(Workshop on "Advances in NASA Relevant, Minimally Invasive Instrumentation," Jet Propulsion Laboratory, Pasadena, CA, 1984)

A PORTABLE BATTERY FOR OBJECTIVE, NON-OBTRUSIVE MEASURES OF HUMAN PERFORMANCE

Robert S. Kennedy, Ph.D.
Essex Corporation
Orlando, Florida

ABSTRACT

A need exists for a standardized battery of human performance tests in order to measure the effects of various treatments. The present paper reports on progress in such a program, funded jointly by NASA and Navy. Three batteries are available which differ in length (7.5; 15; 30 minutes), and the number of tests in the battery (3; 10; 15). All tests are implemented on a portable, lap-held, briefcase-size microprocessor (NEC PC 8201A). Performances measured include information processing, memory, visual perception, reasoning, motor skills, etc. Current programs are underway to determine norms, reliabilities, stabilities, factor structure of tests, comparisons with marker tests, apparatus suitability, etc. Rationale for the battery is provided.

INTRODUCTION

We originally set out to standardize a battery of human performance tests in response to a Navy requirement to study the effects of ship motion on humans. The focus of that program centered on repeated measures because nearly all studies of the effects to humans of exotic environments follow such a paradigm. Because of this, two statistical properties of tests received more attention in our program than in those reported by others: The two properties we studied were stability and reliability. Validity and factor structure, often examined first by others, have been left until later in the program. We continue to argue that this is the correct emphasis because without the first two properties, the second two cannot be meaningfully determined.

The results of that program, called PETER (Performance Evaluation Tests for Environmental Research), were reported in a series of 90 publications (cf., Harbeson, Bittner, Kennedy, Carter, & Krause, 1983, for a complete list). A recent review reported on 114 tests and considered 30 suitable for incorporation into a battery (Bittner, Carter, Kennedy, Harbeson & Krause, 1984). The criteria considered important for such a battery are listed in Table I. The results of the good tests appear in Table II.

Everyone ordinarily concurs that stability and reliability are important issues in testing, but it is not always evident to what extent. What follows is our rationale for selecting these two as our focus.

Table 1. Definition of Task Features

FEATURE	DEFINITION
NAME	Name of the task or measure as used in the literature.
FACTOR	The factor(s) assessed by the measure as identified in the literature or by judgments of the authors.
DOMAIN	Characterization of the domain(s) of assessment of the capability as cognitive, perceptual (including sensory), or motor.
ADMINISTRATION TIME	The typical testing time for a measure; this includes all testing time required to obtain a score. (e.g., components of a derived score)
TYPE OF ADMIN.	Identification of task as individually or group administered.
TOTAL STABILIZATION TIME IN MINUTES (DIFFERENTIAL)	The total stabilization time is the amount of elapsed experimental time (whether massed or distributed) required for mean, variance, and differential (correlational) stabilization. (The amount of elapsed practice time required for Differential Stabilization alone is in parentheses).
RELIABILITY EFFICIENCY (3 minutes)	The differentially stabilized reliability normalized to a 3 minute administration. Normalization to 3 minutes was by the Spearman-Brown Equation (Bittner & Carter, 1981; Winer, 1971).
REFERENCES	Cited in order are the relevant stability study, the original source of the measure, and occasionally other significant references.

Reliability: If performances between subjects differ on tests, those differences may be due to unforeseen, uncontrolled and perhaps unrelated issues, in which case the between subject differences are considered to be error. Alternatively, there may be differences in capability, in which case they are considered true, and if these differences can be measured, they can improve the precision of the statistic which is employed in studying the potential effects of treatments.

For example, the equation below is one of the well known variants on students' t (Winer, 1971) for measuring the differences in means (X) over two independent groups:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2}{N} + \frac{SD_2^2}{N}}} \qquad (1)$$

Moreover, for the special case where N is equal in the two administrations, this equation is sometimes written:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{N}}} \qquad (2)$$

And when, in addition to equal N, the variances are equal over the two occasions or administrations, the equation may be simplified to:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2SD^2}{N}}} \qquad (3)$$

The question these statistics (Equations 1, 2 and 3) permit one to answer is whether the obtained difference between two means X_1 and X_2 (say one group "with" and one "without" the drug) is likely to have occurred by chance. The way we decide is by forming a ratio of the DIFFERENCES (numerator) to the ERROR TERM (denominator). If the difference is many times the error, we infer the difference is not likely to be chance. If the ratio is small, then the converse. When the cost of being wrong is high, we take steps to improve our precision by increasing sample size or we select measures of behavior which exhibit small between subject differences because both of these serve to reduce the denominator. Also, practice usually will reduce between subject differences (variances) too. However, in most cases of human performance measurement, a great deal of the differences between subjects are not ERROR and they are large. People differ along behavioral dimensions.

