SURROGATE MEASURES: A PROPOSED ALTERNATIVE IN HUMAN FACTORS ASSESSMENT OF OPERATIONAL MEASURES OF PERFORMANCE

Robert S. Kennedy, Norman E. Lane, and Lois A. Kuntz
Essex Corporation
1040 Woodcock Road, Suite 227
Orlando, Florida 32803

ABSTRACT

In human factors studies, operational performance measures have de facto validity, but they also tend to have very low numerical reliability (r=.10-.30). We have proposed the use of surrogate measures as an alternative to direct assessment of operational performance for purposes of screening agents (e.g., drugs, environmental stress, etc.), particularly when the surrogate can be empirically validated against the operational criterion. The focus is on cognitive (or throughput) performances in humans as opposed to sensory (input) or motor (output) measures, but the methods should be applicable for development of batteries which will tap input/output functions.

For several years, we have been developing under NASA, NSF, and U.S. Army sponsorship, a menu of performance tasks implemented on a battery-operated portable microcomputer. Currently, 21 tests are available on a disk for microcomputer presentation. These tasks are reliable and become stable in minimum amounts of time, appear sensitive to some agents, comprise constructs related to actual job tasks, and are easily administered in most environments. We will review our progress with this program, and describe implications for human factors engineering studies in environmental stress.

INTRODUCTION

The presence of related environmental stressors in space necessitates the generation of a standardized human factors testing tool in order to detect subtle differences in the integrity of human performance and welfare. This testing tool could be used to predict premonitory onset of decrements in performance, physiology, mood, and behavior before such stressors effect operational efficiency; to enhance identification of susceptible personnel, to explore the possibility of resistance training, and to monitor the neurological status of persons subjected to hazards in their occupations, as well as for longitudinal monitoring in connection with regular physical examinations.

Environmental stressors are most often studied with a pre-, per-, post paradigm. This approach makes maximum use of the "each subject serves as his or her own control" philosophy. Repeated-measures studies, where environmental stress has served as an agent which alters performance, include weightlessness (Nicogossian & Parker, 1982), high altitude (Fowler, Paul, Porlier, Elcombe, & Taylor, 1985), temperature (Ellis, 1982), toxic chemicals (Guillion & Eckerman, 1985), pharmacological agents (Kohl, Calkins, & Mandell, 1986), pressure at depth (Logie & Baddeley, 1985), physical exercise (Englund, Ryman, Naitoh, & Hodgdon, 1985), sleep loss (Woodward & Nelson, 1974), motion (Kennedy & Frank, 1986), fatigue (West & Parker, 1975), dehydration (Banderet, MacDougall, Roberts, Tappan, Jacey, & Gray, 1984), simulated environments (McComas, 1986), vibration (LaRue, 1965, Hornick & Lefritz, 1966), and many other agents have received research consideration. Organismic and induced states within subjects have also received attention and these have invariably been with repeated-measures designs. Kiziltan (1985), Thorne, Genser, Sing, and Hegge (1983), and Donnell (1969) have examined the effects of sleep deprivation on performance. One of the forerunners of Automated Performance Testing (the Neptune Project: McKenzie, White, and Hartman, 1968), built on earlier work by Fleishman and Ellison (1962), who undertook the identification of factors common to psychomotor tests, and French (1951; French, Ekstrom, & Price, 1963) addressed the identification and measurement of cognitive abilities to produce (McKenzie et al., 1968) a fully functioning battery in the late 1960s.

However, as a practical matter, measures of operational performance are elusive and several problems remain in the assessment of human performance; chronically low retest reliability, instability across days due to learning, wide individual differences of unknown or uncontrolled variation, not knowing what to measure, etc. Reviews of the older literature concerning assessment of operational performance document the unreliability of performance measures (Lane, 1986). This problem is not a data acquisition

problem and does not disappear in our era of increasing technology, where it is now possible to measure nearly every variable we wish, 1000 times per second, and save all or play it back at will. The problem is a "what to measure" problem and is partly related to human variablity. The two metric issues which are central to an understanding of this problem are concerns with "stability" and "reliability."

