AD-A056 047 NAVAL AEROSPACE MEDICAL RESEARCH LAB DETACHMENT MICHO--ETC F/G 5/10
THE DEVELOPMENT OF A NAVY PERFORMANCE EVALUATION TEST FOR ENVIR--ETC(U)
1978 R S KENNEDY, A C BITTNER

UNCLASSIFIED TM-77-01 NL

| OF |
AD
A056 047

END
DATE
FILMED
8-78
DDC

LEVEL II

(14) Technical Memorandum TM-77-01

(6) THE DEVELOPMENT OF A NAVY PERFORMANCE EVALUATION TEST FOR ENVIRONMENTAL RESEARCH (PETER)

(11)

(16) (PROJECT NO. F51524)

(19) MF51524.008

(9) Technical memo;

NRDTEPP/M-78
PE 62755N

By

(10) Robert S. KENNEDY, Alvah C. Bittner, Jr

Chief Human Performance Division and Officer-in-Charge
Naval Aerospace Medical Research Laboratory Detachment

and

Alvah C. Bittner, Jr.
Human Factors Engineering Branch
Point Mugu Test Center

(11) 1978

(12) 17p.

Presented at
Productivity Enhancement: Personnel performance assessment in Navy systems at
Navy Personnel Research and Development Center, San Diego, CA, October 1977

THE DEVELOPMENT OF A NAVY PERFORMANCE EVALUATION TEST FOR ENVIRONMENTAL RESEARCH (PETER)

Robert S. Kennedy, CDR MSC USN
Head Human Performance Division and Officer-in-Charge
Naval Aerospace Medical Research Laboratory Detachment
and
Alvah C. Bittner, Jr.
Human Factors Engineering Branch
Point Mugu Test Center

ABSTRACT

The basic problem with performance testing in exotic environments is the general unwillingness of investigators to take the time to standardize a test battery. Many other problems exist and are obvious to all who have tried to measure performance under usual and unusual environmental conditions. It is the purpose of this paper to set forth some of the problems that have grown out of our experiences and which we feel have not been extensively commented upon in the research literature, and also to describe our plan for solution.

Preface

The present plan is a simple one: The literature will be searched for human performance tasks which have been shown to degrade under motion (vibration and ship motion), during thermal exposure, and under pressure. The performances that meet these first criteria will be categorized as cognitive (decision making, information processing, judgment), motor (tracking, reaching), etc., and a taxonomy of performances will be developed. Additionally, each performance task will be evaluated in the following way: 20 subjects will be tested 10 times (5 days/week for 2 weeks) to determine three types of reliability: internal consistency, the accuracy and sensitivity to separate individuals, and the stability of this accuracy and sensitivity over repeated testing. Performances on these tasks will be compared to scores on other tests of mental functions. Progress to date will be reported.

The National Aeronautics and Space Administration, the Advanced Research Project Agency, the Navy (via the Office of Naval Research), and the Bureau of Medicine and Surgery have funded several studies (see Kennedy, 1977 for a review) which have nearly all made very similar points regarding the standardization of a performance test battery for assessment of environmental stressors. In the main, test batteries have been proposed, particularly factor analyzed batteries, but rarely have normative data been collected and never have practice effects been studied effectively.

The original title for the present paper was very broad and included all Navy R & D concerning performance. We intend, however, merely to present how the Naval Aerospace Medical Research Laboratory Detachment plans to research the general area, with specific application to our interests in the effects of ship motion or performance. It should be noted that, in addition to the human performance R & D already presented at this symposium by various members of the Navy

Personnel Research and Development Center, complementary programs also exist within the Engineering Psychology Programs of the Office of Naval Research and within the Human Effectiveness Programs of the Naval Medical Research and Development Command.

INTRODUCTION

Casual observation over several years of performance testing and a comprehensive reading of over 400 "human performance studies" in hyperbaria (see Bachrach & Kennedy, 1977, for a review) suggest that there is a need for future studies into the standardization of a human performance test battery.