Although the size of the sample would also serve to reduce the size of the error term, this option is not always available. In studies of environmental stress and drugs, indeed, it is often impossible, and probably unethical, to expose large groups to the treatments. In these cases, for economy and precision, we usually follow a repeated measures design and each subject serves as his own control. In such a case, the t statistic uses the equation below.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2 - 2r_{12}SD_1SD_2}{N}}} \qquad (4)$$

Note that much of this equation is the same as before. The one addition is the covariance term and this indicates that you may reduce the error term proportional to how well they are correlated over the two exposures. However, it is not obvious to what extent the error term may be reduced but two examples will suffice.

Again, if we assume that the variances are equal (NOT NECESSARILY SMALL!), then the equation simplifies to:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2SD^2 - 2SD^2(r_{12})}{N}}} \qquad (5)$$

Then, if r_{12} = 0.00, the equation returns to the t test (cf., Equation 3 above) that we used for examining the differences in <u>two different groups.</u>

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2SD^2}{N}}} \qquad (6)$$

Alternatively, if the retest reliability for r_{12} = 1.00, then the equation simplifies to:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{0}{N}}} \qquad (7)$$

And in this case, the ERROR TERM approaches zero and thus the obtained differences will be true and significant when they occur. This effort, we believe, provides the best opportunity for obtaining sensitive nonintrusive measures of human performance that we know of.

Stability: We consider that tests must exhibit stability of means, standard deviations or variances and of correlation (cf., Bittner & Carter, 1981; for a review). To be considered stable, means over sessions should be level or asymptotic and, provided that other criteria are met, may also show a regular and predictable trend. Standard deviations should be constant or they may increase proportional to mean increase. Correlations over sessions, to be considered stable, should be constant with no change due to increasing separation of trials. It would not do to take important steps to obtain high reliabilities and then have them change over trials or sessions. When the latter change occurs, it is considered to be an example of "superdiagonal form" (Humphreys, 1960, Jones, 1970, 1972, 1980) and the task is rated unstable. The consequences of such an occurrence are that in the extreme case (retest correlations decrease to zero over trials - Kennedy, Bittner & Harbeson, 1981) the capacity or ability which is being measured disappears and a new one takes its place. In the less extreme case the factor structure of the test shifts to some extent. So far as we can tell, no other attempt at test battery standardization set stability of the correlations as a requirement. If correlations change by becoming lower: a) the task will be insensitive to change **and** b) even if it were to change, you wouldn't know what it tested.

For more information about methods for stability analysis, see Bittner, 1979; and for sophisticated treatments see Jones, Kennedy & Bittner (1981) and Steiger (1980).

Several items emerged from the PETER program in addition to the 30 so-called "golden hits" (Table II). We discovered or rediscovered outcomes that others had reported elsewhere, although not widely.

A. Difference scores

Difference scores have been reported to have poorer reliability than the primary scores from which they are derived (Cronbach & Furby, 1970). Carter and Krause (1983) demonstrated algebraically that slope scores are a form of difference scores, as are percents, ratios and other derived scores. They (Carter & Krause, 1983) then went on to show empirically that slope scores in several experiments within the PETER program possess very low reliabilities, if they are present at all. Bittner et al (1984) reported that derived scores fared significantly poorer ($P<.01$) than other types of scores in the 100+ tests evaluated. Many of the information processing tasks so popular these days employ a slope score as an index of performance. Some of these are advocated as potential indicants of individual traits or capabilities of individuals, and it is implied they may be useful in selection. This advocacy is probably ill-advised. Tests which have been indicted because they contain such scores include Stroop tests, Steinberg's tests, Neisser's tests, reaction time (e.g., Hick's Law) and others. Slope scores do show group differences. For example, the color-word condition on the Stroop test (Harbe-