Stability. A development program was undertaken several years ago by the Navy to evaluate the repeated-measures stability of paper-and-pencil based human performance tests. This program (Performance Evaluation Tests for Environmental Research [PETER]) began in 1977 (Kennedy & Bittner, 1977) and has provided specific paradigms and methods for the evaluation and selection of performance tests applicable to repeated-measures research (Carter & Kennedy, 1980). Typically, alternate forms of a test would be repeatedly administered to the same group of subjects (i.e., 1 to 15 trials). These data would then be analyzed for three types of stability -- Means, Standard Deviations, and Correlations (Kennedy, Bittner, Harbeson, & Jones, 1981).

Reliability. For the past nine years studies of military flying performance in a sophisticated flying trainer, the Visual Technology Research Simulator (VTRS), have been conducted (Lintern, Nelson, Sheppard, Westra, & Kennedy, 1981). In this research tool, the ability to hold many confounding variables constant, the precise data acquisition systems available, and the size of the retest reliability from that which is available in the field setting for the same task changes very little. For example, single carrier landing approach performances on the simulator had retest reliabilities of .23 to .32 for the mean of four trials; air-to-ground bomb miss distance reliabilities were lower, slightly above .20 for the mean of eight trials.

As has been mentioned (Lane, Kennedy, & Jones, 1986), not enough attention has been paid to the reliability of criteria or dependent variables in experimental studies. The consequence of this omission can be seen in the well known correction for attenuation formula reported by Guilford (1954) and symbolized as:

$$R_t = \frac{r_{xy}}{(r_{xx})\ (r_{yy})^{1/2}}$$

where r_{xy} is the predictive validity, r_{xx} is the reliability of the predictor, r_{yy} is the reliability of the criterion, and R_t is the true relationship. Without good reliability in x or y, the greatest possible relationship between the two is limited. For measuring operational performance this relationship is considerable. Although automated systems will permit significant amounts of data to be recorded in real time,we do not believe that the use of automated systems per se will rectify this problem.

Insufficient attention to reliability can lead to reduction of statistical power, higher sample size requirements, cost of experiments, and when hazard is involved, other problems. Without attention to reliability, the outcome can be misinterpreted. Utilizing the correction for attenuation formula (above) often changes conclusions, but not always negatively. For instance, the true predictive validities of operational criteria from paper-and-pencil aptitude tests are often misinterpreted because of criterion unreliability. Again, using our example, an operational reliability may be improved from .20 to .30 even at great expense, but predictor reliability might go from .70 to .90 with much less investment. The alliance in the denominator suggested to us a focus on developing highly reliable measure sets, separate from the operational criteria but highly similar to the criteria in skill requirements. If the measure sets correlate well with the criteria, and behave similarly under changing task conditions, perhaps they could be used for the criteria. The focus should be on developing highly reliable measurement sets, separate from the operational criteria but highly similar to the criteria in skill requirement.

We do not believe that meaningful examinations of environmental effects on human performance can be undertaken unless these performances can be adequately measured. A possible solution would be surrogate measures. Surrogate measures are those which are related or predictive of a construct of interest but are not direct measures. In our plan these are composed of tests or batteries that exhibit five characteristics: 1) stable so that the "what is being measured" is constant; 2) correlate with the performance construct; 3) sensitive to the same factors that would affect performance as the performance variable would; 4) more reliable than field measures; 5) involve minimal training time.

Surrogate measures differ from conventional performance measures in that tests need not involve operations in common with the performance measures, only components or factors in common. They also differ from "synthetic" or "job samples" because the surrogate takes so little practice and is easy to score. Given the great difficulty of obtaining reliable enough field measures to carry out stressor-sensitivity studies on the operational task itself, the case for using a surrogate is strong. Large portions of the variance in extremely complex tasks can be predicted from performance on relatively simple tests. An external test (or battery), though it cannot be as "valid" as the measure itself from a practical standpoint, may tap more of the true variance of the field performance because its reliability is much greater.

BACKGROUND

Performance Evaluation Tests for Environmental Research. In order to study the metric properties of existing performance tests, the Navy evaluated over 140 tests and tasks in a repeated-measures paradigm where a small group of subjects were examined repeatedly on alternate forms of the tests over a fifteen day period. The full report of the work of that program appears in over 90 publications, most of which are listed in Harbeson, Bittner, Kennedy, Carter, and Krause (1983) and fully reviewed and summarized in Bittner, Carter, Kennedy, Harbeson, and Krause, (1986). Only about one fourth of the tests were worthy according to the criteria mentioned above.

