SANDIA REPORT
SAND2007-5906
Unlimited Release
Printed September 2007

Experimental Methods to Validate Measures of Emotional State and Readiness for Duty in Critical Operations

Louise M. Weston

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2007-5906
Unlimited Release
Printed September 2007

Experimental Methods to Validate Measures of Emotional State and Readiness for Duty in Critical Operations

Louise M. Weston
Reliability Assessment and Human Factors
lmwesto@sandia.gov
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-0830

Abstract

A recent report on criticality accidents in nuclear facilities indicates that human error played a major role in a significant number of incidents with serious consequences and that some of these human errors may be related to the emotional state of the individual. A pre-shift test to detect a deleterious emotional state could reduce the occurrence of such errors in critical operations. The effectiveness of pre-shift testing is a challenge because of the need to gather predictive data in a relatively short test period and the potential occurrence of learning effects due to a requirement for frequent testing. This report reviews the different types of reliability and validity methods and testing and statistical analysis procedures to validate measures of emotional state. The ultimate value of a validation study depends upon the percentage of human errors in critical operations that are due to the emotional state of the individual. A review of the literature to identify the most promising predictors of emotional state for this application is highly recommended.

ACKNOWLEDGMENTS

The author thanks the following individuals for their review and comments on the report and for their additional contributions as follows.

Kathleen V. Diegert, Manager of Organization 12335, for providing program structure and for input on the scope of the project.

Courtney C. Dornburg for participating in interviews with subject matter experts and providing input on the scope of the project.

James C. Forsythe for developing the LDRD proposal and obtaining funding for this project.

Elaine M. Hinman-Sweeney for program guidance and management.

Tony J. Kreuch, Organization 3334, Clinical Psychologist/HRP Designated Psychologist and Employee Assistance Program (EAP) Coordinator for providing valuable information on the HRP Program at SNL, information on emotional states and their measurement, potential test populations, and study limitations.

CONTENTS

FIGURES

TABLES

NOMENCLATURE

ANOVA	Analysis of Variance
ANS	Autonomic Nervous System
AUDIT	Alcohol Use Disorders Identification Test
CPT	Continuous Performance Test
DOE	Department of Energy
EEG	Electroencephalograph
EMG	Electromyograph
EOG	Electrooculograph
HR	Heart Rate
HRP	Human Reliability Program
HRV	Heart Rate Variability
HSB	Human Studies Board
IRB	Institutional Review Board
MMPI	Minnesota Multiphasic Personality Interview
NPP	Nuclear Power Plant
PI	Principal Investigator
RF	Russian Federation
SNL	Sandia National Laboratories
UK	United Kingdom

1. INTRODUCTION

Many industries have some form of Human Reliability Program (HRP) to assure that workers in critical positions are psychologically and physically capable of performing their assignments in a reliable, safe, and secure manner. The assurance of human reliability is especially important in high consequence operations (such as Nuclear Power Plant and Security positions) where the safety and security of the individual and the local and surrounding populations may be placed at risk. Typically HRPs consist of physiological and psychological testing prior to placement in the job assignment as well as reevaluations at regular intervals.

At Sandia National Laboratories (SNL), the HRP is a DOE program for individuals in positions with access to certain materials, nuclear explosive devices, facilities, and programs.[1] The goal is to identify individuals who may have impaired judgment due to substance abuse, a physiological or psychological disorder, or a particular circumstance. The program consists of medical and psychological evaluations performed annually, Minnesota Multiphasic Personality Interview (MMPI) assessments performed triennially, and random drug and alcohol screens. A physical examination, drug and alcohol testing, and psychological evaluation are performed annually. The psychological evaluation consists of a semi-structured interview and mental status examination. Two self-report measures are used in the annual psychological evaluation. These are the OQ 45 Symptom Checklist that assesses current stress level (scores in depression and anxiety, interpersonal functioning and social role) and the Alcohol Use Disorders Identification Test (AUDIT) that is a 10-item screen for alcohol problems (hazardous and harmful alcohol use and possible dependence); both of these self-report tests are well-validated measures of current psychological status. The triennial psychological measure, the MMPI-2, is used to assess psychological functioning and fitness for duty. MMPI-2 norms are available for a variety of occupations that are involved in high-consequence operations such as airline pilots and security personnel. One advantage of the MMPI-2 is that it has a set of validity scales that are used to determine whether or not the individual is responding honestly and/or consistently.

In addition to the testing, HRP participants have supervisory reviews and DOE personnel security reviews. On a continuous basis, HRP participants are required to report physical or psychological conditions that require medication or treatment, behaviors or conditions that may affect reliability, prescription medications, non-prescription medications that may affect performance, and unusual behavior in other HRP-certified individuals. Furthermore, managers are required to observe HRP-certified staff daily and to report any unusual behavior. One other requirement is an eight-hour abstinence rule that prohibits individuals in specified HRP positions from consuming alcohol eight hours prior to scheduled work.

[1] Dr. Tony Kreuch, Clinical Psychologist/HRP Site Designated Psychologist, provided information on the SNL HRP.

1.1. Need for Determining Human Reliability Prior to Critical Operations

DOE-sponsored HRPs contain physical and psychological measures to detect potential sources of unreliability at regular periods, but the time between assessments can be relatively long. For example, with the exception of the random alcohol and drug screening, the SNL HRP requires individual assessment only once annually. A year is a relatively long period of time during which changes could occur in the physical or psychological condition of the individual. Although there is a continuous requirement for self-reporting of various events that could affect performance, individuals are on the honor system. Some individuals may be reluctant to report events that could affect their jobs or damage their image with their management or peers. In addition, self-reports and the assessment of their significance is a subjective process, possibly prone to both false negative and false positive errors.

In other cases, the individual may not even be aware of a developing problem. A number of physical diseases are relatively symptom-less and could progress without the knowledge of the individual. Similarly, an individual could be going through a life stress such as a death in the family or divorce that causes insomnia. The individual may not be aware that the sleep loss is affecting his or her behavior, especially if the insomnia has persisted for some time. Then there are very transient conditions such as staying up late with a sick family member or an argument with a peer that could cause temporary unreliable behavior that the individual may not think necessary to report. Although managers and other HRP participants are also required to report unusual behavior in an HRP-certified individual, there may not be any observable behaviors.

More important than the relatively long time in between assessments is the fact that the assessments are not tied to periods of critical operations. A recent report of criticality accidents at nuclear facilities indicates that the human element significantly contributed to many of the accidents (Ref. 1). Those criticality accidents are discussed in the following section. Accurate human reliability assessments of the individual prior to the performance of any high-consequence operation could identify unreliable states to management, leading them to restrict activities of these personnel, thus reducing the occurrence of human-caused accidents.

1.2. Summary of Criticality Accidents

A recent report by U.S. and Russian Federation (RF) Nuclear Criticality Specialists summarizes criticality accidents occurring in nuclear facilities from 1945 through 1999 (Ref. 1). The accidents occurred in the U.S., the RF, the United Kingdom (UK), and Japan. There were a total of 60 accidents, 22 of which occurred in process facilities and 38 during critical experiments or operations with research reactors.

In the process facilities, the operations involve fissile materials and physical and administrative controls are used to prevent critical or near-critical events from occurring. The criticality events were, therefore, unexpected. The operators in the process facilities are usually not trained in criticality physics.

The Nuclear Criticality Specialists reviewed each of the 22 process incidents and identified the causes of the accidents. In some cases, very serious consequences resulted, including 9 fatalities and amputated limbs in 3 survivors. Overall, the accidents had multiple and different causes, rather than a single type of cause. However, the human element played a significant role in the majority of the accidents. *Failure to follow procedures was the most commonly reported human error.* In some cases, individuals deliberately did not follow procedures. For example, in one case a shift supervisor deliberately violated procedures on two occasions and also changed the log to reflect acceptable conditions when the conditions violated regulations. As a second example, a shift supervisor reentered a building against the instructions of a radiation control supervisor; the shift supervisor lost his life. Other actions that involved human deficiencies were failures to notice abnormal conditions, communications errors, and inadequate supervisory monitoring of operations. Other causes of the accidents included training, equipment, and process deficiencies.

The relative contribution of the emotional state of the individual to the process accidents was not directly addressed by the report. However, emotional state could have contributed to some of the human deficiencies identified. On the other hand, monitoring of the emotional state will not reduce those accidents caused primarily by factors such as inadequate training, procedures, or equipment. Statistics cited in the report indicate that training, procedural, and equipment deficiencies have played an important role in the process accidents. Between the late 1950s and middle 1960s, there was about one accident per year in the RF and the U.S. This accident rate dropped by a factor of 10 to a rate of 1 accident per 10 years since the middle 1960s. Reasons given for the decline include:

1. Lessons learned.
2. Increase in management attention to criticality safety.
3. Presence of dedicated staff to control criticality hazards.
4. Documentation of critical mass data and operational good practices.

In the critical experiment and research reactor facilities, the personnel plan to achieve near-critical and critical configurations, and the operating personnel are usually experts in criticality physics. The planned operations are performed under shielded or remote conditions. As found with the process accidents, serious consequences occurred in many of these accidents (fatalities in 8 of the 38 accidents), and human error played a significant role in many of these accidents (Ref. 1, 2, 3). *Failure to follow procedures was the most commonly reported human error. Other significant factors identified were inadequate training and "overconfidence" (Ref. 2, 3).* Some of these errors could have been caused by unreliable behavior related to the emotional state of the individual. Emotional conditions identified were "too much excitement, or absent-mindedness, or total lack of concentration and automatism in performance of critical operations."

In conclusion, the analyses of the criticality accidents at nuclear facilities suggest that a method to detect a deleterious emotional state prior to the performance of critical operations could reduce the occurrence of some accidents with serious safety and security consequences.

2. OBJECTIVE AND SCOPE OF TEST METHODS

The overall objective of this report is to present a set of experimental principles and methods to assess the reliability and validity of measures to predict a deleterious emotional state and personnel readiness for duty in high-consequence operations. Generally, test reliability refers to the consistency or the degree to which the test yields the same results on repeated measurements. Although there are different types of validity, a general definition of validity is the extent to which the test measures what it is intended to measure (Ref. 4, 5). The concepts of test reliability and validity are discussed at length in Sections 4 and 5. Test design variables, issues and approaches are presented in Section 6.

In this report, a deleterious emotional state is defined as a condition of the individual that is hypothesized to increase the likelihood of unreliable behavior. Unreliable behavior may result in various types of human error as discussed in the following section.

2.1. Types of Human Unreliability or Human Error

In the SNL HRP, Human Reliability is defined as "the ability to adhere to security and safety rules and regulations." Swain (Ref. 6) defines Human Reliability as the "probability of successful performance of a mission" and Human Error as "an out-of-tolerance action, where the limits of tolerable performance are defined by the system." Both individual and work factors contribute to human error. Examples of individual factors are level of training, personality type, innate capability, and emotional state. Examples of work factors are procedures, equipment design, and task load.

Human errors can be categorized as unintentional or intentional (Ref. 6). As shown in Figure 1, an unintentional error is when an individual inadvertently performs an incorrect action or forgets to perform a step or procedure. Intentional errors are of two types. An individual can make a *mistake* by performing an action that is incorrect when the individual thinks the action is correct. In the second type of intentional error, the individual *intentionally violates the procedures* by not performing the procedure as written. The reason for the violation could be to save time or because the individual thinks a different procedure is superior for some reason. In the latter cases, the individual does not expect undesirable consequences to result.

