability and Safety Analysis Data Handbook. New York, John Wiley & Sons, Inc.
21. Gertman, D., H. Blackman, et al. (2005). "The SPAR-H Human Reliability Analysis Method." NUREG/CR-6883. U.S. Nuclear Regulatory Commission, Washington, DC.
22. Griffith, C. and S. Mahadevan (2011). "Inclusion of Fatigue Effects in Human Reliability Analysis." Reliability Engineering & System Safety **96**(11): 1437-1447.
23. Griffith, C. (2013). Inclusion of Fatigue Effects in Human Reliability Analysis, Ph.D. Dissertation, Vanderbilt University, Nashville, TN.
24. Groth, K. and A. Mosleh (2012). "Deriving Causal Bayesian Networks from Human Reliability Analysis: A Methodology and Example Model." Journal of Risk and Reliability **226**(4): 361-379.

25. Groth, K. and L. Swiler (2013). "Bridging the Gap between HRA Research and HRA Practice: A Bayesian Network Version of SPAR-H." Reliability Safety Engineering & System **115**: 33–42.
26. Grubbs, F. (1969). "Procedures for Detecting Outlying Observations in Samples." Technometrics, **11**(1): 1-21.
27. Hannaman, G., A. Spurgin, et al. (1984). Human Cognitive Reliability Model for PRA Analysis, NUS-4531. Electric Power Research Institute (EPRI), Palo Alto, CA.
28. Hollangel, E. (1998). Cognitive Reliability and Error Analysis Method - CREAM. Oxford, Elsevier Science.
29. Hursh, S. (1998). "Modeling Sleep and Performance within the Integrated Unit Simulation System (IUSS)." Technical Report Natick/TR-98/026L. Science and Technology Directorate, Natick Research, Development and Engineering Center, United States Army Soldier Systems Command, Natick, MA.
30. Jewett, M., D. Dijk, et al. (1999). "Dose-Response Relationship between Sleep Duration and Human Psychomotor Vigilance and Subjective Alertness." Sleep **22**(2): 171-179.
31. Jones, C., J. Dorrian, et al. (2006). "Self-Awareness of Impairment and the Decision to Drive after an Extended Period of Wakefulness." Safety Science **23**(6): 1253 - 1263.
32. Kirwan, B. (1994). A Guide to Practical Human Reliability Assessment. Bristol, PA, Taylor & Francis.
33. Kobbeltvedt, T., W. Brun, et al. (2005). "Cognitive Processes in Planning and Judgments under Sleep Deprivation and Time Pressure." Organizational Behavior and Human Decision Processes **98**: 1-14.
34. Kushida, C., et al. (2005) "Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005." Sleep **28**, 499-521.
35. Lamond, N. and D. Dawson (1999). "Quantifying the Performance Impairment Associated with Fatigue." Journal of Sleep Research **8**(4): 255–262.
36. Lichor, J., E. Nuotto, et al. (1989). "The Effect of Sleep Deprivation on Psychomotor Performance in Anesthesia Residents." Anesthesia Analgesia **68**.
37. Lieberman, H., G. Bathalon, et al. (2005). "Severe Decrements in Cognition Function and Mood Induced by Sleep Loss, Heat, Dehydration, and Under-nutrition during Simulated Combat." Biological Psychiatry **57**(4): 422-429.
38. Lipsey, M. and D. Wilson (2001). Practical Meta-Analysis. Thousand Oaks, CA, Sage Publications.
39. Lois, E., V. Dang, et al. (2009) "International HRA Empirical Study—Phase 1 Report: Description of Overall Approach and First Pilot Results from Comparing HRA Methods to Simulator Performance Data." NUREG/AI-0216 Vol. 1. U.S. Nuclear Regulatory Commission, Washington, DC.
40. Mahadevan, S. (2007). "Bayesian Updating of PSF Effects and HRA Estimates," in "The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study," (Editors: B. Hallbert, A. Kolaczowski, E. Lois), NUREG/CR-6949, U.S. Nuclear Regulatory Commission, Washington, D.C.
41. Mitler, M., M. Carskadon, et al. (1988). "Catastrophes, Sleep, and Public Policy: Consensus Report." Sleep **11**: 100–109.
42. Monk, T. (1997). "Circadian Rhythms in Human Performance and Mood under Constant Conditions." Journal of Sleep Research **6**: 9-18.