Microcomputerized vs. Paper-and-Pencil Versions of Tests. Although in the past much of the early work in the field of environmental effects made use of paper-and-pencil based tests, or tests which made use of modest apparatus (Bittner et al., 1986; Harbeson et al., 1983), the wide availabilty of low-cost, high-speed computer systems has encouraged psychologists to transfer to such apparatus in their laboratory studies of human performance measurement. In recent years there has been widespread interest in computerized performance tests. The Army (Thorne, Genser, Sing, & Hegge, 1985), Navy (Kennedy, Wilkes, Lane, & Homick, 1985; Kiziltan, 1985; McComas, 1986), Air Force (O'Donnell, 1981; Christal, 1981; Payne, 1982; Shingledecker, Acton, & Crabtree, 1983), and Environmental Protection Agency (Guillion & Eckerman, 1985) have active programs. These programs constitute valuable resources for the research and development of a computerized testing system.

The Automated Performance Test System (APTS) Background. The tests of the NASA battery have been implemented on a NEC PC8201A portable laptop computer and is now called the Automated Performance Test System (APTS). The 8201A was selected because of the amount of onboard memory available (64K bytes), and the low cost of the unit and peripherals (approximately $850.00). The display screen consisted of 240 x 64 pixel (40 characters by 8 lines) liquid crystal display (LCD) with adjustable contrast control. The unit is lightweight (3.8 pounds) and durable. After several tosses down a flight of 22 noncarpeted stairs, the only damage to the NEC was a crack in the housing, four keys fell off, and one horizontal line on the LCD was lost. The NEC also meets minimum requirements for approval for flight on the Space Shuttle.

All tests are written in the BASIC software language. Many functions such as prompting for input, converting lower case letters to upper case, test timing, and response timing were common to all the tests. Assembly language programs were written to perform these common functions, thereby providing more room in memory for data storage and the tests themselves. With careful utilization of the Read Only Memory (ROM) routines written by Microsoft, 10 different tests could be stored in the onboard memory, and enough memory was available to allow the data of 40 subjects to be stored for later off-line storage.

Since the initial implementation of the test battery on the NEC, the IBM Personal Computer has become an industry standard. The widespread use of the IBM and compatibles has made it the "hands-on" favorite in all sorts of laboratories. More and more requests for IBM compatible software has prompted us to convert our test battery to a format which makes it usable on IBM and compatible equipment. Indeed, today it is possible to purchase an IBM clone with 640K bytes memory, two 360K byte floppy drives, keyboard, monitor, serial and parallel interface for the same price as that of the NEC. For the same price, more capability, increased data storage, faster timing increments and better instructions are possible!

We are committed to maintaining the portability aspect of the test battery. Because of this commitment, we have selected the Zenith Data Systems ZFL-181 as the current host of the portable assessment battery. The 181 contains 640K onboard memory, two 720K byte 3.5 inch floppy drives, serial and parallel interfaces, an RGB interface, and 80 characters by 25 line super twist, backlit LCD display, and is completely IBM PC compatible. The batteries are capable of powering the unit with both drives running and the brightness control set on high for 4.2 hours. APTS - Usage. The test battery has been used in a variety of environments ranging from a classroom setting to the cockpit of jet airliners. This versatility provides the experimenter with a multitude of options with respect to the APTS usage. Because of the portability of both systems, on-the-spot testing, rather than in the laboratory, is possible.

Advantages to computer administered testing include: 1) standardized testing conditions leading to higher reliabilities; 2) reduced variability between test procedures and administrators which enhances comparison of results between similar studies; 3) accurate and objective response scoring eliminating unintelligible responses, improper scoring, and subjective interpretation; 4) complete automation of all testing, scoring, and data collection procedures resulting in a reduction of problems associated with lost or misplaced data; 5) utilization of a variety of response measures such as speed and latency; 6) presentation of complex and innovative stimuli involving a variety of sensory modalities; 7) capabilities for precise timing and control of stimulus materials; 8) immediate scoring of responses with easy access to data for rapid analysis or feedback to the subject or administrator; 9) automatic data storage with capabilities for handling quantities of diverse data over repeated trials, with large N's; 10) self-administration of interesting and challenging materials resulting in increased subject motivation and reduced boredom; 11) increased convenience and efficiency in data collection reducing the need for highly skilled

professionals or psychological technicians; 12) portability of the system with the accompanying advantages of reduced size and weight; and 13) adaptive testing, where difficulty level changes with performance, which can shorten testing time.