In our opinion, the persons who initiated the experiments requiring performance testing in exotic environments were generally persons who became involved originally because of a primary interest in the environment rather than in the performance. (Within "environment" we include unusual sensory stimulations, drugs, fatigue, and even learning, as well as motion sickness, hyperbaria, etc.) Thus, we feel that, frequently, several criteria were employed (often trading back and forth among them) in the selection of tasks for inclusion in a battery to be assembled. These criteria have included the following:

1. Literature findings that were recollected, probably because the results of tests were unusual.

2. What colleagues and friends had done.

3. What demonstration experiments were performed in experimental psychology laboratory during their student days.

4. Chapter headings in Woodworth and Schlosberg (1954) and other standard texts.

5. Equipment left behind in the storage room of the laboratory by their predecessors.

6. That which could be quickly and easily assembled from clever ideas, (the so-called toy gadget approach).

7. Stock items from apparatus companies.

8. Logistic limitations forced by the environment or project (e.g., small, inexpensive, no tubes, portable, nonmagnetic, self-scored, no sparks, self-administered, battery powered, and rugged).

9. Similar to the work done by real-world persons.

10. A relatively basic kind of skill is involved; that is, learning theoretically SHOULD be able to be accomplished quickly.

11. Less often, performances could be expected to be disrupted on the task in this environment.

We believe that the criteria listed above have been employed often enough to assemble batteries so that these criteria are worth citing. It should also be noted, however, that, typically, a test battery was generally an ad hoc response to the imminent availability of an environmental condition, whether the environment was a hurricane (Kennedy, Moroney, Bale, Gregoire, & Smith, 1970), a rotating room (Guedry, Kennedy, Harris, & Graybiel, 1964; Fregly & Kennedy, 1965; Kennedy, Tolhurst & Graybiel, 1965), or a deep dive. Thus, long-range planning frequently is not possible. In summary, it is felt that performance test batteries are often assembled for largely practical reasons, on short notice, by persons whose major interest is not performance testing. To alleviate these problems we have combined, in tabular form, what we consider the traditional, important criteria for test construction along with the practical aspects concerning operational performance assessment. These criteria are summarized in Tables 1 - 4. In addition, other problems with performance test battery construction exist.

1. What performance tests are designed to measure

Although this distinction is not generally made, it is implicit that performance testing is undertaken for two main purposes: first, to be able to make some statement about the integrity of the organism, and second, to determine whether an environment interacts with an organism's ability to do a particular kind of work (cf. Table 3). In this paper, the first purpose will be called "CNS status," and the second, "effectiveness of a system's output." Examples of tests designed for the former purpose include reaction time, digit span, tremor, electroencephalogram, speed of tapping, and CFF. Examples of the latter include an underwater pipe puzzle, a sonar monitoring task, Morse code tests, and speech intelligibility tasks. Frequently, both types of tasks are included in a single experiment into the environment's effect on man and without regard to the distinction made above. The advantage of the latter approach is that the system's concept is used and the translation to real-activities is direct. (Also, subject cooperation is usually better.) The disadvantage is that no general principles are adduced and the application of the findings holds only for the stimulus condition employed. For instance, tracking studies with CRT displays have been conducted for many years and very few general rules have resulted (Adams, 1961). The major disadvantage of the first approach (index of an organism's integrity) is that they depend heavily upon the knowledge of the validity of the task. If only face validity is available, other considerations (money, size, apparatus, and availability) must be used to justify inclusion. If face validity is not evident, then justification is very tenuous.

The distinction made between these two strategies is subtle, but it is also real, and its existence complicates the results of many studies. This is chiefly due to the fact that the two approaches require different research philosophies, although the ultimate aim of both approaches is similar: namely, prediction (i.e., an ability to account for 100 percent of the variance).

The first approach comes directly from experimental psychology and usually follows an analysis of variance model. Thus, the numerous tests in a test battery are designed to sample all of the skills (factors) of the organism. The implication is that, if the full range of human abilities is tested, one can generalize the findings and apply them to other circumstances (e.g., subjects, treatments, etc.). This approach depends heavily upon following the principles of test

construction: (1) norms, (2) reliabilities, (3) validities, (4) factors tested, (5) effects of practice, and (6) individual differences. If all these principles were satisfactorily fulfilled, it would be possible to employ the test in an exotic environment and account for all the main effects of such an environment on human performance. For example, if it were known that hand dynamometry correlated perfectly with all other kinds of voluntary skeletal muscle output, and the Harvard Step Test (Kennedy & Hutchins, 1971) with all cardiac muscle output, then it would not be necessary to use other tests of these functions. The difficulty, of course, is that neither of these tests correlates sufficiently. Additionally, other "more psychomotor" tasks are even less clear-cut with regard to what they are measuring (i.e., validities). However, the problem does not end here. Reliabilities of a test battery--any test battery--are not completely known. No norms (expected values) are available on a sizable population, particularly when practice effects are concerned. However, factor analyses studies (e.g., those of Fleischman) have been completed for some samples.[1]

