FINAL REPORT

DEVELOPMENT OF MICROCOMPUTER-BASED MENTAL ACUITY TESTS FOR REPEATED-MEASURES STUDIES

Prepared by:
R. S. Kennedy, R. L. Wilkes,
D. R. Baltzley, & J. E. Fowlkes
Essex Corporation
1040 Woodcock Road, Suite 227
Orlando, FL 32803
(407) 894-5090

Prepared for:

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Contract No. NAS9-17326

January 25, 1990

ACKNOWLEDGMENTS

Funding for this research was from the National Aeronautics and Space Administration Contract No. NAS9-17326 and National Science Foundation Grant ISI-8521282.

The authors are indebted to many of good will who participated in this effort. First, the subjects, and there were over 500 in the various experiments; then to the sponsors -- Dr. Frank Kutyna of the National Aeronautics and Space Administration and Dr. Joseph L. Young of the National Science Foundation. To M. G. Smith, a special vote of thanks for the programming of all tests and subsequent rendering of the data for analyses. In addition, the co-authors of our reports in this series, Dr. Gene Rugotzke, Dr. Janet J. Turnage, Skip Gillam, Dr. William P. Dunlap, Dr. Norman E. Lane. Research assistants, Renee M. Boost, Leilani DeSaram, Mary K. Osteen, and Lois A. Kuntz also deserve special mention.

ABSTRACT

The purpose of this report is to detail the development of the Automated Performance Test System (APTS), a computer battery of mental acuity tests that can be used to assess human performance in the presence of toxic elements and environmental stressors. There were four objectives in the development of APTS. First, the technical requirements for developing APTS followed the tenets of the classical theory of mental tests which requires that tests meet set criteria like stability and reliability (the lack of which constitutes insensitivity). To be employed in the study of the exotic conditions of protracted space flight, a battery with multiple parallel forms is required. The second criteria was for the battery to have factorial multidimensionality and the third was for the battery to be sensitive to factors known to compromise performance. A fourth objective was for the tests to converge on the abilities entailed in mission specialist tasks.

A series of studies is reported in which candidate APTS tests were subjected to an examination of their psychometric properties for repeated-measures testing. From this work, tests were selected that possessed the requisite metric properties of stability, reliability, and factor richness. In addition, studies are reported which demonstrate the predictive validity of the tests to holistic measures of intelligence. Finally, nine sensitivity studies have been conducted where sensitivity of APTS subtests to stressors, agents, and treatments has been demonstrated. The last sensitivity performed in this program, described in detail, entailed calibrating changes on APTS' subtests to blood alcohol level. A report exists dealing with a task analysis of mission specialist work and indexes APTS to these elements. From the experimental work described in this report, sponsored jointly by NASA, NSF, and Essex internally, a well-studied menu of 40 APTS tests is now available. These tests will run on several versions of laptop portables and desk top personal microcomputers. In addition, there are short (< 10 min.), medium (10-15 min.) and longer (> 15 min.) batteries available with factor loadings and predictive validities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

A need exists for microcomputer-based performance batteries to examine the environmental and toxic stresses encountered in space exploration. Such devices should have sound psychometric properties (stability, reliability), be portable and rugged, and permit repeated testing by self-administration with a modicum of training to the subjects.

There are several potential advantages of microcomputer implementation of performance tests (e.g., standardized presentation may lead to improved comparability of tests, higher test reliabilities may result due to more accurate control of stimulus material, performance testing may be performed in innovative modes, fewer errors in data transfer may be realized, and there is the potential for new assessment paradigms and perspectives for understanding of human performance). However, establishing reliability and validity of newly developed microcomputer tests has lagged far behind both the use and marketing of such tests. Well-established principles for constructing and validating tests have been virtually ignored by software developers and users to date. Elsewhere (Kennedy & Bittner, 1977), the traditional criteria for validity and reliability, along with equipment factors and other measurement issues, have been listed. These criteria had been earlier used to evaluate tests for inclusion in a paper-and-pencil-based performance test battery (cf. Carter, Kennedy, & Bittner, 1981). Farrell (1983) has more recently reminded the psychological community that these guidelines should be followed in constructing microcomputer tests. In addition, he has observed that the "obvious evaluation (of microcomputer tasks) is seldom seen in the literature." Farrell has indicated that the reliability and validity of computer tests should be established prior to use.

Other than the work reported here, a handful of recently published studies has compared automated and manual versions of tests and reported favorable results (Wilson, Thompson, & Wylie, 1982). For instance, manual and "automated" versions of the Raven Progressive Matrices have been compared and a high correlation between the two was reported. It is worthy of mention that the automated test used employed an adaptive approach (Watts, Baddeley, & Williams, 1982; Rock & Nolen, 1982). Standard and automated versions of both Digit Span and the Mill Hill Vocabulary Scale were administered and found to be significantly correlated (Wilson et al., 1982; Watts et al., 1982). A computerized battery of information processing tests was found to have moderate convergent validity, as evidenced by the similar intercorrelations observed between manual and automated batteries (Barrett, Alexander, Doverspike, Cellar, & Thomas, 1982), although some have questioned the factor richness of such batteries (Dunlap, Kennedy, Harbeson, & Fowlkes, 1989; Harbeson, Kennedy, Krause, & Bittner, 1982). However, merely being significantly correlated is insufficient evidence for considering the computer-generated test to be equivalent to the traditional one on which it is based. Annually, a review of the difficulties to be experienced in validation is the topic for sessions at American Psychological Association meetings (e.g., Berger, Shermis, Stemmer, & Anderson, 1988; Giannetti, 1988). An experiment conducted by Krause (1983) illustrates this point: Microcomputer and paper-and-pencil versions of four well-documented cognitive tests were compared in two paper-and-pencil forms and one computer test form.

Reliability correlations for paper-and-pencil tests were significantly and substantially higher than the computer versions when corrected for test lengths.

In recent years there has been widespread interest in computerized performance tests. The Department of Defense, Veterans Administration, Environmental Protection Agency, other agencies, and several universities have active programs. These programs constitute valuable resources for the research and development of a computerized testing system. Selected studies from these programs are reviewed below.

Army - Thorne, Genser, Sing, and Hegge (1985) administered the Performance Assessment Battery (PAB) in a 72-hour sleep deprivation experiment. Eight subjects participated in a laboratory environment under high task load conditions. Performance, mood activation, and physiological measures were taken. The PAB was shown to be sensitive to changes in performance, with all tasks showing similar decrement patterns across time. Banderet and colleagues (Banderet, Shukitt, Walthers, Kennedy, Bittner, & Kay, 1988; Banderet & Burse, 1984; Banderet, MacDougall, Roberts, Tappan, Jacey, & Gray, 1984) are conducting a related program of testing to evaluate the effects of hypoxia.

Air Force - A neurophysiological microprocessor test battery was developed at the Air Force Aerospace Medical Research Laboratory (AFAMRL) to assess the effects of workload on operator performance. Tests are implemented in software to be used in a field environment by nontechnical personnel (O'Donnell, 1981). In addition, a subjective workload scale has also been developed (Reid, Shingledecker, Nygun, & Eggeneier, 1981; Schlegel & Shingledecker, 1985). The Learning Abilities Measurement Program (LAMP) at the Air Force Human Resources Laboratory (AFHRL) is investigating individual differences in cognitive abilities and information processing (Christal, 1981; Payne, 1982). Tests have been programmed on microcomputers in a laboratory with 30 automated testing stations. More recently (Kantor & Bordelon, 1985; Carretta, 1989), psychomotor and cognitive tests have been related to success in aviation training in general and in different pipelines.

Navy - The Performance Evaluation Tests for Environmental Research Program (PETER) (Kennedy & Bittner, 1978) was conducted over a five-year period. The chief outcome of that program was a methodology for determining when, if ever, performance on a test had stabilized, as well as a catalog of stabilized tests (Bittner & Carter, 1981; Bittner, Carter, Kennedy, Harbeson, & Krause, 1986; Jones, 1980; Kennedy, Bittner, Harbeson, & Jones, 1981). Video games were studied (e.g., Jones, Kennedy, & Bittner, 1981) and evaluated (Bittner et al., 1986). The PETER methodology was employed to conduct a sophisticated assessment of the group and individual stability and reliability of the tests. Such an analysis needs to be performed prior to factor analyses in order to evaluate the factor structure and richness across the different tests and relate this to the "what is being measured" of the tests. A short (6-minute) battery of tests implemented on a NEC PC8201A microprocessor (Kennedy, Wilkes, Lane, & Homick, 1985) showed encouraging stability, reliability, and factor structure when four tests were compared for computer versus paper-and-pencil format. Other programs within the Navy include those conducted at their medical research laboratories by Naitoh (1982) and Orr &

Naitoh (1976) in San Diego and by Moeller (Rogers, Noddin, & Moeller, 1982) in New London.

Other - In the Appletox program, sponsored by the Environmental Protection Agency (EPA) at the University of North Carolina, Chapel Hill, Eckerman and his colleagues (Gullion & Eckerman, 1986) developed an automated test battery to detect the effects of toxic substances on human performance. The primary test device is an APPLE II microcomputer. Tests identified by the cognitive experimental paradigm of J.B. Carroll (1980) have been selected for evaluation. More tasks are in process, some data have been collected, and refinement of tasks and technical equipment is ongoing (Eckerman, personal communication, June 1985). Related batteries are found in this country (Baker, Letz, Fidler, Shalot, Plantamura, & Lyndon, 1985; Rosa & Colligan, 1988), and abroad (Hanninen, & Lindstrom, 1979; Logie & Baddeley, 1985; Heslegrave & Angus, 1985).

PROGRAM OBJECTIVES

The overall program objective was to develop a computer-implemented measure of mental acuity that could be used to provide an indication of the onset, duration, and severity of impairment in operational performance which may be due to environmental hazards or toxic chemicals. There were three primary objectives in the development of the battery. The first was to deal with only tests or tasks that could be shown to be psychometrically sound. This required the demonstration of stability of means and standard deviation 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. The second goal was to demonstrate that the battery has factorial multidimensionality and that the subscales cross-correlate with earlier performance tests and other recognized instruments of ability. Finally, it was necessary to demonstrate and document sensitivity to factors known to compromise performance potential in the laboratory and ultimately in real-world situations. Throughout this experimental program to select the "best" tests for an optimal computerized test battery for assessment of environmental effects on skilled behavior and higher level tasks, we have stressed the need for repeated-measures experiments to properly evaluate test stability, reliability, and factorial purity. This report reviews a program of interlocking normative studies which have yielded a menu of tests that demonstrates specific metric features: stability, task definition, reliability efficiency, as well as factor diversity and sensitivity.

METHOD OF APPROACH AND PRINCIPAL ASSUMPTIONS: HISTORY OF THE AUTOMATED PERFORMANCE TEST SYSTEM (APTS)

THE PETER PROGRAM

The chief antecedent to the present work is the Navy's PETER program, mentioned above. The strategy employed in this work followed a repeated-measures paradigm based on classical test theory (Gulliksen, 1950). In the Navy work, the basic objective was to evaluate mental capacity and show whether and to what extent it may have been adversely affected by an agent or treatment. The environment explicitly to be studied was ship motion. The program began in 1976 and was completed in 1981.

Because repeated measures of the same subjects is the usual method for studying such effects, it was reasoned that first it would be necessary to have tests of constructs or capabilities which would have sufficient alternate (parallel) forms. Repeated-measures designs are more efficient and economical than alternate approaches (Winer, 1971) and are ideally suited to experiments with small numbers of subjects. However, sufficient attention has not been paid to the statistical requirements for meaningful interpretation of repeated-measures experiments (Bittner & Carter, 1981; Jones, 1980; Kennedy, et al., 1981). The compound symmetry requirement of the variance-covariance matrix for simple repeated-measures analysis of variance (Winer, 1971) demands that intertrial correlations be unchanging (differentially stable) and variances be homogeneous across baseline repetitions (Bittner, 1979; Jones, 1980; Lord & Novick, 1968). The epsilon correction (Dixon, 1983) can help meet some violations of this statistical assumption, but provides no improvement for concerns with "what is being measured."

In the PETER work and in the APTS work which followed, tests were first subjected to an examination of their psychometric properties for repeated-measures testing (Bittner et al., 1986; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Kennedy, Wilkes, Lane, & Homick, 1985). The cardinal psychometric qualities of tests which are to be employed in such repeated-measures designs are stability and reliability of between-subject variance. In other words, "attribution of effect" requires that the capability being sampled be stable to measurement. It is further helpful if these properties are achievable with an economy of time.

Stability

Repeated-measures studies of environmental influences on performance require stable measures if changes in the treatment are to be meaningfully related to changes in performance (Jones, 1970a). Of particular concern is the fact that a subject's score may differ significantly over time due to measure instability. For example, the Jones two-process theory of skill acquisition (Jones, 1970a,b) maintains that the advancement of a skill involves an acquisition phase in which persons improve at different rates, and a terminal phase in which persons reach or approximate their individual limits. The theory further implies that when the terminal phase is reached scores will cease to deviate despite additional practice. Unless tests have been practiced to this point of differential stability, the determination of change in scores due to practice or some other variable are confounded. For example, in a study of the effects of alcohol, if scores on a performance test remained the same before and after exposure, and if the test were not differentially stable, it would be impossible to determine whether a decline in performance was masked by practice effects or whether there was no treatment effect. Only after differential stability is clearly and consistently established between subjects can the investigator place confidence in the adequacy of his measures and subsequent results.

Table 1 summarizes the criteria for acceptability of tests.

TABLE 1. CRITERIA FOR ACCEPTABILITY OF TESTS

Criterion	Description
STABILITY	The extent to which a constant mixture of human performance capabilities is assessed on each trial of repeated testing. Parallelism of the tests. The means, variances AND the cross session correlations should be stable.
RELIABILITY EFFICIENCIES	The reliability (R) of a stabilized task standardized to a 3-minute administration base.
TIME TO STABILITY	Total amount of elapsed training time which is required to reach stabilization .
CEILING/FLOOR	Range over which the test can test. There should be no narrowing of the between subject differences as occurs when tests have a "top".
FACTOR RICHNESS	The mental faculty assessed by the measure and the diversity of factors measured.
VALIDITY	These include "traditional" types of validity like construct, consensual and predictive, as well as the more practical requirement that the tests be sensitive to a variety of stimuli like toxic agents and environmental stress.

Reliability Efficiency

Test reliability is known to be influenced by test length (Guilford, 1954). Tests with longer administration times and/or more items maintain a reliability advantage over shorter test times. Test length must be equalized before meaningful comparisons can be made. A useful tool for making relative judgments is the reliability efficiency, or standardized reliability, of the test (Kennedy, Carter, & Bittner, 1980). Reliability-efficiencies are computed by correcting the reliabilities of different tests to a common test length by use of the Spearman-Brown prophecy formula (Guilford, 1954, p. 354). Reliability-efficiency not only facilitates judgments concerning different tests, but also provides a means for comparing the sensitivity of one test with the sensitivity of another test.

Stabilization Time

The evaluation of highly transitory changes in performance may be necessary when studying the effects of various treatments, drugs, alcohol, or

environmental stress. Good performance measures should quickly stabilize following short periods of practice without sacrificing metric qualities. As a general rule, good performance measures should always be economical in terms of time. A task under consideration for environmental research must be represented in terms of the number of trials and/or the total amount of time necessary to establish stability. Stabilization time must be determined for the group means, standard deviations, and intertrial correlations (differential stability).

Task Ceiling

If all subjects asymptote at the maximum level of performance, then the task is said to have a ceiling (Jones, 1980). Ceilings are undesirable because they limit discrimination between subjects. When subjects perform equally well, except for random error, between-trial correlations fall to zero.

Factor Richness

Where possible, subtests should be selected that tap independent factors with little or no overlap. Such selection ensures that the overall battery is rich in factor structure while free of unwanted redundancies.

Validity

Good tests are those which are demonstrably valid according to several criteria. For example, they should: be sensitive to agents and stimuli like hypoxia, drugs, and sleep loss; predict other mental test scores and cognitive performances; tap constructs and factors which reflect a theoretical basis; appear on the face to be testing a mental acuity function; etc.

Following these criteria, experiments were conducted at the Naval Biodynamics Laboratory in New Orleans, Louisiana, over a a six-year period. During this time over 140 mental acuity tests from the psychological literature, including measures of cognition, information processing, reasoning, prediction, decision making, memory, and many others were tested in a normal group of subjects over a three-week repeated-measures paradigm. The initial purpose was to demonstrate that the tests were stable; the secondary purpose was to demonstrate that the tests had reliability. The tertiary purpose was to rank order/prioritize the tests according to their efficiency (e.g., time, reliability, and factor structure).

After five years of research a menu of 33 tests was surfaced which could be used for creating a battery of tests to assess mental functions. What that program lacked was that the tests were largely presented in rough "old fashioned" media (e.g., paper-and-pencil, slide projector); sensitivities of the test battery and the tests within the test battery were not known; and no adequate factor analyses had been performed.

AUTOMATED PERFORMANCE TEST SYSTEM (APTS)

In 1983 the National Aeronautics and Space Administration (NASA) provided support to Essex Corporation to continue development of those paper-and-pencil tests and to implement them on a portable, lap-top 2.2 lb., notebook-sized

battery operated microcomputer (the NEC 8201A). For two years that work proceeded and in 1985 the National Science Foundation (NSF) awarded a Phase I of a Small Business Innovative Research Grant. The Phase I was followed in 1986 by a Phase II, in order to continue similar and related work and to broaden the battery to include not only applications for NASA relative to motion sickness preparations, but to be serviceable as a generic test battery, an industry standard for all toxic agents, and for assessment of subject state over a repeated-measures application. The collective goal of the research reported here is the development of a menu of tests embedded in a coherent package of hardware and software which will be useful in repeated-measures studies of the effects of environmental and chemical stressors on human performance.

In the APTS program we conducted a series of interlocking studies. Most of them were intramural, but some were conducted "piggy-back" with other studies and with other agencies. As with the PETER program after which it was fashioned, initially we focused on basic metric issues like stability, reliability, and correlations between tasks using a core test menu. Then we added the practical considerations of subject and experimenter time. Finally, we focused on factor analysis and validity of the tests. Validity began with correlational studies and worked into studies with toxic agents.