43. Monk, T. (2007). "Practical Consequences of Fatigue-Related Performance Failures." Sleep **30**(11): 1402-1403.
44. Mosleh, A. and Y. Chang (2007). "Cognitive Modeling and Dynamic Probabilistic Simulation of Operating Crew Response to Complex System Accidents – Part 1: Overview of the IDAC Model." Reliability Engineering and System Safety **92**: 997–1013. (2007).
45. National Transportation Safety Board (1990). "Practices That Related to the *Exxon Valdez*." M90-23-31, National Transportation Board, Washington, DC.
46. Nilsson, J., M. Soderstrom, et al. (2005). "Less Effective Executive Functioning after One Night's Sleep Deprivation." Journal of Sleep Research **14**: 1-6.
47. Oxstrand, J. (2010). "Human Reliability Guidance – How to Increase the Synergies between Human Reliability, Human Factors, and System Design & Engineering. Phase 2: The American Point of View – Insights of How the US Nuclear Industry Works with Human Reliability Analysis." Technical Report NKS-229. Nordic Nuclear Safety Research Council (NKS).
48. Philip, P., J. Taillard, et al. (2004). "Age, Performance and Sleep Deprivation." Journal of Sleep Research **13**(2): 105–110.
49. Powell, N., K. Schechtman, et al. (2001). "The Road to Danger: The Comparative Risks of Driving while Sleepy." Laryngoscope **111**(5): 887-893.
50. Rajaraman, S., A. Gribok, et al. (2007). "Individualized Performance Prediction of Sleep-Deprived Individuals with the Two-Process Model." Journal of Applied Physiology **104**: 459-468.
51. Reason, J. (1990). Human Error. New York, Cambridge University Press.
52. Reznick, R. and J. Folse (1987). "Effect of Sleep Deprivation on the Performance of Surgical Residents." The American Journal of Surgery **154**: 520-525.
53. Roach, G., D. Dawson, et al. (2006). "Can a Shorter Psychomotor Vigilance Task Be Used as a Reasonable Substitute for the Ten-Minute Psychomotor Vigilance Task." Chronobiology International **23**(6): 1379-87.
54. Robbins, J. and F. Gottlieb (1990). "Sleep Deprivation and Cognitive Testing in Internal Medicine House Staff." Western Journal of Medicine **152**: 82-86.
55. Rosa, R. and M. Bonnet (1993). "Performance and Alertness on 8-h and 12-h Rotating Shifts at a Natural Gas Utility." Ergonomics **36**(10): 1177-1193.
56. Sagaspe, P., M. Sanchez-Ortuno, et al. (2006). "Effects of Sleep Deprivation on Color-World, Emotional, and Specific Stroop Interference and on Self-Reported Anxiety." Brain and Cognition **60**: 76-87.
57. Sanders, A. and W. Reitsma (1982). "The Effect of Sleep-Loss on Processing Information in the Functional Visual Field." Acta Psychologica **51**(2): 149-162.
58. Scott, J., L. McNaughton, et al. (2006). "Effects of Sleep Deprivation and Exercise on Cognitive, Motor Performance and Mood." Physiology and Behavior **87**: 396-408.
59. Sträter, O. and H. Bubb (1998). "Assessment of Human Reliability Based on Evaluation of Plant Experience: Requirements and their Implementation." Reliability Engineering and System Safety **63**(2): 199-219.
60. Storer, J., H. Floyd, et al. (1989). "Effects of Sleep Deprivation on Cognitive Ability and Skills of Pediatrics Residents." Academic Medicine **64**: 29-32.