The second approach is in vogue more now than previously, probably because it emphasizes a systems approach. The statistical model employed is correlation, and in general, single factor studies are conducted. The overall plan is to replicate real-world work and to do it under controlled conditions. The second approach does not depend upon the validity of the task as heavily as the first method, since it, itself, is the work. However, the characteristics of the subjects are critical. It is important, and usually essential, that the subjects be the same kind of people as the real-world workers toward whom the data will be applied. The shortcoming of this strategy is also its chief advantage: the application of the findings from such studies is specific and immediate, but sometimes it is so specific that generalization within the same environment, but with slight differences, may not be possible.

2. Two experimental paradigms

There are two main ways in which to study the effects of the environment on a subject's ability to do work. The first (most often used) uses the subject as his own control and generally follows a pre-, per- and post- paradigm. In the pretest, the subject is practiced on all the tests to be employed in order to arrive at a learning plateau. Then he is placed in the experimental situation to see whether or not it disrupts performance. Posttesting is used to monitor recovery effects, if there are any. There are many problems with this approach. Chiefly, psychomotor performance almost never arrives at a plateau. This is discussed in more detail later in this paper. Asymptotes occasionally are obtained, but these, too, are infrequent. Even on tests where one would expect practice to be accomplished quickly (e.g., reaction time, CFF, tracking visual acuity),[2] the environment itself occasionally causes certain tests to be performed less well while standing during rotation, and is probably also measuring

---

[1] Sinbad (1969) is based on these studies and, when standardized, may be used to obviate some of the problems mentioned above.

[2] The use of signal detection theory (Swet, Tanner, & Birdsall, 1961) as a methodology may be helpful here, but as we all know from the way the 100-yard dash record is continually broken, it is not just a criterion problem. Stated differently, a knowledge of sensory sensitivity, d' (d-prime) separated from the subject's criterion (beta) would refine present knowledge, but d', even carefully and prudently measured, may change with practice.

body sway (Graybiel, Kennedy, Knoblock, Guedry, Mertz, McLeod, Colehour, Miller, & Fregly, 1965). This point will also be discussed later. Post-effects also present difficulties since motivation changes (e.g., end spurt in vigilance) usually attend the imminent completion of an experiment.

The alternative approach: to test "just before" and "just after" the environmental exposure (say a 12-hour overwater ASW flight) has its own problems; namely, the experimenter feels that it is necessary to be aware of the status of the subject during the exposure. If the testing is short (e.g., hand dynamometry), it can be influenced by the bias of a subject and summoning efforts for a "one-shot-deal" so that, often, changes are not obtained even though the subject is frankly tired. If the testing period is long (e.g., treadmill), it can contribute to the fatigue. In addition, lengthy posttests are often unfair to the subject.

3. Assessment of input-integrator-output circuits

The general form of psychological experimentation follows an S-R paradigm, or SOR, where O is for organism (Graham, 1951). Performance testing employs this paradigm particularly when "CNS status" type experiments are conducted. Typically, in these studies the experimenter is mainly interested in whether his treatment (drugs, hypoxia, confinement, magnetic fields) produces any CNS change. So, a stimulus is presented and the output of the organism is monitored for changes. Frequently, however, due account is not taken as to whether the stimulus was adequately received by the receptor (retina, ear, hair cells, etc.) then properly delivered along that nerve pathway; also, whether the output (muscle) pathway is similarly unaffected. For example, during acceleration stress, the lack of oxygen to the retina indicates that signals are not adequately received at the receptor site. This also occurs with the differences obtained in visual performance underwater. The physical conduction of light in air versus water may account for these differences -- most likely the visual signal is just not delivered to the receptor in water as well as in air, so one would not posit CNS changes underwater to account for the poorer visual acuity obtained. At the other end of the nerve-muscle circuit, changes in four-choice reaction time done underwater clearly have the friction of water on the one hand to slow down performance as well as the possible other effects of compression and mixed gases and so, probably, CNS changes cannot adequately be assessed with this task. So, too, past pointing underwater may be different: not because of central involvement, but because of inertial differences on the arm. This is not to imply that such studies should not be undertaken, rather, it behooves the experimenter to indicate where possible which part of the OSR circuit he is testing. Therefore, one must know about the transmission characteristics of light, the dependency of the retina on oxygen, and the viscosity and buoyancy characteristics of water. However, if such tasks are included in batteries that have other tests, (the intention of which is to tap the state of the CNS) when all results are reported together, there is confusion.