It is not uncommon for the development of test batteries to follow from cognitive theories (e.g., Hunter, 1975; Carretta, 1987; Hunt & Pellegrino, 1986; Gullion & Eckerman, 1986; Braune & Wickens, 1985). However, as the theory is modified by new experience, so too may the tests in the battery be modified, and as a result, not only are such programs seldom completed, but often tests will not be continued in subsequent studies and so threads and standardization are lost. Thus, it becomes difficult or impossible to "mark" or "index" findings from early studies to different treatments or dosages which may be collected later.

The approach followed in APTS work employs test theory as an engineering strategy to build a battery from parts. For example, test theory (Allen & Yen, 1979) makes simplifying assumptions such as that Obtained scores are comprised of a True score (T) and an Error score (E) regardless of the context of what they might measure. Test theory further assumes that True scores and Error scores are additive (rather than some other relationship), and that the True score portion of an Obtained score will be correlated with the True score portion when tested again, whereas the Error portion will not because it is nonsystematic or random. If fatigue occurs or learning is still going on (which can occur over repeated administrations of tests) then, in addition to the True score, there are other elements being measured which differ systematically from (i.e., are uncorrelated with) ability on the test. In this case, the "True" score has two systematic parts and the assumptions of the theory are compromised.

Such an approach can easily accommodate hypothetical constructs like "controlled vs. automatic" processing (Ackerman & Schneider, 1984) or "components" (Sternberg, 1979) as they emerge. So when a test is stable, then systematic differences in automaticity, learning, or fatigue are no longer present and the effectiveness of the introduction of treatments or agents may be seen to influence the construct which the True score purports to tap.

Therefore, a critical requirement of tests which are employed in repeated measures applications and within-subject designs, is that the tests be stable, so that alternate forms of the tests be parallel. The requirement for parallel forms is logically necessary for proper interpretation of any loss (or gain) in the performance being measured as being due to a treatment. We believe that in the past when test batteries have been developed, little if any attention was paid to certain areas of test theory, particularly stability. More to the point, we know of no battery which has followed a differential approach, but since within-subject designs are so often the intended application, we believe this is a critical failing of other batteries. Specifically, the argument for differential stability which follows must be addressed:

If individual differences in ability are present, which are not Error, then the retest correlation is proportional to the ratio of True Score to Total Score variance. We therefore require that tests exhibit suitable differential stability before they can be recommended for use in the study of stressors. Additionally, tests which are not reliable lack statistical power and will likely be insensitive to stressor effects.

Because theoretically, environments and treatments can be expected to degrade some performance and not others, a test battery should tap a variety of different mental capacities. Thus, the next purpose is related to an explanation of the factorial diversity of tests. After a sufficient number of tests have been identified which possess stable and reliable metric properties, it then becomes important to determine to what extent they overlap or are unique. Generally, the correlation of each test with each other test, corrected for the attenuation due to the known reliabilities of each test (Spearman, 1904), can be employed to provide such an index. More sophisticated treatments (viz., factor analysis) should also be undertaken.

Finally, we sought to address the important issue of validity because the cardinal requirement of any test or test battery is that it be valid. The manual of standards and practices for tests (American Psychological Association, 1982) suggests that "good" tests should have more than one kind of validity. In the validity phase we sought experimentally to obtain three forms of validity: (1) correlation with other test batteries, (2) construct -- through correlational and factor analysis of subtests within the battery, and (3) predictive -- by showing sensitivity to various agents and treatments. Because such a large literature exists relating scores on holistic measures of intelligence (or IQ) to most forms of academic and job performance, at first we proposed linking the microcomputer tests to holistic measures of intelligence and job performance (e.g., American College Testing (ACT) Test, Armed Services Vocational Aptitude Battery (ASVAB), Wechsler Adult Intelligence Scale-Revised (WAIS-R), and Wonderlic. We are aware that Hunt (1985) eschews predictive validity as a goal in itself, but we believe such knowledge can guide the further development of cognitive theory and the interpretation of tests. Stability of group means and variances are recognized by most developers of test batteries. Yet, when learning occurs at different rates, tests can be differentially unstable until all subjects perform in a parallel fashion over sessions. There is at least one case (McCauley, Kennedy, & Bittner, 1980) where mean and standard deviation stability were obtained quickly and retained for three weeks of testing, but

where learning or strategy shifting occurred so that the relative positions of subjects shifted systematically over sessions. In that experiment, performances one or two days apart were correlated reasonably well with each other ($r > 0.70$) but those as few as four trials apart were barely correlated with each other at all ($r < 0.25$ and sometimes $r = 0.00$). Since individual differences were present, this means that whatever the first test signified (or was correlated with), performance on Day 4 would not be correlated with it or measuring it! Although not always this dramatic, such forms of instability have been found in half of all tests studied in the PETER work (Bittner et al., 1986).

Much of the work in the APTS program (44 articles) is completed and published in the form of conference proceedings, government sponsored technical reports and peer reviewed scientific journals and they are listed in Appendix B. A summary of the results of the metrology and sensitivity studies are described in more detail below. In addition, a sensitivity to alcohol study, the last experiment in this program, is described fully. This study focuses on validation of the best tests in the APTS battery which were administered to subjects with various levels of induced alcohol intoxication. A demonstration disk which contains the menu of "recommended" tests is available from Dr. Robert S. Kennedy, Vice President, Essex Corporation, 1040 Woodcock Road, Orlando, Florida, 32803.

BASIC DATA GENERATED AND SIGNIFICANT RESULTS: METROLOGY STUDIES, SENSITIVITY STUDIES, TASK ANALYSIS STUDY, AND ALCOHOL CALIBRATION STUDY

METROLOGY STUDIES

Study 1

The first study in the APTS work compared the best tests from the PETER program with the same tests implemented on a portable lap-top computer (NEC 8201A) (Kennedy, Wilkes, Lane, & Homick, 1985). A small sample (N=20) received six tests over four sessions and the newly implemented microcomputer-based versions were compared to the old-fashioned paper-and-pencil versions in the same subjects. Microcomputer tests included Grammatical Reasoning, Pattern Comparison, Code Substitution, and the Tapping series. Tapping was substituted to be comparable to Aiming and Trail Making from the PETER series. The other paper-and-pencil versions were implemented to be comparable. The results of that study revealed that all the tests achieved stability very early in practice and the reliability values all exceeded $\underline{r} = .70$ for even very brief (< 3 minutes) periods of performance. The microcomputer versions of tests correlated as high as their reliabilities would allow with the more traditional paper and pencil versions.

Study 2

This study followed the form of Study 1 but expanded on it (Kennedy, Wilkes, Dunlap, & Kuntz, 1987). In addition to evaluating stability and reliability of more tests and over more trials, predictive validity was also examined. Twenty-five subjects were tested over 10 replications on 10 microcomputer tests. The 10 microcomputer tests were concurrently

administered in paper-and-pencil (marker battery) where possible and microcomputer-based versions and compared to scores on the Wechsler Adult Intelligence Scale-Revised (WAIS-R). The WAIS-R was administered by a licensed psychologist. Nine of the 10 microcomputer-based tests achieved stability and were recommended for inclusion into the menu of APTS tests. Correlations between certain microbased subtests and the WAIS identified common variance.

Study 3

In this experiment (Kennedy, Wilkes, Kuntz, & Baltzley, 1988), 18 different tests, including six visual and auditory monitoring tests, and a tracking test (Air Combat Maneuvering) were administered. The tests were self-administered, that is, after an initial practice session the subjects were permitted to test themselves in standardized ways but at nonstandardized times in their homes or in school classrooms. The results showed that performances on 13 out of the 18 tests were stable and reliable, and performances and stabilities were comparable to what had been obtained on the core battery in previous experiments, implying that self-administration was not the major cause of the lack of stability or reliability of some of those tests which did not qualify. At the conclusion of this experiment, there were now 13 tests in the APTS series that were considered to have the minimum reliability and stability characteristics. Additionally, the correlations between the tests again tended to be low, implying that a battery selected from the tests on this menu could provide diverse factor structure (Kennedy, Wilkes, Kuntz, & Baltzley, 1988).

Study 4

The focus of Study 4 (Kennedy, Baltzley, Dunlap, Wilkes, & Kuntz, 1989), which was partly sponsored by the National Science Foundation, was to broaden the test base of APTS and replicate the predictive validity with holistic measures of intelligence which were reported in Study 2 above. A number of subjects (N = 27) received a longer version of the tests administered in Study 2 and all subjects who received these tests were administered a series of IQ-like tests. The global measures of IQ included American College Testing scores which were available from the subjects' school records, a synthetic ASVAB (Steinberg, 1986), a WAIS-R, and a Wonderlic (Wonderlic, 1978). Mental tests have long been used to signal cognitive dysfunction (e.g., Wechsler Adult Intelligence Scale [Wechsler, 1981], Arthur Point Performance Scale [Arthur, 1949], Halstead-Reitan Battery [Reitan & Davison, 1974], etc.), and it has been argued that these tests are more sensitive to subtle decrements in mental ability than clinical neurological tests such as CAT Scan or EEG (e.g., Casson, Siegel, Skarn, Campbell, Tarlau, & DiDomenico, 1984). However, these tests are ordinarily limited to one or two alternate forms and entail individual administration by trained psychometricians requiring heavy investment in technical staff and considerable time must be devoted to data reduction and analysis.

The results of this experiment, which also involved the use of two different microcomputers administered separately (the NEC PC 8201A and the Zenith PC 181) revealed the following outcomes: (1) 13 of the 14 tests achieved sufficient levels of stability and reliability to qualify for

subsequent listings in the menu. (2) performance on many of the tests correlated with IQ measures and approximately 50% of the variance of each of the global tests was explained by combinations of the microcomputer tests. The highest correlations were with ASVAB composite scores, the lowest correlations were with verbal IQ. (3) There was no clear-cut advantage for either computer over all the tests. Some of the tests were more quickly performed on the NEC microcomputer, and some on the Zenith; some were directly comparable. The experimental design was not crossed over between the two systems. (4) It is possible to self-administer these tests and to have them be stable and reliable, even in the absence of a proctor administering tests in a formal laboratory.

Studies 5, 6, and 7

Under contract to the U.S. Army Medical Research and Development Command, the existing tests of the NASA APTS battery were compared to tests from the Tri-Service UTC-PAB (Englund, Reeves, Shingledecker, Thorne, Wilson, & Hegge, 1986). Tests from the PETER program were conducted to ascertain their fulfillment of psychometric and administrative criteria in order to surface additional tests which might be implemented and tested on the APTS. Study 5 (N = 25, trials = 15), Study 6 (N = 25, trials = 15), and Study 7 (N = 25, trials = 10) evaluated the six core APTS and 15 PAB tests. The findings reveal that all six APTS tests and 10 out of 15 PAB tests were considered to be stable and sufficiently reliable to be qualified for use in an APTS criterion-based performance test battery. That is, stability is achievable in less than 10 minutes total practice per test and the reliability is greater than .707 for three minutes of testing. From these three studies there were now 20 acceptable tests proposed on NEC and Zenith systems. In general, the metric properties reveal good reliabilities, good stabilities, and low intercorrelations implying multifactor test battery prospect. Further details on these studies may be found in Kennedy, Turnage, and Osteen (in press).

Study 8

At this stage in the development of APTS, there had been no factor analysis, although correlational analyses in small samples with multiple replications provided guidance in estimates of factor structure and richness. However, it was decided that a large scale (more than 100 subjects) study was required to delineate the diversity of constructs assessed with the menu of tests thus far surfaced. Under NSF sponsorship, 11 tests were therefore selected -- seven from the APTS series and four from the UTC-PAB which, on the basis of content and their previous correlations, particularly from Studies 5, 6, and 7, suggested that they would be largely orthogonal. These were administered three times to each of 108 Central Pennsylvania college students (48 males and 60 females) and marked against the Wonderlic Personnel Test. Factor analyses, which were carried out on each administration, yielded three consistent factors: a spatial/numerical factor on which Pattern Comparison (APTS) loaded most heavily, a verbal factor of which Grammatical Reasoning (APTS) loaded most heavily, and a motor factor defined by the Tapping tests (APTS). Based on these results a core battery could include Pattern Comparison, Grammatical Reasoning, Math Processing, and Tapping, and the Preferred and Nonpreferred (but not the Two-Finger) Tapping tests. This battery provides three well identified factors, one verbal, another

spatial/numerical, and the third motor, and which might be usefully augmented, especially in operational situations, by Code Substitution and Choice Reaction Time tests, both from the APTS battery, but which were not evaluated in this experiment. Manikin (APTS) is another recommended test for augmentation of the core battery because it is known to measure a different factor from IQ (Kennedy, Baltzley, Turnage & Jones, 1989).

Study 9

Another factor analysis was conducted with a slightly larger pool of tests and sponsored jointly by NASA and AMRDC. One hundred college students from the Orlando area received five administrations of 23 tests from the recommended list which surfaced from experiments 5, 6 and 7, and which reflected on the factor analysis of Experiment 8. This study confirmed the results of Study 8: all the tests appeared stable within 3-4 sessions and reliabilities exceeded $r = .707$ as would have been predicted from their previous development findings in Experiments 1-8. Additionally, the factor analysis revealed consistent factors (Lane & Kennedy, 1988). Although factor labelling involves an element of risk with respect to the "true" content of the factor, a synthesis of factor and correlational analyses across a series of studies suggests the following interpretation. There are least three important factors in the APTS tests that consistently recur in various studies (even in early trials), and a fourth factor that emerges at or around the trial at which most tests are stable. (1) Motor Speed - speed of response execution, particularly those for which the "rules" are simple and output is in part dependent on how rapidly responses can be entered. (2) Symbol Manipulation/Reasoning - involves a "generalized" ability to reason abstractly through the application of rules rather than the learning or remembering of the rules themselves. (3) Cognitive Processing Speed - reflects the extent to which defined rules governing generation of response alternatives for a particular test have been learned through practice and can be used progressively more rapidly. (4) Response Selection Speed - the speed with which responses can be selected from the generated set of response alternatives.

SENSITIVITY STUDIES

Study 10

Two sensitivity experiments with APTS have been conducted under hypoxic conditions; the first by scientists of the US Air Force and the second by the US Army Institute for Environmental Medicine using Essex scientists for test administration and analysis. The results were concordant. There was a definite cognitive performance decrement with sustained periods at simulated altitudes of 23,000 feet (Kennedy, Dunlap, Banderet, Smith, & Houston, 1989) and with abrupt, short periods at 27,000 feet (Schifflett, personal communication). However, motor performance remained essentially unchanged in both studies. This finding is not surprising and is consistent throughout the remaining sensitivity studies. Perhaps motor performances, which are the simplest and most well-practiced, may require a very large effect to disrupt them.

Study 11

In a NASA-sponsored study (Kennedy, Odenheimer, Baltzley, Dunlap, & Wood, 1990), with high doses of motion sickness drugs (scopolamine 1.0 mg, amphetamine 10 mg), all of the scores for both motor and cognitive tests changed in a rational direction; ANOVA revealed that Pattern Comparison was significantly poorer with scopolamine and that amphetamine significantly increased Nonpreferred Hand Tapping (a motor skill test). There was a trend toward increased scores on Short-term Memory (an item recognition test). The study further showed an interaction of scopolamine and dexedrine with Two-Hand Tapping.

Study 12

An experimental preparation (drug X) and an over-the-counter antihistamine (Benadryl) were compared in a double-blind study. The general findings were that the subjects treated with the antihistamine had a significant drop in performance over the placebo condition and the experimental drug effect was less than the antihistamine and greater than placebo (Essex Corporation, 1988).

Study 13

At the Fred Hutchinson Cancer Research Center at the University of Washington, patients who were receiving bone marrow transplants and chemoradiotherapy treatments were studied (Parth, Dunlap, Kennedy, Lane, & Ordy, 1989). In this study the tests of the basic NASA APTS battery were administered, along with other tests, to both a patient population and controls. 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 battery as a whole was strikingly effective in detecting performance shifts in patients and significantly differentiating patients from controls throughout the therapy period. Greater discrimination was apparent in the complex cognitive measures (i.e., Code Substitution) than in the "motor" (i.e., Tapping). Discrimination was present for both accuracy and latency measures, although effects were stronger for accuracy performance.

Study 14

A number of subjects were sleep deprived for one night at the U.S. Naval Postgraduate School in Monterey, California. Statistically significant effects on Code Substitution were observed, but only nonsignificant directionally appropriate changes on the other tests were obtained (Kiziltan, 1985).

Study 15

In this study, 400 Navy pilots were tested before and after their exposure to a flight simulator (Kennedy, Fowlkes, Lilienthal, & Dutton, 1987). There were differing amounts of motion sickness experienced by the pilots. None of the pilots exhibited any loss in performance during post-testing when compared to pretest performances, although when compared to a control group who were not exposed to motion during the pre/post-testing, the increase in performance ordinarily expected due to learning in two sessions was not seen in the experimental group.

Study 16

At the Ames Research Center at Moffett Field, CA, a number of subjects were exposed to long-term bed rest. In general, learning curves continued over the entire period of exposure and there did not appear to be significant losses in performance (Deroshia, in press).

Study 17

Eighteen subjects were voluntarily placed in a cave in Bari, Italy and otherwise isolated. The subjects were monitored night and day through telecommunication systems, but were otherwise unaware of the time of day or the day of the week or the period of their exposure. They were tested periodically with the NASA APTS battery. Over the course of a month, isolated and deprived of natural light and cues of time and day, their performances generally revealed slight learning curves throughout the period of exposure. There were no evidences of a loss in performance. A control group was not available for comparison. A Mood Adjective Checklist revealed a substantial drop through the course of the study followed by a rapid return to "normal" levels a day prior to the termination of the experiment. The time course of the mood effect was in marked contrast to the stable performance curves. These results led to an interview with a NASA physician on an Italian television show to present the findings of no performance decrements but substantial motivational/emotional swings.

Study 18

Under NASA sponsorship (Calkins, 1989), 10 subjects in double-blind fashion were exposed to 48-hour periods of halon gas in concentrations of 20 ppm. There were small but identifiable differences in performance between the two conditions with halon conditions generally being poorer.