TABLE 2: GOOD**

NAME	FACTOR	DOMAIN	ADMIN TIME (MIN)	ADMIN TYPE	TOT STAB TIME IN MINUTES (DIFF)	REL EFF MIN IAB	REFERENCES
AIMING	AIMING: FINE EYE-HAND COORDINATION (FLEISHMAN & ELLISON, 1962)	P M	2	G	30(30)	0.87	KRAUSE & WOLDSTAD (1983); FLEISHMAN & ELLISON (1962)
ARITHMETIC: VERTICAL ADDITION	NUMBER FACILITY (N) (EKSTROM ET AL., 1976)	C	4	G	48(8)	0.90	BITTNER, CARTER, KRAUSE, KENNEDY, & HARBESON (1983); CARTER & SBISA (1982)
ASSOCIATIVE MEMORY: NUMBER CORR: LIST 1	ASSOCIATIVE MEMORY (MA) (EKSTROM ET AL., 1976)	C	2.5	G	20(20)	0.65	CARTER & KRAUSE (1982); UNDERWOOD ET AL. (1977); KRAUSE & KENNEDY, 1980
ATARI● AIR COMBAT MANEUVERING	PURSUIT TRACKING (KENNEDY, BITTNER & JONES, 1981)	P M	2.25	I	135(135)	0.63	JONES, KENNEDY, & BITTNER (1981); KENNEDY, BITTNER, HARBESON, & JONES (1982)
ATARI● ANTIAIRCRAFT	UNKNOWN	P M	2.25	I	126(126)	0.67	JONES & KENNEDY (IN PRESS) WITH ADAPTATIONS
CHOICE REACTION TIME: 1-CHOICE	SIMPLE REACTION TIME (DONDERS, 1868)	P	5.0	I	35(35)	0.58	KRAUSE & BITTNER (1982); TEICHNER & KREBS (1974)
CHOICE REACTION TIME: 4-CHOICE	CHOICE REACTION TIME (DONDERS, 1868)	P	5.0	I	50(50)	0.80	KRAUSE & BITTNER (1982); TEICHNER & KREBS (1974)
CODE SUBSTITUTION	MEMORY ASSOC.(MA) PERCEPTUAL SPEED (P)(EKSTROM ET AL., 1976)	C P	2.0	G	16(16)	0.84	PEPPER, KENNEDY, BITTNER, & WIKER (1980); WECHSLER (1981)
FLEXIBILITY OF CLOSURE	CLOSURE, FLEXIBILITY OF (CF) (EKSTROM ET AL., 1976)	P	3	G	9(9)	0.88	BITTNER, ET AL. (1983); MORAN & MEFFORD (1959)
GRAMMATICAL REASONING	REASONING, LOGICAL (RL) (EKSTROM ET AL., 1976)	C	1.5	G	18(18)	0.93	BITTNER, ET AL. (1983); CARTER, KENNEDY, & BITTNER (1981); BADDELEY (1968)
GRAPHEMIC AND PHONEMIC ANALYSIS: SENSE/ NONSENSE	READING SPEED (BARON & MCKILLOP, 1975)	C	8	G	16(16)	0.66	HARBESON, KENNEDY, KRAUSE, & BITTNER (1982A); BARON & MCKILLOP (1973); ROSE & FERNANDES (1977)
LETTER CLASSIFICATION: NAME	RETRIEVAL FROM LTM & MATCHING (POSNER & MITCHELL, 1973)	C	12	G	84(84)	0.55	HARBESON, ET AL. (1982A); POSNER & MITCHELL (1973); ROSE & FERNANDES (1977)
LETTER CLASSIFICATION: CATEGORY	RETRIEVAL FROM LTM & MATCHING [POSNER & MITCHELL, 1973	C	11	G	121(121)	0.69	HARBESON, ET AL. (1982A); POSNER & MITCHELL (1973); ROSE & FERNANDES (1977)