Despite the advantages to microbased testing, it was our view that in order to avoid unknown influences from medium differences it was necessary to compare the "good" paper-and-pencil tests of the PETER program to serve as markers in building a battery (or menu) of computer-based performance tests. In addition to the issues related to reliability and stability previously mentioned, the studies cited above also call attention to the necessity for careful preparation and evaluation of research tools during development. For example, Moran and Mefford (1959) identified the need for comparability of alternate forms. Repeated measurements must possess certain characteristics to be meaningful and easily and clearly interpretable (American Psychological Association, 1974; Jones, 1972; Lord & Novick, 1968). To summarize the characteristics, the statistical requirements for easily interpretable results of repeated-measures include level or linearly increasing means, level variances, and differential stability (Bittner et al., 1986).

Objectives of the Automated Performance Test System. In summary, the philosophy of our approach to performance test development involves three different goals. The first is to deal with only tests or tasks that can be shown to be psychometrically sound. This requires that we demonstrate stability of means and standard deviations within few administrations, and most important, that correlational stability, the stability of trial-to-trial intercorrelations, be shown to occur quickly and with high test-retest prescreening correlations (i.e., reliability). The second goal will be to demonstrate that the battery has factorial multidimensionality and that the subscales cross-correlate with earlier performance tests and other recognized instruments of ability. Then it is necessary to demonstrate and document sensitivity to factors known to compromise performance potential in the laboratory and ultimately real-world situations. Lastly, the tasks must be shown to be predictive of the types of work performed in the real world.

PSYCHOMETRIC STUDIES

NASA I. Originally, for proof of concept, we began our first testing under NASA sponsorship using the methodology of stability and reliability with a microbased computer (Kennedy, Wilkes, Lane, & Homick, 1985). Twenty subjects were tested over four replications using paper-and-pencil versions as well as the computerized version of six tests. All tests achieved stability within the four test sessions, reliability efficiencies were generally high ($\underline{r}$ >.707 for 3-minute testing), and the computerized tests were largely comparable to the paper-and-pencil version from which they were derived. The tests that were evaluated for inclusion in this experiment were Grammatical Reasoning, Pattern Comparison, Code Substitution, and the Tapping series, tests which had largely proven their metric properties in paper-and-pencil versions earlier in the PETER work. As these tests all exhibited stability and reliability within our proposed standards, all were proposed for further testing.

NASA II. In this study, in addition to evaluating stability and reliability of the tests, predictive validity was also examined. Twenty-five subjects were tested over significantly more replications (10) and tests (11) than previously. The 11 tests were concurrently administered in paper-and-pencil (marker battery) and microcomputer-based versions and compared to the Wechsler Adult Intelligence Scale (WAIS). Nine of the 11 microcomputer-based tests achieved stability. Reliabilities were generally high, with $\underline{r} \geq .77$ for 3 minutes of testing for the recommended tests. Cross-correlations of microbased tests with traditional paper-and-pencil versions and indices of stability suggest equivalency between the tests in the different modes. Correlations between certain microbased subtests and the WAIS identified common variance.

NASA III. In this experiment, we administered 21 different tests, including six short-term memory tests which had not been administered before. Air Combat Manuvering, Pattern Comparison, and Reaction Time Four Choice took the longest of the original battery to stabilize, but all tests stabilized by trial 5; the memory tests took a little longer and with only modest reliabilities.

ARMY I & II. Currently, under contract to the U.S. Army Medical Research and Development Command, we are evaluating other tests from the tri-service Performance Assessment Battery (PAB) (Englund, Reeves, Shingledecker, Thorne, Wilson, & Hegge, 1986) along with our existing menu of tests for potential inclusion into a menu of tests for the battery. This menu would permit investigators to customize a battery to their specific needs. These studies, while completed, are still in draft form. In general, the tests from the NASA battery performed better than candidate PAB tests despite differences in test administration. The best PAB tests were Recall, Mathematical Processing, and Matrix Rotation, with average reliabilities of 0.60, 0.63, and 0.71, respectively. In contrast, Item Order, Memory Search, and Successive Pattern Comparison average reliabilities were 0.43, 0.49, and 0.44, respectively. The best NASA battery tests were Grammatical Reasoning, Simultaneous Pattern Comparison, and Manikin tests, with average reliabilities of 0.84, 0.79, and 0.95, respectively. The two Reaction Time tests were also good measures, although they are highly intercorrelated. Perhaps just the 4-Choice version, which is the more reliable of the two, should be used in the future. Tapping tests, which were used in both test batteries, exhibited consistently high levels of stability and reliability. What is most intriguing, however, is the lack of significant overlap of Tapping tests with other tests, indicating their relatively pure nature. Because of the ease of administering the

measure and its independence from perceptual and cognitive tests, it may be a measure of motivation. Further testing should address this issue.