SUMMARY OF 18 APTS STUDIES

From this experimental work, a well-studied menu of 40 APTS tests is now available. These include 23 tests which surfaced originally from the U. S. Navy's PETER program (Bittner et al., 1986), and 17 related tests from the tri-service UTC-PAB program (Englund et al., 1986). These were combined into a menu and evaluated in a series of interlocking studies. These tests will run on several versions of laptop portables and desk top personal microcomputers. In the various studies listed above, the menu of tests has been shown to be stable, reliable, and factorially rich (Lane & Kennedy, 1988). They can also be self-administered and scored (Kennedy, Wilkes, Kuntz & Baltzley, 1988;). In addition to demonstrating predictive validity to holistic measures of intelligence (Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Kennedy, Baltzley, Turnage & Jones, 1989), nine sensitivity studies have been conducted where validity to stressors, agents, and treatments have been demonstrated. In addition, other tests under development, vision tests, a mood questionnaire, a metacognitive self-efficacy inventory (McCombs, Doll, Baltzley, & Kennedy, 1986) and a motion sickness questionnaire (Lane et al., 1988) are also available.

TASK ANALYSIS STUDY

A surrogate approach to human performance testing has been proposed (Lane, Kennedy, & Jones, 1986). This approach suggests that if tests of the same mental faculties (as are in operational performance) can be shown to change with treatments, one might infer that the operational performance might also be degraded.

As part of the development, we sought a technique that would permit comparison of abilities tested by APTS tests and the requirements for those abilities in various NASA mission specialist tasks was sought. To follow this strategy, two goals needed to be accomplished: (1) a metrically sound battery of tests needed to be developed, and (2) the tests in the battery needed to be compared to the elements of jobs performed by mission specialists.

In the 18 studies reviewed above, the APTS is shown to possess a menu of cognitive and motor tasks. What were next required were a task analysis of NASA mission specialist jobs in order to link the APTS tests to operational performance.

For this purpose, Dr. R. Jeanneret, a well-known analyst of jobs, was enlisted to conduct a task analysis of 14 NASA mission specialist jobs and then compare those abilities to abilities tested by the various APTS tests.

A generic position was selected for study. This position, the job of Aerospace Payload Specialist, covered the range of anticipated duties of astronauts and others assigned to a space station. For this effort, the task was decomposed following the approach of the Position Analysis Questionnaire (PAQ) (Jeanneret, 1988). The PAQ is perhaps the most widely used example of such an analysis instrument which has the capability to describe jobs in mental attributes and the development of the PAQ was originally sponsored by ONR. The PAQ is a structured job analysis questionnaire that can be used for analyzing jobs of many different types. It consists of six major divisions: (1) information input, (2) mental processes, (3) work output, (4) relationships with other persons, (5) job context, and (6) other job characteristics.

The preliminary results of the PAQ analysis yielded a set of behavioral job dimensions which characterized the content of these positions and permitted estimation of requirements for effective job performance. These elements are shown to converge with APTS test factors in matrix form (Jeanneret, 1988, pp. 38-39) and tabular form (pp. 41-42).

This document, and the two comparison works, may be employed to plan experimental work regarding human performance changes of relevance to National Aeronautics and Space Administration.

ALCOHOL CALIBRATION STUDY

The field studies described above indicate that a microbased human performance battery is available for identifying the effects of environmental and toxic stressors. It was our view that the next step in development was to conduct a precisely regulated laboratory validation study, designed to

accurately calibrate treatment levels relative to APTS subtest score changes. It was anticipated that calibration findings from such a study would 1) provide future researchers with a known standard for estimating possible effects of various treatments as well as aiding in the selection of appropriate subtests, and 2) provide quantitative insight into the sensitivity of the battery as well as specific subtests. It was the purpose of this study to index performance deficit against a well-known and well-researched treatment and to compare the results to a placebo condition. For this work we selected various Blood Alcohol Concentrations (BACs) of small (.05% BAC), medium (.10% BAC), and large (.15% BAC) dosages.

Methods

Subjects

Subjects were male students, 21 years of age or older, attending Casper College, Casper, Wyoming. A total of 33 students were initially briefed regarding the study. Twenty-seven of those addressed volunteered for participation. From those volunteering, a pool of acceptable candidates was established. Acceptable candidates were those indicating some, but not excessive, experience with alcohol, no past history of chronic dependency of any type, good general health, and indications of low risk for future alcohol-based problems. The typical subject identified himself as having "moderate" previous experience with alcohol (Calahan Volume-Variability Scale $\underline{M}$ = .36, SD = .24), and at low risk for future problems with alcohol (Iowa Scale of Preoccupation with Alcohol [median category = 5 and range = Categories 5 to 3]). Students indicating problem family histories of chemical abuse/dependency and/or past personal histories of chemical abuse/dependency were advised not to participate. Initially, 21 subjects were randomly selected from the pool to participate. One of the subjects selected for participation elected to withdraw from the study; a second subject was unable to complete data collection requirements during one of the five sessions; a third subject was dropped due to questionable analysis results. The 18 subjects completing the study ranged in age from 21 to 35 (M=24.6, SD=3.9) with weights from 134 to 235 pounds (M=183.3, SD=32.6).

Experimental Design

Subjects were randomly assigned to a series of three Blood Alcohol Concentration (BAC) treatments and one placebo condition over four separate testing sessions. Each subject was tested for performance decrements at 0.00, and approximately 0.05, 0.10, and 0.15 BACs. Order effects were controlled through the serial use of Latin Square randomization techniques (Edwards, 1985, pp. 289-290). Each subject served as his own control with performance measures completed both prior to and after the treatment. Blood alcohol concentrations were closely monitored by breath testing procedures until prescribed treatment levels were attained. Corresponding whole blood, blood sera, and urine measures were then obtained. Double-blind procedures were employed across all testing sessions to control for experimenter and subject expectancy effects. The independent and dependent variables are described below.

1. Independent Variable-Blood Alcohol. Blood alcohol concentration was manipulated by administering alcoholic drinks mixed from orange, tomato or fruit juice and 80-proof alcohol (95% alcohol) with 2.5 ml drops of rum extract floated on top. Eight drinks were premixed for each subject, with weight and BAC treatment condition determining the proportions of alcohol. Proportions of grain alcohol and juice were combined to raise a subject's BAC slightly above the targeted level, permitting monitoring on the descending limb of the BAC curve. The amount of grain alcohol in milliliters was calculated using a condensed version of the Widmark Equation: 1 ml of Grain Alcohol = (200/190) (30) (0.13) (weight in pounds with target BAC + 0.05). If the assigned treatment target = 0.00% (placebo preparation), then no grain alcohol was used.

2. Independent Variables-Blood Alcohol Concentration. Breath monitoring for BAC was initiated approximately 30 minutes after a subject had finished drinking. Monitoring continued until breath analysis demonstrated that the BAC had stopped increasing. Other BAC dependent measures were then introduced and included samplings for whole blood, blood serum, and urine analyses. Breath alcohol was analyzed using two Intoximeter 3000 breath test units. These are computer controlled instruments commonly used in Wyoming law enforcement agencies which operate on the principle of nondispersive infrared molecular absorption. The blood and urine samples were analyzed by the Wyoming Chemical Testing Program Laboratory in Cheyenne, Wyoming. One blood sample was gently mixed and analyzed for the amount of alcohol in whole blood and the second sample was allowed to clot and the serum was analyzed. The blood, serum, and urine were analyzed using the State-approved procedure, except results were reported to three places.

3. Dependent Variables-Human Performance. Human performance was assessed with the Automated Performance Test System (APTS) (Essex, 1986). Development of the APTS was based on the concepts and empirical findings of the Performance Evaluation Test for Environmental Research (PETER) program (Bittner et al., 1986), and is comprised of three subsystems: (1) hardware, (2) test programs, and (3) system control. The APTS provides for microbased repeated measures of human performance while under the influence of various environmental or experimental agents. The reliability, stability, factor structure, and sensitivity of the measures are discussed inter alia and hardware specifications appear in the Apparatus section.

4. Dependent Variables-Field Sobriety. Field sobriety was assessed by a trained police officer, using standard procedures for administering and scoring the Gaze Nystagmus, Walk-and-Turn, and One-Leg Stand. Each measure was separately derived at the time of assessment and subjects could obtain total scores of 6, 10, and 7 respectively. These results are to be reported separately.

Materials

Various paper-and-pencil and computer software materials were employed in screening and assessing the individual subjects. These materials are identified and discussed below:

1. Personal Information Questionnaire (PIQ). The PIQ was specifically developed for use in this study. The questionnaire assesses the personal characteristics and histories of potential research subjects. The information provided partial basis for the selection of students into the final subject pool. Relevant information concerning weight and general health were addressed. The PIQ was administered once.

2. Current Health State Questionnaire (CHSQ). The CHSQ questionnaire was specifically developed for use in this study. The questionnaire assesses a subject's state of health immediately prior to the administration of an experimental alcoholic treatment. Information collected with the CHSQ facilitated alcohol treatment preparations and identification of subjects not currently fit for participation. The CHSQ was administered prior to each experimental session, for a total of five replications.

3. Iowa Scale of Preoccupation with Alcohol (IS). The IS, developed by Mulford and Miller (1961), consists of 12 behaviorally defined statements scaled to distinguish two levels of drinking behavior. Three self-descriptive statements are associated with each of the first four levels of drinking behavior. The fifth level is reserved for individuals not responding affirmatively to items associated with the previous four levels. Subjects respond to the IS by indexing statements applying to them. Agreement with any two items within a level identified a subject as to "type of drinker." Subjects identifying levels I and II are classified as "alcoholic drinkers" (Mulford et al., 1961, p. 28). Subjects identifying as levels III and IV are simply classified as drinkers. The IS was employed in determining the potential risk associated with participation in the study. The scale was administered once in conjunction with the PIQ and was an important measure in eliminating candidate participants from inclusion in the subject pool.

4. Cahalan Volume-Variability Scale (V-V). The V-V (Cahalan, Cisin, & Crossley, 1969) assesses alcohol consumption. Assessment was based on students' self-report of the quantity, frequency, and variability of alcoholic beverage consumption over a standard period of time. Subjects respond to the V-V scale by indicating how often, and how much, they consumed of wine, beer, liquor or any type of alcoholic beverage. The average daily volume is estimated by multiplying the frequency of consumption of each beverage by the estimated quantity of the beverage per occasion. Variability for each of the three volume groups is established by subdividing each volume group according to the number of drinking occasions per month (Cahalan et al., 1969, pp. 213-215). Based on the average daily volumes, as well as daily variabilities in alcohol consumption, individuals are classified according to eight identifiers ranging from "High Volume, High Maximum" to "Abstainee." The V-V was administered once in conjunction with the PIQ and IS, and was an important measure in eliminating totally inexperienced and extremely heavy users of alcohol from inclusion in the subject pool.

5. APTS Subtests. The subtests selected for inclusion in the APTS battery have been researched and commercially developed by Essex Corporation, Orlando, Florida. Each subtest had been previously evaluated relative to repeated-measures selection criteria. The subtests have demonstrated reliabilities $\geq$ 0.707, with mean, standard deviation, and differential stability achievable in 8 to 12 minutes of practice (Kennedy, Lane, Wilkes, &

Homick, 1985; Wilkes, Kennedy, Dunlap, & Lane, 1986; Kennedy, Baltzley, Wilkes, & Kuntz, 1989). Collectively, the nine subtests have been demonstrated to identify four separate factors (cf., Kennedy, Baltzley, Turnage, & Jones, 1989) including: motor speed; symbol manipulation/reasoning; cognitive processing speed; and speed of response selection. Table 2 indicates the subtest order, practice, trial, and battery time.

TABLE 2. HUMAN PERFORMANCE SUBTEST ORDER, PRACTICE, TRIAL, AND BATTERY TIME

Subtasks in Order of Battery Presentation	Trials/ Battery	Practice Time	Trial Time	Total Task Time in a Battery Less Practice
PHT	2	10[a]	10	20
GR	1	30	150	150
MP	1	30	180	180
CS	1	30	150	150
PC	1	30	150	150
MK	1	30	150	150
STM	1	30	150	150
RT	1	30	90	90
NPT	2	10	10	20
Totals		230	1040	1060

[a] All times reported in seconds
PHT = Preferred-Hand Tap
GR = Grammatical Reasoning
MP = Math Processing
CS = Code Substitution
PC = Pattern Comparison
MK = Manikin
STM = Short-Term Memory
RT - Reaction Time-4 Choice
NPT = Nonpreferred-Hand Tap

a. Tapping (two tests: PHT and NPT). Tapping tests are motor skills/performance tasks that may be placed throughout the test battery, serving as a check against interfering factors during battery administration (e.g., boredom). The participant is required to press the indicated keys as fast as he or she can with either the Preferred (PHT) or Nonpreferred (NPT) hand. Preferred-hand Tap and NPT each require two, 10-second trials with PHT the first test in the battery and NPT the last test in the battery.

Performance is based on the number of alternate key presses made in the allotted time. In a recent study (Kennedy, Wilkes, Lane, & Homick, 1985), tapping was described as a psychomotor skill assessing factors common to both Aim and Spoke. Tapping has been highly recommended for inclusion in a repeated-measures microcomputer battery (Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Wilkes, Kennedy, Dunlap, & Lane, 1986).

b. Grammatical Reasoning (GR). The GR test (Baddeley, 1968) requires the participant to read and comprehend a simple statement about the order of two letters, A and B. Five grammatical transformations on statements about the relationship between the letters or symbols are made. The five transformations are: (1) active versus passive construction, (2) true versus false statements, (3) affirmative versus negative phrasing, (4) use of the verb "precedes" versus the verb "follows," and (5) A versus B mentioned first. There are 32 possible items arranged in random order. The subject's task is to respond "true" or false," depending on the verity of each statement with performance scored according to the number of transformations correctly identified. Grammatical Reasoning is presented as one, 150-second trial of testing. The task is described as measuring "higher mental processes" with reasoning, logic, and verbal ability, important factors in test performance (Carter, Kennedy, & Bittner, 1981). According to Bittner, Carter, Kennedy, Harbeson, and Krause (1986), GR "assesses an analytic cognitive neuropsychological function associated with the left hemisphere." Previous studies with GR, identified in Bittner, Carter, Kennedy, Harbeson, and Krause (1986), have indicated that the task is acceptable for use in repeated-measures research. Recent field testing with a microcomputer version of the task (Kennedy, Wilkes, Lane, & Homick, 1985; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Wilkes, Kennedy, Dunlap, & Lane, 1986) have resulted in strong recommendations for inclusion of GR in repeated-measures microcomputer test batteries.

c. Mathematical Processing (MP). Mathematical Processing (Shingledecker, 1984) includes arithmetical operations as well as value comparison of numeric stimuli. The participant performs one to three addition or subtraction operation(s) in a single presentation. Then, a response is made indicating whether the obtained total is greater or less than a prespecified value of five. The problems are randomly generated using only numbers 1 through 9. There are response deadlines for the problems corresponding to the demand characteristic of the test. Mathematical Processing is presented as one 180-second trial of testing.

d. Code Substitution (CS). The CS test (Ekstrom, French, Harmon, & Dermen, 1976) is a mixed associative memory and perceptual speed test with visual search, encoding, decoding, and rote recall, important performance factors. The computer displays nine alpha characters across the top of the screen and beneath the corresponding digits 1 through 9. The subject's task is to associate the digits with the alpha characters and to repeat the assigned digit code when presented with alpha characters. Code Substitution is presented as one, 150-second trial of testing. Previous studies of CS (Pepper, Kennedy, Bittner, & Wiker, 1980) have indicated that the task is acceptable for use in repeated-measures research. Recent field testing with a microbased version of the task (Kennedy, Dunlap, Wilkes, & Lane, 1985) further confirmed the acceptability of this tool.

e. Pattern Comparison (PC). The PC task (Klein & Armitage, 1979) is accomplished by the subject examining two patterns of asterisks that are simultaneously displayed on the screen. The participant is required to determine if the patterns are the same or different and respond with a corresponding "S" or "D" key press. Patterns are randomly generated with similar and different pairs presented in random order. Pattern Comparison is presented as one, 150-second trial of testing. According to Bittner, Carter, Kennedy, Harbeson, and Krause (1986), PC "assesses an integrative spatial function neuropsychologically associated with the right hemisphere." A review of PC studies (Bittner et al., 1986) indicated that the task is acceptable for use in repeated-measures research. Recent field testing with a microcomputer adaptation of the task (Kennedy, Wilkes, Lane, & Homick, 1985; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Wilkes, Kennedy, Dunlap, & Lane, 1986) resulted in strong recommendations for inclusion of PC in repeated-measures microcomputer test batteries.

f. Manikin (MK). This performance test (Benson & Gedye, 1963) involves the presentation of a simulated human figure in either a full-front or full-back facing position. The figure is shown to have two easily differentiated hand-held patterns. One of the two patterns is the matched pair to a pattern appearing below the figure. The subject's task is to determine which hand of the figure holds the matching pattern and respond by pressing the appropriate microprocessor key. Pattern type, hand associated with the matching pattern and front-to-back figure orientation, are randomly determined. Manikin is presented as one, 150-second trial of testing. The MK test is a perceptual measure of spatial transformation of mental images and involves spatial ability (Carter & Woldstad, 1985). Bittner et al. (1986) recommended the use of the MK test when latency scores are reported, and Kennedy et al. (1985) identified the MK test for inclusion in microcomputer repeated-measures batteries.

g. Short-Term Memory (STM). The STM (Sternberg, 1966) involves the presentation of a set of four letters for one second (Positive set), followed by a series of single letters presented for two seconds (probe letters). The subject's task is to determine if the probe letters accurately represent the positive set and respond with the appropriate key press. Subject response is recorded from the two buttons (T=true) (F=false) on the keyboard. Performance is based on the number of probes correctly identified. Short-Term Memory is described as a cognitive-type task which reflects short term memory scanning rate (Bittner et al., 1986). Previous research with the task (Carter, Kennedy, Bittner, & Krause, 1980; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Wilkes et al., 1986) has indicated that STM is acceptable for use in repeated-measures research.

h. Reaction Time-Four-Choice. The RT test (Donders, 1969) involves the presentation of a visual stimulus and measurement of a response latency to the stimulus. The subject's task is to respond as quickly as possible with a keypress to a simple visual stimulus. On this test, four boxes are displayed and a short tone signals a "change" in the status of one of the boxes. One of the boxes visually changes and the subject responds as rapidly as possible with a keypress beneath the box. Reaction Time is presented as one, 90-second trial of testing. Simple reaction time has been described as a perceptual

task responsive to environmental effects (Krause & Bittner, 1982), and has been recommended for repeated-measures research (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985).