61. Swain, A. and H. Guttman (1983). "A Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications." NUREG/CR-1278. U.S. Nuclear Regulatory Commission, Washington, DC.
62. Swain, A. (1987). "Accident Sequence Evaluation Program Human Reliability Analysis procedure." USNRC-NUREG/CR-4772. U.S. Nuclear Regulatory Commission, Washington, DC.
63. Swann, C., G. Yelland, et al. (2006). "Chronic Partial Sleep Loss Increases the Facilitory Role of a Masked Prime in a Word Recognition Task." Journal of Sleep Research **15**: 23-29.
64. Thomas, M., H. Sing, et al. (2000). "Neural Basis of Alertness and Cognitive Performance Impairments during Sleepiness: Effects of 24-h of Sleep Deprivation on Waking Human Regional Brain Activity." Journal of Sleep Research **9**: 335-352.
65. US Nuclear Regulatory Commission (2000). "Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA)." NUREG-1624. US Nuclear Regulatory Commission, Washington, DC.
66. Venkatraman, V., Y. Chuah, et al. (2007). "Sleep Deprivation Elevates Expectation of Gains and Attenuates Response to Losses following Risky Decisions." Sleep **30**: 603-609.
67. Whaley, A., D. Kelly, et al. (2011) "SPAR-H Step-by-Step Guidance." INL/EXT-10-18533, Rev. 2, Idaho National Laboratory, Idaho Falls, Idaho.
68. Webb, W. (1985). "Experiments on Extended Performance: Repetition, Age, and Limited Sleep Periods." Behavior Research Methods, Instruments, & Computers **17**(1): 27-36.
69. Wesensten, N., W. Killgore, et al. (2005). "Performance and Alertness Effects of Caffeine, Dextroamphetamine, and Modafinil during Sleep Deprivation." Journal of Sleep Research **14**: 255-266.
70. Williams, H. and A. Lubin (1967). "Speeded Addition and Sleep Loss." Journal of Experimental Psychology **73**: 313-317.
71. Williams, J. (1986). HEART (Human Error Analysis and Reduction Technique) - A Proposed Method for Assessing and Reducing Human Error. 9th Advances in Reliability Technology Symposium, University of Bradford, U. K.
72. Yoo, S., P. Hu, et al. (2007). "A Deficit in the Ability to Form New Human Memories without Sleep." Nature Neuroscience **10**: 385-392.

Figure 1: Probability $P(T > k)$ for Reaction Time ($k = 500$)

Figure 2: Probability Ratios (P_T/P_C) for Reaction Time ($k = 500$)

Figure 3: Probability Ratios (P_T / P_C) for Reaction Time vs. Hours of Wakefulness (Approach 1)

Figure 4: Probability Ratios $P(T > C)/0.5$ for Reaction Time vs. Hours of Wakefulness (Approach 2)

Figure 5: Mean Probability Ratios and 95% Confidence Intervals of the Mean for Reaction Time ($k = 500$) (Approach 1)

Figure 6: Reaction Time vs. Hours of Wakefulness

Figure 7: Probability Ratios (P_T / P_C) for Number of Lapses vs. Hours of Wakefulness (Approach 1)

Figure 8: Probability ratios $P(T > C)/0.5$ for Number of Lapses vs. Hours of Wakefulness (Approach 2)

Figure 9: Mean Probability Ratios and 95% Confidence Interval of the Mean for Number of Lapses ($k = 1$) (Approach 1)

Figure 10: Number of Lapses (Test) vs. Hours of Wakefulness

Figure 11: Probability Ratios $P(T > C)/0.5$ for Accuracy vs. Hours of Wakefulness (Approach 2)

Figure 12: Mean Probability Ratios and 95% Confidence Interval of the Mean for Accuracy (Approach 2)

Figure 13: Accuracy Data (Test Condition Values) vs. Hours of Wakefulness

Table 1: THERP Table 20-14, Estimated HEPs in detecting stuck locally operated valves (Swain and Guttman, 1983)

Given that a locally operated valve sticks as it is being changed or restored,* the operator fails to notice the sticking valve, when it has:

Item	Potential Errors	HEP	EF
(1)	A position indicator ** only	0.001	3
(2)	A position indicator ** and a rising stem	0.002	3
(3)	A rising stem but no position indicator**	0.005	3
(4)	Neither rising stem nor position indicator**	0.01	3

* Equipment reliability specialists have estimated that the probability of a valve sticking in this manner is approximately 0.001 per manipulation, with an error factor of 10.

** A position indicator incorporates a scale that indicates the position of the valve relative to a fully opened or fully closed position. A rising stem qualifies as a position indicator if there is a scale associated with it.