It would be useful to other investigators if results of experiments were reported relative to that part of the circuit which is being tested. This cannot be done in all cases, but it is possible to improve present reporting practices. Perhaps if we intellectually remove the known physical environmental effects from the periphery (nerve and muscle), we may be left with the finding that motivation

and the partial pressure of oxygen in the brain are the chief contributors to performance decrement under all conditions. The above criticism does not apply to the "systems output" type of studies which take no position regarding where in the circuit the problem occurs. Rather, their sole purpose is to determine whether an interaction of environmental condition occurs on people doing work. It is proposed that "CNS status" be used as a term to be contracted with "input/output quality" types of studies, whereby the former would deal with throughput changes due to the environment and the latter would address the physical aspects of the environment on man.

4. Practice effects

In a significant but not widely referenced paper, Bradley (1962) reported the persistence of sequence effects during psychomotor testing. Virtually all who study performance over many sessions have obtained similar findings. As was mentioned earlier, the investigator usually performs baseline pretesting before placing the subjects in the environment. Often, many trials are given (in one study, 7 days of testing) in an effort to have performance asymptotic "so that the pimple on the line can be more easily seen."[3] What is usually obtained is the well-known learning curve, which may, but does not always, asymptote. The problem with this approach is obvious, but there is another less obvious problem; that is, performance on a task after many trials is probably no longer an index of the same activity or place in the CNS that it was initially.

Studies by Ades and Raab, 1949, on the Kluver Bucy Syndrome (cited in Bachrach and Kennedy, 1977) illustrate the latter point where animals with certain portions of their brains removed were able to perform a visual discrimination task about as well as unoperated animals; however a simularly operated group was never able to learn this task.

Moreover, it is well known from the learning literature that, with extended practice, subjects overlearn, and when something is overlearned, it becomes more resistant to extinction. Therefore, for performance testing in exotic environments, if intensive practice is given on the tests prior to their use in the experimental environment, two factors appear inevitable: (1) the work is not an index of what it was at first, and (2) disruption of performance becomes very difficult. An example of this is as follows: move the index (first) and ring (third) fingers preferred hand together with the palms resting on a flat surface. Then move the second and fourth fingers together. Then, alternate 1 and 3, then 2 and 4, etc. Everyone can do this work, but it requires far more concentration for the average person than for a person who frequently plays the piano. The investigators believe that control for this activity is exerted high in the cortex for nonpianists, but has perhaps been shunted to a lower center in the CNS in practiced pianists. If the above is similar to what occurs in performance testing studies, the implications are obvious.

Because of the problems listed above, the following approach is planned: We feel that the approach is innovative, but it will draw heavily on the research literature for the initial selection of tests to be included for further study.

---

[3]Radloff, 1971, personal communication.

Those tests will be selected from the literature that meet criteria in one of the following areas: (1) demonstrated sensitivity to either thermal, motion, or hyperbaric environments by exhibiting degraded performances, (2) diagnostic capability (i.e., brain-damaged individuals have been found to perform differently from a normal population), and (3) measurement capability of a parameter of human information processing. After initial selection of the tests, the most promising will be subjected to further tests. The test and equipment attributes of each test will be viewed from the standpoint of the following factors ranked in general order of importance: (1) reliability (e.g., test-retest, alternate form, between and within administrations), (2) validity (e.g., predictive, context, construct, diagnostic-concurrent, fact), (3) other practical test factors (range of capability levels covered, sensitivity, transportability, efficiency), (4) equipment factors (e.g., availability, equipment reliability, transformability, safety, economy). Those tests that demonstrate a high level of adequacy on the above criteria will comprise an experimental battery. Performances on this battery will be compared to performances on a factor pure (e.g., Sinbad) battery to determine uniqueness of factors. Paper and pencil tests of cognitive functions (e.g., Bender-Gestalt, Guillford-Zimmerman) as well as well-standardized intelligence tests (e.g., Wais, Ravens, Stanford-Binet, Reitan, Halstead, Wunderlich) will be administered to this same population to further delineate and validate the factors obtained.

The first test that we have selected for further study is the so-called Beeper reviewed by Kennedy and Bruns (1975). The reasons for selecting this test originate partly from the literature review and partly from the study of acceleration stress by the NAS/NRC Committee on Bio-Astronautics, who convened a working group headed by Robert Galambos to discuss and report on principles and problems of performance testing. Using criteria based largely on earlier suggestions of Broadbent (1953), a performance test battery was proposed that would have general and specific applications.