Apparatus

The technical equipment/instrumentation used is discussed below:

1. NEC PC 8201A. Microcomputer testing was conducted with eight NEC PC8201A microprocessors. The NEC 8201A is configured around an 80C85 microprocessor with 64K internal ROM containing BASIC, TELCOM, and a TEXT EDITOR. RAM capacity may be expanded to 96K onboard, divided into three separate 32K banks. An RS-232 interface allows for hook-up to modem, to a CRT or flat-panel display, to a "Smart" graphics module, to a printer, or to other computer systems. Visual displays are presented on a 8-line LCD with 40 characters per line. Memory may be transferred to 32K modules with independent power supplies for storage and mailing. The entire package is lightweight (3.8 lbs), compact (110W X 40H X 130D mm), and fully portable with rechargeable nickel cadmium batteries permitting up to four hours of continuous operation. Table 3 abstracts the technical features of the system which are more fully described in NEC (1983) and Essex (1985).

TABLE 3. NEC 8201A TECHNICAL SPECIFICATIONS

FEATURES	SPECIFICATIONS
SIZE	30 CM (11 IN) X 22 CM (8.25 IN) X 6 CM (2.5 IN). 1.7 KG (3.8 LBS)
CPU	80C85 (CMOS VERSION OF 8085) WITH 2.4 MHZ CLOCK
ROM	32K (STANDARD) -- 128K (OPTIONAL)
RAM	24K (STANDARD) -- 96K (OPTIONAL)
KEYBOARD	67 STANDARD (10 FUNCTIONS, 4 CURSOR DIRECTIONAL AND 58 ADDITIONAL)
DISPLAY	19 CM (7.5 IN) X 5.0 CM (2.0 IN) WITH REVERSE VIDEO OPTION. MAY BE CONFIGURED AS EITHER A 240 X 62 ELEMENT MATRIX OR 40 CHARACTERS X 8 LINE DISPLAY
INTERFACES	1 PARALLEL (CENTRONICS COMPATIBLE) AND 3 SERIAL (RS232C AND 6 & 8 PIN BERG) JACKS
POWER SUPPLY	4 AA NONRECHARGEABLE BATTERIES, OR RECHARGEABLE NICKEL-CADMIUM PACK, OR AC ADAPTER 50/60 Hz @ 120 VAC, OR EXTERNAL BATTERY SYSTEMS (e.g., 8 AMP HR)

2. Intoximeter Model 3000. The Intoximeter Model 3000 (Intoximeter, Inc., 1987) is a gas chromatograph device that determines alcohol concentrations in the blood by analysis of breath. A breath sample is collected and moved through a tubular column by a flow of carrier gas to an analyzer. The analyzer employs the well-established principles of nondispersive infrared (NDIR) molecular absorption. Each absorption bands at frequencies unique to the compound. The position of these absorption bands do not change. However, the strength of a given absorption band will vary in direct relation to the change in the number of molecules within a fixed path. The analyzer uses a narrow band pass interference filter to isolate an absorption band at 3.39 microns, which is one of the strong absorption bands for alcohol. A heated element sends infrared energy through a two-chambered gas sample cell of fixed path length. With no absorbing gas in the sample half of the cell, the energy of the sample beam is ratioed against the energy passing through the reference half of the cell. The ratio is used to set and establishes the zero set point. The presence of alcohol in the sample cell will absorb some of the sample beam energy. The amount of energy attenuated is proportional to the number of alcohol molecules in the sample cell. The analyzer then transmits the proportional concentrations of alcohol to a graphic recorder. The results are both printed and displayed in a digital readout of BAC.

Data Collection

One week prior to the first experimental session subjects were instructed in the use of the microbased performance battery and required to practice the battery for a minimum of six trials. This prior training ensured that the subjects were familiar with the microbased testing procedures and were practiced to asymptotic levels. Subsequent analysis of practice data indicated that all subjects achieved asymptotic levels on all battery subtests.

Data collection was scheduled over a six-week period with experimental sessions conducted on Friday evenings beginning at 5:00 P.M. and ending at approximately 12:00 Midnight. Sessions were held at the Evansville, Wyoming, Police Department Headquarters and subjects were transported to and from this location. Following data collection subjects were returned to a controlled college housing environment where they were required to spend the remainder of the evening. A total of five sessions were employed in completing the various aspects of the study. The primary purpose of Session #1 was subject training and procedure familiarization and refinement. Sessions #2-#5 started two weeks after Session #1 and were devoted to performance assessment under varying (or no) amounts of alcohol. Subjects were requested not to ingest alcohol or other drugs for 24 hours prior to and following each experimental session. Subjects were also requested to eat a typical noon meal on data collection days, but abstain from further eating until the conclusion of the experimental session.

1. Training - Session #1. The primary purpose of Session #1 was to familiarize and train subjects in the study protocol and methods. Initially, subjects completed one replication of the microbased battery while under direct supervision. Subjects were then given practice with the alcohol consumption procedures, breath analysis techniques, and blood and urine

sampling methods. During training each subject consumed an alcoholic drink premixed to raise the BAC to 0.10. Alcohol consumption was carried out in a group setting and typically transpired over a 30- to 50-minute period. Subjects were encouraged to finish their drinks as rapidly as possible. Breath analysis, blood sampling, microbased performance assessment, field sobriety testing, and urine sampling followed within one to one-and-one-half hours.

2. Experimental Sessions #2-#5. Upon arrival at the experimental site, subjects completed the CHSQ which was then assessed for subject suitability for research participation. In particular, body weight, health status, prior alcohol or drug consumption, and drink mix preference were noted. Any subject indicating alcohol consumption in the previous 24-hour period was breath tested with the Intoximeter 3000. Two successive batteries of the microbased performance tests were then completed. Responses for each subject were inspected for anomalies or departures from testing protocol and, if needed, corrective action was taken. Microbased performance testing directly prior to the administration of alcohol ensured that each subject was well practiced and performing at asymptotic levels, as well as establishing pretreatment subtest performance (i.e., base rate data). Secondly, the obtained data provided a pretreatment base rate for subtest performance.

In a group setting an alcohol or placebo drink was consumed over a 3- to 5-minute period. Order of treatment application had been previously randomly determined for each subject and was known only by the study personnel preparing the drinks. Microbased performance assessment and field sobriety testing were supervised or conducted by study personnel unaware of the assigned treatment levels. Collectively, the study procedures ensured that both data collectors and subjects were equally blind to a subject's treatment status.

Following consumption of the drink, BACs were periodically monitored with the Intoximeter 3000. Breath monitoring required approximately 45 minutes for a typical subject. When the breath BAC reached aymptotic level or was on the descending limb of the BAC curve, other data collection procedures were initiated. In order of occurrence these measures consisted of the immediate drawing of two 10 cc vials of blood (one red, one gray), bladder voiding, microbased testing, field sobriety testing, urine sampling, and a final breath test. The entire data collection period was timed for each subject and was typically under 55-minutes' duration. Whole blood, blood sera, and urine BAC measures were employed both as a reliability check against measured breath BACs and for intermeasure comparisons. Furthermore, the timing of each subject, in conjunction with pre/post-breath testing, facilitated the interpolation of a subject's BAC at any point during the data collection process. Upon completing all data obligations, subjects were provided with a dinner including nonalcoholic beverages and returned to college housing.

The procedures described above, common to Experimental Sessions #2-#5, are presented in chronological order in Figure 1. Uniform application of these procedures insured the internal consistency of experimental treatments and controls.

Upon waking the following morning, subjects were required to self-administer one battery of the microbased performance tests. This "hangover" measure was completed by 10:00 A.M. and typically occurred within 8 to 12 hours of the previous pretreatment microbased measure. Corresponding blood, urine, and breath measures were not taken at this time and subjects were not assessed for total sleep or the use of hangover home remedies such as aspirin, water drinking, etc.

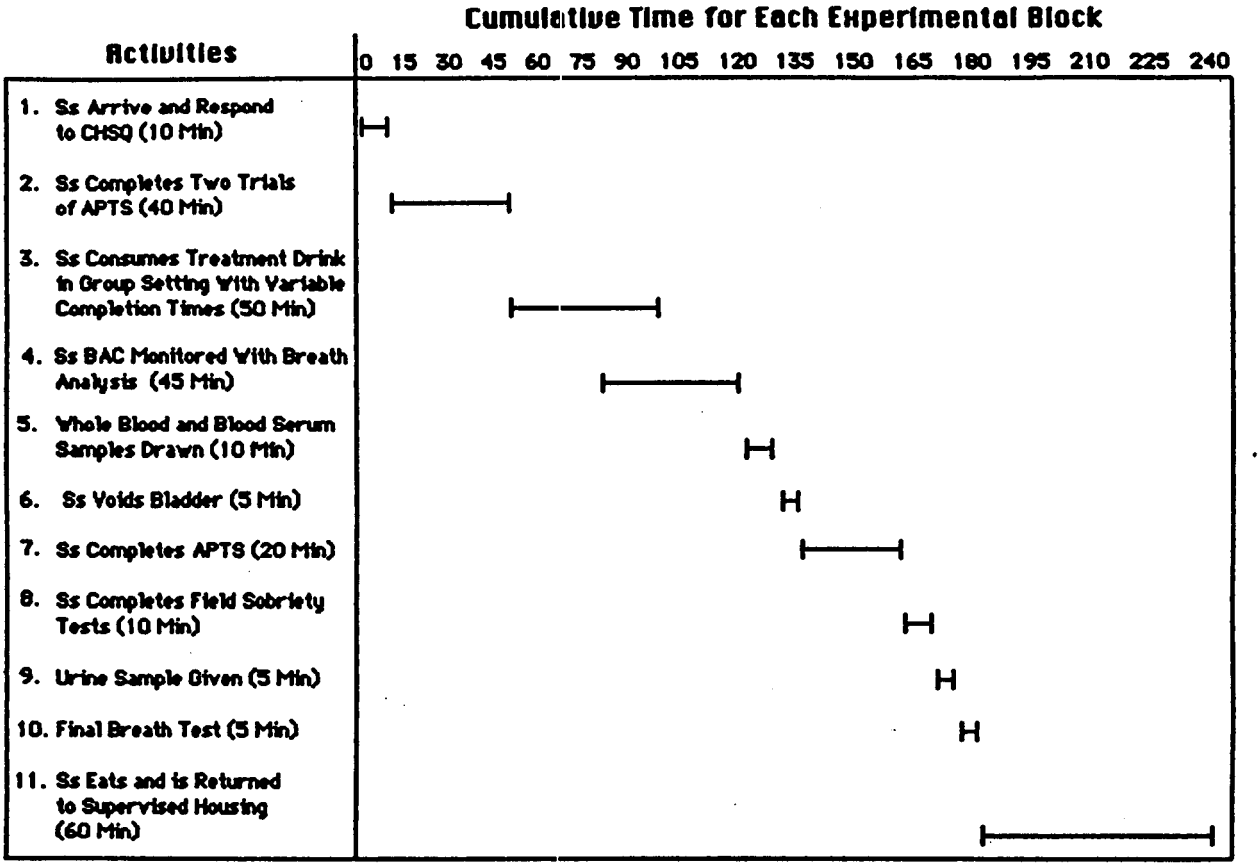

Figure 1. Time-line representing the chronological application of procedures during Experimental Sessions #2--#5

Results

Statistical analyses were performed using measures from two distinct sources: the Automated Performance Test System and the physiological indicants of alcohol concentration. The performance test battery variables included number correct, average response latency, and percent correct scores for each test administered. Alcohol concentration measures consisted of whole blood, blood sera, urine, and two breath measures (initial and final), all recorded at each blood alcohol treatment which was administered (i.e., from 0.0 to 0.15 BAC). For experimental purposes we considered the alcohol concentrations as the independent variable and sought to determine the effect on the individual APTS subtest measures.

1. The Criterion Measures

Table 4 lists descriptive statistics (means and standard deviations) of the alcohol concentration measures as recorded for each treatment. These results reflect the experimental procedure in which the initial breath test was obtained approximately 30 minutes after a subject had finished drinking and in order to catch the alcohol concentration level on the descending limb. Therefore, monitoring continued using breath analysis until the concentration had stopped increasing and achieved desired levels. Then the other indicants of concentration were taken. The results show that whole blood and sera have equal or higher concentration levels than urine measures and all biochemical assays are generally higher than the initial breath measures implying that they are indexing concentrations which are slightly later in the metabolic process. All final breath measures, recorded after the experimental procedures were completed, provide the lowest values for blood alcohol level. Therefore breath measures, if anything, underestimate blood alcohol concentrations and to a less extent this is also true of urine concentrations.

TABLE 4. DESCRIPTIVES FOR PHYSIOLOGICAL MEASURES

Variable	Mean	Standard Deviation
Alcohol level = 0.00		
Whole Blood	.00	.00
Blood Serum	.00	.00
Urine	.00	.00
Initial Breath Test*	.00	.00
Final Breath Test**	.00	.00
Alcohol level = 0.05		
Whole Blood	.06	.01
Blood Serum	.06	.01
Urine	.06	.01
Initial Breath Test	.05	.01
Final Breath Test	.05	.01
Alcohol level = 0.10		
Whole Blood	.11	.02
Blood Serum	.11	.02
Urine	.10	.02
Initial Breath Test	.10	.01
Final Breath Test	.09	.02
Alcohol level = 0.15		
Whole Blood	.16	.02
Blood Serum	.15	.02
Urine	.15	.02
Initial Breath Test	.14	.02
Final Breath Test	.13	.02

* Initial Breath Test was the breath test taken closest to the desired BAC and prior to the other physiological measures.
** Final Breath Test was the breath test taken after all the other physiological measures had been taken.

Correlations among the four measures of blood alcohol concentration are found in Table 5 for the three administrations of alcohol. The placebo indicants (not shown) contained zero correlations. It may be seen that all correlations among methods are positive. Moreover, some correlations are very high even though they constitute correlations which were calculated WITHIN A TREATMENT LEVEL where substantial range restriction can be expected to have been created by using the initial breath levels to bring all subjects to the same treatment level (i.e., all started their procedures when given levels [viz., .05; .10; .15] were reached). In general the correlations were higher with increased dosages, presumably aided by the increased variability at these higher levels and this relationship is clearly seen in Figures 2 (a-f) where scatter plots of the different methods appear. This presentation of the data shows graphically the relationship of the different measures although it should be recognized that it combines WITHIN and BETWEEN sources of variance in a single presentation.

Because of the positive, high intramethod correlations of the four physiological indicants of alcohol concentration (Table 4), and because the method with the highest intertask correlation was presumed to be the most valid and also likely to be the most reliable, we selected that method (whole blood) for further development and data analysis.

TABLE 5. INTERCORRELATIONS AMONG PHYSIOLOGICAL VARIABLES WITHIN A GIVEN ALCOHOL DOSAGE LEVEL

0.05 BAC

	Blood	Serum	Urine	IBreath	FBreath
Blood		0.98	0.56	0.95	0.70
Serum			0.55	0.91	0.61
Urine				0.51	0.50
IBreath					0.72

0.10 BAC

	Blood	Serum	Urine	IBreath	FBreath
Blood		0.99	0.94	0.91	0.67
Serum			0.92	0.90	0.66
Urine				0.86	0.67
IBreath					0.77

0.15 BAC

	Blood	Serum	Urine	IBreath	FBreath
Blood		0.99	0.90	0.90	0.89
Serum			0.91	0.91	0.88
Urine				0.75	0.77
IBreath					0.91

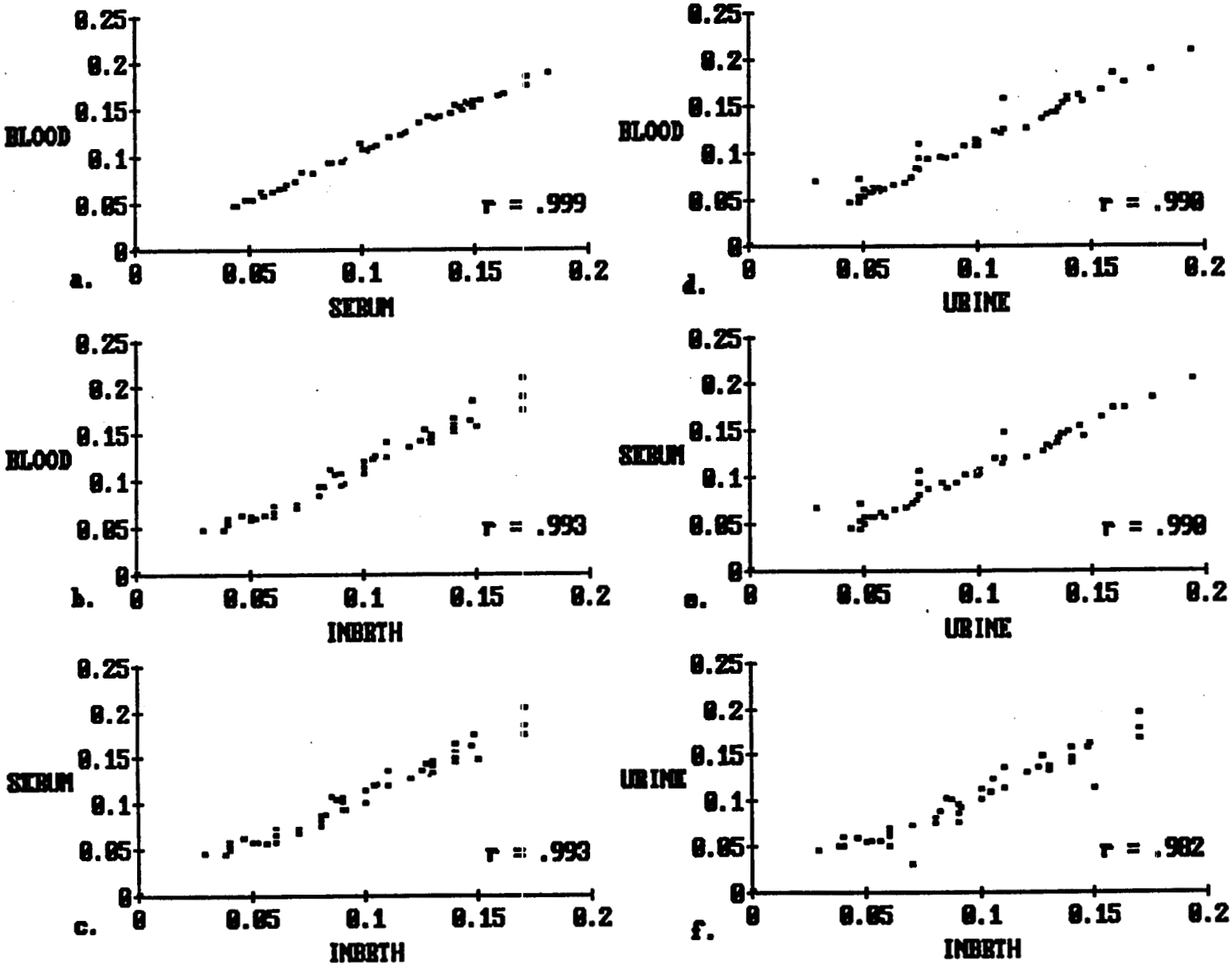

Figure 2. Scatterplots of individual alcohol concentration measures.