NSF II. A factor analysis was conducted on 11 selected tests from the PAB, and NASA performance test batteries were administered three times to each of 108 Central Pennslyvania college students (48 males and 60 females). The Wonderlic Personnel Test was administered just before the first and just after the last administration of the performance tests. The test-retest reliability of the Wonderlic in the total sample was .78, which yields a Spearman-Brown estimated reliability for the sum of the two Wonderlic scores, the "combined" Wonderlic, of .88. The multiple R in the total sample between the combined Wonderlic as criterion and Grammatical Reasoning (NASA) and Math Processing (PAB) as predictors ranged between .48 and .55 on the three test administrations. Factor analyses carried out on each administration yielded three consistent factors: a spatial/numerical factor on which Pattern Comparison (NASA) loads most heavily, a verbal factor of which Grammatical Reasoning (NASA) loads most heavily, and a motor factor defined by the Tapping tests (NASA). Based on these results we would recommend a core battery consisting of Grammatical Reasoning (NASA), Mathematical Processing (PAB), Pattern Comparison (NASA), and the Preferred and Nonpreferred (but not the Two-Finger) Tapping tests. This battery provides a good short estimate of IQ, based on Grammatical Reasoning and Mathematical Processing and three well identified factors, one verbal, another spatial/numerical, and the third motor. This core battery can be usefully augmented, especially in operational situations, by Code Substitution and Choice Reaction Time tests, both from the NASA battery. Manikin (NASA) is another recommended test for augmentation because it measures a different factor from IQ.

SENSITIVITY STUDIES

Altitude. Until recently, lack of an adequate human performance research tool has resulted in the employment of a variety of techniques, methods, and measures that limit systematic comparisons across altitude studies. Such limitations have delayed the development of a cohesive body of knowledge regarding human performance at altitude. Measurement and data collection inadequacies have further contributed to research difficulties. While highly controlled studies systematically relating sustained exposure to human performance are largely lacking, we believe that exceptions are beginning to appear (cf. e.g., Banderet & Burse 1984; Banderet, Benson, McDougall, Kennedy, & Smith, 1984).

The NASA battery has been tested at simulated altitude by scientists of the U.S. Air Force, and the U.S. Army Institute for Environmental Medicine (Banderet et al., 1984). The initial results show a definite cognitive performance decrement with sustained periods at altitudes of 23,000 feet, and with abrupt, short periods at 27,000 feet. However, motor performance remained essentially unchanged. An important point to note is that typical measures of performance would not have detected the effect altitude had on the mental capabilities of the participants.

Drugs. With regular doses of certain motion sickness drugs, virtually all of the scores for both motor and cognitive tests changed in a theoretically rational direction in studies conducted by Dr. Charles Wood at Louisiana State University Medical School. That is, amphetamine scores increased and scopolamine scores decreased over placebo. A simple ANOVA revealed no significant outcomes (other than that Pattern Comparison, one of the APTS tests, scores appeared to be significantly poorer with hyoscine). The within-subject variables were scopolamine and dexedrine, arranged factorially in a totally within-subject design (a more powerful approach). The results indicate that amphetamine significantly increased Nonpreferred Hand Tapping (a motor skill test) and there was a trend for increased scores on the Sternberg (an item recognition test). This would mean there were more "hits" or that latency improved. There was not a significant effect of scopolamine on Preferred-Hand Tapping. The study further showed an interaction of scopolamine and dexedrine with Two-Hand tapping. Though not statistically significant, overall it appears that scopolamine facilitates performance more when dexedrine is also present then it does without dexedrine.