**Complete reference citations are contained in Bittner et al. (1984).

TABLE 2: GOOD (CONTINUED)**

NAME	FACTOR	DOMAIN	ADMIN TIME (MIN)	ADMIN TYPE	TOT STAB TIME IN MINUTES (DIFF)	REF 3 MIN RELIAB	REFERENCES
MANIKIN TEST: LOG. LATENCY	SPATIAL TRANSFORMATION (EGAN, 1978)	P	7	I	14(14)	0.79	CARTER & WOLDSTAD (1982); READER, BENEL, & RAHE (1981)
MINNESOTA RATE OF MANIPULATION: TURNING	MANUAL DEXTERITY (FLEISHMAN & ELLISON, 1962)	M	2-4	I	10(10)	0.64	CARTER, STONE, & BITTNER (1982); SCHOENFELDT (1972)
PATTERN COMPARISON: NUMBER CORRECT MINUS NUMBER INCORRECT	SPATIAL ABILITY (KLEIN & ARMITAGE, 1979)	P	2	G	18(18)	0.93	SHANNON, CARTER, & BOUDREAU (1983); KLEIN & ARMITAGE (1979); CARTER & SBISA (1982)
PERCEPTUAL SPEED	PERCEPTUAL SPEED (PS) (EKSTROM ET AL., 1976)	P	2.5	G	23(15)	0.86	BITTNER, CARTER, KRAUSE ET AL.(1982); MORAN & MEFFORD (1959)
SEARCH FOR TYPOS IN PROSE: MEDIAN DETECTION TIME	READING SPEED	P	6	I	54(54)	0.65	SHANNON ET AL. (1983); CARTER & KRAUSE (1983)
SPOKE CONTROL (C) TASK	SPEED ARM MOVEMENT (FLEISHMAN & ELLISION, 1962)	M	0.67 APPROX	G	1(1)	0.95	BITTNER, LUNDY, KENNEDY, & HARBESON (1982)
STERNBERG ITEM RECOGNITION: POSITIVE SET 1	SHORT TERM MEMORY SCAN (STERNBERG, 1966, 1975)	C	3	I	18(18)	0.70	CARTER, KENNEDY, BITTNER, & KRAUSE (1980); STERNBERG (1969, 1975)
STERNBERG ITEM RECOGNITION: POSITIVE SET 4	SHORT-TERM MEMORY SCAN (STERNBERG, 1966, 1975)	C	3	I	15(9)	0.80	CARTER ET AL. (1980); CARTER & KRAUSE (1982); STERNBERG (1969, 1975)
STROOP: COLOR WORDS (CW)	MIXED	C P	0.5	G	1.5(1.5)	0.97	HARBESON, KRAUSE, KENNEDY, & BITTNER (1982B)
TRACKING: CRITICAL	TRACKING, CRITICAL (JEX, MCDONNELL & PHATAK, 1966)	P M	1	I	100(100)	0.60	DAMOS, BITTNER, KENNEDY, HARBESON, & KRAUSE (1984); JEX, MCDONNELL & PHATAK (1966)
TRACKING: DUAL CRITICAL	TRACKING, CRITICAL & DUAL FACTOR? (DAMOS ET AL., 1981)	P M	1	I	100(100)	0.50	DAMOS, BITTNER, KENNEDY, & HARBESON (1981)
VISUAL CONTRAST SENSITIVITY: METHOD OF INCREASING CONTRAST	CONTRAST SENSITIVITY FUNCTION: 1, 2, 4, 8, 16 cpd (GINSBURG & EVANS, 1982)	P P P P P	3 3 3 3 3	I I I I I	<1(<1) <1(<1) <1(<1) <1(<1) <1(<1)	0.51 0.52 0.74 0.75 0.53	GINSBURG, BITTNER, KENNEDY, HARBESON (1983); GINSBURG & EVANS (1982)
WORD FLUENCY	WORD FLUENCY (FW) (EKSTROM ET AL., 1976)	C	3	G	<1(<1)	0.79	CARTER, CURLEY, & STYER (IN PRESS)

**Complete reference citations are contained in Bittner et al. (1984).

son, Krause, Kennedy & Bittner, 1982) virtually always has a greater latency than the black-and-white word or color block conditions. The reliabilities of these basic scores are in the range of .90, but their differences have reliablities which are essentially zero.

B. Power from replications

Another methodological finding within the PETER program had to do with the tradeoffs between sample size and test-retest reliability in the special cases of repeated measures where variances are constant. If one uses "Student's t" formula, where each subject serves as his own control, great power is obtained by having high test-retest correlations. This issue is described well in the paper by Carter, Kennedy & Bittner, (1981) where a nomogram is available (Figure 1) to permit the tradeoff of sample size for reliability of test scores to obtain iso-precision of significance. If one is dealing with ability measurement, and one is faced with a repeated measures design in an unusual environment, it is ordinarily difficult to increase the sample size beyond some value and 12 or 15 is not an uncommon upper limit. Sharpening the t-test is ordinarily thought to be best effected by minimizing between-subject variance or increasing sample size. A third way is by maximizing the test-retest reliability. The latter can be more economical than increasing sample size, and if hazard is involved is probably more ethical. A fourth method is replication (Dunlap, Bittner, & Jones, 1983).