2. Automated Performance Test System (APTS) Measures

Descriptive statistics were initially reviewed for all performance test battery variables and Appendix A contains a complete listing of means and standard deviations for each performance test for number correct (NC), response latency (RL), and percent correct (PC) at the various blood alcohol levels and for the four different test periods. Because all tests are administered for fixed time periods, response latency and percent correct scores are essentially transforms of the number correct score and would be expected to produce similar (cf., Turnage, Kennedy & Osteen, 1987), albeit less powerful, descriptions of the same results. Test means for number correct are rationally and metrically most defensible and were selected for characterizing the findings.

Figure 3 shows the time-course stability of the nine performance tests for the two pretest trials on each of the four experimental days and prior to receiving either alcohol or placebo that session. It may be seen that improvement over sessions was gradual and less than 10% over all sessions.

Figure 4 shows the mean performances during the time when the alcohol dosage was at the four prescribed levels for the nine performance tests. Graphically, it may be seen that all 0.05 BAC mean performances are lower than all placebo performances for all tests, all 0.15 BAC mean scores are lower than all 0.10 BAC performances; most 0.10 BAC mean scores are lower than 0.05 BAC scores; the greatest change was found for the .15 BAC level. In order to maximize statistical power, the performance data obtained for the four experimental trials, were analyzed in a repeated measures ANOVA framework. Each subject was considered to have received exposures to four treatment levels (.00, .05, .10, and .15 BAC). To enhance interpretability each test was analyzed separately yielding a total of nine analyses for APTS. Two problems attend such a strategy: 1) the individual blood alcohol concentrations may contain additional predictive power and 2) multiple comparisons do not provide protection for the type I error rate; that is, by testing the same sample over and over we increase the chance of finding a difference where none exists. Analyses to account for the first problem are covered below. To offset the increase in type I bias, we selected a higher than usual alpha level for acceptance of significance, in this case a cut-off of .001 was adopted.

The resulting probability values for the APTS tests are shown to the side of each curve in Figure 2a-i. Only one test did not show a significant decrement from placebo at our selected alpha level - Grammatical Reasoning. The remaining APTS tests were significant in excess of $P < .001$ level.

Figure 5 follows the organization of figure 4, but depicts performances the morning after a full night's sleep following the four dosages of alcohol. In this data set, the subjects were tested in the "hangover" phase before being released. There is a general downward trend for the tests. While none of these relations is statistically significant, and what changes there are, are small, nonetheless only one of the twenty-seven "morning after" test scores (Tapping at .05) is as high as the placebo condition.

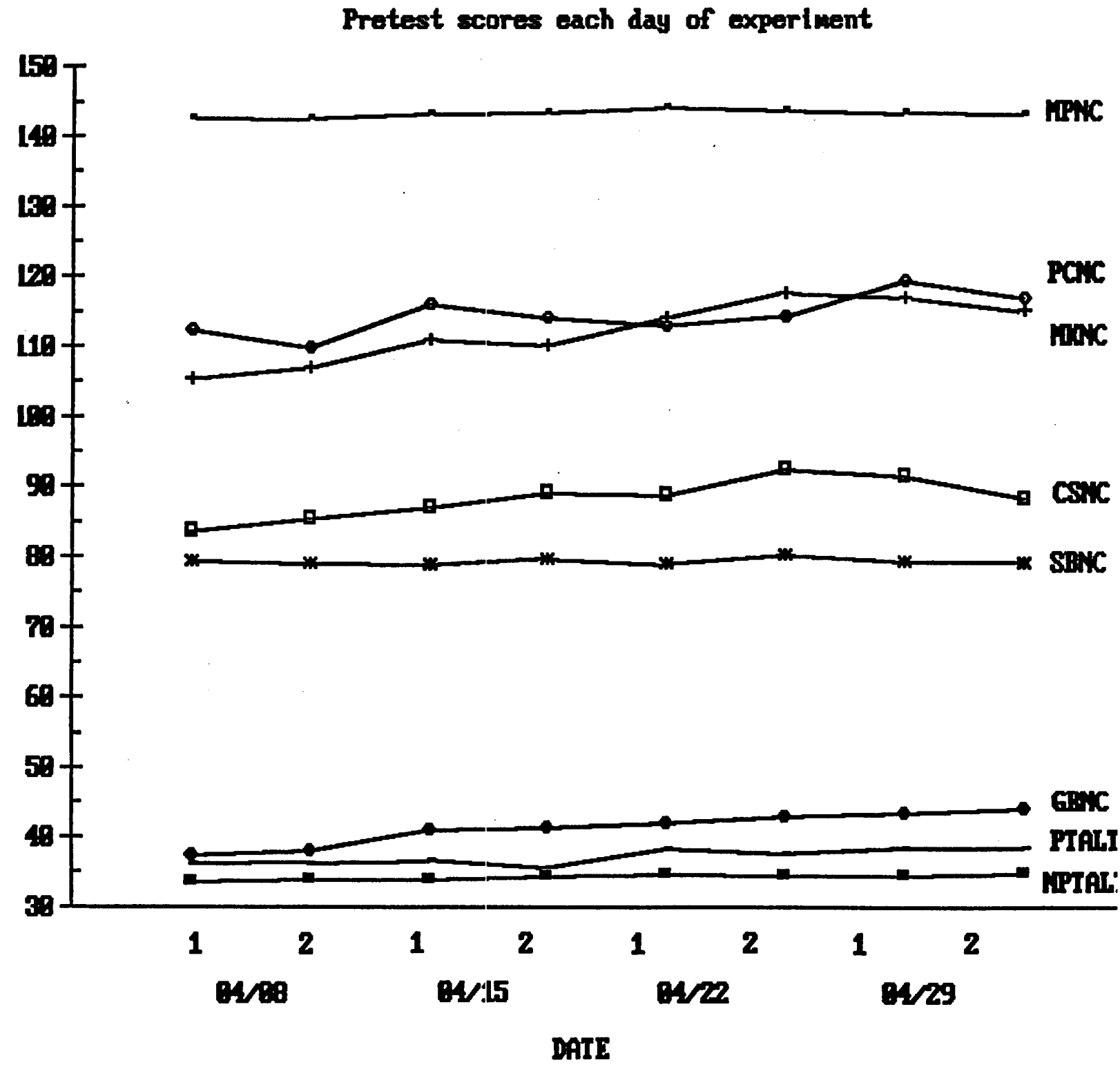

Figure 3. Time-course changes over four experimental sessions in pretest scores.

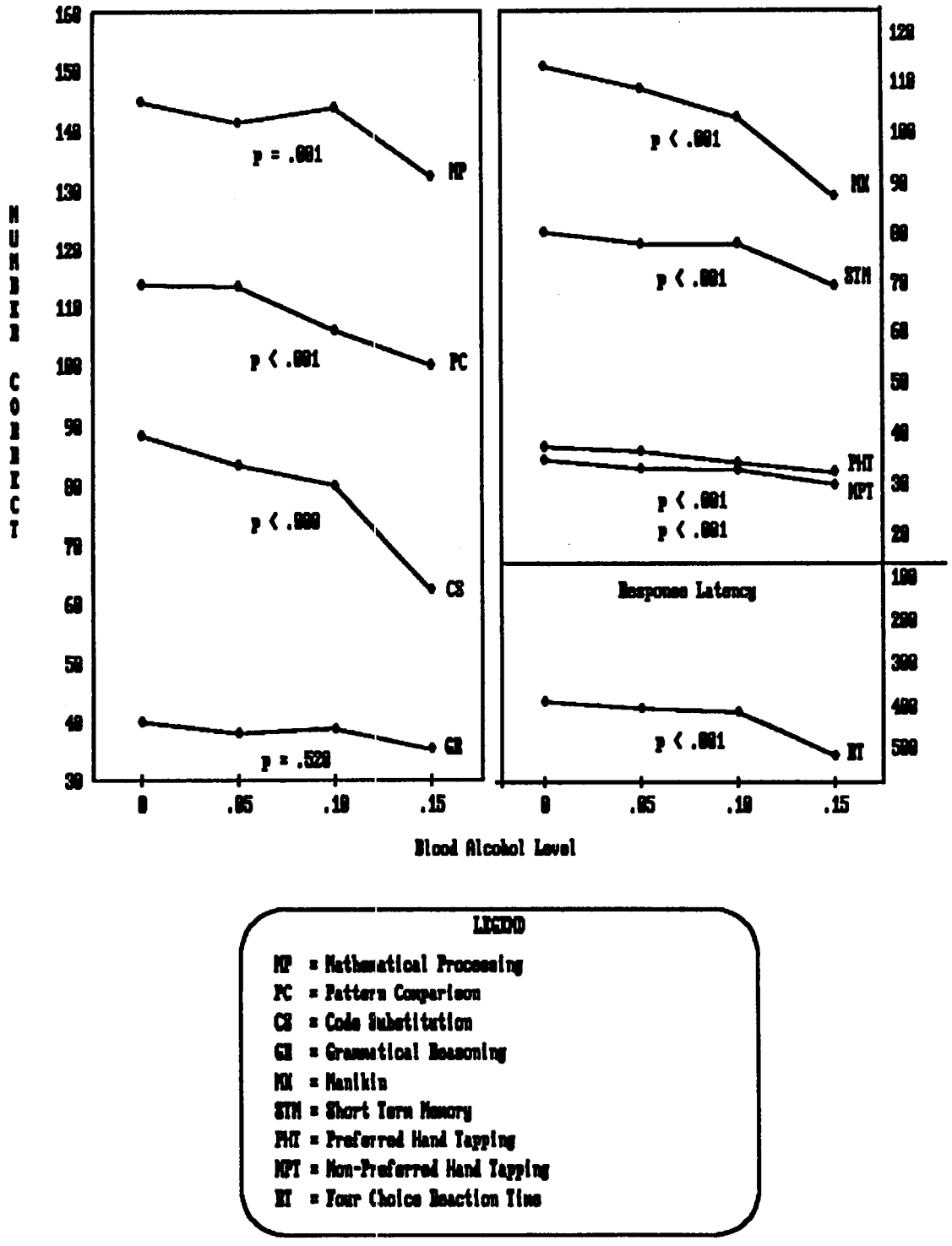

Figure 4. Effects of three graded dosages of alcohol compared to placebo for nine microcomputer tests.

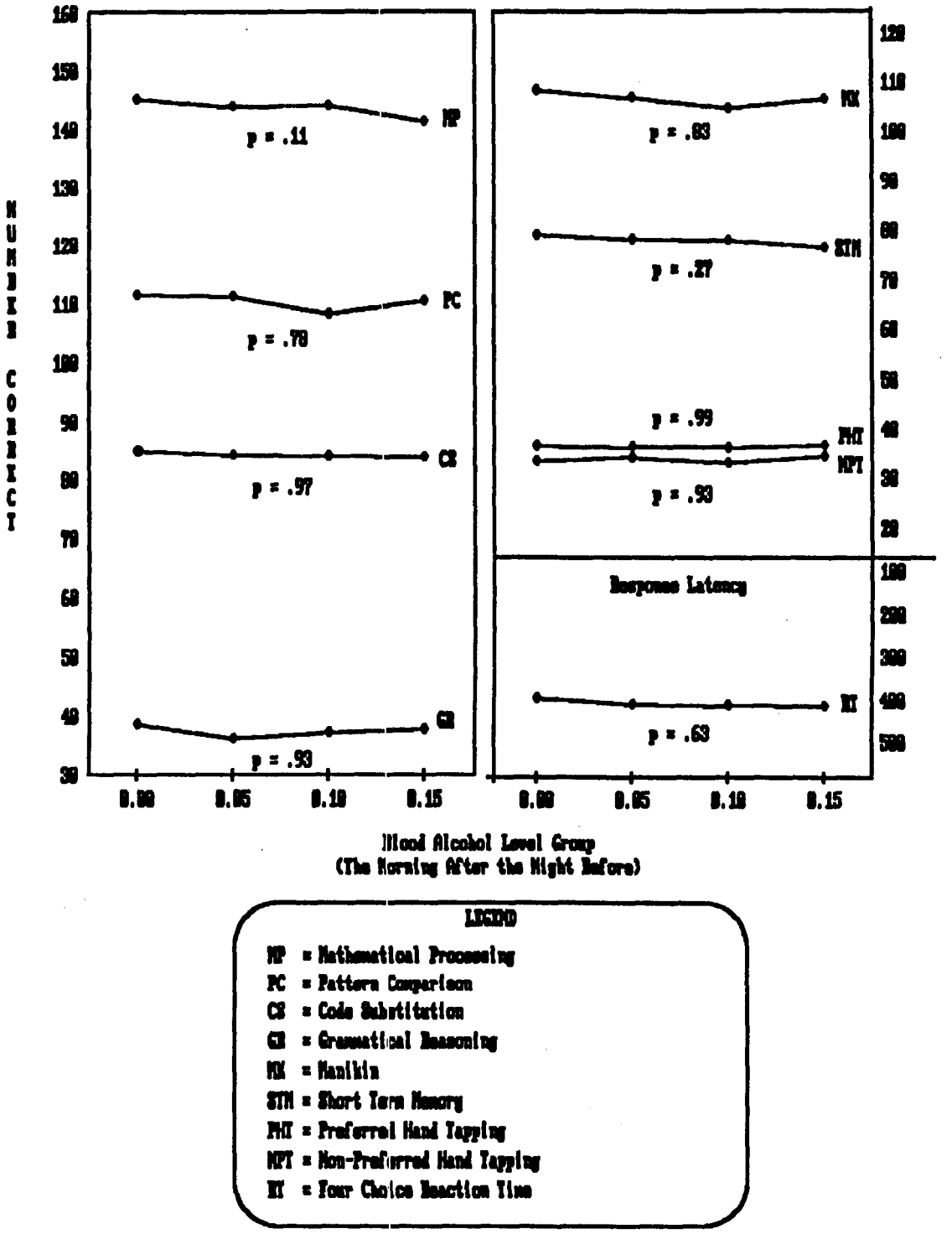

Figure 5. Performance 8-12 hours after graded dosages of alcohol ingestion for nine performance tests.

As mentioned above, the high correlations between the different alcohol concentration methods (Table 5) implied there may be additional precision retained at each treatment level which would be lost if everyone were considered part of a fixed treatment condition and simple ANOVA or MANOVA analysis methods were applied. To avoid this difficulty we sought a method of analysis which would retain the strength of the performance x alcohol concentration relationship within each subject, and also normalize, somewhat, the data for the different levels of ability of the various subjects on the different tests. For the analysis EACH subject's score for EACH test was regressed against (correlated with) HIS alcohol concentration (whole blood) value measured at the time of his performance. Thus, an individual predictive validity (in the form of a Pearson product moment correlation coefficient) was obtained for all tests and for all subjects. After converting the obtained correlations to Fisher's Z, the group average of these predictive validities was calculated. The obtained value was returned to a Pearson correlation and it was considered that this average correlation over all the subjects would index the strength of the predictive relationship between the two measures. Likewise, the average correlation of all tests for each subject, over all the measures, could also be used to index the effectiveness of alcohol as a stimulus for that subject. This latter technique we felt would be a useful method with individual cases in subsequent fitness-for-duty applications.

Table 6 shows correlations of the nine tests rank ordered by strength of relationship. It may be seen that performances on Code Substitution, Manikin, the two Tapping tests and Reaction Time followed most closely the increasing blood alcohol levels. This ordering is consistent with the ANOVA (Figure 4) where Grammatical Reasoning was not shown to reveal a statistically significant decrement with alcohol dosage. (One aberrant subject caused the results for Grammatical Reasoning to be atypical of previous experience. This is elaborated later.)

TABLE 6. AVERAGE CORRELATION (WITHIN SUBJECTS) BETWEEN APTS MEASURES AND BLOOD ALCOHOL LEVELS

Test Name	Correlation
1. Code Substitution	-.742
2. Manikin	-.728
3. Reaction Latency	.626
4. Preferred Tapping	-.621
5. Sternberg	-.590
6. Nonpreferred Tapping	-.558
7. Math Processing	-.540
8. Pattern Comparison	-.534
9. Grammatical Reasoning	-.291

For purposes of hypothesis gathering, we sought to determine whether there was increased precision available from combining tests into a single score. Carrying this correlational analysis further, each subject's score for each

performance test was regressed (correlated) with each other test score within each of the four treatment conditions in order to create a cross-task correlation matrix according to the same approach as was followed in obtaining the correlations between each test and the blood alcohol concentration. From this matrix a standard SPSS backward multiple regression solution was undertaken and cut-offs selected at the $p < .50$ level.

These results showed that three tests (Code Substitution, Reaction Time and Grammatical Reasoning) produced a multiple correlation of $r = 0.79$ which after adjustment for shrinkage still accounted for 54% of the variance. This relation, while strong, produced complex-to-explain beta weights for the Grammatical Reasoning variable, possibly because of correlations between Grammatical Reasoning and Reaction Time which are not shared with the criterion and possibly because of the aberrant subject. A multiple correlation was therefore calculated omitting Grammatical Reasoning and the result produced a correlation of $r = 0.76$ which, adjusting for shrinkage, accounts for 52% of the variance.

Because of this positive finding we elected to combine scores from several tests to derive a combination score which might add information to each test studied singly. Although the battery of tests is factorially diverse, previous studies have also shown that a general factor accounts for approximately half the variance in performance. Therefore, we set out to combine the scores further for the individual subject. It should be noted that the experimental design, while suited for evaluating the tests in the battery, was not optimal for providing individual scores for subjects since each subject received a different order of administration in order to counterbalance sequence effects. Therefore, an averaging technique was required.