Serious incidents in high-consequence operations can also be caused by malevolent behavior, which are actions performed for the purpose of sabotage. This could be an act by an insider with the intent of causing some harmful effect. Malevolent acts are not included in the definition of human error, since such acts are not error, but planned actions to achieve some type of undesirable outcome (Ref. 6).

The purpose of outlining the different types of human error is to point out that a deleterious emotional state may be more relevant to some types of human error than others. For example, unintentional human errors appear to be more likely to be related to emotional state than intentional human errors. For example, sleep deprivation or depression may cause the individual to be less alert and more likely to commit an unintentional incorrect action or forget to perform a

step. On the other hand, it is less clear how an individual's emotional state could contribute to some intentional errors. A case in point is where the individual deviates from procedures because the individual thinks another procedure is adequate or even superior. The latter may be more related to training deficiencies or the individual's personality type, neither of which may result in a detectable deleterious emotional state, especially when measured with a short-duration test instrument. Similarly, a malevolent insider may not display signs of a deleterious emotional state such as distress; by the process of self-selection, a malevolent insider may be a personality type that is more immune to showing signs of distress. The measures used to predict a deleterious psychological state, referred to as the Predictor Measures, are discussed in the following sections.

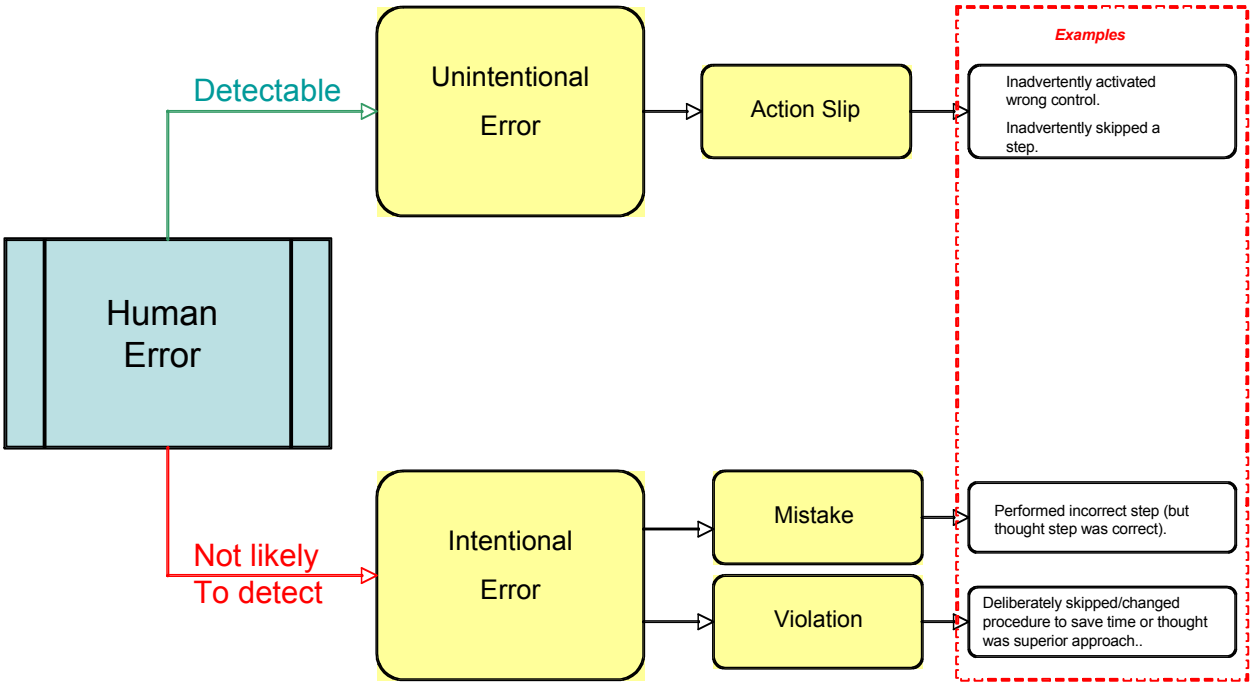

Figure 1. Types of Human Error (Based on Swain, Ref. 6).

2.2. Predictor Measure Requirements and Issues

There are a number of requirements that must be met for the Predictor Measures for deleterious emotional state to be useful in this application. A basic requirement is that the measures must be administered prior to the performance of any high-consequence operation. For example, in a nuclear facility, testing may be required prior to each shift, resulting in daily testing for operations personnel. A problem that this raises is a potential *learning effect* when some types of tests are given repeatedly and with relatively short inter-test intervals. The learning effect refers to a change in the individual's responses in accuracy, content, or speed with repeated measurement that is due to the testing itself. For example, if the test is a self-report, the individual could learn the acceptable responses or the responses that permit continued performance of duties with repeated testing. Thus, with repeated administration, the test would

be expected to become less predictive of the true emotional state of the individual. As a second example, if the test is a knowledge-based test, the individual could learn the correct responses over time and again the test would be expected to be less predictive over time. To some extent, use of alternative forms would mitigate the learning effect. Alternate forms are equivalent measures of the same thing. Alternate forms would not, however, be a viable solution when testing is required relatively often, since the number of alternate forms would be large. Some types of measures would be expected to be less subject to a learning effect such as physiological measures or ability/skill type measures than self-report or knowledge-based tests. Therefore, test selection is limited by the susceptibility of the test to learning effects when testing is required relatively frequently.

Another basic requirement for this application is that the test is relatively short in duration, i.e., 15 minutes or less testing time (Ref. 3). As will be discussed in a subsequent section of the report, the short test time may raise issues for test reliability as well as for validity due to the relatively small sampling of the physiological or cognitive behavior being obtained. Therefore, there may be limitations on how short the test can be to produce reliable and valid results.

Other desirable characteristics noted in a recent Russian proposal for nuclear facility pre-shift monitoring are as follows (Ref. 3):

1. The test can be group administered so that multiple operators can be tested simultaneously.
2. The test can be objectively scored.
3. The test is relatively simple. It does not require sophisticated instrumentation, extensive measurements, or complex data processing algorithms.
4. The test is not intimidating to personnel.
5. The test is flexible, allowing adjustment of normal limits for an individual over time.
6. A medical assistant or staff member with appropriate training can administer the test.

The Russian proposal noted one important limitation associated with pre-shift emotional state testing. Because the prediction period may be relatively long (a shift of operations), the condition of the individual could change at some point after testing has been completed. For example, there could be events that occur after the pre-shift testing period that could produce a deleterious emotional state in the individual. Along those lines, one can imagine that an event could happen while performing operations during the shift that upset the individual. To predict performance at such times, a measure of emotional stability or resiliency (a personality trait) is recommended in that proposal in addition to the measure of emotional state.

2.3. General Types of Predictor Measures

A deleterious emotional state is defined as a condition that is hypothesized to increase the likelihood of unreliable behavior. When defining emotional state, many people think of negative conditions or events such as sleep deprivation or depression. However, positive events could potentially also result in unreliable behavior. For example, an individual could be distracted due to a life event such as an upcoming marriage or birth of a child. Therefore, the goal of the

predictor measures is to detect an emotional condition in the individual that is indicative of a possible performance deficiency.

In a recent report, the U.S. military reviewed methods to continuously monitor the physical and mental status of individuals in field environments (Ref. 7). This type of monitoring is needed in combat situations to indicate impending performance deficiencies so that such deficiencies could be corrected or individuals could be reassigned to maintain readiness and proficiency at an acceptable level. The military is interested in a wide range of physical and cognitive deficiencies, some of which are relevant to the current problem and will be discussed later in this report. Both physiological and behavioral measures are proposed. The behavioral measures could be one of two types, self report or performance measures. Although self-report measures are easy to obtain and some have been found useful in certain situations, there are a number of limitations with them that make them questionable for our purposes. The physiological measures are especially important in situations where the behavioral manifestations of some underlying physiological condition are time dependent and may not be evident in behavioral tests in the pre-shift test period.

2.4. Limitations of Validation Methods

The intent of this test plan is to present principles and methods to validate any type of instrument that measures an emotional condition assumed to increase the likelihood of unreliable behavior. As indicated earlier, the emotional condition could be negative (some form of emotional distress) or positive (excitement over a positive event such as an upcoming marriage). *However, the focus of this report will be on forms of emotional distress, since this state is of highest interest.*

As suggested earlier, the eventual value of the testing effort is limited by the nature of the error that we are attempting to prevent as well as the characteristics of the individual responsible for the error. One limitation is that the error is associated with the emotional state of the individual and not other causes. As noted earlier, there are many types of human error that are caused by conditions other than the state of the individual, including deficiencies in management, training, procedures and equipment. Therefore, detecting a deleterious emotional state will not be cost effective if this type of cause is not a significant factor in unreliability in critical operations. The second limitation pertains to the type of error. As indicated earlier, unintentional errors are more likely to be related to emotional state than are intentional errors. Thirdly, these procedures are not applicable to the actions of a malevolent insider or individuals with certain personality profiles.

3. DEFINITIONS AND PREDICTOR MEASURES OF EMOTIONAL DISTRESS

One purpose of this section is to define the types of emotional distress that might be important to monitor for the prediction of unreliable behavior in high consequence operations. The theory underlying the predictor measures and the types of predictor measures that show some promise for measuring these types of emotional distress are also discussed.

3.1 Relevant Types of Emotional Distress

To define the relevant emotional states, the SNL HRP Site Designated Psychologist was interviewed. This individual is a Clinical Psychologist who is responsible for the initial and continuing psychological evaluation of HRP participants.[2]

Two important constraints are important when defining the emotional states. First, there are two basic types of individual characteristics, *state and trait characteristics*. State characteristics are temporary internal fluctuations in an individual, such as being depressed or angry. Trait characteristics, on the other hand, are relatively permanent or long-standing internal characteristics, such as being outgoing or friendly. States are generally measurable with a short test instrument, whereas, traits are not. For example, Antisocial Personality Disorder (also referred to as psychopathy, sociopathy, and dyssocial personality disorder) is characterized by the traits of deceit and manipulation (Ref. 8, 9, 10). Because of the complex and multidimensional nature of personality disorders (and with traits in general), psychologists typically utilize comprehensive tests such as the MMPI-2 that are well-validated with such diagnostic groups rather than shorter, "screening" instruments that often do not have tests of validity and are more subject to error. Therefore, the relevant emotional states that can be monitored with short test instruments are *only those that reflect state characteristics.*

The second constraint has been noted earlier. *The unreliable behaviors that we may be able to predict are those that are unintentional.* Many safety-related incidents fall into this category. Intentional unreliable behaviors (such as deliberate and planned actions) are unlikely to be predicted through the monitoring of emotional state. Actions of a malevolent insider would fall into the latter category.

The question that the HRP Site Designated Psychologist was asked was: *"What are the types of emotional states that might be important to monitor to predict safety unreliability and which are measurable using a short test time (15 minutes)?"* The responses were as follows:

1. Alcohol disorder.
2. Anger impulsivity.
3. Anxiety.

[2] Dr. Tony Kreuch provided information.

4. Depression.
5. Mood instability.
6. Psychotic disorder.
7. Stress.
8. Suicidal tendency.

Research indicates that sleep deprivation and fatigue are additional conditions that can significantly affect mood as well as cognitive performance (Ref. 7).

3.2 Potential Predictors of Emotional Distress

The Committee on Metabolic Monitoring for Military Field Applications recently published an overview of the current status of our capabilities of monitoring the physical and mental status of military personnel in field situations (Ref. 7). The results of a workshop and an extensive literature review were published in their report and will form the basis for much of the information in this section.