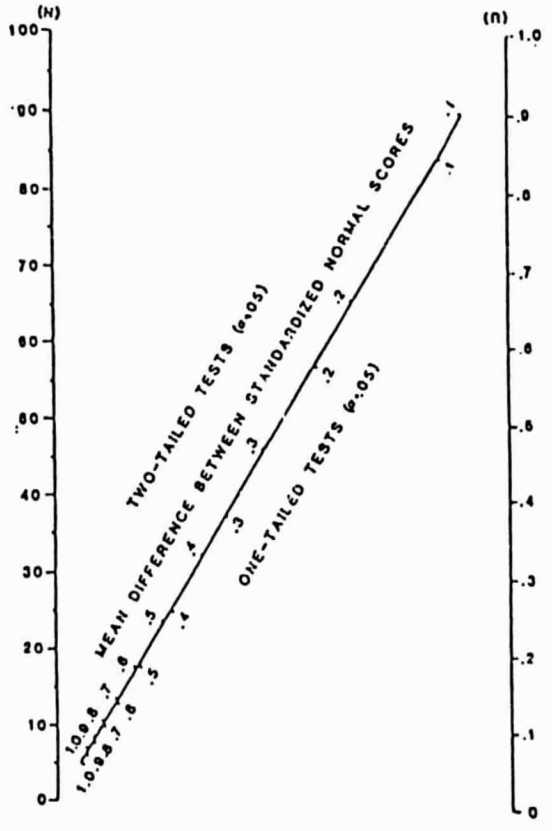

Figure 1. Nomogram relating sample size (N), intertrial correlation (R), and the smallest significant (p = .05) difference (D).

C. Convergence of factor structure

A third issue (not completely studied but important in our judgment) emerged when we evaluated families of tests. A "memory" family, a "video-game" family, a "search and target acquisition" family, an "information processing" family, and a "cognitive" family all were studied. In these studies it appeared as though fewer factors were available late in practice than earlier. That is, the factor structure resembled what was to be expected from the reports which were in the literature only during the initial practice on the tasks. Once the tasks reached full stability, the number of factors appeared to converge and be fewer than earlier in practice. For example, four factorially different tests from the Underwood battery (Underwood, Boruch & Malmi, 1977) were administered to the same population over three weeks (Harbeson, Krause, & Kennedy, 1981). After dropping the tests that had no reliability at all and/or that did not stabilize a single large factor seems capable of describing the performances that result. A not dissimilar finding occurred when a series of information processing tasks was studied. Again this outcome was obtained after dropping unstable or unreliable task's scores. Of those that remained, it was not uncommon for one factor to be able to be used to characterize all performance. This was never pursued adequately in the PETER program and should be followed-up because it is possible that given adequate practice to achieve stability on a series of tasks, one may find that fewer factors are necessary than during the early stages of acquisition. This result could have a profound effect on primary and secondary selection, as well as other forms of testing.

D. Repeated Measures

We were prepared to find long-term practice effects from other work we had done, so we began with the idea that it might take many replications to obtain stability. We originally set up for 10 session studies, and lengthened that to 15 (i.e., three weeks). As work progressed in the PETER Program, four issues emerged related to extended practice: 1) improvements persist over many and sometimes all sessions; 2) they are often very large; 3) they are not limited to tests of SKILL but occur in ABILITY tests (e.g., cognitive and information processing) too; and 4) the improvement often occurs at different RATES for different persons. This latter led us to the quest for "differential stability."

E. Differential Stability

"Differential stability" emerges from a notion offered by Jones in the early 70s. Picking up on ideas discussed by Humphreys (1960), Jones (1970, 1972) suggested that when practice occurs, performance improves, and not always at the same RATE for all subjects. Therefore, some people acquire skill rapidly and others acquire it less rapidly. Moreover, TERMINAL skill levels are not necessarily predictable from subjects' original performance (or intercept), nor from the rate at which they acquire

terminal levels of performance. From work in the PETER program we now recognize that the two-process theory (Jones, 1970), which had been developed to handle data in the area of repeated measures of SKILL acquisition, extends to memory, cognition, information processing, and probably all human performances. Thus initial scores on ABILITY tests and on SKILL tasks may not be perfectly correlated with terminal levels, nor with the rate at which the terminal levels are reached.