We looked into Broadbent's report for ideas relative to the common problems of motion and acceleration stress and of exotic environments in general. Recommendations were also included for the use of tasks which are: "(a) work paced; (b) require vigilance; (c) over a long period of time; and (d) during which there is uncertainty in the stimulus display" (p. 22):

1. Laboratory norms on six different versions of this task for each of the approximately 100 college graduate males are available, as well as relationships to personality and other subject variables (e.g., hours of sleep) for these persons.

2. Neurophysiological correlates (vestibular nystagmus) of performance were shown.

3. Practice effects appear small on the three-channel auditory version and are known for the three-channel visual version.

4. The test can be group-administered.

5. It is relatively simple and inexpensive to construct.

6. There are many possibilities for constructing alternate forms.

7. Task difficulty can be controlled largely by instructions.

8. Latency of response within broad limits (namely, 1-2 seconds) is generally not a factor and so the task can appropriately be used even when environmental variables can interact physically with response speed (e.g., underwater).

9. Stimulus recording is binary and therefore is mechanically simple. Further, the regularity of the stimuli makes a scoring relatively easy and relatively independent of where on the magnetic tape a session begins.

10. Proportion measures are essentially linear (R .95) with absolute measures (namely, hits) and, therefore, direct comparisons can be made over different tasks.

11. Unlike many other vigilance tasks, many signals and responses occur and so individual time-line analyses are possible.

12. The results suggest that performance on forms of this task may be age-related.

The approach we have utilized includes the daily administration (15 minutes) of the Beeper for 2 weeks to study the reliability of the test in three ways: internal consistency, the accuracy and sensitivity to separate individuals, and stability of this accuracy and sensitivity over repeated testings.

We feel that this approach will serve as a model for future tasks to be included in our battery. At this writing, data are being collected, however the study is not completed. These results should be available at the meeting in October.

Table 1

Equipment Factors

Factors	Definition	References	Comments
Availability	Equipment software and hardware for presenting tasks, receiving responses, recording, scoring and integrating should be acquirable without excessive delays.	Alluisi (1967, 1969); Reilley & Cameron (1968); Kennedy (1971); Theologus et al. (1973)	Rose (1974) has suggested paradigm "reproducibility" (frequency with which a task has been studied) as a criteria for selection. Certainly selection of tasks most readily available in psychological laboratories would insure maximum cross laboratory availability. Some paper-and-pencil tasks rate high on this factor.
Equipment Reliability	Equipment software and hardware must be sufficiently reliable to permit sustained use for lengthy durations, i.e., have a high expected "mean time between failure." (MTBF)	Alluisi (1967, 1969); Theologus et al. (1973)	A method of checking hard and software states--proper or improper functioning--is a necessity for PTB tasks.
Transformability	Tasks can be adapted for administration in various environments of interest without seriously altering measurement capability.	Reilly & Cameron (1968)	Environments of interest could include "shirt sleeve laboratory," exotic environs (e.g., underwater), or field conditions. Portability (Rose, 1974) and potential for group administration (Kennedy, 1971) are valued elements.
Safety	Equipment should not present a potential health or safety hazard to subjects, and equipment must not be vulnerable to damage by stressed subjects.	Theologus et al. (1973)	This is the most important feature of any battery.
Economy	Costs for acquisition of equipment hard and software, administration, scoring, interpretation and maintenance should be reasonable.	Alluisi (1967, 1969); Reilley & Cameron (1968); Kennedy (1971); Theologus et al. (1973)	Temporal and monetary costs are important, albeit able to be traded off. Equipment based batteries have not been extensively applied or developed because of costs being excessive. Less expensive and sophisticated batteries would encourage standardizations of tasks in the literature.