A brief discussion of the underlying theory might be useful to better understand the potential predictors of emotional distress. It is assumed that metabolic processes form the foundation for the human's adaptation and response to the environment. As such, these regulatory processes are thought to be the earliest indicators of the state of the individual and most sensitive to changes in the state of the individual. The individual (and team members) may not even be aware of the development of an unreliable condition due to excessive or prolonged stress, sleep deprivation, or fatigue. Therefore, understanding these regulatory functions could lead to significant predictors or *biomarkers* of the physiological and cognitive status of the individual. It is hoped that by developing technologies to measure these biomarkers in a reliable fashion, we will be able to predict impending failures in the human's adaptive response and avoid, mitigate, or correct those failures prior to serious operational consequences. The military considers two types of biomarkers, physiological and behavioral, to assess the status of the individual.

In the military scenario, there are many areas of concern that do not apply to the current problem. For example, bone and muscle metabolism, kidney function, and hydration are highly important to assess the status of individuals in sustained combat operations, but not likely to be as important for individuals in the civilian high-consequence shift operations being considered here. However, there are three areas of mutual concern:

1. Physiological Biomarkers of Stress
2. Physiological Biomarkers of Cognitive Status
3. Behavioral Biomarkers of Operational Status

The major findings in each of the three research areas are summarized in the subsequent subsections.

3.2.1 Physiological Biomarkers of Stress

Stressors are defined as adjustive demands that are placed on an individual. Stress results from an inability to adequately cope with stressors. The adjustive demands can be either positive such as an upcoming childbirth ("eustress") or negative such as a death in the family ("distress"). Although both eustress and distress can be problematic, distress is typically considered to be more damaging (Ref. 10). There are both psychological (stress perception) and physiological effects of stress. The physiological effects include the release of neurotransmitters and hormones to regulate the responses of the immune and other systems. The physiological effects form the basis for the biomarkers of interest.

One such biomarker is heart rate variability (HRV). The Autonomic Nervous System (ANS) has two major branches, the sympathetic and parasympathetic system. The sympathetic system is associated with energy mobilization and responsible for the "fight or flight" response to a stressor. The parasympathetic branch is responsible for restorative functions to bring the body back to homeostasis. Normally, these two systems are in a dynamic balance. Sympathetic activity is generally higher during daytime and parasympathetic activity increases during nighttime. ANS imbalance results when one of the branches dominates over the other branch. One cause of ANS imbalance is persistent negative emotions; in such cases, the sympathetic system dominates over the parasympathetic system. The imbalance results in energy demands on the system that become excessive when the imbalance is prolonged. This is supported by research showing an association between ANS imbalance (typically sympathetic dominance) and pathological conditions (Ref. 7). HRV is considered to be an indication of ANS balance. Specifically, high HRV indicates parasympathetic control (Ref. 7, 11). Increased parasympathetic activity is associated with more pronounced acceleration and deceleration and more variable intervals between heartbeats, or higher HRV. In addition, research indicates an association between decreased HRV and various physical and psychological disease conditions, including cardio-vascular disease, diabetes, anxiety, depression, hostility, and post-traumatic stress disorder (Ref. 7).

A second potential biomarker of stress is the level of cortisol. Cortisol is a hormone that is released by the adrenal gland and functions to make energy stores available throughout the system. Cortisol levels rise almost immediately after a stressful event. The effects of cortisol are adaptive when it is acutely released, but they may become harmful when cortisol is chronically released. Effects reported with chronic release include insulin resistance, osteopenia/osteoporosis and excessive fear (Ref. 7). Results, however, appear to be more variable with this biomarker than with HRV. For example, research indicates that psychological stimuli can result in an increase or a decrease in cortisol level. Factors such as the characteristics of the stress event and the individual's perception of the stress event could contribute to this variability.

3.2.2 Physiological Biomarkers of Cognitive Status

There are a number of research areas that are relevant to assess cognitive status in high consequence operations as follows:

1. Sleep Quantity and Quality Measures
2. Heart Rate Measures
3. Eye Movement and Electroencephalogram (EEG) Measures

Each of these will be discussed in subsequent subsections.

3.2.2.1. Sleep Quantity and Quality Measures

Sleep deprivation can produce a variety of effects on human performance. There are many variables that determine the size of the effect including the degree of sleep loss as well as task and individual variables. Generally, research indicates the following possible outcomes (Ref. 7):

1. Attention lapses
2. Carelessness
3. Degraded verbal communication skills
4. Impaired judgment
5. Mood disturbances
6. Motivational decrements
7. Perceptual disturbances
8. Reduced physical endurance
9. Short-term memory loss
10. Slower reaction times

Some of the important task variables are the length of the task, whether or not there is performance feedback, task complexity, short-term memory requirements, task familiarity, and task boredom. In addition, there are large individual differences in vulnerability to performance effects from sleep loss. Inter-individual variability has been suggested as one reason for the relatively inaccurate prediction of performance decrements from sleep loss.

In addition to the quantity of sleep, the quality of sleep is also important. Research indicates that frequent sleep interruptions can have as much of an effect on performance as reduced hours of sleep.

Although there is sufficient evidence to suggest that sleep deficiencies are important factors in mood and performance, physiological measurement methods require lengthy recordings and are not suitable for a pre-shift measurement. The available methods are:

1. Polysomnographic recordings that involve the collection of EEG, electromyographic (EMG), and electrooculographic (EOG) data from skin-mounted electrodes. Typically, recordings are taken overnight in a sleep laboratory.
2. Actigraphy. Actigraphs are battery-powered wrist units that permit recording in non-laboratory settings.

Although the actigraphs provide a portable measure of sleep/wakefulness periods that are reported to be reasonably accurate when compared to polysomnographic recordings, it is also not suitable for our application because of the length of the recording that is required.

3.2.2.2. Heart Rate Measures

There are two measures that are predictive of cognitive performance, heart rate (HR) and HRV.

Research indicates an association between HR and the level of cognitive demand in several operational contexts (Ref. 7). For example:

1. HR has been associated with performance in flying combat missions and flying surface-attack training missions.
2. In flying, HR discriminates between pilot and co-pilot roles and between lead and wing positions.
3. Increased HR was associated with increased task difficulty in flight simulators.

Earlier, it was noted that HRV appears to be an important measure for detecting an imbalance in the ANS, and it is a biomarker for stress. The Committee on Metabolic Monitoring for Military Field Applications made the following conclusion:

"Measuring heart-rate variability should be considered as an accurate, sensitive, and noninvasive way to measure the relative activity of the sympathetic and parasympathetic nervous systems."

Research also suggests that HRV is associated with task demands in operational contexts (Ref. 7). For example, HRV was associated with task demands in simulated flight tasks and HRV declined during take off and landing when compared to cruise segments of flying. Several laboratory studies suggest that HRV is a significant predictor of cognitive performance. For example, in dental phobics, higher HRV was associated with faster reaction times on tasks containing threat-related words. Research has also shown that naval cadets with higher HRV perform better at tasks involving working memory than cadets with lower HRV.

Laboratory studies also suggest that HR to some extent, but most notably HRV, is an index of self-regulatory strength (Ref. 11). *Self-regulatory strength is important in inhibition of impulses, control of emotions, decision-making, and persistence at difficult tasks.* Like muscle strength, self-regulatory strength is subject to fatigue, and individuals differ in their baseline levels of self-regulatory strength. *Higher HRV is hypothesized to be associated with higher self-regulatory strength.* The proposed relationship between HRV and self-regulatory strength has some theoretical support since the brain structures involved in self-regulation and ANS regulation overlap. For a number of tasks, HRV was found to be higher in tasks requiring high self-regulation (or inhibition) than tasks requiring low self-regulation. In addition, individuals with high baseline HRV persisted at solving difficult/impossible anagram problems (requiring self-regulation) longer than did those with lower baseline HRV.

3.2.2.3. Eye Movement and Electroencephalogram (EEG) Measures

There are two other categories of measures that show some promise in predicting cognitive status (Ref. 7). However, both require significant instrumentation and longer recording periods, and evidence on their capability to predict operational performance is weaker than for measures discussed earlier.

Both saccadic eye movements and eye blinks are measures that have been studied to detect fatigue. Saccadic eye movements are the movements of the eye from one focal point to another. Research in this area suggests that fatigue can produce longer saccade latency (time between stimulus and saccade), reduced saccade velocity, and reduced accuracy (undershooting or overshooting the target). Smaller-amplitude eye blinks and reduced eye blink frequencies are also associated with fatigue.

EEG has been monitored through a variety of scalp locations using small electrodes. There are some studies showing that EEG measurements are sensitive to sleep deprivation and fatigue. However, the relationships between EEG measurements and operational performance are weak, except in cases of extreme sleepiness (Ref. 7). In addition, the EEG presents major measurement issues and more research on obtaining reliable and sensitive measures is needed.

3.2.3 Behavioral Biomarkers of Operational Status

The behavioral biomarkers of operational status fall into two categories, *cognitive skills tests and self-assessment.* In the military application, cognitive skills testing is difficult for a number of reasons, such as the large number of different job specialties that would need to be represented and the difficulties of testing in the field with rapidly changing environments (Ref. 7). However, this type of testing is applicable to fixed environments representative of many critical operations facilities.

The SNL HRP Site Designated Psychologist indicated that the relevant cognitive skills that could be measured are attention, working memory (ability to mentally shift and multi-task), distractibility, mental flexibility, processing speed, and short-term memory. Some tests that could be used to measure these skills include a performance test to evaluate sustained attention/distractibility (such as Connors' Continuous Performance Test (CPT)), parts of the Wechsler Scales for working memory, and the Trails Test for Mental Flexibility. However, most measures for these cognitive skills require formal administration and are probably not practicable for pre-shift testing. The CPT is an exception, since it could potentially be set up on a laptop computer and self-administered.

A job sample approach is commonly used in industry for the selection of personnel (Ref. 4, 13). In developing the job sample, it is critical that the tasks and actual job conditions for the different specialties are adequately represented. Driving tests are familiar examples of job sample tests. In applying the job sample approach to predicting human error in critical operations, emphasis could be given to those aspects of job performance that are critical in preventing accidents. This type of approach has been recommended in a Russian proposal for nuclear facility pre-shift

monitoring (Ref. 3). A "Professional Efficiency Test" is described in which the skills required of the job are represented. Emphasis is placed upon skill testing, rather than information testing, to minimize the learning effect.

Self-assessment or self-report is widely used for screening and selection of personnel, in part because of the ease of administration. There are a number of self-report tests that could be administered to assess emotional distress. Some examples are listed in Table 1. As indicated earlier, two of these, the OQ 45 Symptom Checklist and AUDIT, are used in the annual SNL HRP psychological assessment. There are a number of problems associated with these tests. First, there is the problem of a learning effect when the test is administered often and with short inter-test intervals. Most of these tests do not have alternate forms to mitigate the learning effects. The second problem is that multiple tests might be required to measure multiple states of interest, and this could result in a lengthy pre-shift test time.