More importantly, it follows that the terminal level of performance may provide a better index of the true ability (potential, capability, capacity, penchant, tendency, proclivity, talent) of the individual than performance earlier in practice. Therefore, if treatments (environments, chemicals) are introduced, their effect can be better observed as changes in performance from such a baseline. Obviously this approach has implications for selection and training research too. A possible criticism of the PETER program is that it concentrated all its energies on the RELIABILITY (stability and sensitivity) of tests and never got around to studying the VALIDITY. To some extent this is true, because although all the tests which were studied had already demonstrated their validity elsewhere (cf. Carter, Kennedy & Bittner, 1980), since adequate attention had not previously been paid to stability in other efforts, it is problematic whether the previously found validities were indeed valid. However, the few validity studies conducted in the program showed that a subset of the tests are sensitive to ship motion (Wiker, Kennedy, McCauley & Pepper, 1979) vibration (Guignard, Bittner & Harbeson, 1983), altitude (Bandaret, personal communciation, 1984) and visual kinematics (Kennedy, Ricard, Bittner & Frank, 1984).

Until we began the PETER program, concerted efforts at repeated measures studies had not appeared with any regularity in the recent literature (Forrester, 1984). Yet it is only with such a paradigm that certain critical questions about abilities can be answered. In my opinion, our most important contributions were the focus on stability and reliability. By stability we mean "differential stability", and we called the reliability of test scores "task definition." It was an "individual differences" approach.

In previous programs of test battery development, attention was paid to stability of means (average scores) and to a lesser extent to the stability of standard deviations or variances. We added the requirement that the cross session correlations must be constant because of Jones' work (1970, 1972, 1980). We know of no other battery development effort where such a requirement was formally stipulated. This is not different from the need for symmetry of the variance covariance matrix, which is recognized to be necessary for repeated measures ANOVA (Winer, 1971), although in my experience some investigators incorrectly expect that control groups or large samples or something else will make this problem go away. Therefore, we attempted to show whether a test was stable or not by showing that it met minimum require-

ments for mean, standard deviation and cross-session correlational stability. Differential stability is not just statistical frou-frou. Lack of it implies that what is being measured is changing in unknown ways.

Automated Performance Test System (APTS)

We have begun development of an integrated performance measurement and assessment system. It includes hardware (NEC PC8201A) and software which has the capability for data storage/retrieval including offline storage of data collected within the system. This system is fully portable and we believe somewhat rugged, but have not tested to what extent. It is a self-contained, battery operated (dry cell), notebook sized, 64K internal RAM, with a self-contained display resolution of 240 x 64 elements. The measurement response time is 4.0 milliseconds. The bundled performance measurement software interfaces with a desk top (or hand-held) printer. The software is being designed by M.G. Smith, who in addition to serving as chief trouble-shooter for NTEC's Human Factors' computer laboratory, is also the Essex' Orlando Head of Systems. Thus far, we have 15 tests/tasks games/questionnaires on the microprocessor. All are automatically scored and registered. A cartridge can be inserted to off-load a subject's scores, thus leaving the testing device in the field for continued use.

The tests include: Grammatical Reasoning, Code Substitution, a Video Game, Speed of Tapping (3 forms), Arithmetic, Tower of Hanoi, Fitts' Histoforms, Dynamic Visual Acuity, Motion Sickness History Questionnaire, Mood Adjective Checklist, Motion Sickness Symptomatology, Pattern Comparison Manikin Test, Sternberg's Test, and Simple and Choice Reaction Time.

Thus far we have used forms of the battery before and after simulator hops at three sites and the tests appear to be at least as sensitive as postural equilibrium and subject reports. We are continuing our development under NASA sponsorship, and have some efforts under way comparing paper and pencil with microprocessor presentations of stimuli. The Navy has begun to use it at Warminster (NADC) before and after spin-test work with F/A 18 simulations on the centrifuge. The USAF (Aeromedical Research Laboratory at Brooks AFB) has begun a program to study performances using these devices at simulated altitudes, and Louisiana State University Medical Center is using a version to study motion sickness drug effects. We are in the process of adding some memory, information processing, spatial perception, and visual function tests.