Table 2

Reliability Factors

Factor	Definition	References	Comments
Test-Retest Reliability	Correlation established by administration of the same test to the same individuals on two different occasions.	Alluisi (1969) Grodsky (1967) Kennedy (1971) Theologus et al. (1973)	Experimenters using PTBs frequently administer the same task to subjects a large number of times. This has been shown in the literature to frequently result in changes in the nature of what is being measured and low correlations between early and later trials (cf., Woodrow, 1938 a & b; Fleishman and Hempel, 1955; Parker & Fleishman, 1960; Parker & Fleishman, 1961; Parker, 1964). Test-Retest reliabilities need to be determined over numbers of trials task will be administered.
Alternate-Form Reliability	Correlation established by administration of two "equivalent forms" of the same test (measuring same aspect of performance but of different questions or items) on two different occasions.	Theologus et al. (1973) Teichner (1974)	Alternate form reliability is appropriate for tasks with elements which lose potency when exposed to subjects. Also note that comment for Test-Retest applies with alternate forms which are employed over substantial numbers of trials.
Internal Consistency Reliability	Correlation estimate of the homogenity of a task's item scores established on a group of individuals on one occasion.	Theologus et al. (1973)	Note comment for Test-Retest has implication for internal consistency estimates a point delineated by Thorndike (1949).
Between Test Administrators Reliability	Correlation established by administration and scoring of the same or equivalent form of a task to the same individuals by two administrators.	Teichner (1974)	This reliability has not been of interest in most developments of PTBs, although the "experimenter effect" has a long history in experimental research (cf., Rosenthal, 1961).
Within Test Administrators Reliability	Correlation established by administration and scoring of the same or equivalent form of a task by the same administrator on two different occasions.	Teichner (1974)	This is the special case under which most Test-Retest and alternate forms reliabilities are established.

Table 3

Validity Factors

Factor	Definition	References	Comments
Predictive Validity	Correlation between operator performance on a task (or tasks) and future criterion performance or status.	Alluisi (1967, 1969) Grodsky (1967) Theologus et al. (1973)	"Real world" performance is a concern of experimenters who optimize this criteria vs. diagnosis of performance which is concerned with concurrent diagnostic validity.
Concurrent (Diagnostic) Validity	Correlation between test score and a diagnostic criterion status obtained at approximately the same time.	Teichner (1974)	Teichner (1976) stresses diagnostic aspects or tasks for assessment of subjects internal status (e.g., Is a nervous system dysfunction present?)
Content Validity	Extent to which a task or task battery covers a representative sample of the behavior domains to be measured.	Alluisi (1967, 1969) Reilley & Cameron (1968)	Related to content validity are the concepts of battery "generality" or "comprehensiveness" given as criteria by Theologus et al. (1973), Rose (1974) and Teichner (1974). These concepts stress that a battery should encompass as many critical aspects as possible while minimizing redundancy.
Construct Validity	Extent to which a test may be said to measure a "theoretical construct" or trait where theoretical construct or trait is established by convergence of information from a variety of sources.	Theologus et al. (1973) Rose (1974)	Rose (1974) has particularly stressed this concept by his emphasis on well used "paradigms" from experimental psychology with correlational and factor analysis as methods of convergence.
Factor Validity	Extent to which factor analysis indicates task as identifying or correlating with a factor.	Reilly & Cameron (1968) Theologus et al. (1973) Rose (1974)	Factor analysis has additional use in the assessment of the amount of redundancy in a battery.

Table 3 (Cont.)

Validity Factors

Factor	Definition	References	Comments
Face Validity	Extent to which test "looks valid" to subjects who take it, experimenters, or other observers	Alluisi (1967, 1969) Grodsky (1967) Reilly & Cameron (1968) Theologus et al. (1973)	Alluisi (1967, 1969) and Grodsky (1967) both stress need of face validity to insure subjects feel tasks are relevant and are motivated. Theologus et al. (1973), however, stresses the need of "...face validity to permit subjective generalization of effects...to the effects... on a 'real world' task..." Attempts to measure task face validity have not been reported. Briefing on importance of tasks vs. "face validity" method of motivating subjects have not been reported in literature although used as research strategy (e.g., by Cross & Bittner, 1969).