Table 1. Self Report Tests for Measuring Emotional Distress[3]

Type(s) of Emotional State/Conditions	Applicable Self-Report Measure
Alcohol problems	AUDIT (10 items) (Ref. 15)
Both state and trait anger /impulse control	State-Trait Anger Expression Inventory (57 items) (Ref. 16)
Anxiety	State-Trait Anxiety Inventory (40 items) (Ref. 17)
10 clinical problem domains (negative affect, hostile control, acting out, suicidal thinking, health problems, alienation, psychotic features, alcohol problem, social withdrawal, anger control)	Personality Assessment Screener (22 items) (Ref. 18)
Current stress level	OQ 45 Symptom Checklist (Ref. 19)
Current stress level	Perceived Stress Scale (14 items) (Ref. 20)

The Committee on Metabolic Monitoring for Military Field Applications reports that some self-report measures have shown relationships with behavior in some circumstances (Ref. 7). These include subjective ratings of alertness, fatigue, sleep quality, sleepiness, and mood states. A notable example of the successful use of self-reports is in sports medicine, where self-report and peer-report measures have often been found to be superior to physiological measures in predicting physical performance. However, self-report measures suffer from numerous problems in our intended application as follows:

1. There is a potential learning effect when such tests are administered relatively frequently and with short inter-test intervals.
2. In chronic conditions, the individual may adapt to the state and not be aware of a deteriorating condition. For example, it is reported that sleep-deprived individuals cannot

[3] Information provided by SNL HRP Site Designated Psychologist.

accurately assess their level of sleepiness and fatigue after the first day or two of sleep deprivation (Ref. 7).

3. Motivational factors and peer and supervisor pressures can affect self-reports, especially if an individual will be removed from job duties as a consequence of reporting a condition or situation.

4. TEST RELIABILITY

Test reliability is defined as the degree to which scores for a set of individuals are consistent on repeated measurements of the same characteristic on the same individuals (Ref. 4, 5, 13). It indicates the degree to which individual differences in test scores can be attributable to true differences among individuals in the characteristic being measured and to chance or random error. Random error occurs in all measurements to some degree. It is unsystematic error that is equally likely to occur above and below the true value and with the same magnitudes. Therefore, the average of the positive and negative error is expected to be zero in the long run. The sources of random error include measurement error and uncontrolled or chance variations in the testing environment or within the individual during the repeated measurements.

The correlation coefficient (reliability coefficient) is used to express the consistency or degree of association between sets of scores. Reliability is computed across individuals. Therefore, high reliability would mean that if an individual has the highest score on one measure of a characteristic, that individual would be among the highest scorers on the second measure of the characteristic. An individual's measurements will not be identical on multiple measurements, but they will be relatively consistent. Generally, the greater the consistency of the measurements, or the less the random error in the measurements, the higher is the reliability coefficient.

There are four different types of test reliability. They differ in their testing requirements and sources of error variance and, therefore, have different strengths and limitations. Each of these types of reliability is discussed below.

The primary importance of estimating the reliability of a predictor measure is that reliability places a limitation on the size of the validity coefficient (discussed in Section 5). Basically, a test cannot have high *validity* if it doesn't have *high reliability*.

4.1 Test-Retest Reliability

The simplest and most obvious form of reliability is *Test-Retest Reliability*. It is determined by repeating the same test on all individuals on two different occasions. The reliability coefficient is the correlation between the two sets of scores for all individuals on the two test administrations. The error variance in this situation is associated with random differences in the test environment and within the individual from testing at two different times (Time Sampling).

A critical variable in Test-Retest Reliability is the interval of time between the two test administrations. Test-retest reliability declines as the inter-test interval increases (Ref. 4, 13). This is most likely due to the occurrence of non-random fluctuations that occur over longer periods of time. Therefore, the time between the testing sessions should always be reported along with the reliability coefficient. In addition, any intervening experiences of the test population, such as changes in education or employment, should also be reported. Short inter-test intervals

on the order of two weeks to one month (never more than 6 months) are recommended (Ref. 4, 5, 13).

There are several limitations associated with this form of reliability. The first is the requirement for two test administrations and the uncertainty of the appropriate inter-test interval. The second limitation is that the method is subject to learning or practice effects, i.e., responses on the first administration of the test affect the responses on the second administration of the test. One can anticipate that the shorter the inter-test interval, the greater the memory of the questions/tasks and responses from the first administration, and perhaps the greater the learning effect. Therefore, the two measurements are not independent and the learning factor could result in an inflated reliability coefficient. The size of this effect will depend upon the nature of the test. For example, self-report and knowledge-based measures may be more susceptible to the learning effect than physiological or skill-based behavioral measures.

4.2 Alternate Forms Reliability

In *Alternate Form Reliability*, two comparable forms of the same test are developed. The two forms could be administered in immediate succession (Immediate Administration) or could be separated by a period of time (Delayed Administration). The reliability coefficient in this case is the correlation between the scores for all individuals on the two forms of the test.

In the Immediate Administration Procedure, there is only one source of error variance, variation due to different items on the two forms of the test (Content Sampling). In the Delayed Administration Procedure, both time sampling and content sampling contribute to error. As with Test-Retest Reliability, the inter-test time interval is an important variable that should be reported.

One of the difficulties with this type of reliability procedure is the difficulty of developing two forms that are truly parallel. Nominally, the two forms should contain the same number of items and have the same range of content, with equivalent difficulty levels and time limits. A second limitation with Alternate Forms Reliability is a potential learning effect. Although the use of alternate forms (rather than a single form) will reduce the learning effect, it does not eliminate it. This could result in an inflated reliability coefficient. As with Test-Retest Reliability, the size of the learning effect will depend upon the nature of the test (self-report or knowledge-based tests versus physiological or ability-based tests).

4.3 Split-Half Reliability

In *Split-Half Reliability*, two scores are obtained for each person on one test, by splitting the test into comparable halves. The reliability coefficient is the correlation between the scores for all individuals on the two halves of the test. Without any adjustment, this coefficient would be the reliability for only a half test. This is important because test consistency is positively correlated with test length (Ref. 4, 13). Formulas are available to adjust the reliability coefficient for the

full test length. The only source of error variance in this type of reliability is from content or item sampling.

One variable in Split-Half Reliability is the method used to obtain two comparable halves of the test. Obviously, a division into the first half and second half would not be advisable due to potential differences in difficulty level as well as other effects such as practice and fatigue. Typically, the procedure used is to divide the test into odd and even halves.

4.4 Internal Consistency Method

In the fourth type of reliability, the *Internal Consistency Method*, the reliability coefficient is based upon the average inter-item correlation. Thus, this reliability measure is based on all of the test items. Only one test administration is required for this procedure.

This type of reliability has two sources of error variance. One source is content sampling (as in alternate form and split-half reliabilities). The second type is the heterogeneity of the behavior being sampled by the items. The more heterogeneous the items, the lower the reliability. Internal Consistency is also a function of the number of test items, with reliability increasing as the number of test items increase.

4.5 Summary of Reliability Methods

The characteristics of the four reliability methods are summarized in Table 2. As shown, the methods differ in their test requirements, their sources of error variance, and their limitations. The choice of method will depend in part upon a consideration of practical factors such as the feasibility of multiple test sessions and the availability of alternate forms. In addition, the nature of the characteristic being measured is important. For example, the susceptibility of the measurement to a learning effect will be a function of the type of measurement. As noted earlier, self-report and information-based tests may be more susceptible to a learning effect than physiological and skill-based tests.

The reliability coefficient that results from these measures represents the percentage of score variance attributable to individual differences in the characteristic being measured by the test. For example, a reliability coefficient of .80 indicates that 80% of the variance is attributable to individual differences, and 20% of the variance is due to sources of error variance.

There are a number of group factors that affect the size of the reliability coefficient. One important factor is the range of the individual differences of the group used to measure the characteristic of interest. For example, if all members of a group have similar scores score on an ability test, the ability test scores for that group cannot correlate very highly with any other set of scores for that group. Homogeneous groups will have lower reliability coefficients than will heterogeneous groups (Ref. 13). Other characteristics of the group such as age, gender, educational level, and occupation are also important. Therefore, reliability coefficients should be

accompanied by a description of the group used to determine them. Furthermore, the sample used to estimate reliability should have characteristics similar to the population of interest.

The primary importance of estimating the reliability of a predictor measure is that reliability places a limitation on the size of the validity coefficient (discussed in the following section). Basically, a test cannot have high *validity* if it doesn't have *high reliability*. Although the adequacy of the size of the reliability coefficient is somewhat situation-dependent, generally, it is desirable to have a coefficient of .80 or greater (Ref. 4, 5).

Table 2. Characteristics of Different Types of Test Reliability

Type of Test Reliability	Test Requirements	Sources of Error Variance	Limitations
Test-Retest	One test form. Two test sessions.	Time sampling.	Two test administrations required. Size of inter-test interval is a variable. Likely/potential learning effect.
Alternate Forms – *Immediate Administration*	Two test forms. One test session.	Content sampling.	Potential learning effect. Two comparable test forms required. *
Alternate Forms – *Delayed Administration*	Two test forms. Two test sessions.	Time sampling. Content sampling.	Two test administrations required. Size of inter-test interval is a variable. Potential learning effect. Two comparable test forms required. *
Split-Half	One test form. One test session.	Content sampling.	Two comparable halves required. * Reliability coefficient could vary with different methods of developing two halves.
Internal Consistency	One test form. One test session.	Content sampling. Content heterogeneity.	Reliability coefficient varies with degree of item heterogeneity.

* The alternate forms or split halves are designed so that the range of content is equivalent.

5. TEST VALIDITY

Test validity is defined as the extent to which the test measures what it is intended to measure (Ref. 4, 5, 12, 13). Test reliability is a necessary, but not sufficient condition for validity. A test can have high reliability (consistent scores) and low validity (little or no relationship to the characteristic of interest). Validity does not have meaning apart from its purpose. A test may have high validity for one purpose (such as predicting party affiliation) and low validity for another purpose (such as predicting voting behavior). Validity issues occur when there is some source of nonrandom error. Nonrandom error refers to the presence of one or more factors that have a biasing effect on the measurement. The nonrandom error exists in the test measure in addition to (or instead of) the characteristic of interest and in addition to random error.

There are four different types of test validity. Basically, all types are measures of the relationships between test scores and some other observable facts about the characteristic of interest. The different types of validity differ in the questions that they address, their analysis procedures, their applications, and their limitations.

5.1 Face Validity

Face validity is not validity in the technical sense, but is included here for completeness. It refers to whether or not a test superficially appears to measure what it is designed to measure. The opinions of the test takers, test administrators, and other technically untrained observers are used to determine face validity. Face validity is a desirable feature of a test, since seemingly irrelevant items on a test may elicit poor cooperation and unreliable performance from the test takers. Therefore, this is a subjective rather than objective form of validity.

5.2 Content Validity

Content validity answers the question of whether or not a test adequately represents the domain that it is designed to measure. Content validity procedures are frequently used in achievement tests. However, the procedures can also be used to develop criterion or performance tests for other forms of validity procedures. In constructing a test, all of the relevant topics and processes of the behavioral domain must be specified as well as the relative importance or weight of each. Topics/processes with higher importance require more test items than those that with lesser importance. In addition, it is equally important that the test does not contain items for areas that are irrelevant to the domain.

Construction of the test may involve examination of relevant course materials or a job analysis by subject matter experts. Based on these analyses, topics and processes are defined as well as their relative importance, and the number of items in each area is determined.