(NOTE: Retarding the learning process may also be a sensitive indicant in its own right, but such a question is different from the question of whether performance per se is disrupted).

REFERENCES

Bandaret, L. Personnel communication, 1984.

Bittner, A. C., Jr. Statistical tests for differential stability. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Santa Monica, CA, 1979, 541-545. Also Research Report No. NBDL-81R010, Naval Biodynamics Laboratory, New Orleans, September 1981, 10-14. (NTIS No. AD A111086)

Bittner, A. C., Jr., & Carter, R. C. Repeated measures of human performance: A bag of research tools. (Research Report No. NBDL-81R011) Naval Biodynamics Laboratory, New Orleans, 1981. (NTIS No. AD A113954)

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. Performance Evaluation Tests for Environmental Research (PETER): The good, bad, and ugly. Proceedings of the 28th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors Society, in press.

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. Selection of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors Society, 1980, 320-324. Also Research Report No. NBDL-81R008, Naval Biodynamics Laboratory, New Orleans, July 1981, 1-7. (NTIS No. AD A111296)

Cronbach, L. J., & Furby, L. How should we measure change - or should we? Psychological Bulletin, 1970, 74, 68-70.

Dunlap, W. P., Jones, M. B., & Bittner, A. C., Jr. Average correlations versus correlated averages. Bulletin of the Psychonomic Society, 1983, 21, 213-216.

Forrester, W. E. Publication trends in human learning and memory: 1962-1982. Bulletin of the Psychonomic Society, 1984, 22, 92-94.

Guignard, J. C., Bittner, A. C., Jr., & Harbeson, M. M. Current research at the U.S. Naval Biodynamics Laboratory on human whole-body motion and vibration (NBDL-83R008). Naval Biodynamics Laboratory, New Orleans, LA: July 1983. (A A138367)

Harbeson, M. M., Krause, M., & Kennedy, R. S. The comparison of memory tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors Society, 1980, 349-353. Also (Research Report No. NBDL-80R008), Naval Biodynamics Laboratory, New Orleans, July 1981, 34-40. (NTIS No. AD A111296)

Harbeson, M. M., Krause, M., Kennedy, R. S., & Bittner, A. C., Jr. The Stroop as a Performance Evaluation Test for Environmental Research. Journal of Psychology, 1982, 111, 223-233.

Harbeson, M. M., Bittner, A. C., Jr., Kennedy, R. S., Carter, R. C., & Krause, M. Performance Evaluation Tests for Environmental Research (PETER): Bibliography. Perceptual and Motor Skills, 1983, 57, 283-293.

Humphreys, L. G. Investigations of the simplex. Psychometrika, 1960, 4, 313-323.

Jones, M. B. A two-process theory of individual differences in motor learning. Psychological Review, 1970, 77, 353-360.

Jones, M. B. Individual differences. In R. N. Singer (Ed.), The psychomotor domain. Philadelphia: Lea and Febiger, 1972.

Jones, M. B. Stabilization and task definition in a performance test battery (Final Report on Contract No. N0023-79-M-5089, Monograph No. NBDL-M001). Naval Biodynamics Laboratory, New Orleans, October 1980. (NTIS No. AD A099987)

Jones, M. B., Kennedy, R. S., & Bittner, A. C., Jr. A video game for performance testing. American Journal of Psychology, 1981, 94, 143-152.

Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. An engineering approach to the standardization of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design and Research Association (EDRA), Charleston, SC; March 1980. Also (Research Report No. NBDL-82R004), Naval Biodynamics Laboratory, New Orleans, November 1981, 1-7. (NTIS No. AD A111180)

Kennedy, R. S., Ricard, G. L., Bittner, A. C., Jr., & Frank, L. H. Effects of simulator visual kinematics on human performance. Proceedings of the Annual SAFE Conference, Las Vegas 1984.

Steiger, J. H. Tests for comparing elements of a correlation matrix. Psychological Bulletin, 1980, 87, 245-251.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. The composition of episodic memory. Evanston, IL: Northwestern University, 1977. (NTIS No. AD-040-696)

Wiker, S. F., Kennedy, R. S., McCauley, M. E., & Pepper, R. L. Susceptibility to seasickness: Influence of hull design and steaming direction. Aviation, Space, and Environmental Medicine, 1979, 50, 1046-1051.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

This project was partially funded under Contract N61339-81-C-0105 and NAS 9-16982.