Chemoradiotherapy Treatments. From the University of Washington, Dr. Parth has been studying patients who are receiving bone marrow transplants and chemoradiotherapy treatments. In this study, the tests of the basic NASA battery were administered, along with other tests, to both a patient population undergoing chemotherapy subsequent to bone marrow transplants and to a control population of sibling donors. Four replications of the battery were given spaced over one year, including prior to transplant therapy, during therapy, and in a follow-up examination. The primary purpose of NASA's use was to determine battery sensitivity to physiological stressors different from those examined in previous studies. The battery as a whole was strikingly effective in detecting performance shifts in patients and significantly differentiating patients from controls throughout the two therapy test periods. Greatest discrimination was apparent in the complex cognitive measures (i.e., Code Substitution) than in the "motor" (i.e., Tapping). Discrimination was present for both accessory and latency measures, although effects were stronger for accuracy performance.

Study with Sleep Loss. Two different studies of sleep loss have been conducted. In the first study, Kiziltan (1985) at the U.S. Naval Postgraduate School in Monterey, California, observed statistically effects on Code Substitution but obtained only directional changes (nonsignificant) on the other tests following one night without sleep. Another study was performed with the NASA battery tests at Ames Research Center in

Moffett Field, California. The experiment lasted 41 days, 30 of which was the bedrest phase. The results of this study revealed modest or no change on most tests.

In summary, for the past few years our research efforts have concerned study and identification of reliable performance measurement instruments for exotic environments. Under the sponsorship of NASA, NSF, and Army MRDC a menu of performance tasks implemented on a battery-operated portable microcomputer has been developed. These measures differ from conventional performance measures in that tests need not involve operations in common with the performance measures, only components/ factors in common. The tests also exhibit higher reliabilities ($r \geq .70$) than traditional performance measures ($r = .10-.30$). Currently, 21 tests are available on a disc for microcomputer presentation. These tasks are reliable and become stable in minimum amounts of time, appear sensitive to some agents, comprise constructs related to actual job tasks, and are easily administered and scored. Collectively these tests are known as the Automated Performance Test System (APTS). In numerous experiments the APTS has been shown to be a stable and reliable indicator of performance. If a person performs in a predictable manner and an intervening factor is introduced that may have an adverse effect on performance (i.e., zero gravity, stress) it may be detected by the APTS. Using a stable, sensitive, battery of performance tests would be analogous to taking a person's temperature, blood pressure, or weight. If administered on a daily basis it would be a form of record keeping that would show whether a person's performances were being affected by the environment or factors such as fatigue or workload. The APTS tests cognitive factors related to job performance and is therefore more predictive of performance than traditional methods of respiration, heart rate, blood pressure, et cetera.

REFERENCES

American Psychological Association, STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTS, American Psychological Association, Washington, DC, 1974.

Banderet, L. E., Benson, K. P., McDougall, D. M., Kennedy, R. S., & Smith, M., "Development of cognitive tests for repeated performance assessment," PROCEEDINGS OF THE 26TH MILITARY TESTING ASSOCIATION CONFERENCE, Munich, Germany, 1984, pp. 375-380.

Bandaret, L. E., & Burse, R. L., "Cognitive performance at 4500m simulated altitude," PROCEEDINGS OF THE 92ND ANNUAL AMERICAN PSYCHOLOGICAL ASSOCIATION, Toronto, Canada, 1984, August.

Banderet, L. E., MacDougall, D. M., Roberts, D. E., Tappan, D., Jacey, M., & Gray, P., "Effects of dehydration or cold exposure and restricted fluid intake upon cognitive performance," PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, COMMITTEE ON MILITARY NUTRITION RESEARCH WORKSHOP, National Academy of Sciences, Washington, DC, 1984.

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M., "Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 measures," PERCEPTUAL AND MOTOR SKILLS, 63, 1986, pp. 683-708.

Carter, R. C., & Kennedy, R. S., "Selection of Performance Evaluation Tests for Environmental Research," PROCEEDINGS OF THE 24TH ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY, Los Angeles, CA, 1980, pp. 320-324.

Christal, R. E., "The need for laboratory research to improve the state-of-the-art in ability testing," NATIONAL SECURITY INDUSTRIAL ASSOCIATION, DoD CONFERENCE OF PERSONNEL AND TRAINING FACTORS IN SYSTEMS EFFECTIVENESS, San Diego, CA, 1981.

Donnell, J., "Performance decrement as a function of total sleep loss and task duration," PERCEPTUAL AND MOTOR SKILLS, 29, 1969, pp. 711-714.