Empirical (experimental) procedures may also be used to further examine the content validity of a test. For example, if age-related changes in performance are expected on an achievement test, the scores on the test could be analyzed as a function of age to determine if the changes are as expected. Similarly, when the procedures are used to construct criterion tests such as job samples, the types of errors made on the test could be compared with errors committed on the job. To determine whether the test includes irrelevant factors, test scores can be compared with scores on a test for the irrelevant factor. For example, if ability to read instructions is considered an irrelevant factor on a mechanical skills test, scores on the mechanical skills test could be compared with scores on a reading comprehension test to see if they are uncorrelated.

Content validity is valuable in many situations, but it has a several limitations. Although useful for proficiency tests, content validity procedures are inappropriate for more abstract domains such as aptitudes and personality where the domains are more difficult to define. Secondly, there is no agreed-upon objective criterion for determining whether a test has sufficient content validity.

5.3 Construct Validity

As noted above, content validity is not appropriate for abstract theoretical concepts where the domains cannot be easily defined. Tests of abstract concepts can be developed using *construct validation procedures*. The distinguishing feature of construct validation is that it is theory-based. Construct validation answers the question of whether or not a test adequately measures a theoretical construct or trait. Examples of such constructs are intelligence, anxiety, self-esteem, and neuroticism.

Construct validation procedures assess whether a test measure is related to other measures in a manner that is consistent with hypotheses derived from theory. The first step is to define the construct, its expected characteristics and relationships to other variables, and its theoretical relationships to other constructs or traits. The next step is to empirically assess the hypothesized characteristics and relationships. Methods that could be used to empirically demonstrate construct validation are:

1. Analysis of test scores as a function of important variables to validate expected relationships. For example, if scores were expected to increase with age, analysis of scores as a function of age would validate that hypothesis.
2. Analysis of the correlation between test measures on the current test and scores on other tests that purport to measure the same construct. This is especially useful where the current test is an alteration of another test to achieve some benefit, as when a shortened version of a test is needed for a particular application.
3. Analysis of the correlation between test measures on the current test and scores on other tests that measure traits that are hypothesized to be *unrelated* to the concept being measured.
4. Factor analysis procedures to identify the major and common traits (factors) associated with the test scores as well as the relative importance (weight) of the factors.

5. Examination of the internal consistency of the test scores to determine whether the test scores are homogeneous. Homogeneity is taken as evidence of a common construct or trait. For example, if a test has several subtests, scores on each subtest could be correlated with the total test score; those subtests with low correlations with the total score are removed from the test.
6. Empirical studies to determine whether the effects of variables on test scores correspond to theoretical expectations. For example, to validate a test that is designed to measure anxiety, individuals could be assessed with the anxiety test both before and after exposure to a stressful or demanding task. The expectation would be that scores on the test would be higher after than prior to performance of the stressful task.

Construct validity is demonstrated, not by one piece of evidence, but by multiple pieces of evidence from hypothesis testing. The more complicated the concept, the more hypotheses that can be generated, and the more evidence that is required to demonstrate validation.

5.4 Criterion-Related Validity

Criterion-related validity is what is classically meant by the term validity. It answers the question of whether or not a test can predict some behavior that is external to the test. Unlike construct validity, criterion-related validity is predominantly empirical in nature and is atheoretical. Although there is some theory involved in choosing the criterion measures, the emphasis is on the empirical demonstration of the predicted relationship.

The external behavior, the criterion, can be either future behavior or present behavior. *Predictive criterion validity* refers to the assessment of the relationship between a test measure and future behavior. *Concurrent criterion validity* refers to the assessment of the relationship between a test measure and the behavior at the same point in time.

The major task for demonstrating criterion-related validity is the development of the criterion measures. In *predictive validity*, measures such as performance in training or actual job performance are often used. In this approach, individuals are given the test, and then test scores are compared with criterion measures that are obtained at a later date. One disadvantage of this approach is the amount of time it takes to collect sufficient criterion behavior. Other problems that have been reported with this approach include (Ref. 4, 13):

1. It is difficult to obtain criterion data that are reliable and comprehensive.
2. The number of employees is often too small for a statistical study.
3. There is a restricted range through pre-selection for the job. Only those individuals who are hired are tested. A restricted range on the predictor or and/or the criterion scores will lower the correlation between the predictor and the criterion measure (Ref. 4, 13).

Concurrent criterion validity procedures are used as a substitute for predictive validity procedures often because they are more practicable. In some cases, tests are administered to individuals who already have criterion data available. For example, training records, existing scores on aptitude tests, existing measures on similar tests, or existing measures of job

performance could be used. Another method is to use the contrasted groups approach where group membership is the criterion. For example, in the development of some personality tests, psychiatric diagnosis has been used to establish the validity of the tests. Yet another approach is to develop a behavioral test and test the individuals on the behavioral test at the same time that they are tested on the predictor measure. One such criterion measure would be to develop a job sample test. To develop this type of test, a job analysis would be required where critical aspects of the job are identified and sampled in the criterion test.

A major challenge for criterion-related validity is the difficulty of developing the criterion measures. Often the criterion behavior is highly complex as in predicting job success. Job success is dependent upon many factors, some of which may be difficult to represent in a job sample test. The reliability and content validity of the criterion measure itself is as important as the predictor measure. Evidence comes from studies showing wide variations in the correlation of a single type of test with criteria of job proficiency (Ref. 4). Some of the variation is due to variation in the predictor measures and differences in the groups tested. However, a major part of the variability is also due to differences in the criterion measures of job proficiency. Other studies show that criterion-related validity varies over time when predicting job performance. One explanation for this is that the traits required for successful performance vary with the amount of experience in the job (Ref. 4).

5.5 Summary of Validity Methods

The characteristics of the four validity measures are summarized in Table 3. As shown, the methods differ in the questions they are designed to answer, the analysis procedures used to assess validity, their potential applications and their limitations. In spite of these differences, the different forms of validity are not totally distinct. For example, some of the procedures for establishing content and criterion-related validity can be used to assess construct validity. As a second example, content validity is important in the construction of all tests, both predictor and criterion tests.

In evaluating these methods, construct validity procedures appear to be more widely applicable to the social sciences. Content validity is not applicable to abstract domains, and there is no objective criterion for assessing whether validity is sufficient or not. The difficulty of defining the criterion measures, especially in abstract concepts such as personality, limits the usefulness of criterion-related validity. Construct validity procedures, on the other hand are widely applicable, and validity can be objectively determined. Construct validity procedures can be applied to investigate the validity of criterion measures used in criterion-referenced validity, and it is an alternative approach to assess validity in situations where criterion measures are not available.

The validity coefficient, most often discussed with reference to criterion-referenced validity, is the objective metric. In criterion-referenced validity, the coefficient is defined as the correlation between a test score and a criterion measure. Factors that affect the size of the validity coefficient are similar to those already discussed in the section on reliability. As in the case of reliability, there are a number of group factors that could affect the validity coefficient. These

include group homogeneity, and characteristics of the group such as age, gender, educational level, and occupation. Therefore, validity coefficients should be accompanied by a description of the group used to determine them. In addition, the sample used to estimate validity should have characteristics similar to the population of interest. Another factor that can affect the size of the validity coefficient is the passage of time. As was also noted earlier in the discussion of criterion-related validity, the validity coefficient could change over time because of changes in conditions such as the experience level of the individuals.

The Pearson Product-Moment Correlation Coefficient is often used to determine the relationship between a predictor and criterion measure when both variables are continuous. However, this statistic has several assumptions that must be met for its accuracy. One important assumption is that there is a linear relationship between the predictor and the criterion measure. A second important assumption is homoscedasticity or equal variability about the regression line for all values of the predictor variable. The square of the correlation coefficient represents the proportion of variance in the scores attributable to variance in the predictor variable. Thus, if the validity coefficient is .80, the proportion of variance accounted for by the predictor is 64%. Finally, an important question is how high a validity coefficient should be. The validity coefficient should, of course, be statistically significant at some acceptable level of significance (.05 or .01). Beyond that, the adequacy of the size of the validity coefficient depends upon the particular application. Much higher correlations are required to accurately predict an individual's exact criterion score than if the intent is to determine if an individual will exceed a predictor cutoff score (Ref. 4).

Table 3. Characteristics of Different Types of Test Validity

Type of Test Validity	Question Addressed	Analysis Procedure(s)	Potential Applications	Limitations
Face Validity	Does the test appear to measure what it is designed to measure?	Collect opinions of technically untrained observers.	Applies to all tests.	Not a technical form of validity.
Content Validity	Does the test measure a representative sample of the domain to be measured?	Subject matter experts define the topics and processes in the domain from course descriptions or job analyses. Subject matter experts judge adequacy with which the test represents the domain. Empirical procedures to determine if test scores are consistent with expected relationships to other variables.	Achievement tests. Occupational tests. Criterion tests.	Inappropriate for abstract domains. No objective criterion for adequate validity.
Construct Validity	Does the test measure a theoretical construct or attribute?	Define the construct, its hypothesized relationships with other variables and with other constructs or traits. Empirical procedures to determine if hypothesized relationships to variables and other constructs are as expected. Factor analysis to determine primary and common factors in the test. Measures of internal consistency to determine if test items are homogeneous.	Measurement of abstract theoretical concepts or traits such as anxiety or neuroticism.	Requires theoretical framework.

Table 3 (Cont). Characteristics of Different Types of Test Validity

Type of Test Validity	Question Addressed	Analysis Procedure(s)	Potential Applications	Limitations
Criterion-Related Validity	*Predictive Validity:* Does the test predict future behavior?	Determine criterion behaviors: Performance in Training. Actual job performance.	Personnel Selection And Classification. Screening tests.	Difficult to establish criterion measures, especially for abstract concepts.
	Concurrent Validity: Does the test predict current behavior?	Determine criterion behaviors: Training records. Achievement tests. Existing measures of job performance. Other similar test measures. Job sample tests. Contrasted groups approach where group membership is the criterion. Empirical procedures to determine if test scores are correlated with criterion measures.		

6. TEST DESIGN

The test design information in this section is most directly applicable to the measurement of reliability and criterion-referenced validation, although much of it can be applied to content and construct validity as well. One purpose of this section is to summarize the types of predictor variables, independent variables, and dependent or criterion variables. A second purpose is to discuss the potential subpopulations to consider for application of a contrasted groups approach to reliability and validity assessment. Thirdly, various statistical comparisons and procedures for reliability and validity assessment and for development of a predictor cutoff score are outlined.

6.1 Predictor Variables

Potential predictor variables for emotional distress have been discussed in earlier sections of this report and will only be summarized here. Basically there are several options as follows:

1. *Physiological Measures* – Potential measures include physiological biomarkers of stress and physiological biomarkers of cognitive status discussed earlier. From the limited information reviewed, HRV appears to have the most research showing construct validation for the detection of emotional distress. For example, HRV research suggests that HRV is an indicator of ANS imbalance and has been associated with various distressful conditions such as anxiety, depression, hostility and post-traumatic stress disorder (Ref. 7). In addition, studies indicate that HRV is also associated with task demands in operational settings and is an indicator of self-regulatory strength (Ref. 11). An advantage of the physiological measures is that they are generally considered to be less subject to learning effects than behavioral measures.

2. *Performance Measures* – One category of performance measures is cognitive skills testing. Examples include measures of attention, working memory, distractibility, process speed, and short-term memory. Another type of test that could be used is a job-sample test. An example of the latter is a "Professional Efficiency Test" recommended in a Russian proposal for pre-shift monitoring of nuclear facility personnel (Ref. 3).