Table 2: SPAR-H Predefined PSFs

SPAR-H Predefined PSFs	
Time Available	Procedures
Stress	Ergonomics
Complexity	Fitness-for-Duty
Experience and Training	Work

Table 3: Studies Reporting Effect of Sleep Deprivation on Performance Measures

Study #	Reference	Performance Measure	Sample size	Experimental Conditions	Task Type
1	Caldwell et al., 2003	L	15	RM	PVT
2	Chee and Choo, 2004	A, R	13	RM	SRT
3	Choo et al., 2005	A, R	12	RM	SRT
4	Dinges and Powell, 1988	R	17	RM	VRT
5	Doran et al., 2001	R	15	RM	PVT
6	Drake et al., 2001	R	12	RM	PVT
7	Drummond and Brown, 2006	A, R	32	RM	PVT
8	Frey et al., 2004	A, L, R	25	RM	PVT
9	Jewett et al., 1999	L, R	32	2GC	SRT
10	Jones et al., 2006	R	32	RM	PVT
11	Kobbeltvedt et al., 2005	R	21	2GC	Military Task
12	Lamond and Dawson, 1999	R	22	RM	PVT
13	Lichter et al., 1989	R	6	RM	VRT
14	Lieberman et al., 2005	R	31	RM	RT
15	Monk, 1997	R	17	RM	VVT
16	Nilsson et al., 2005	R	11	2GC	SRT
17	Philip et al., 2004	L	10	RM	SRT
18	Powell et al., 2001	L, R	8	RM	PVT
19	Rajaraman et al., 2007	L, R	48	RM	PVT
20	Reznick and Folse, 1987	R	12	RM	Purdue Pegboard
21	Roach et al., 2006	R	16	RM	PVT
22	Robbins and Gottlieb, 1990	A, R	23	RM	TTT
23	Rosa and Bonnet, 1993	R	8	RM	RT
24	Sagaspe et al., 2006	A, R	12	RM	RT
25	Sanders and Reitsma, 1982	R	12	RM	RT
26	Scott et al., 2006	R	6	RM	SRT
27	Storer et al., 1989	R	26	RM	TTT
28	Swann et al., 2005	A, L	12	RM	PVT
29	Thomas et al., 2000	A, R	17	RM	PVT
30	Venkatraman et al., 2007	R	26	RM	RT
31	Webb, 1985	R	6	RM	RT
32	Wesenten et al., 2005	R	12	RM	PVT
33	Williams and Lubin, 1967	R	8	RM	RT
34	Yoo et al., 2007	A, R	14	2GC	Memory Task

Note: Performance measures investigated are: accuracy (A), number of lapses (L), and reaction time (R).

Experimental Design Conditions are: repeated measures (RM) and two group comparisons (2GC)

Task Types: Psychomotor vigilance task (PVT); simple reaction time (SRT); visual vigilance task (VVT); total time on task (TTT); reaction time (RT); and visual reaction time (VRT)

Table 4: Number of Data Points per Probability Ratio Calculation for $k = 250, 500, 750$, and 1000

	$k = 250$	$k = 500$	$k = 750$	$k = 1000$
No. of Probability Data Points	249	249	249	249
No. of Points with $P_T = 0$ or $P_C = 0$	139	150	172	177
No. of Probability Ratio Data Points	110	99	77	72
No. of Outliers Removed	2	15	6	8
No. of Probability Ratios Considered	108	84	71	64

Note: The results in this table are based on data from studies 2-16, 18-27, and 29-34.

Table 5: Probability Ratios for Reaction Time: Means and Standard Deviations

Threshold Value	< 20 hrs	20 to 40 hrs	40 to 60 hrs	60 to 80 hrs	> 80 hrs
Approach 1					
$k = 250$	1.075 (0.28)	1.152 (0.44)	1.042 (0.24)	1.139 (0.07)	1.156 (0.08)
$k = 500$	1.157 (0.87)	1.367 (1.03)	1.468 (0.40)	1.744 (0.20)	1.766 (0.14)
$k = 750$	1.226 (0.47)	2.229 (1.59)	2.511 (1.19)	3.489 (0.20)	3.482 (0.14)
$k = 1000$	1.581 (1.22)	4.703 (3.29)	5.536 (3.60)	8.596 (3.60)	8.384 (0.48)
Approach 2					
$P(T > C)/0.5$	1.024 (0.59)	1.331 (0.66)	1.639 (0.63)	1.634 (0.38)	1.537 (0.38)

Note: Standard deviations are shown within parentheses.