It will be recalled (Figure 3) that a slight improvement in pretest means occurred over the four sessions and in some cases, the main effect of a treatment (e.g., 0.05 BAC) may have been less than the overall improvement on a test over sessions. Therefore, in order to obtain a maximally stable score for each subject, all eight pretests, over four sessions (Figure 2) were summed to form a baseline score for each person. Obtained scores under each treatment condition were then compared (i.e., divided) by this baseline in order to determine what proportion of baseline performance was retained in each treatment. Figure 6 shows the means of the nine tests after this transformation for the period after alcohol (or placebo) administration and Figure 7 shows similar data for the morning after administration of treatments. Note that baseline means average about 1.0 and each treatment reveals largely monotonic reductions proportional to dosage for the two conditions although the effects are substantially larger the night before than the morning after.

Next the score for each subject on each test was added to create a series of combination scores using more or less of the data. The rule used for adding test scores was dictated by the order (i.e., strength) of the correlations found in Table 6 and each score shows the effect of adding one more test. It was found that five tests (Code Substitution, Manikin, Reaction Time, Pattern Comparison, Sternberg) produce a combination score of $r = 0.56$, accounting for 32% of the variance and addition of the remaining tests produces no accretion.

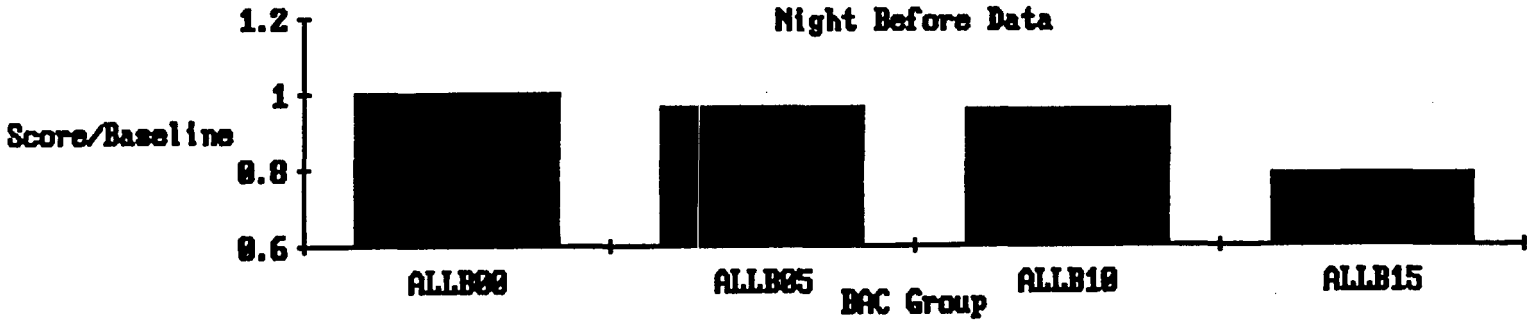

Figure 6. Combination scores for four alcohol treatments (including placebo) reflected as proportion of 8 pretest baselines.

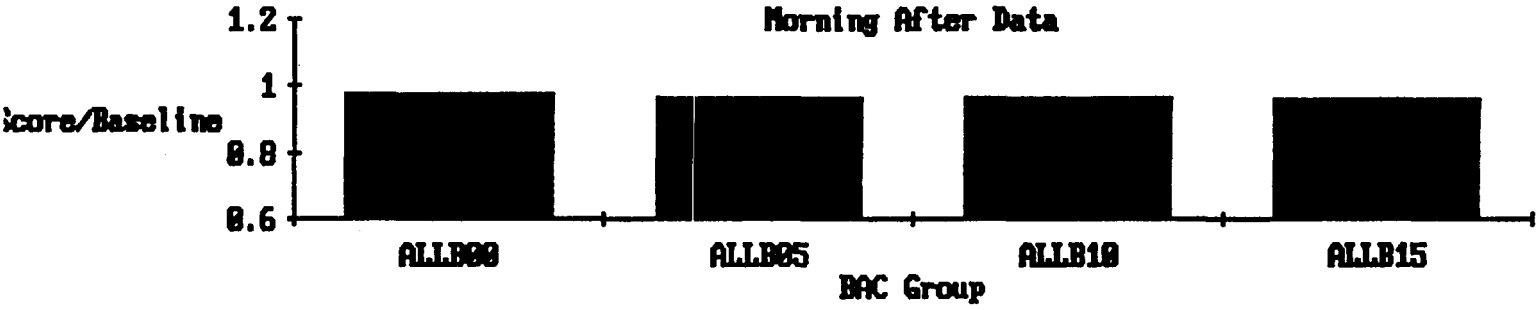

Figure 7. Combination scores for 8-12 hour period after four alcohol treatments (including placebo) reflected as proportion of 8 pretest baselines.

Table 7 contains the individual subjects listed according to the average correlation of their individual performances against alcohol. A high correlation implies that performances became increasingly degraded the higher the dosage of alcohol. Conversely, a low correlation would mean that alcohol was relatively less effective on an individual. It may be seen that not all subjects were equally affected by alcohol. Indeed, one subject had a average correlation between blood alcohol and performance over nine tests which was positive and on one test in particular (Grammatical Reasoning) had a correlation of $\underline{r} = 0.90$! This rather sobering outcome implies a less than suitable subject for the purpose of this experiment and for post hoc "what if" queries we elected to drop him and rerun the data above. The results are as expected. All tests, including Grammatical Reasoning, are statistically significant and the remaining relations, covered elsewhere in this section, are essentially the same as was reported with the "bad" subject, but are now more regular and more highly significant. This finding points out quite clearly the importance of data screening, particularly with small samples and repeated measures. We believe that further development of this metric is warranted.

TABLE 7. AVERAGE CORRELATIONS OF EACH SUBJECT'S PERFORMANCE WITH OBTAINED BLOOD ALCOHOL LEVEL OVER NINE TESTS

Subject Number	Correlation
4680	-.966
5608	-.936
0334	-.901
9758	-.860
9600	-.859
8881	-.823
4220	-.818
8373	-.780
1481	-.758
1716	-.739
3566	-.720
1950	-.712
7758	-.682
6503	-.633
4649	-.546
0220	-.379
4222	-.342
8624	-.282

LIMITATIONS AND SUGGESTED ADDITIONAL EFFORT

It was seen that there were individual differences in resistance to alcohol, and there is strong inference that these differences would be reliable if they were tested again. Using this technique to operationally define "resistant" subjects, the performance tests became dramatically more sensitive when the three most resistant subjects were dropped. (At least one of the subjects was apparently faking a low baseline performance.) We believe that further development and study of such techniques is warranted for use in fitness-for-duty testing.

Although the sample size employed in this study was satisfactory for the purpose of validating the battery, the regularity of the group data suggest that with larger samples it should be possible to calibrate tests against the alcohol as a standard marker stimulus, and when that performance deficit is matched one can use this information as advisory information for purposes of establishing exposure limits. To illustrate how such an analysis might work, we have converted the data from the present study and the data from five others to "percent reduction from baseline". In four of these studies, each subject was employed as his/her own control, and in one (drugs) a control group was employed. In Figure 8 we have normalized the mean effects obtained in the present experiment (including the one which was not statistically significant) in order to demonstrate how a dose equivalency analysis might work.

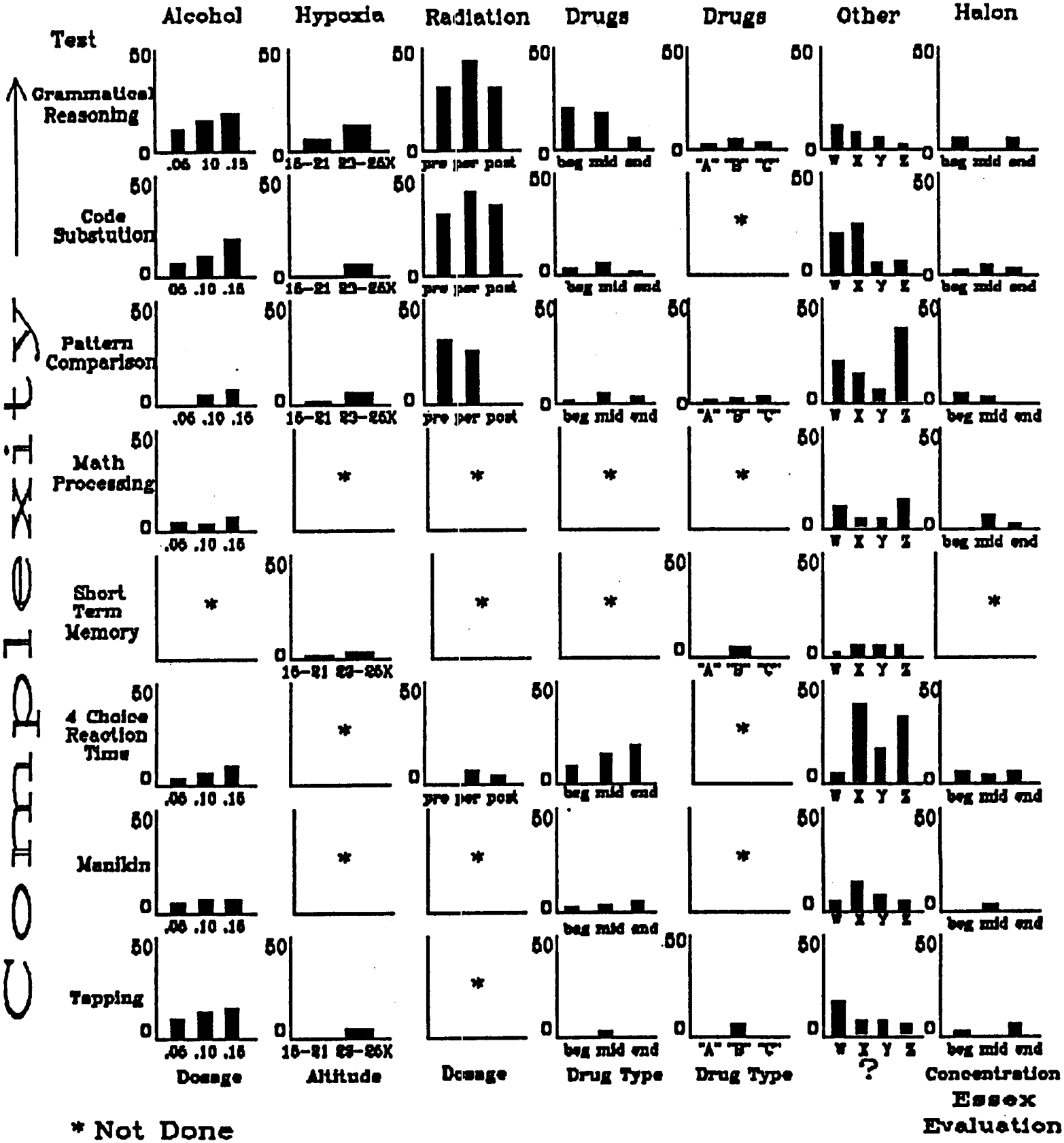

Figure 8. Dose equivalency: A proposed methodology for indexing toxic agents and treatments using the same tests.

The data include: (1) three blood alcohol levels, .05, .10, .15, (2) simulated altitude at 15-20K, at 23-25K; (3) motion sickness drugs, scopolamine, and a combination of scopolamine and dexedrine; (4) effects of chemoradiotherapy, reported as an average decrement across treatment; and (5) two antihistamines. It may be seen that the blood alcohol levels based against placebo show an orderly loss to performance from .05 to .15. We suggest that this relation be used as a preliminary marker to index other comparable effects calculated as percentage of baseline. This approach is advocated for providing guidance regarding strength of relationships and "dose equivalency," not for statistical testing. It is well known that percentages (Turnage et al., 1987) lack sufficient statistical power and are generally to be avoided. Percent decrement is basically a difference score between ratios and tends to be unstable. Therefore, it is not an optimal metric. However, a percent decrement score does allow for the comparison of diverse studies across a common metric.

When these rational and experimentally well-controlled data from an alcohol study are used to "calibrate" or mark the other results, it would appear that the chemoradiotherapy treatments (Parth, Lane, Dunlap, Chapman, Kennedy, & Ordy, 1988) exhibit the strongest effect, although we also know (not shown) that this effect recovers when the subjects who survived the treatment were tested 12 months later. Note also that while scopolamine alone has a slight (and mildly significant) effect (Kennedy, Wood, Graybiel, & McDonough, 1986), when scopolamine is combined with amphetamine this effect is lessened. The altitude study (Banderet, Shukitt, Crohn, Kennedy, Smith, Houston, & Bittner, 1987) shows a similar effect and even at the highest altitude obtained (23-25,000 feet, the approximate height of Mt. Everest), the effect is no stronger than we found with 2-3 drinks of alcohol (i.e., .05-.10 BAC). Although the data are too sparse to conclude confidently, the pattern of the changes is illustrative of what conclusions which may be possible with a larger data base; for example, the more complex mental tests (e.g., Code Substitution and Grammatical Reasoning) appear to be most sensitive; 4-Choice Reaction Time, a response speed measure, also appears sensitive. Whether other treatments will show the same effect or not is problematic and awaits further study. We believe a completely filled matrix of tests X agents X dosages X mental factor would be extremely useful. This is the rationale behind proposing similar testing for fitness-for-duty decisions as well.

CONCLUDING REMARKS

Mental tests can provide an indication of the onset, duration, and severity of impairment in operational performance which may be due to environmental hazards or toxic chemicals. The advent of microcomputers can expand the potential for assessment over paper-and pencil media by permitting more rapid, diverse, and accurate assessment of capabilities. Suitability requirements for such test materials include satisfying metric criteria and practical factors. This paper reviewed a program of several interlocking normative studies which have yielded a menu of tests demonstrates specific metric features: stability, task definition, reliability efficiency, and factor diversity. Throughout this experimental program to select the "best" tests for an optimal computerized test battery for assessment of environmental effects on skilled behavior and higher level tasks, we have stressed the need

for repeated-measures experiments to properly evaluate test stability, reliability, and factorial purity.

From this work, sponsored jointly by NASA, NSF, and Essex internally, we now have short (< 10 min.), medium (10-15 min.) and longer (> 15 min.) batteries available with factor loadings and predictive validities from correlations with holistic measures of intelligence. Previously validation was available in the form of extramural sensitivity studies (drugs, sleep loss, mixed gas, simulated altitude, and chemoradiotherapy). The present alcohol study described in this report adds additional validation data for the medium length battery (nine tests) in the form of statistical and graphic changes in performance with increasing dosages of alcohol.

Although the total number of tests available in the menu seems relatively large, it should be noted that the tests taken together tap only a limited number of dimensions. Factor analyses indicate that the 40 tests contain no more than five, and possibly as few as three factors, and that most (80% to 90%) of the reliable variance in the battery is present in the first three dimensions. (The "exact" dimensionality of the battery depends to some extent on how a factor is defined and how "important" a factor should be before it is considered "real." There is also a tendency for the factor pattern to change as practice on the tests continues.) Because the number of factors is so small relative to the number of tests, using more than six to eight selected tests adds very little to the information obtained, while materially complicating administration of the battery. Therefore, when time permits we have proposed the use of nine tests, seven of which assess cognitive acuities and two (the Tapping series) motor skills.

While all tests appear valid, some of them appeared more sensitive than others. Code Substitution, Manikin, and Choice Reaction Time are good bets for a short battery. The first three have also been used in other environments (Kennedy, Odenheimer, Baltzley, Dunlap, & Wood, 1989; Kennedy, Dunlap, Banderet, Houston, & Smith, 1989) with success. From the standpoint of these tests it would appear that greater changes occurred in cognitive function between the placebo and .05 level than between the .05 and .10 level. However, the greatest reduction in performance occurred between .10 and .15, and the relatively abrupt nature of this change implies that sharp cut-offs in cognitive performance occur at that point, and future studies should focus on this breakpoint and explore its functional shape since it has important implications for agencies with regulatory responsibilities. Not surprisingly, this breakpoint coincides with the legal limit on driving while intoxicated (DWI) or driving under the influence (DUI) used in most of the United States. The present study did not surface any evidence which suggests that this is an inappropriate break point, at least from a measurement standpoint.

REFERENCES

Ackerman, P. L., & Schneider, W. (1984, April). Ramifications of practice effects for selection and training: A new approach to individual differences assessment. Paper presented at the First Annual Mid-Central Ergonomics/Human Factors Society, Cincinnati, OH.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks Cole Publishing.

American Psychological Association (1982). Ethical principles to the conduct of research with human subjects. Washington, DC: Author.

Arthur, G. (1949). The Arthur adaptation of the Leiter international performance scale. Journal of Clinical Psychology, 5, 345-349.

Baddeley, A. D. (1968). A three-minute reasoning test based on grammatical transformation. Psychonomic Science, 10, 341-342.

Baker, E. L., Letz, R. E., Fidler, A. T., Shalot, S., Plantamura, D., & Lyndon, M. (1985). A computer-based neurobehavioral evaluation for occupational and environmental epidemiology: Methodology and validation studies. Neurobehavioral Toxicology & Teratology, 7, 369-377.

Banderet, L. E., & Burse, R. L. (1984, August). Cognitive performance at 4500m simulated altitude. Proceedings of the 92nd Annual American Psychological Association. Toronto, Canada.

Banderet, L. E., MacDougall, D. M., Roberts, D. E., Tappan, D., Jacey, M., & Gray, P. (1984). Effects of dehydration on cold exposure and restricted fluid intake on cognitive performance. Proceedings of a Workshop: Predicting Decrements in Military Performance Due to Inadequate Nutrition (pp. 69-79). Washington, DC: National Academy Press.

Banderet, L. E., Shukitt, B. L., Crohn, E. A., Kennedy, R. S., Smith, M. G., Houston, C. S., & Bittner, A. C., Jr. (1987). Cognitive performance and subjective responses during prolonged ascent to 7600m (25,000 ft) simulated altitude. Manuscript submitted for publication.

Banderet, L. E., Shukitt, B. L., Kennedy, R. S., Bittner, A. C., Jr., & Kay, G. G. (1988, November). Psychometric properties of three condition tasks with different response requirements. Proceedings of the 30th Annual Meeting of the Military Testing Association, Arlington, VA.

Barrett, G. V., Alexander, R. A., Doverspike, D., Cellar, D., & Thomas, J. (1982). The development and applications of a computerized information-processing test battery. Applied Psychological Measurement, 6, 13-29.