3. *Self-Assessment Measures* – There are a number of self-report measures that could be used to assess emotional state (see Table 1). The advantage of this type of measure is that it is easy and quick to administer. Although self-report measures have been shown to be useful in a variety of areas, there are a number of problems associated with these tests for the application at hand. Self-report measures are subject to a learning effect when administered often and with short inter-test intervals. Secondly, motivational factors and peer and supervisor pressures can affect them, especially in an application where the outcome may be that an individual is removed from duty that day. Additionally, in an individual with a chronic condition, the person may not even be aware of a deteriorating condition due to adaptation to the condition.

In choosing the type of measure, the constraints for this application must be considered. In particular, the requirement for a relatively short test time (15 minutes or less) and relatively frequent test administration (e.g., daily) play major roles. For example, it is quite possible that a physiological or behavioral measure may not be reliable unless sampled for periods much longer than the 15-minute constraint. In addition, some tests are more susceptible to a learning effect than others. Generally, physiological and ability based measures may be less subject to learning effects than knowledge-based and self-report measures.

Other desirable features for predictor tests are that the tests can be group administered, objectively scored, are relatively easy to set up and score, permit adjustment of cutoff scores over time, and can be administered by a medical assistant or staff member with appropriate training (Ref. 3).

6.2 Independent Variables

There are a number of independent variables that may affect reliability as well as validity assessments. The first variable is the duration of measurement. Some predictor measures may require longer periods of measurement for reliability and may not be suitable for the current application. Although a measurement time of 15 minutes or shorter is desirable, it would be worthwhile to obtain measurements for longer periods of time (e.g., 60 minutes), so that reliability and validity can be examined as a function of this variable. In this manner, the optimum measurement duration can be determined, which may be shorter or well beyond the operational requirement.

A second variable is the time of day that the measurement is taken. This variable is important for some physiological measures that have cyclical patterns or have circadian rhythms.

A third variable is the duration or stage of the emotional state. For example, in some cases, effects may not be observable until the condition has persisted for some time and has become chronic. In other cases, it could be argued that the emotional disruption may be greatest at an early stage of the situation, such as in a divorce or death in the family.

A fourth variable may be the type of emotional state. For example, the type of distress (e.g., depression, anxiety) or type of life stress (e.g., financial hardship, divorce) may have a bearing on the predictor measure that is most effective and/or the extent and type of the behavioral effect.

In addition, there are many group variables that can affect the size of the reliability and validity coefficients. One important factor is group homogeneity; homogeneous groups will have lower reliability and validity coefficients than will heterogeneous groups (Ref. 13). In addition, there are a number of group factors such as age, gender, ethnic background, educational level, and job experience that may affect the size of the coefficients. These factors are often referred to as Moderator Variables or characteristics of individuals that affect the predictive power of a test instrument. For example, an examination of the coefficient for different subgroups may reveal a high coefficient for some subgroups and negligible ones for others. If the validity coefficient is

computed across all subgroups of individuals, a validity coefficient may be too low to be of much practical significance.

Therefore, reliability and validity coefficients should be accompanied by descriptions of the groups used to determine them. In addition, the samples used to estimate reliability and validity should have characteristics similar to the population of interest.

It may not be possible to control for the effects of many of these variables due to the large number of potential conditions that could result. However, variables (such as duration of stressful event, type of stress, group characteristics) should be recorded so that such variables can be considered in statistical analyses.

6.3 Dependent Variables

As with the predictor measures, the dependent or criterion measures may be one of the following types:

1. *Physiological Measures* – An example would be correlating a physiological predictor measure with another physiological measure (criterion variable) that is thought to measure the same characteristic. This might be done to determine if a shorter test (predictor) could substitute for the recognized standard (and longer) test for that characteristic. This would be an example of construct validity.
2. *Performance Measures* – These measures could be cognitive tests (e.g., measures of attention, working memory, distractibility, process speed, short-term memory), job-sample tests, training performance or on-the-job performance.
3. *Self-Assessment Measures* – Self-assessment measures could also be used, although the limitations of these tests as discussed earlier would apply.

The dependent measures listed above could be used in the process for determining content, construct, or criterion-related validity.

Often, not enough attention is given to the development of the performance or criterion measures. In addition to demonstrating reliability, a performance test should be examined for its content validity, the extent to which the test adequately measures the domain of interest. For example, in developing a job-sample test that represents behaviors likely to be disrupted by emotional distress, task variables are especially important. One type of task variable is whether the task involves skill, rule, or knowledge-based behaviors. Skill-based behaviors are highly rehearsed automatic behaviors that primarily involve stored patterns of behavior (e.g., manipulating a control). Rule-based behaviors require more conscious effort and involve the use of stored or written rules (e.g., calibration task). Knowledge-based behaviors are the most complex, often involving unfamiliar tasks where considerably more cognition is required for the operator to make decisions (Ref. 6). Swain (Ref. 6) suggests that all tasks that require some mental involvement (rule-based and knowledge-based tasks) are likely to be degraded under stress. As a second example, greater disruption from emotional distress would be expected under

higher than lower workload conditions. Therefore, it is important that any performance test captures those aspects of behavior that are most likely to be affected by the emotional state.

In the situation where on-the-job performance is the criterion behavior, classification of the type of error is essential. As discussed earlier, there are many sources of error, some of which are related to an individual's emotional state and some of which are related to other types of deficiencies such as training, equipment, and procedural deficiencies. Therefore, an attempt to correlate a predictor with any and all errors might result in a validity coefficient that is of little practical significance. It would be worthwhile to keep a log of the conditions surrounding the error to attempt to determine its primary or root cause. In addition to the type of error, the time that the error was made is also important. As indicated earlier, there may be a long time between the pre-shift measurement of emotional state and the time when the error occurred. Events could have occurred during the shift to change the emotional state of the individual.

One other variable for many of the criterion measures is the test duration. As discussed with predictor measures, the test time for the criterion measure may affect the reliability and validity of the study. Therefore, criterion test duration is a variable that should be considered in the test design as well.

6.4 Test Populations for Contrasted Groups Approach

A major difficulty with the validation of a predictor of emotional state is finding relevant populations for study. A common approach to the validation of predictors of complex and abstract traits, such as personality, is the Contrasted Groups Approach. This approach is highly relevant to this application, since emotional distress is both complex and abstract.

In the Contrasted Groups Approach, group membership is used as the criterion. The scores of one group of individuals are compared to the scores of another group of individuals who are assumed to differ on the characteristic of interest. For example, in the development of some personality tests, psychiatric diagnosis has been used as evidence of test validity and has also been used in the actual selection of test items. In this application, the goal would be to identify groups of individuals who are undergoing some form of emotional distress (criterion group) and compare their responses to an unselected population (standard group).

Sources for the "emotionally distressed" group might be Outpatient Counseling Centers where the participation from individuals who are undergoing counseling for adjustment to life stresses or for conditions such as depression or anxiety might be obtained. Outpatient Counseling Centers from within the company would provide the best source of individuals with the appropriate demographics. The "normal" or standard group could also be obtained from within the company. This would represent an unselected group. A questionnaire could be administered to individuals in the standard group to obtain information on the presence of sources of emotional distress (see Table 1 for types of self-report measures). If sufficient subjects cannot be obtained from within the company, outside companies with similar occupations/demographics could also be considered. In using the Contrasted Groups Approach, group variables must be

considered. At a very minimum, gender and age must be adequately sampled to represent the reference population (Ref. 7).

Both the reliability as well as the construct or criterion validity of the predictor measure could be assessed in this manner. The importance of the Contrasted Groups approach is that it will demonstrate whether a predictor measure shows consistent results and whether it discriminates between the major groups of interest. If a predictor measure does not discriminate between the criterion and standard group, then it will not likely be a predictor of behavior that is assumed to be associated with emotional distress of the type for which the criterion group was selected.

It may be possible to form a criterion and standard group of individuals using a questionnaire approach or a self-report test (see Table 1 for potential self-report measures). Using the questionnaire or self-report test information, individuals could then be classified into the criterion and standard group of individuals. As a word of caution, individuals classified as "emotionally distressed" in this manner may not be as extreme as individuals who have been selected using the Contrasted Groups Approach where individuals are undergoing counseling for emotional distress.

6.5 Statistical Comparisons and Procedures

In this section, statistical comparisons and procedures to achieve the following goals are discussed:

1. Determine the reliability and validity of a predictor measure using the contrasted groups approach.
2. Determine the relationship between the predictor measure and the criterion measure.
3. Develop a cutoff score for the predictor measure.

Each area is discussed in the following subsections.

To illustrate the procedures, a hypothetical example will be used where appropriate. The example will be the detection of transient depression from a life event. The predictor that will be used is HRV, since there is a theoretical basis for this and research indicates an association between decreased HRV and depression (Ref. 7). The criterion will be errors in Nuclear Power Plant (NPP) operations that could be associated with emotional state.

6.5.1 Reliability and Validity of Predictor Measure: Contrasted Groups Approach

The first step in a validation procedure is to determine the reliability of the predictor measure for the groups of interest. The major groups of interest are the criterion depressed group and the standard (no reported depression) group. The groups could be obtained as indicated above in the Contrasted Groups Approach, preferably from within the company. Reliability coefficients would be calculated for both the Criterion and Standard Group and subgroups as possible (gender and age variables). Subgroups are empirically identified groups by age, gender or other demographic factors. Since the predictor is physiological (HRV), split-half reliability or test-

retest reliability could be used. To examine the effect of the duration of test measurement, HRV could be monitored for an extended duration (1-2 hours). The effect of the length of the test interval used to compute HRV could then be estimated.

Given that the reliability coefficients are satisfactory for the groups and subgroups of interest, the next step would be to determine whether the predictor measure discriminates between the individuals in the criterion and standard groups. Statistical comparisons using analysis of variance (ANOVA) could be performed to determine whether HRV differs significantly for criterion and standard groups of individuals as expected. In these analyses, group factors such as gender and age, would be examined to the degree possible. The overall objective is to determine whether there are significant differences between the criterion and standard populations in HRV across various group variables. Such an analysis could indicate, for example, that HRV discriminates the criterion and standard category, but only for males or only for a certain age group. The effect of the length of the measurement period to arrive at the HRV measure would also be examined in these analyses.

Information on the reliability and discriminating power (validity) of the predictor measure may be available from existing studies of criterion and standard groups. The relevance of existing information will depend upon the similarity of the group characteristics in existing studies to those of the population of interest. Application of existing studies to the reliability and validity of the predictor measure and population under current consideration must be justified.

6.5.2 Relationship Between the Predictor Measure and the Criterion Measure

Once the predictor measure has been shown to demonstrate reliability and adequate discrimination between criterion and standard populations, the next step is to estimate the degree of association between the predictor and the ultimate criterion. The ultimate criterion in this hypothetical example is the frequency of certain types of errors on NPP tasks, i.e., errors that could be associated with an emotional state, in this case transient depression. The major groups of interest are the criterion group (depressed group) and the standard (no reported depression) group.

There are a number of approaches that can be taken to estimate *concurrent criterion-related validity*, the relationship between a predictor and current behavior. One of the simplest approaches would be to determine if there is a relationship between the predictor measure and another physiological or behavioral measure that has already demonstrated criterion-related validity. In this case, the new test would have some advantage over existing tests such as being simpler and/or quicker to administer. Individuals for this validation testing could be obtained as indicated earlier in the Contrasted Groups Approach, preferably from within the company. The individuals would need to have characteristics that are similar to the population of interest, including the type and amount of training and experience.