Ellis, H. D., "The effects of cold on the performance of serial choice, reaction time and various discrete tasks," HUMAN FACTORS, 24, 1982, pp. 589-598.

Englund, C. E., Reeves, D. L., Shingledecker, C. A., Thorne, D. R., Wilson, K. P., & Hegge, F. W., "Unified Tri-service Cognitive Performance Assessment Battery (UTC-PAB)," REPORT NO. 86-1, U.S. Army Research and Development Command, Fort Dietrick, MD, 1986.

Englund, C. E., Ryman, D. H., Naitoh, P., & Hodgdon, J. A., "Cognitive performance during successive sustained physical work episodes," BEHAVIOR RESEARCH METHODS, INSTRUMENTS & COMPUTERS, 17, 1985, pp. 75-85.

Ferris, S. H., Crook, T., Clark, E., McCarthy, M., & Rae, D., "Facial recognition memory deficits in normal aging and senile dementia," JOURNAL OF GERENTOLOGY, 35, 1980, pp. 707-714.

Fleishman, E. A., & Ellison, G. D., "A factor analysis of fine manipulative tests," JOURNAL OF APPLIED PSYCHOLOGY, 46, 1962, pp. 96-105.

Fowler, B., Paul, M., Porlier, G., Elcombe, D. D., & Taylor, M., "A reevaluation of the minimum altitude at which hypoxic performance decrements can be detected," ERGONOMICS, 28, 1985, pp. 781-791.

French, J. W., "The description of aptitude and achievement in terms of rotated factors," PSYCHOMETRIC MONOGRAPHS NO. 5., 1951.

French, J. W., Ekstrom, R. B., & Price, L. A., "Manual for kit of reference tests for cognitive factors," RESEARCH CONTRACT NONR-2214-00, Educational Testing Service, Princeton, NJ, June, 1963.

Guilford, J. P, PSYCHOMETRIC METHODS (2nd ed.), McGraw-Hill, New York, 1954, pp. 400-402.

Guillion, C. M., & Eckerman, D. A., "Field testing for neurobehavioral toxicity: Methods and methodological issues," BEHAVIORAL TOXICOLOGY, Z. Annau (Ed.), Johns Hopkins Press, Baltimore, MD, 1985.

Harbeson, M. M., Bittner, A. C., Jr., Kennedy, R. S., Carter, R. C., & Krause, M., "Performance Evaluation Tests for Environmental Research (PETER): Bibliography," PERCEPTUAL & MOTOR SKILLS, 57, 1983, pp. 283-293.

Hornick, R. J., & Lefritz, N. M., "A study and review of human response to prolonged random vibration," HUMAN FACTORS, 8, 1966, pp. 481-492.

Jones, M. B., "Individual differences," THE PSYCHOMOTOR DOMAIN, R. N. Singer (Ed.), Lea & Febiger, Philadelphia, PA, 1972, pp. 107-132.

Jones, M. B., Kennedy, R. S., & Bittner, A. C., Jr., "A video game for performance testing," AMERICAN JOURNAL OF PSYCHOLOGY, 94, 1981, pp. 143-152.

Kennedy, R. S., & Bittner, A. C., Jr., "The development of a Navy performance evaluation test for environmental research (PETER)," PRODUCTIVITY ENHANCEMENT: PERSONNEL PERFORMANCE ASSESSMENT IN NAVY SYSTEMS, L. T. Pope & D. Meister (Eds.), Navy Personnel Research & Development Center, San Diego, CA, 1977. (AD A111180)

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B., "Perspectives in performance evaluation tests for environmental research: Collected papers," RESEARCH REP. NO. NBDL-80R004, Naval Biodynamic Laboratory, New Orleans, LA, 1981. (AD A111180)

Kennedy, R. S., & Frank, L. H., "A review of motion sickness with special reference to simulator sickness," 65th Annual Meeting of the Transportation Research Board, Washington, DC, 1986.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L., "Preliminary evaluation of microbased repeated-measures testing system, REPORT NO. EOTR-85-1, Essex Corporation, Orlando, FL, 1985.

Kiziltan, M., "Cognitive performance degradation on sonar operated and torpedo data control unit operators after one night of sleep deprivation," Unpublished master's thesis, Naval Postgraduate School, Monterey, CA, 1985.