Table 4

Ability Range, Sensitivity, Trainability and Efficiency Factors

Factor	Definition	References	Comments
Range of Ability Levels Covered	Extent to which differing subject populations (varying in background, developmental level, training, etc.) can be tested.	Alluisi (1967, 1969) Reilly & Cameron (1968) Teichner (1974)	Although pointed out as important, this factor has not been given much study.
Sensitivity	Extent to which test reflects effects of conditions of study.	Alluisi (1967, 1969) Grodsky (1967) Reilly & Cameron (1968) Theologus et al. (1973) Rose (1974) Teichner (1974)	Alluisi (1967, 1969), Grodsky (1967), and Theologus et al. (1973) emphasize sensitivity to effects only to magnitude experienced in operational situation. Reilly & Cameron (1968) define sensitivity as extent to which conditions are likely to influence performance. Teichner (1974), however, discusses it in terms of quickness of detecting dysfunctions. Sensitivity $S = (M_1 - M_2)/SD_1^2 + SD_2^2)$, where M_i and SD_i^2 are the mean and variance under condition "i" appears a more useful metric for purposes such as Teichner (1974).
Trainability	Asymptotic levels of performance should be attainable with the selected tasks after a minimum of training except where tasks are selected to measure changes in this function per se.	Alluisi (1967, 1969) Kennedy (1971) Theologus et al. (1973) Rose (1974)	To date studies to insure "asymptotic levels of performance" have not been accomplished for non-leanring tasks. Development of tasks to measure different types of learning per se is very lacking though, as Teichner (1974) point out, the most sensitive test will have "minimal practice" as a characteristic. Learning tasks appear to have high potential for future PTBs and should be more fully studied.
Efficiency	Importance of test's contribution with respect to cost, time and effort of implementation.	Reilly & Cameron (1968) Theologus et al. (1973) Teichner (1974)	Contributions of tasks in terms of cost, time and effort appear to be accomplishable by appropriate analysis. The reliability of a task for one minute of study ($r_{1:1}$), for example, can be estimated by, $r_{1:1} = r_{tt}/(t + (1 - t)\, r_{tt})$, where r_{tt} is the observed reliability for t minutes of observation.

REFERENCES

Adams, J. A. Human tracking behavior. Psychological Bulletin, 1961, 58, 1, 55 - 79.

Alluisi, E. A. Optimum uses of psychobiological sensorimotor and performance measurement strategies. Human Factors, 1975, 17 (4), 309 - 320.

Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. Human Factors, 1976, 9 (4), 375 - 384.

Bachrach, A. J. Psychological research: An introduction. (2nd ed.) New York: Random House. 1965.

Bachrach, A. J. & Kennedy, R. S. Psychological performance testing under water and pressure: Problems and prospects. Bethesda, Md.: U.S. Naval Medical Research Institute, 1977. (In press).

Bradley, J. V. Studies in research methodology. III. The persistence of sequential effects despite extended practice. Wright-Patterson Air Force Base, Ohio, MRL Technical Document 62-60, June 1962.

Broadbent, D. Noise, paced performance, and vigilance tasks. British Journal of Psychology, 1953, 44, 295 - 303.

Cross, K. A. & Bittner, A. C., Jr. Accuracy of altitude, roll angle, and pitch angle judgments as a formation of size of vertical contact analog display. Point Mugu, CA: Naval Missile Center, January 1969 (PM-69-2).

Fleishman, E. A., & Hemple, W. E., Jr. The relation between abilities and improvement with practice in a visual discriminatory task. Journal of Experimental Psychology, 1955, 49, 301-312.

Fregly, A. R., & Kennedy, R. S. Comparative effects of prolonged rotation at 10 rpm on postural equilibrium in vestibular normal and vestibular defective human subject. Aerospace Medicine, 36, 12, 1965.

Guedry, F. E., Jr., Kennedy, R. S., Harris, C. S., & Graybiel, A. Human performance during two weeks in a room rotating at three rpm. Aerospace Medicine, 35, 11, 1964.

Graham, C. H. Visual perception. In S. S. Stevens (Ed.) Handbook of experimental psychology. New York: Wiley & Sons, 1951.

Graybiel, A., Kennedy, R. S., Knoblock, E. C., Guedry, F. E., Jr., Mertz, W., McLeod, M. E., Colehour, J. K., Miller, E. F., II, & Fregly, A. R. Effects of exposure to a rotating environment (10 rpm) on four aviators for a period of twelve days. Aerospace Medicine, 36, 8, 1965.

Grodsky, M. A. The use of full scale mission simulation for the assessment of complex operator performance. Human Factors, 1967, 9 (4), 341 - 348.

Kennedy, R. S. Individual differences in auditory vigilance performance on the band-pass ability (B-PA) test: Some theoretical considerations. Presented at Human Factors Society annual meeting, San Francisco, CA, October 1970.

Kennedy, R. S. A sixty-minute task with 100 scoreable responses. Naval Aerospace Medical Center, Pensacola, Florida, NAMI-1045, 1968.

Kennedy, R. S. A performance assessment in exotic environments: A flexible, economical, and standardized vigilance test. Paper presented at the Fifteenth Annual Human Factors Society Meetings, New York, October 1971.