A second approach is to develop a behavioral criterion test that could be administered to the criterion and standard groups and correlated with the HRV predictor measure. Procedures discussed for content validity would be useful to develop a job-sample type of test that could be

delivered in a NPP simulator. First, a job analysis by subject-matter experts would be required to identify the skills and tasks that might be associated with the emotional state of the individual. For example, emphasis might be placed on rule and knowledge-based tasks, since those types of tasks require mental processes that are expected to degrade under stress (Ref. 6). In addition, the relative importance of the various skills and tasks would need to be determined so that more important areas are sampled more heavily than less important areas. For example, those skills and tasks that are required more often or those that can lead to more costly safety and/or security consequences might be given higher importance weights. Once a test is constructed, empirical procedures could be used to provide evidence on the content validity of the test. For example, performance on the test might be expected to improve with increasing levels of experience, a relationship that can be studied. In addition, errors committed on the job-sample test could be correlated with errors committed on the job. Once the test has sufficient content validity, scores on the job-sample test could be correlated with scores on the predictor measure to establish the degree of association. The groups could be obtained as indicated above in the Contrasted Groups Approach, preferably from within the company. The individuals for the study would need to have characteristics that are similar to the population of interest, including the type and amount of training and experience.

Another possible method to examine *concurrent criterion-related validity* will be briefly discussed. In this paradigm, a stressor such as a task stress (e.g., insolvable problems) is introduced to create an "emotionally distressed" condition in individuals. Pre-stress and post-stress measures of the predictor and criterion (job-sample) measures are then compared as evidence of validity. However, there are several weaknesses associated with this type of approach. First, it is uncertain whether task stress would be a sufficient stressor when compared to the level of stress associated with significant life events. Secondly, a paradigm where individuals are exposed to stressors may raise Human Studies Board (HSB) concerns.

Another approach would be to estimate *predictive criterion-related validity* by correlating the predictor score (HRV) with on-the-job performance using a sample of individuals from the relevant industry and population of workers. There are a number of requirements associated with this approach as follows:

1. First, the predictor measure such as HRV would need to be recorded prior to each shift for each individual.
2. Secondly, it would be advisable to obtain independent verification of the emotional state (depressed or not) if possible. Although self-report measures could be used (see Table 1), such measures are not suitable for repeated measurement due to a likely learning effect. Independent verification, although advisable, is not required if evidence for the relationship between the HRV predictor and the emotional state (depressed or not) has already been established using the Contrasted Groups Approach.
3. The third requirement is that a log must be kept of the performance for each individual on each shift. Information on the errors committed and the root cause for each is necessary. Determination of root cause is required to rule out sources of error that are not associated with the emotional distress (depression) of the individual. Examples are errors due to faulty procedures, equipment, and training.

4. In addition to the error, the time that the error occurred is also important, since many hours may have elapsed between the HRV measure and the occurrence of the error.

One problem with this approach is that an individual may have a value for the predictor measure (HRV) that, according to group norms, potentially indicates emotional distress. The dilemma is whether this individual should be permitted to perform duties as usual or not. A second problem with this approach is that each individual's performance must be tracked on each shift. Observers can be used to obtain this information, however, the presence of the observer or the operator's knowledge that he or she is being observed may change the behavior. Other weaknesses typically associated with industrial studies are that the sample sizes are too small for statistically significant results, and often the correlations will be lowered by restricted range problems since the operators are already a selected group. The small sample problem may be addressed by repeating the study at other facilities having similar personnel and tasks. The major problem with this approach, however, is the amount of time that will be required to obtain adequate data for analyses. It may take excessive time for sufficient instances of both an emotional state such as depression to occur as well as sufficient errors to be able to perform a meaningful statistical analysis.

6.5.3 Development of Predictor Cutoff Scores

There are two basic methods to determine cutoff scores. Cutoff scores can be determined using groups norms or individual baselines. Each will be discussed below.

In the group norm method, the distributions of the scores on the predictor measure (such as HRV) are determined for a criterion (emotionally distressed) and standard (not emotionally distressed) groups. Criterion and standard groups of individuals could be operationally defined by group membership as defined by the Contrasted Groups Approach. The Criterion Group consists of individuals who are undergoing counseling, whereas the Standard Group consists of unselected individuals from the population of interest. As an assurance, individuals in the unselected group could be given a questionnaire to determine if they are undergoing some type of life stress or condition such as anxiety or depression. Also available are scores for the Criterion and Standard Group on the criterion measure such as performance on a job-sample test. A cutoff score on the criterion measure is also determined. The choice of the predictor cutoff score will depend upon the acceptable levels of various types of errors, as shown below.

Figure 2 shows a hypothetical distribution of predictor scores and criterion scores for a sample size of 100 where the correlation between the two variables is approximately .70 (Ref. 13). The criterion cutoff score, indicated by the bold horizontal (blue) line, indicates the minimum value of the criterion score for job success. The bold vertical (red) line indicates the predictor cutoff score. There are four possible outcomes indicated in the table, two are correct decisions (valid acceptance and valid rejection) and two are incorrect decisions (false acceptance and false rejection). Decreasing one type of error (such as false acceptance) results in an increase in the other type of error (false rejection). The choice of predictor cutoff score is a value judgment, based upon the relative importance of the two types of error. For example, in high-consequence operations, the predictor cutoff should be set high enough to exclude most possible false

acceptances. Other factors that may be important in determining the predictor cutoff score are the number of available personnel and the number of personnel required for the day's activities.

There are a number of potential problems with the group norm method for this application. First, group data may be too imprecise unless information is available on the relevant subpopulations of interest. As indicated earlier, variables such as gender, age, educational level, job experience level and cultural background may be important determinants of an individual's response. Secondly, there may be large individual differences in the predictor measures such that an individual's baseline physiological or behavioral biomarker values may fall outside the normal range. Individuals may not only differ in their average responses, but also in their response variability. Thirdly, some measures may have temporal components such as circadian rhythms or other types of variation that are specific to the individual. For these reasons, an approach based on an individual baseline is suggested.

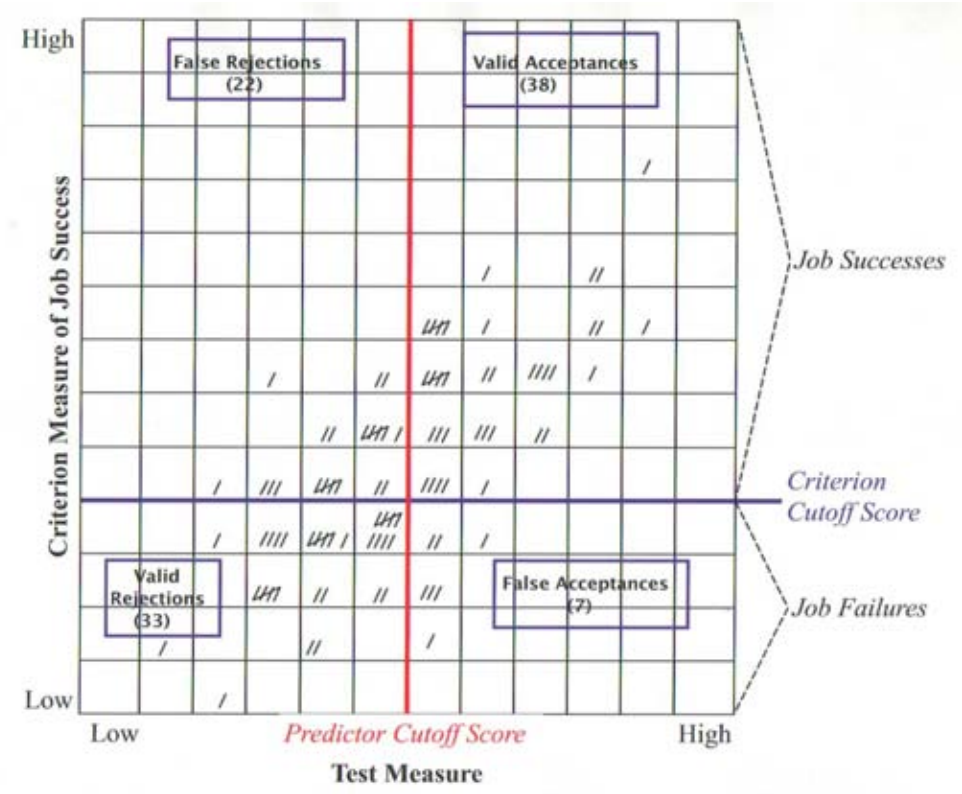

Figure 2. Hypothetical Bivariate Distribution of Predictor Test Scores and Criterion Measure of Job Success (N = 100, Correlation = .70) (Ref. 13).

The Committee on Metabolic Monitoring for Military Field Applications suggests a two-phase approach (Ref. 7). Initially, protocol data of normal and abnormal ranges could be used. Subsequently, individual baseline measurements and predictor cutoff scores should be developed. Some test batteries set predictor cutoff scores based upon the standard deviation

from the individual's average predictor score, such as 1.5 standard deviations to indicate an alert condition and 2 standard deviations to indicate a safety-critical change (Ref. 7).

When the criterion behavior is complex, more than one test may be needed to accurately predict the behavior. A test battery consists of multiple tests to predict a single criterion behavior (Ref. 13). Often the different tests measure different aspects of behavior or different characteristics. The question that arises is how the information from the different tests is used to arrive at a decision. One method is multiple regression where the individual's predicted criterion score is based on an individual's scores on all tests in the battery. The multiple regression equation is based upon the partial correlation of each test with the criterion behavior as well as the intercorrelations among the tests. Each test is assigned a weight with the highest weight being assigned to the test with the highest validity and the least amount of overlap with the other tests in the battery. The specific regression weights are dependent upon the group characteristics of the sample tested.

Another approach to analyzing a test battery is to use multiple cutoff scores, consisting of a minimum score for each test. In this instance, only individuals who meet the cutoff scores on all tests are accepted. This approach is preferred where meeting cutoff criteria on all of the tests is essential. In multiple regression, an individual may obtain an acceptable total score even with a low score on one of the tests, because a high score on one test in the battery can compensate for a low score on another test. Multiple regression, however, will lead to more accurate decisions in cases where the individual need not have all of the skills or characteristics in the test battery. A combination of procedures can be used in such cases. The multiple cutoff score approach is used followed by the multiple regression approach for those who meet all of the essential test criteria.

7. INSTITUTIONAL REVIEW BOARD (IRB) REQUIREMENTS

There are a number of requirements for human subject research that must be met prior to undertaking the reliability and validation processes defined in this document. Every institution that conducts research with federal support is required to have an assurance with the Office of Human Research Protections that defines the regulations and ethical principles it will employ to protect research subjects. Responsibility for reviewing all proposed human research and authority to approve or disapprove such research lies with Institutional Review Boards (IRBs). At SNL, the IRB is called the Human Studies Board (HSB). The stated purpose of the SNL HSB is "to assure that risks to human research subjects are minimized and reasonable in relation to the anticipated benefits, and to protect the rights and welfare of research subjects in accordance with applicable federal regulations, state laws, DOE directives and SNL policy." A general review of the SNL HSB procedures follows.

The HSB process applies to all human subject research activities that use SNL funds, facilities, or personnel. The first step in the HSB process is to determine whether the proposed activity is human subjects research. The following two definitions must apply for a study to be human subjects research:

1. Research is a systematic investigation that contributes to generalizable knowledge. Information is generalizable when new information is added to an existing body of knowledge or is applied to different populations or settings.
2. A human subject is a living person from whom an individual gets either data through intervention or interaction with the person or identifiable private information or materials.