Benson, A. J., & Gedye, J. L. (1963). Logical processes in the resolution of orientation conflict (Rep. 259). Farnborough, United Kingdom: Royal Air Force Institute of Aviation Medicine.

Berger, C. F., Shermis, M. D., Stemmer, P. M. Jr., & Anderson, G. E. (1988, August). Microcomputers for psychological research. Paper presented at the meeting of the American Psychological Association, Atlanta, GA.

Bittner, A. C., Jr. (1979). Statistical tests for differential stability. Proceedings of the 23rd Annual Meeting of the Human Factors Society (pp. 541-545). Santa Monica, CA: Human Factors Society.

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

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1986). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-708.

Braune, R., & Wickens, C. D. (1985). The functional age profile: An objective decision criterion for the assessment of pilot performance capacities and capabilities. Human Factors, 27(6), 681-693.

Cahalan, D., Cisin, I. H., & Crossley, H. M. (1969). American drinking practices: A national study of drinking behavior and attitudes (Monograph No. 6). Rutgers Center of Alcohol Studies, Rutgers University, New Brunswick, NJ.

Calkins, D. S. (1989). Results of performance testing. Paper presented at the 60th Annual Scientific Meeting of the Aerospace Medical Association (Halon 1301 Panel), Washington, DC.

Carretta, T. R. (1987, September). Basic attributes test (BAT) systems: Development of an automated test battery for pilot selection (AFHRL TR-87-9). Brooks Air Force Base, TX: Air Force Systems Command. (NTIS No. AD A185649)

Carretta, T. R. (1989). USAF pilot selection and classification systems. Aviation, Space, and Environmental Medicine, 60, 46-49.

Carroll, J. B. (1980, April). Individual difference relations in psychometric and experimental cognitive tasks (Rep. No. 163). Chapel Hill, NC: The L. L. Thurstone Psychometric Laboratory, University of North Carolina. (NTIS No. AD A086057)

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. (1980). Selection of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 320-324). Santa Monica, CA: Human Factors Society.

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. (1981). Grammatical reasoning: A stable performance yardstick. Human Factors, 23, 587-591.

Carter, R. C., Kennedy, R. S., Bittner, A. C. Jr., & Krause, M. (1980). Item recognition as a performance evaluation test for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 340-344). Santa Monica, CA: Human Factors Society.

Carter, R. C., & Wolstad, J. C. (1985). Repeated measurements of spatial ability with the Manikin test. Human Factors, 27(2), 209-219.

Casson, I. R., Siegel, D., Skarn, R., Campbell, E. A., Tarlau, M., & DiDomenico, A. (1984). Brain damage in modern boxers. Journal of the American Medical Association, 251, 2663-2667.

Christal, R. E. (1981). The need for laboratory research to improve the state-of-the-art in ability testing. Paper presented at the National Security Industrial Association, DOD Conference on Personnel and Training Factors in Systems Effectiveness, San Diego, CA.

Deroshia, C. (in press). The effect of exercise and training upon performance and mood during antiorthostatic bedrest. Moffett, CA: Ames Research Center.

Dixon, W. J. (1983). BMDP statistical sofware. Los Angeles, CA: University of California at Los Angeles.

Donders, F. C. (1969). On the speed of mental processes. (Translated by W. G. Koster). Acta Psychologica, 30, 412-431.

Dunlap, W. P., Kennedy, R. S., Harbeson, M. M., & Fowlkes, J. E. (1989). Difficulties with individual difference measures upon recent cognitive paradigms. Applied Psychological Measurement Journal. (In press).

Eckstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976, August). Manual for kit of factor-referenced cognitive tests (Office of Naval Research Contract No. N00014-71-C-0117). Princeton, NJ: Educational Testing Service.

Edwards, A. L. (1985). Experimental design in psychological research. New York: Harper & Row.

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

Essex Corporation. (1985, January). Isoperformance from disparate combinations of practice, selection, and equipment (SBIR proposal Phase I, Topic No. 204, Air Force). Orlando, FL: Author.

Essex Corporation. (1986). Automated Performance Test System (Brochure). Orlando, FL: Essex.

Essex Corporation (1988). Unpublished evaluation report. Orlando, FL: Author.

Farrell, A. D. (1983). When is a computerized assessment system ready for distribution? Computers in Psychiatry/Psychology, 5, 9-11.

Giannetti, R. A. (1988, August 15). Psychological assessment by computer: computerized adaptive acquisition of self-report psychosocial history data. Paper presented at the meeting of the American Psychological Association, Atlanta, GA.

Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York: McGraw Hill.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Gullion, C. M., & Eckerman, D. A. (1986). Field testing for neurobehavioral toxicity: Methods and methodological issues. In Z. Annau (Ed.), Neurobehavioral toxicology. Baltimore, MD: Johns Hopkins University.

Hanninen, H., & Lindstrom, K. (1979). Behavioral test battery for toxicopsychological studies used at the Institute of Occupational Health in Helsinki (2nd ed.). HElsinki, Finland: Institute of Occupational Health.

Harbeson, M. M., Kennedy, R. S., Krause, M., & Bittner, A. C., Jr. (1982). Repeated measures of information processing. Proceedings of the 26th Annual Meeting of the Human Factors Society (pp. 818-822). Seattle, WA: Human Factors Society.

Heslegrave, R. J., & Angus, R. G. (1985). The effects of task duration and work-session location on performance degradation induced by sleep loss and sustained cognitive work. Behavior Research Methods, Instruments, & Computers, 17(6), 592-603.

Hunt, E. (1985). Science, technology, and intelligence (Tech. Rep. No. 9). Arlington, VA: Office of Naval Research, Personnel and Training Research Programs.

Hunt, E. B., & Pellegrino, J. (1986). Testing and measures of performance. Proceedings of the 27th Annual Meeting of the Psychonomic Society (pp. 385). New Orleans, LA.

Hunter, D. R. (1975). Development of an enlisted psychomotor/perceptual test battery (AMFHRL-TR-7560). Brooks Air Force Base, TX: Air Force Human Resources Laboratory. (NTIS No. AD A020544)

Intoximeters, Inc. (1987). Intoximeter 3000 Supervisors Manual. 1901 Locust Street, St. Louis, MO 63103.

Jeanneret, P. R. (1988). Position requirements for space station personnel and linkages to portable microcomputer performance assessment (EOTR). Orlando, FL: Essex Corporation.

Jones, M. B. (1970a). A two-process theory of individual differences in motor learning. Psychological Review, 77(4), 353-360.

Jones, M. B. (1970b). Rate and terminal processes in skill acquisition. American Journal of Psychology, 83(2), 222-236.

Jones, M. B. (1980). Stabilization and task definition in a performance test battery (Final Rep., Contract N00203-79-M-5089). New Orleans, LA: U.S. Naval Aerospace Medical Research Laboratory. (AD A099987)

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

Kantor, J. E., & Bordelon, B. S. (1985). The USAF pilot selection and classification program. Aviation, Space, and Environmental Medicine, 56, 258-261.

Kennedy, R. S., Baltzley, D. R., Dunlap, W. P., Wilkes, R. L., & Kuntz, L. A. (in preparation). Microcomputer-based tests for repeated-measures: Metric properties and predictive validities. Orlando, FL: Essex Corporation.

Kennedy, R. S., Baltzley, D. R., Turnage, J. J., & Jones, M. B. (1989). Factor analysis and predictive validity of microcomputer-based tests. Perceptual and Motor Skills, 69, 1059-1074.

Kennedy, R. S., Baltzley, D. R., Wilkes, R. L., & Kuntz, L. A. (1989). Psychology of computer use: IX. A menu of self-administered microcomputer-based neurotoxicology tests. Perceptual and Motor Skills, 68, 1255-1272.

Kennedy, R. S., & Bittner, A. C., Jr. (1977). The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In L. T. Pope & D. Meister (Eds.), Productivity enhancement: Personnel performance assessment in Navy systems (pp. 393-408). Naval Personnel Research and Development Center, San Diego, CA. (NTIS No. AD A056047)

Kennedy, R. S., & Bittner, A. C., Jr. (1978). Progress in the analysis of a Performance Test for Environmental Research (PETER). Proceedings of the 22nd Annual Meeting of the Human Factors Society (pp. 29-35). Detroit, MI. (NTIS No. AD A060676)

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B. (1981). Perspectives in Performance Evaluation Tests for Environmental Research (PETER): Collected papers (Research Rep. No. NBDL-80R004). New Orleans, LA: Naval Biodynamics Laboratory. (NTIS No. AD A111180)

Kennedy, R. S., Carter, R. C., & Bittner, A. C., Jr. (1980). A catalogue of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 344-348). Los Angeles, CA.

Kennedy, R. S., Dunlap, W. P., Banderet, L. E., Smith, M. G., Houston, C. S. (1989). Cognitive performance deficits in a simulated climb of Mount Everest: Operation Everest II. Aviation, Space, and Environmental Medicine, 60, 99-104.

Kennedy, R. S., Dunlap, W. P., Jones, M. B., Lane, N. E., & Wilkes, R. L. (1985). Portable human assessment battery: Stability, reliability, factor structure, and correlation with tests of intelligence (Tech. Rep. No. EOTR 86-4). Orlando, FL: Essex Corporation.

Kennedy, R. S., Dunlap, W. P., Wilkes, R. L., & Lane, N. E. (1985). Development of a portable computerized performance test system. Proceedings of the 27th Annual Conference of the Military Testing Association (pp. 107-112). San Diego, CA: Navy Personnel Research and Development Center.

Kennedy, R. S., Fowlkes, J. E., Lilienthal, M. G., & Dutton, B. (1987). Postural and psychomotor performance changes in Navy pilots following exposures to flight simulators (NTSC TR-87-010). Orlando, FL: Naval Training Systems Center.

Kennedy, R. S., Jones, M. B., Baltzley, D. R., & Turnage, J. J. (1989). Factor and regression analysis of a microcomputer-based cognitive test battery. Orlando, FL: Essex Corporation.

Kennedy, R. S., Odenheimer, R. C., Baltzley, D. R., Dunlap, W. P., & Wood, C. D. (1990). Differential effects of scopolamine and amphetamine on microcomputer-based performance tests. Submitted for publication, Aviation, Space, and Environmental Medicine.

Kennedy, R. S., Turnage, J. J., & Osteen, M. K. (in press). Performance of performance tests: Comparison of psychometric properties of 30 tests from two microcomputer-based batteries. Manuscript submitted for publication, Human Performance.

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., & Kuntz, L. A. (1987, October). Microbased repeated-measures performance testing and general intelligence. Paper presented at the 29th Annual Conference of the Military Testing Association, Ottawa, Ontario, Canada.

Kennedy, R. S., Wilkes, R. L., Kuntz, L. A., & Baltzley, D. R. (1988, October). A menu of self-administered microcomputer-based neurotoxicology tests (EOTR 88-10). Orlando, FL: Essex Corporation.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L. (1985). Preliminary evaluation of a microbased repeated-measures testing system (Tech. Rep. No. EOTR 85-1). Orlando, FL: Essex Corporation.

Kennedy, R. S., Wood, C. D., Graybiel, A., & McDonough, R. B. (1986). Side effects of some antimotion sickness drugs as measured by psychomotor test and questionnaires. Aerospace Medicine, 37, 408-411.

Kiziltan, M. (1985). 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.

Klein, R., & Armitage, R. (1979). Rhythms in human performance: 1-1/2 hour oscillations in cognitive style. Science, 204, 1326-1328.

Krause, M. (1983). Paper-and-pencil and computerized performance tests: Does the medium make a difference? New Orleans, LA: Naval Biodynamics Laboratory.

Krause, M., & Bittner, A. C., Jr. (1982). Repeated measures on a choice reaction time test (Rep. No. NBDL-82-R006). New Orleans, LA: Naval Biodynamics Laboratory. (NTIS No. AD A121904)

Lane, N. E., & Kennedy, R. S. (Eds.). (1988, May). Users manual for the Essex Automated Performance Test System (APTS) (Tech. Rep. No. EOTR 88-5). Orlando, FL: Essex Corporation.

Lane, N. E., Kennedy, R. S., & Jones, M. B. (1986). Overcoming unreliability in operational measures: The use of surrogate measure systems. Proceedings of the 30th Annual Meeting of the Human Factors Society (pp. 1398-1402). Dayton, OH: Human Factors Society.

Lane, N. E., & Kennedy, R. S. (1988, June). A new method for quantifying simulator sickness: Development and application of the simulator sickness questionnaire (SSQ) (Tech. Rep. EOTR-88-7). Orlando, FL: Essex Corporation.

Logie, R. H., & Baddeley, A. D. (1985). Cognitive performance during simulated deep-sea diving. Ergonomics, 28(5), 711-746.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McCauley, M. E., Kennedy, R. S., & Bittner, A. C., Jr. (1980). Development of Performance Evaluation Tests for Environmental Research (PETER): Time estimation. Perceptual & Motor Skills, 51, 655-665.

McCombs, B. L., Doll, R. E., Baltzley, D. R., & Kennedy, R. S. (1986). Predictive validities of primary motivation scales for reenlistment decision making (Contract No. MDA903-86-C-0114). Alexandria, VA: Army Research Institute. (NTIS No. AD A187247)

Mulford, H. A., & Miller, D. E. (1961). An index of alcoholic drinking behavior related to the meanings of alcohol. Journal of Health and Human Behavior, 2, 26-31.

Naitoh, P. (1982). Chronobiologic approach for optimizing human performance. In F. M. Brown & R. C. Gaeber (Eds.), Rhythmic aspects of behavior (pp. 41-103). Hillsdale, NJ: Erlbaum.

NEC Home Electronics (USA). (1983). NEC PC-8201A users guide. Tokyo: Nippon Electric Co., Ltd.

O'Donnell, R. D. (1981). Development of a neurophysiological test battery for workload assessment in the U.S. Air Force. Proceedings of the International Conference on Cybernetics and Society (pp. 398-402). IEEE, Atlanta, GA: Air Force Aerospace Medical Research Laboratory, Human Engineering Division, Wright-Patterson Air Force Base, OH.

Orr, W. C., & Naitoh, P. (1976). The coherence spectrum: An extension of correlation analysis with applications to chronobiology. International Journal of Chronobiology, 3, 171-192.

Parth, P., Dunlap, W. P., Kennedy, R. S., Lane, N. E., Chapman, R., & Ordy, J. M. (1989). Motor and cognitive testing of bone marrow transplant patients after chemoradiotherapy. Perceptual and Motor Skills, 68, 1227-1241.

Parth, P., Lane, N. E., Dunlap, W. P., Chapman, R., Kennedy, R. S., & Ordy, J. M. (1988). Cognitive deficits resulting from chemoradiotherapy in bone marrow transplant patients. Orlando, FL: Essex Corporation.

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

Pepper, R. L., Kennedy, R. S., Bittner, A. C., Jr., & Wiker, S. F. (1980). Performance Evaluation Tests for Environmental Research (PETER): Code substitution test. Proceedings of the 7th Psychology in the DoD Symposium (pp. 451-457). Colorado Springs, CO: USAF Academy.

Reid, G. B., Shingledecker, C. A., Nygun, T. E., & Eggeneier, F. T. (1981). Development of multidimensional subjective measure of workload. Proceedings of the International Conference on Cybernetics and Society (pp. 403-406). Atlanta, GA: Air Force Aerospace Medical Research Laboratory, Human Engineering Division, Wright-Patterson Air Force Base, OH.

Reitan, R. M., & Davison, L. A. (Eds.). (1974). Clinical neuropsychology: Current status and applications. New York: Halstead.

Rock, D. L., & Nolen, P.A. (1982). Comparison of the standard and computerized versions of the Raven Coloured Progressive Matrices Test. Perceptual and Motor Skills, 54, 40-42.

Rogers, W. H., Noddin, E. M., & Moeller, G. (1982). The effect of the thermal conditions of training and testing on the performance of motor tasks measuring primary manual abilities (Rep. No. 983). Groton, CT: Naval Submarine Medical Research Laboratory.

Rosa, R. R., & Colligan, M. J. (1988). Long wordays versus restdays: Assessing fatigue and alertness with a portable performance battery. Human Factors, 30(3), 305-317.

Schlegel, R. E., & Shingledecker, C. A. (1985). Training characteristics of the criterion test set workload assessment battery. Paper presented at the 29th Annual Meeting of the Human Factors Society, Baltimore, MD: Human Factors Society.

Shingledecker, C. A. (1984). A task battery for applied human performance assessment research (Tech. Rep. No. AFAMRL-TR-84). Dayton, OH: Air Force Aerospace Medical Research Laboratory.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

Steinberg, E. P. (1986). Practice for the armed services test. New York: Acco Publishing Co.

Sternberg, R. J. (1979). The nature of mental abilities. American Psychologist, 34 (3), 214-230.

Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.

Thorne, D. R., Genser, S. G., Sing, H. C., & Hegge, F. W. (1985). The Walter Reed Performance Assessment Battery. Neurobehavioral Toxicology & Teratology, 7, 415-418.

Turnage, J. J., Kennedy, R. S., & Osteen, M. K. (1987). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations. Orlando, FL: Essex Corporation.

Watts, K., Baddeley, A., & Williams, M. (1982). Automated tailored testing using Raven's matrices and the Mill Hill vocabulary tests: A comparison with manual administration. International Journal of Man-Machine Studies, 17, 331-344.

Wechsler, D. (1981). WAIS-R manual: Wechsler Adult Intelligence Scale-revised. San Antonio, TX: The Psychological Corporation.

Wilkes, R. L., Kennedy, R. S., Dunlap, W. P., & Lane, N. E. (1986). Stability, reliability, and cross-mode correlation of tests in a recommended 8-minute performance assessment battery (Tech. Rep. No. EOTR 86-4). Orlando, FL: Essex Corporation.

Wilson, S. L., Thompson, J. A., & Wylie, G. (1982). Automated psychological testing for the severely physically handicapped. International Journal of Man-Machine Studies, 17, 291-296.

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

Wonderlic, E. F. (1978). Wonderlic personnel test manual. Northfield, IL: Wonderlic.