Kohl, R. L., Calkins, M. A., & Mandell, A. J., "Arousal and stability: The effects of five new sympathomimetic drugs suggest a new principle for the prevention of space motion sickness," AVIATION, SPACE, AND ENVIRONMENTAL MEDICINE, 57, 1986, pp. 137-143.

Lane, N. E., "Skill acquisition curves and military training," IDA PROFESSIONAL PAPER NO. P-1945, Institute for Defence Analyses, Alexandria, VA, 1986.

Lane, N. E., Kennedy, R. S., & Jones, M. B., "Overcoming unreliability in operational measures: The use of surrogate measure systems," PROCEEDINGS OF THE 30TH ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY, Santa Monica, CA, 1986, pp. 1398-1402.

LaRue, M. A., "The effects of vibration on accuracy of a positioning task," JOURNAL OF ENVIRONMENTAL SCIENCES, 8, 1965, pp. 33-35.

Lintern, G., Nelson, B. E., Sheppard, D. J., Westra, D. P., & Kennedy, R. S., "Visual Technology Research Simulator (VTRS) human performance research: Phase III," NAVTRA-EQUIPCEN 78-C-0060-11, Naval Training Equipment Center, Orlando, FL, 1981.

Logie, R. H., & Baddeley, A. D., "Cognitive performance during simulated deep-sea diving," ERGONOMICS, 28, 1985, pp. 731-746.

Lord, F. M., & Novick, M. R., STATISTICAL THEORIES OF MENTAL TEST SCORES, Addison-Wesley, Reading, MA, 1968.

McComas, L. A., "Effects of simulated ship motion on the performance of underway Officer of the Deck," Unpublished master's thesis, Naval Postgraduate School, Monterey, CA, 1986.

McKenzie, R. E., White, D. D., & Hartman, B. O., "Neptune: A multielement task system for evaluating human performance, RESEARCH REP. NO. SAM-TR-69-25, U.S.A.F. School of Aerospace Medicine / Aerospace Medical Division, Brooks Air Force Base, TX, October, 1969.

Moran, L. J., & Mefford, R. B., "Repetitive psychometric measures," PSYCHOLOGICAL REPORTS, 5, 1959, pp. 260-275.

Nicogossian, A. E., & Parker, J. F., SPACE PHYSIOLOGY AND MEDICINE, NASA / Scientific & Technical Information Branch, Houston, TX, 1982.

O'Donnell, R. D., "Development of a neurophysiological test battery for workload assessment in the U.S. Air Force," PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON CYBERNETICS SOCIETY, Atlanta, GA, 1981, pp. 398-402.

O'Neal, M. R., & Cannon, T. R., "Refractive error change at the United States Air Force Academy Class of 1985," AMERICAN JOURNAL OF OPTOMETRY & PHYSIOLOGICAL OPTICS, 64, 1987, pp. 344-354.

Payne, D. L., "Establishment of an experimental testing and learning laboratory," 4th International Learning Technology Congress & Exposition of the Society for Applied Learning Technology, Orlando, FL, February, 1982,

Shingledecker, C. A., Acton, W. H., & Crabtree, M. S., "Development and application of a criterion task set for workload metric evaluation," SAE TECHNICAL PAPER SERIES NO. 831419, Warrendale, PA, 1983.

Thorne, B., Genser, S., Sing, H., & Hegge, F., "Plumbing human performance limits during 72 hours of high task load," THE HUMAN AS A LIMITING ELEMENT IN MILITARY SYSTEMS, Defense and Civil Institute of Environmental Medicine, Toronto, Ontario, Canada, 1983.

Thorne, D. R., Genser, S. G., Sing, H. C., & Hegge, F. W., "The Walter Reed Performance Assessment Battery," NEUROBEHAVIORAL TOXICOLOGY AND TERATOLOGY, 7, 1985, pp. 415-418.

West, V., & Parker, J. F., "A review of recent literature: Measurement and prediction of operational fatigue," ONR REPORT NO. 201-067). Office of Naval Research, Arlington, VA, 1975. (NTIS No. AD A008405)

Winer, B. J., STATISTICAL PRINCIPLES IN EXPERIMENTAL DESIGN (2nd ed.), McGraw-Hill, New York, 1971.

Woodward, D. P., & Nelson, P. D., "A user oriented review of the literature on the effects of sleep loss, workrest schedules, and recovery on performance, ONR REPORT NO. ACR 206, Office of Naval Research, Washington, DC, 1974. (NTIS No. AD A009778)