Given that the definitions above apply, the next step is to determine the type of review that is required. The types are:

1. Exempt – There are six categories of research that are exempt. As an example, research conducted in established educational settings involving normal educational practices is exempt. The HSB conducts a preliminary review to determine if the research is exempt.
2. Expedited - Research that presents no more than minimal risk to the subjects. "No more than minimal risk" means that the probability and degree of harm in the research is no greater than what the individual ordinarily experiences in everyday life. In addition, the research must be one of a defined set of research categories; examples are some surveys, and research on individual or group characteristics and behavior. The HSB chair, a designated voting member, or a group of voting members of the HSB can conduct this level of review.
3. Full Review – This is the highest level of review and requires a convened meeting of a quorum of the HSB members.

The principal investigator (PI) at SNL is required to take training to understand PI responsibilities and HSB procedures, and to prepare a protocol review package for HSB review. The review package includes an abstract, research protocol, informed consent forms, and any

subject recruitment material. Once a protocol has been approved through expedited or full review, it must be reviewed again at least every twelve months.

An important part of any human subject research is the intended subject population. There are a number of vulnerable populations that require special attention for their protection. Vulnerable populations consist of individuals who are less able to protect themselves and are more susceptible to coercion or undue influence. Coercion refers to the use of a threat of harm or force to control another person. Undue influence refers to using a position of power to control another person. Common examples of vulnerable populations are children, prisoners, pregnant women, handicapped or mentally disabled persons, and economically or educationally disadvantaged populations.

The worker population is a central population of interest in the reliability and validation of measures of emotional distress. For a number of reasons, the worker population is also considered a vulnerable population. One type of vulnerability is referred to as "paycheck vulnerability." Employers or unions may encourage worker participation, and failure to participate may threaten an employee's career. In some cases, workers may be discouraged from participation or told to respond in a particular manner. In addition, the study findings themselves may threaten an individual's career. Furthermore, a researcher may have access to an individual's confidential records and the release of that information could have negative consequences on the employee. These vulnerabilities need to be specifically addressed by measures to assure individual consent without coercion and measures to protect the confidentiality of data and records.

Full details regarding IRB requirements and reviews may be found in 10 CFR 745, the federal "common rule" for the protection of human research subjects (Ref. 14).

8. SUMMARY AND CONCLUSIONS

Many industries have HRPs to assure that individuals in critical positions are psychologically and physically capable of performing their assignments in a reliable manner. HRPs are especially important in industries with high consequence operations where the safety and security of the local and surrounding populations may be at risk. Typically HRPs consist of physiological and psychological testing prior to placement in the job assignment and at regular intervals.

One shortcoming of current HRPs is that often there are relatively long intervals in between scheduled assessments. Although there are continual requirements for self-reporting and reporting by HRP peers and supervisors for unusual behaviors in HRP-certified individuals, such reporting may not be reliable for a number of reasons. HRP-certified individuals may be reluctant to report events that could affect their jobs or damage their image with management or peers or they may not be aware of a developing problem. In addition, there may not be any behaviors that are observable by peers or supervisors.

A second shortcoming of current HRPs is that the assessments are not tied to periods of critical operations. The importance of assessment prior to the performance of critical operations is suggested by a recent report of 60 criticality incidents that have occurred in nuclear facilities. An examination of these incidents indicated that serious consequences (fatalities and dismemberments) occurred in many of the incidents and further that human error played a significant role in many of them. Some of these errors could have been related to the emotional state of the individual. This suggests that a method to detect a deleterious emotional state prior to the performance of critical operations could reduce the occurrence of some accidents with serious safety and security consequences.

There are some limitations on the types of errors that can be prevented by pre-shift testing. First, errors that are primarily caused by factors unrelated to the emotional state of the individuals such as deficiencies in equipment, procedures, management, or training will not be affected. Secondly, the type of error is important. Unintentional errors (such as inadvertently activating the wrong control or skipping a step) are more likely to be related to the emotional state of the individual than intentional errors (such as deliberately changing a procedure). Furthermore, malevolent behaviors are not likely to be prevented by pre-shift testing.

There are a number of practical constraints for pre-shift testing that are important in the choice of predictor measures. One requirement is that the test can be administered relatively frequently. This requirement presents issues for tests that are subject to learning effects. In addition, although state characteristics (temporary fluctuations in an individual) can be detected with short test instruments, trait characteristics (relatively permanent personality characteristics) are not likely to be detected. A second requirement is that the test time is relatively short (15 minutes or less). The short test time may present issues for reliability as well as validity.

In a recent review of the literature, the military has reported several potential predictors of emotional distress that could be studied in the current application. These consist of physiological

biomarkers of stress, physiological biomarkers of cognitive status, and behavioral biomarkers of operational status. One of the more promising predictor measures for stress and cognitive status is HRV. Definitions for various types of reliability and validity and methods to assess them have been outlined in this report. One of the most challenging problems in applying the procedures is the selection of relevant test populations. Ideally, two general groups are needed, a criterion group and a standard group. The criterion group is the "emotionally distressed" group and the standard group is a group of unselected (those not reporting emotional distress) individuals. Sources for the criterion group could be counseling centers within or outside the company where participation from individuals who are undergoing counseling for adjustment to life stresses or for conditions such as depression or anxiety could be obtained. The standard group could also be obtained from within or outside the company. Group variables, such as age, gender, educational level, occupation, and years experience are important in assessing both the reliability and validity of the predictor test.

Statistical analysis procedures basically involve three stages. The first is a comparison of scores for criterion and standard groups to determine if the predictor measure discriminates between the two primary populations. Group variables must also be represented in these comparisons to make sure that the test discriminates across group variables such as age and gender. The second stage is a comparison of the predictor scores with the criterion measure. The criterion measure could be a current assessment of behavior as measured with a job sample test or future assessment of behavior such as on-the-job performance. The final stage involves the development of the cutoff score for the predictor. The predictor score could be developed using group norms or individual baselines. Ultimately, the use of individual baselines is suggested due to individual differences in the average baseline scores as well as potential individual differences in variability. Additionally, some measures have temporal characteristics that make individual baselines even more necessary.

Finally, this type of research is subject to review and approval by IRBs. Because the worker population is considered a vulnerable population, measures are required to assure that informed consent has been obtained from the participants and that the confidentiality of the data and records will be maintained.

The eventual value of this testing effort depends upon a number of factors. Most important is the percentage of errors in high-consequence operations that are related to the emotional state of the individual. Although there is some suggestion from the review of criticality incidents in nuclear facilities that these types of errors could exist, there are many types of errors that are caused by factors other than the emotional state of the individual, including deficiencies in management, training, procedures, and equipment. Therefore, detecting a deleterious emotional state may not be cost effective if this type of cause is not a significant factor in unreliability in critical operations. The potential high consequences of the event could, however, provide sufficient justification for this type of program even if the relative occurrence of such errors is low.

Another factor is how well a pre-shift test will predict over the shift interval. For example, if a shift interval is 7 hours, a measure taken at the beginning of the shift may have limited value. There are a number of events that could occur during the shift to change the emotional status of the individual. Retesting midway through the shift could be considered if schedule and time

permit. Alternatively, some measures such as physiological ones, could perhaps be taken continuously during the course of the shift.

In conclusion, follow-on research should include a literature review of predictor measures for emotional state. The information in this report is based on a very limited review of the area. Factors specific to this application such as the short test time and relatively frequent test administration are important in this review in addition to information on group and individual variation.

9. REFERENCES

1. McLaughlin, T.P., S.P. Monahan, N.L. Pruvost, V.V. Frolov, B.G. Ryazanov, and V.I. Sviridov, A Review of Criticality Accidents, LA-13638, Los Alamos National Laboratory, Los Alamos, NM, May 2000.
2. Vorontsov, S.V. Analysis of the Causes of Research Reactor and Criticality Facility Accidents, Russian Federal Nuclear Center Institute of Experimental Physics (VNIIEF), Sarov, Russia, 2006.
3. Vorontsov, S.V., Selection of a Baseline Methodology for Pre-Shift Monitoring of the Operating Personnel at Nuclear Research Facilities, Russian Federal Nuclear Center Institute of Experimental Physics (VNIIEF), Sarov, Russia, 2007.
4. Anastasi, A. Psychological Testing. Fourth Edition. 1976, New York: Macmillan Publishing Co., Inc.
5. Edward, G.C. and R.A. Zeller. Reliability and Validity Assessment, 1990, Thousand Oaks CA: Sage Publications, Inc.
6. Swain, A.D. and H.E. Guttmann. Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications, SAND80-0200, Sandia National Laboratories, Albuquerque, NM, August 1983.
7. Institute of Medicine of the National Academies. Committee on Metabolic Monitoring for Military Field Applications. Standing Committee on Military Nutrition Research Food and Nutrition Board. Monitoring Metabolic Status. Predicting Decrements in Physiological and Cognitive Performance. 2004, Washington, D.C.: The National Academies Press.
8. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. 2000. Arlington, VA: American Psychiatric Association.
9. Desk Reference to the Diagnostic Criteria from DSM-IV-TR. 2000. Arlington, VA: American Psychiatric Association.
10. Butcher, J.M., S. Mineka, and J.M. Hooley, Abnormal Psychology, Thirteenth Edition. 2007, New York: Pearson Education, Inc.
11. Segerstrom, S.C. and L. S. Nes, Heart Rate Variability Reflects Self-Regulatory Strength, Effort, and Fatigue. Psychological Science, 2007, 18(3): p. 275-281.
12. Cronbach, L.J. Essentials of Psychological Testing. Third Edition. 1970, New York: Harper & Row, Publishers.
13. Anastasi, A. and S. Urbina. Psychological Testing, Seventh Edition. 1997, New Jersey: Prentice Hall.
14. Department of Energy, Protection of Human Subjects, Title 10, Chapter III, Part 745.
15. Babor, T.F., J.C. Higgins-Biddle, J.B. Saunders, and M.G. Montiero. The Alcohol Use Disorders Identification Test: Guidelines of use in primary care (2nd. Ed.). 2001, Washington, DC: World Health Organization Department of Mental Health & Substance Dependence.
16. Spielberger, C.D., State-Trait Anger Expression Inventory: Professional Manual. 1996, Odessa, FL: Psychological Assessment Resources, Inc.
17. Spielberger, C.D., Manual for the State-Trait Anxiety Inventory. 1983, Mountain View, CA: Consulting Psychologists Press.

18. Morey, L.C., Personality Assessment Screener: Professional Manual. 1997, Odessa, FL: Psychological Assessment Resources, Inc.
19. Lambert, M.K. and G.M. Burlingame, Administration & Scoring Manual for the OQ 45.2. 1996, Stevenson, MD: Professional Credentialing Services.
20. Cohen, S., T. Kamarck and R. Mermelstein, A global measure of perceived stress. Journal of Health and Social Behavior, 1983, 24, 385-396.

DISTRIBUTION

1	MS0370	T.G. Trucano	01411
1	MS0830	K.V. Diegert	12335
1	MS0830	C.C. Dornburg	12335
3	MS0830	L.M. Weston	12335
1	MS1032	T.J. Kreuch	03334
1	MS1188	J.C. Forsythe	06341
1	MS1188	J.S. Wagner	06341
1	MS1374	E.M. Hinman-Sweeney	06723
2	MS9018	Central Technical Files	8944
2	MS0899	Technical Library	4536
1	MS0123	D. Chavez, LDRD Office	1011