Table 6: Number of Data Points per Probability Ratio Calculation for Number of Lapses

	$k = 1$	$k = 3$	$k = 6$	$k = 9$
No. of Probability Data Points	68	68	68	68
No. of Points with $P_T = 0$ or $P_C = 0$	8	9	12	12
No. Probability Ratio Data Points	60	59	56	56
No. of Outliers Removed	1	4	7	40
No. Probability Ratios Considered	59	55	49	16

Note: The results in this table are based on data from studies 1, 8, 9, 17-19, and 28.

Table 7: Probability Ratios for Number of Lapses Means (Standard Deviations) over 20 Hr. Time Blocks

	< 20 hrs	20 to 40 hrs	40 to 60 hrs	60 to 80 hrs	> 80 hrs
Approach 1					
$k = 1$	0.937 (0.25)	1.120 (0.47)	0.897 (0.05)	0.937 (0.12)	0.998 (0.24)
$k = 3$	0.979 (0.58)	1.713 (1.25)	1.090 (0.07)	1.095 (0.07)	1.015 (0.06)
$k = 6$	1.583 (1.70)	1.769 (0.69)	2.692 (0.21)	2.666 (0.23)	2.339 (0.17)
$k = 9$	2.317 (2.63)	2.978 (3.79)			
Approach 2					
$P(T > C)/0.5$	1.122 (0.48)	1.400 (0.38)	1.528 (0.116)	1.591 (0.19)	1.657 (0.34)

Note: Standard deviations are shown in parentheses.

Table 8: Mean Probability Ratios and Standard Deviations for Accuracy (Approach 2)

Approach 2	< 20 hrs	20 to 40 hrs	40 to 60 hrs	60 to 80 hrs	> 80 hrs
P(T > C)/0.5	0.851 (0.27)	1.081 (0.27)	1.466 (0.29)	1.696 (0.01)	1.236 (0.42)

Note: The results in this table are based on studies 2, 3, 7, 8, 22, 24, 28, 29, and 34.

Table 9: PSF Multipliers in SPAR-H for Diagnosis

PSF	Description	PSF Level	Multiplier
Available Time	Time available to complete task	Inadequate time	P(failure) = 1
		Barely adequate time	10
		Nominal time	1
		Extra time	0.1
		Expansive time	0.01
Stress	Stress level of operator	Extreme	5
		High	2
		Nominal	1
Complexity	Complexity of task	Highly Complex	5
		Moderately Complex	2
		Nominal	1
		Obvious diagnosis	0.1
Experience/Training	Operator's experience level	Low	3
		Nominal	1
		High	0.5
Procedures	Existence and clarity of documented procedures	Not Available	50
		Incomplete	20
		Available, but poor	5
		Nominal	1
		Diagnosis system oriented	0.5
Ergonomics	Interaction between operator and equipment	Missing/Misleading	50
		Poor	10
		Nominal	1
		Good	0.5
Fitness for Duty	Mental and physical state	Unfit	P(failure) = 1
		Degraded fitness	5
		Nominal	1
Work Processes	Organizational factors	Poor	2
		Nominal	1
		Good	0.8

Table 10: PSF Multipliers in SPAR-H for Action

PSF	Description	PSF Level	Multiplier
Available Time	Time available to complete task	Inadequate time	P(failure) = 1
		Barely adequate time	10
		Nominal time	1
		Extra time	0.1
		Expansive time	0.01
Stress	Stress level of operator	Extreme	5
		High	2
		Nominal	1
Complexity	Complexity of task	Highly Complex	5
		Moderately Complex	2
		Nominal	1
Experience/Training	Operator's experience level	Low	3
		Nominal	1
		High	0.5
Procedures	Existence and clarity of documented procedures	Not Available	50
		Incomplete	20
		Available, but poor	5
		Nominal	1
		Diagnosis system oriented	0.5
Ergonomics	Interaction between operator and equipment	Missing/Misleading	50
		Poor	10
		Nominal	1
		Good	0.5
Fitness for Duty	Mental and physical state	Unfit	P(failure) = 1
		Degraded fitness	5
		Nominal	1
Work Processes	Organizational factors	Poor	5
		Nominal	1
		Good	0.5