Kennedy, R. S., & Bruns, R. A. Consideration for the utilization of a flexible economical, vigilance test to assess performance in exotic environments. Presented at the October 1975 Aerospace Medical Panel Specialists' Meeting, Ankara, Turkey.

Kennedy, R. S., & Hutchins, C. W. Relationships between physical fitness, endurance, and success in flight training. Naval Aerospace Medical Center, Pensacola, Florida, NAMI-1088, in press, 1971.

Kennedy, R. S., Moroney, W. F., Bale, R. M., Gregoire, H. G., & Smith, D. C. Motion sickness symptomatology and performance decrements occasioned by hurricane penetrations in C-121, C-130, and P-3 Navy aircraft. Aerospace Medicine, 43, 1235 - 1239, 1972.

Kennedy, R. S., Tolhurst, G. C., & Graybiel, A. The effects of visual deprivation on adaptation to a rotating environment. NSAM-918. NASA Order No. R-93. Pensacola, FL: Naval School of Aviation Medicine, 1965.

Kennedy, R. S. PETER for Mentation Mensuration. A point paper, Naval Aerospace Medical Research Laboratory Detachment, New Orleans, LA, December 1977. (In press).

Parker, J. F., Jr. & Fleishman, E. A. Ability factors and component performance measures as predictors of complex tracking behavior. Psychological Monographs, 1960, 74 (16, Whole No. 503).

Parker, J. F., Jr. & Fleishman, E. A. Use of analytical information concerning tasks requirements to increase effectiveness of skilled training. Journal of Applied Psychology, 1961, 45, 295-303.

Parker, J. F., Jr. Use of an engineering analogy in the development of tests to predict tracking performance. The Matrix Corporation. (Office of Naval Research Contract No. ONR-3065(00)). February 1964.

Reilly, R. E. & Cameron, B. J. An integrated measurement system for the study of human performance in the underwater environment. ONR Contract N0014-67-C-0410, December 1968.

Rose, A. M. Human information processing: An assessment and research battery. University of Michigan, Technical Report No. 46.

Rose, A. M. Human information processing: An Assessment and research battery. Ann Arbor, MI., University of Michigan, Doctoral dissertation, 1974, also published as AFOSR-PR-74-1372 (AD-785-411).

Rosenthal, R. Experimental outcome--orientation and the results of the psychological experimentation. Psychology Bulletin, 1963, 61, (6), 405-442.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. Decision processes in perception. Psychological Review, 1961, 68, 301 - 340.

Teichner, W. H. Quantitative models for predicting human visual/perceptual/motor performance. New Mexico State University/Office of Naval Research - Technical Report 74-3, Las Cruces, New Mexico, October 1974.

Theologus, G. C., Wheaton, G. R., Mirabella, A., Brakler, R. E., & Fleischman, E. A. Development of a standardized battery of performance tests for the assessment of noise stress effetts, NASA CR 2149, Washington, D. C., 1973.

Thorndike, R. L. Personnel selection: Tests and measurements techniques. New York: Reilly, 1949

Woodrow, H. The effect of practice on groups of different initial ability. Journal of Educational Psychology, 1938, 29, 268-278(b).

Woodrow, H. The relation between abilities and improvement with practice. Journal of Educational Psychology, 1938, 29, 215-230(a).

Woodworth, R. S. & Schlosberg, H. Experimental psychology. New York: Henry Holt & Co., 1954.

ABOUT THE AUTHOR

Robert S. Kennedy has been an aerospace experimental psychologist since he entered the Navy in 1959. He received an MA in experimental psychology from Fordham University in 1959 and a Ph.D. from the University of Rochester in 1972. His previous military experience includes two tours in the Aerospace Psychology Division at the Naval Aerospace Medical Institute, Pensacola, Florida, where he conducted research on vestibular function, motion sickness, vigilance, and habituation in exotic environments; one tour at the Behavioral Sciences Department at the Naval Medical Research Institute, Bethesda, Maryland, working on the Man-in-the-Sea program; one tour at the Pacific Missile Test Center, Point Mugu, California; and one tour at the Air Development Center, Warminister, Pennsylvania, where he worked mainly on the development, test, and evaluation of airborne weapons systems from the standpoint of human factors engineering. Presently, he is the Officer-in-Charge of the Naval Aerospace Medical Research Laboratory Detachment working on human performance mensuration in unusual environments, specifically ship motion.