APPENDIX A

SUMMARY TABLES FOR ALCOHOL STUDY

Descriptives - Means

Test	0.00 BAL A	B	C
Grammatical Reasoning			
Number Correct	42.22	40.00	38.50
Response Latency	3078.0	3200.0	3491.0
Percent Correct	89.45	88.91	89.80
Mathematical Processing			
Number Correct	143.83	144.83	145.28
Response Latency	6400.0	6500.0	6800.0
Percent Correct	96.39	97.03	97.11
Code Substitution			
Number Correct	88.44	88.28	84.78
Response Latency	1528.0	1546.0	1621.0
Percent Correct	97.84	98.65	97.89
Pattern Comparison			
Number Correct	115.72	113.83	111.44
Response Latency	931.0	962.0	987.0
Percent Correct	95.34	94.87	94.42
Manikin			
Number Correct	112.00	113.00	108.39
Response Latency	1075.0	1087.0	1140.0
Percent Correct	95.87	97.38	96.68
Short-Term Memory			
Number Correct	80.61	80.11	78.89
Response Latency	640.0	640.0	660.0
Percent Correct	97.10	96.87	96.00
Preferred Hand Tapping			
Number of Alternate Taps	35.81	37.08	36.39
Non-Preferred Hand Tapping			
Number of Alternate Taps	34.42	34.50	33.25
Reaction Time			
Response Latency	395.0	391.0	397.0

Descriptives - Means

Test	0.05 BAL		
	A	B	C
Grammatical Reasoning			
Number Correct	40.94	37.89	37.50
Response Latency	3129	3558	3768
Percent Correct	86.20	88.48	90.68
Mathematical Processing			
Number Correct	143.61	141.28	143.61
Response Latency	.67	.68	.69
Percent Correct	96.03	95.73	96.18
Code Substitution			
Number Correct	89.28	83.33	84.39
Response Latency	1540	1646	1619
Percent Correct	98.39	97.85	97.79
Pattern Comparison			
Number Correct	111.83	113.61	111.61
Response Latency	959	937	966
Percent Correct	94.64	93.71	94.05
Manikin			
Number Correct	113.67	108.50	107.39
Response Latency	1081	1145	1140
Percent Correct	97.21	95.47	96.06
Short-Term Memory			
Number Correct	79.61	77.78	78.44
Response Latency	637	654	692
Percent Correct	96.05	94.46	96.96
Preferred Hand Tapping			
Number of Alternate Taps	38.25	36.17	36.33
Nonpreferred Hand Tapping			
Number of Alternate Taps	34.75	32.78	34.25
Reaction Time			
Response Latency	413	406	413

Descriptives - Means

Test	0.10 BAL A	B	C
Grammatical Reasoning			
Number Correct	42.18	38.71	37.06
Response Latency	3198	3484	3756
Percent Correct	90.83	88.21	90.56
Mathematical Processing			
Number Correct	143.65	143.76	144.00
Response Latency	.65	.68	.70
Percent Correct	96.00	96.26	97.31
Code Substitution			
Number Correct	91.59	80.06	83.94
Response Latency	1485	1735	1648
Percent Correct	98.86	97.24	98.70
Pattern Comparison			
Number Correct	113.53	106.00	108.29
Response Latency	945	1018	1035
Percent Correct	94.86	93.20	96.36
Manikin			
Number Correct	112.24	102.88	104.53
Response Latency	1116	1209	1203
Percent Correct	96.85	95.68	97.21
Short-Term Memory			
Number Correct	79.06	77.82	77.76
Response Latency	636	672	690
Percent Correct	95.05	95.44	96.50
Preferred Hand Tapping			
Number of Alternate Taps	36.41	34.00	36.06
Nonpreferred Hand Tapping			
Number of Alternate Taps	33.65	32.53	33.00
Reaction Time			
Response Latency	394	414	413

Descriptives - Means

Test	0.15 BAL A	0.15 BAL B	0.15 BAL C
Grammatical Reasoning			
Number Correct	40.72	35.22	37.72
Response Latency	3209	3559	3742
Percent Correct	88.05	76.65	91.11
Mathematical Processing			
Number Correct	140.94	132.28	141.22
Response Latency	.67	.73	.67
Percent Correct	95.14	91.32	94.99
Code Substitution			
Number Correct	85.28	62.33	83.72
Response Latency	1595	1999	1672
Percent Correct	97.16	87.72	98.36
Pattern Comparison			
Number Correct	113.50	100.11	110.78
Response Latency	941	1040	975
Percent Correct	92.70	89.84	93.29
Manikin			
Number Correct	111.17	87.44	106.28
Response Latency	1116	1355	1175
Percent Correct	96.15	90.18	95.89
Short-Term Memory			
Number Correct	78.50	69.22	76.50
Response Latency	640	743	671
Percent Correct	94.92	91.29	93.61
Preferred Hand Tapping			
Number of Alternate Taps	37.06	32.17	36.69
Nonpreferred Hand Tapping			
Number of Alternate Taps	34.42	22.56	34.33
Reaction Time			
Response Latency	405	485	415

Descriptives - Standard Deviations

Test	0.00 BAL		
	A	B	C
Grammatical Reasoning			
Number Correct	11.14	10.20	13.14
Response Latency	819.53	749.16	953.41
Percent Correct	6.34	8.19	4.11
Mathematical Processing			
Number Correct	4.36	5.19	3.21
Response Latency	.08	.09	.08
Percent Correct	2.57	3.27	2.06
Code Substitution			
Number Correct	15.69	16.32	18.63
Response Latency	275.06	271.08	321.74
Percent Correct	2.25	2.66	2.51
Pattern Comparison			
Number Correct	15.08	20.71	21.03
Response Latency	199.57	253.05	283.82
Percent Correct	3.94	3.16	4.46
Manikin			
Number Correct	19.18	20.74	20.21
Response Latency	233.16	247.48	281.12
Percent Correct	2.86	2.85	2.61
Short-Term Memory			
Number Correct	4.13	4.48	4.64
Response Latency	93.58	87.16	110.47
Percent Correct	3.06	2.92	3.10
Preferred Hand Tapping			
Number of Alternate Taps	7.44	7.03	5.89
Nonpreferred Hand Tapping			
Number of Alternate Taps	9.35	8.54	8.37
Reaction Time			
Response Latency	44.59	44.68	57.29

Descriptives - Standard Deviations

Test	0.05 BAL		
	A	B	C
Grammatical Reasoning			
Number Correct	13.14	14.50	15.32
Response Latency	924.81	1234.54	1283.67
Percent Correct	6.81	7.48	4.93
Mathematical Processing			
Number Correct	4.55	9.79	4.65
Response Latency	.08	.11	.09
Percent Correct	3.14	3.66	2.98
Code Substitution			
Number Correct	18.04	16.82	16.93
Response Latency	335.38	338.54	314.61
Percent Correct	1.62	1.79	2.28
Pattern Comparison			
Number Correct	14.17	16.73	17.61
Response Latency	188.74	232.51	221.96
Percent Correct	3.80	4.76	3.44
Manikin			
Number Correct	21.94	24.93	19.48
Response Latency	253.91	333.94	257.96
Percent Correct	2.30	3.49	4.02
Short-Term Memory			
Number Correct	5.16	6.30	6.24
Response Latency	87.23	117.58	141.47
Percent Correct	3.53	4.20	2.79
Preferred Hand Tapping			
Number of Alternate Taps	7.98	7.36	6.17
Nonpreferred Hand Tapping			
Number of Alternate Taps	9.30	9.40	7.89
Reaction Time			
Response Latency	50.71	47.79	64.71

Descriptives -- Standard Deviations

Test	0.10 BAL		
	A	B	C
Grammatical Reasoning			
Number Correct	14.25	15.07	14.67
Response Latency	907.26	1031.47	1167.86
Percent Correct	5.69	7.78	7.02
Mathematical Processing			
Number Correct	4.40	4.49	5.34
Response Latency	.08	.08	.09
Percent Correct	3.05	2.36	2.28
Code Substitution			
Number Correct	15.82	19.89	16.94
Response Latency	251.38	405.56	337.27
Percent Correct	1.42	2.14	1.23
Pattern Comparison			
Number Correct	15.80	17.10	17.61
Response Latency	182.31	230.51	230.66
Percent Correct	4.26	5.27	2.92
Manikin			
Number Correct	24.37	22.30	20.42
Response Latency	346.39	309.84	312.09
Percent Correct	2.54	3.49	2.92
Short-Term Memory			
Number Correct	4.96	5.50	4.78
Response Latency	103.33	106.61	116.93
Percent Correct	2.83	2.84	4.88
Preferred Hand Tapping			
Number of Alternate Taps	7.59	7.97	7.53
Nonpreferred Hand Tapping			
Number of Alternate Taps	9.07	8.04	7.56
Reaction Time			
Response Latency	40.49	62.48	61.80

Descriptives - Standard Deviations

Test	0.15 BAL		
	A	B	C
Grammatical Reasoning			
Number Correct	13.59	16.06	14.70
Response Latency	850.80	1492.43	1254.79
Percent Correct	8.34	12.40	5.62
Mathematical Processing			
Number Correct	8.09	10.28	8.86
Response Latency	0.10	0.14	0.09
Percent Correct	4.87	6.48	5.44
Code Substitution			
Number Correct	16.90	19.15	19.45
Response Latency	338.52	382.67	423.07
Percent Correct	3.15	19.18	1.73
Pattern Comparison			
Number Correct	21.02	14.25	19.94
Response Latency	292.72	242.16	311.95
Percent Correct	6.52	6.78	8.58
Manikin			
Number Correct	23.75	19.37	23.50
Response Latency	329.19	422.38	338.61
Percent Correct	4.15	12.88	6.16
Short-Term Memory			
Number Correct	5.92	15.46	6.23
Response Latency	115.41	148.68	137.77
Percent Correct	4.70	7.03	3.52
Preferred Hand Tapping			
Number of Alternate Taps	8.04	6.56	7.98
Nonpreferred Hand Tapping			
Number of Alternate Taps	9.32	31.24	7.77
Reaction Time			
Response Latency	83.94	202.84	80.82

APPENDIX B

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

Kennedy, R.S., Baltzley, D.R., Turnage, J.J., & Jones, M. B. (1989). Factor analysis and predictive validity of microcomputer-based tests. Perceptual and Motor Skills, 69, 1059-1074.

Kennedy, R.S., Baltzley, D.R., Dunlap, W.P., Wilkes, R.L., & Kuntz, L.A., (1989, May). Microcomputer-based tests for repeated-measures: Metric properties and predictive validities (EOTR-89-02). Orlando, FL: Essex Corporation.

Kennedy, R.S., Baltzley, D.R., Lane, N.E., & Jones, M.B. (1989). A strategy for modeling combat related performance decrements: Dose equivalency. Presented at the MORIMOC II Workshop: Human Behavior and Performance as Essential Ingredients in Realistic Modeling of Combat. Alexandria, VA: Center for Naval Analyses.

Kennedy, R. S., Turnage, J. J., Price, H. E., & Lane, N. E. (1989). Human issues in plant operations and management. Transactions of the 14th Biennial Conference on Reactor Operating Experience Plant Operations: The Human Element (Suppl) No. 1 to Vol. 59: ISSN:0003-018). American Nuclear Society.

Kennedy, R. S., Wilkes, R. L., & Rugotzke, G. B. (1989). Quantifying toxic effects with microbased performance testing. Presentation at the American Academy of Forensics, Las Vegas, NV.

Kennedy, R. S., Dunlap, W. P., Bandaret, L. E., Smith, M. G., & Houston, C. S. (1989). Cognitive performance deficits in a simulated climb of Mount Everest: Operation Everest II. Aviation, Space, and Environmental Medicine, 60, 99-104.

Parth, P., Dunlap, W. P., Kennedy, R. S., Lane, N. E., & Ordy, J. M. (1989). Motor and cognitive testing of bone marrow transplant patients after chemo-radiotherapy. Perceptual and Motor Skills, 68, 1227-1241.

Kennedy, R. S., Baltzley, D. R., Wilkes, R. L., & Kuntz, L. A. (1989). Psychology of computer use: IX. A menu of self-administered microcomputer-based neurotoxicology tests. Perceptual and Motor Skills, 68, 1225-1272.

Baltzley, D. R., Kennedy, R. S., & Turnage, J. J. (1989). Assessing fitness-for-duty: An alternative to problems associated with drug testing in the workplace. Paper presented at the 1989 Annual Meeting of the Human Factors Society, Denver, CO: Human Factors Society.

*Kennedy, R. S., Dunlap, W. P., & Kuntz, L. A. (1989). Application of a portable automated performance test battery for the study of drugs and driving performance [Abstract]. Second International Symposium on Medicinal Drugs and Driving Performance.

Kennedy, R. S., Wilkes, R. L., & Rugotzke, G. G. (1989). Cognitive performance deficit regressed on alcohol dosage. Paper presented at the 11th International Conference on Alcohol, Drugs, and Traffic Safety. Chicago, IL.

Rugotzke, G. G., Wilkes, R. L., & Kennedy, R. S. (1989). Reliability and validity of blood alcohol concentration related to measured performance decrement. Paper presented at the 11th International Conference on Alcohol, Drugs, and Traffic Safety. Chicago, IL.

Kennedy, R. S., Turnage, J. J., & Dunlap, W. P. (1989). Screening for performance deficits in the military by microcomputerized testing. Paper presented at the 31st Annual Meeting of the Military Testing Association, San Antonio, TX.

Banderet, L. E., Shukitt, B. L., Walthers, M. A., Kennedy, R. S., Bittner, A. & Kay, G. G. (1988, November). Psychometric properties of three addition task with different response requirements. Paper presented at and published in the Proceedings 30th Annual Meeting Military Testing Association, Arlington, VA.

Dunlap, W. P., Kennedy, R. S., Harbeson, M. M., & Fowlkes, J. E. (1988). Difficulties with individual difference measures based upon some componential cognitive paradigms. Manuscript submitted for publication.

Fowlkes, J. E., Kennedy, R. S., Dunlap, W. P., & Harbeson, M. M. (1988, October). A paradigm for the identification of independent cognitive constructs. Paper presented at the 32nd Annual Meeting of the Human Factors Society, Anaheim, CA.

Jones, M. B., Kennedy, R. S., & Baltzley, D. R. (1988). Factor analysis of a microcomputer-based performance battery and its utility in predicting the Wonderlic Personnel Test (EOTR-88-8). Orlando, FL: Essex Corporation.

Kennedy, R.S., & Baltzley, D.R. (1988, June). Factor analysis of a microcomputer-based performance battery and its utility in predicting the Wonderlic personnel test. EOTR 88-88, Orlando, FL: Essex Corporation.

Kennedy, R. S., Baltzley, D. R., & Osteen, M. K. (1988). A microcomputer test battery: Normative data and sensitivity to military stressors. Paper presented at the 30th Annual Military Testing Association Conference, Arlington, VA.

Kennedy, R. S., Baltzley, D. R., Osteen, M. K., & Turnage, J. J. (1988, October). A differential approach to microcomputer test battery development and implementation. Paper presented at the 32nd Annual Meeting of the Human Factors Society, Anaheim, CA.

Kennedy, R.S., Berbaum, K.S., Collyer, S.C., May J.G., & Dunlap, W.P. (1988). Spatial requirements for visual simulation of aircraft at real-world distances. Human Factors, 30(2), 153-161.

Kennedy, R. S., Jones, M. B., & Baltzley, D. R. (1988, February). Empirical demonstration of Isoperformance methodology preparatory to development of an interactive expert computerized decision aid. Final Report submitted to Army Research Institute, Washington, DC.

Kennedy, R.S., Jones, M.B., & Baltzley, D.R. (1988, July). Optimal solutions for complex design problems: Using isoperformance software for human factors trade-offs. Paper presented at Space Operations Automation and Robotics Workshop: Space Application of Artificial Intelligence, Human Factors, and Robotics. Dayton, OH.

Kennedy, R. S., Turnage, J. J., & Lane, N. E. (1988, June). Application of a portable microcomputer mental acuity battery for fitness-for-duty assessment in power plant operations. Paper presented at the IEEE 4th Conference on Human Factors and Power Plants, Monterey, CA.

Kennedy, R. S., Turnage, J. J., & Lane, N. E. (1988). Assessment of fitness-for-duty in power plant operations by a portable microcomputer mental acuity battery. Transactions of the American Nuclear Society, 56, 521-522. (ISSN:000-018X)

Kennedy, R.S., Wilkes, R.L., Kuntz, L.A., & Baltzley, D.R. (1988, November). A menu of self-administered microcomputer-based neurotoxicology tests (EOTR-99-10). Orlando, FL: Essesx Corporation.

Lane, N. E., & Kennedy, R. S. (Eds.). (1988, May). Users manual for the U.S. Army Aeromedical Research Laboratory Portable Performance Assessment Battery (EOTR 88-5). Ft. Rucker, AL: U.S. Army Aeromedical Laboratory.

Turnage, J.J., Kennedy, R.S., Gilson, R.D., Bliss, J.P., & Nolan, M.D. (1988, December 12). The use of surrogate measurment for the prediction of flight training performances. Orlando, FL: Essex Corporation.

Turnage, J. J., Kennedy, R. S., Osteen, M. K., & Tabler, R. E. (1988). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations: Study 3. Orlando, FL: Essex Corporation.

Wilkes, R. L., Kuntz, L. A., Kennedy, R. S., & Tabler, R. E. (1988). Stability and reliability of a menu of performance tests self-administered on a portable microcomputer (NASA Contract NAS9-17326). Houston, TX: NASA Johnson Space Center.

Williams, M., Kennedy, R. S., Baltzley, D. R., May, J. G., & Dunlap, W. P. (1988). Reliability, stability, and cross-task correlations of six visual temporal factor tests. Essex Orlando Technical Report. Orlando, FL: Essex Corporation.

Jones, M. B., Kennedy, R. S., Kuntz, L. A., & Baltzley, D. R. (1987, October). Isoperformance: Trading off selection, training, and equipment variations to maintain the same level of systems performance. Proceedings of the 31st Annual Meeting of the Human Factors Society (pp. 634-637). Santa Monica, CA: Human Factors Society.

Turnage, J. J., & Lane, N. E. (1987, October). The use of surrogate techniques for the measurement of team performance. Proceedings of the 31st Annual Meeting of the Human Factors Society (pp. 638-642). Santa Monica, CA: Human Factors Society.

Wilkes, R. L., Kuntz, L. A., & Kennedy, R. S. (1987, October). Development of a menu of performance tests self-administered on a portable microcomputer. NASA Technical Report under Contract NASA 9-17326. Houston, TX: NASA.

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., & Kuntz, L. A. (1987, October). Microbased repeated-measures performance testing and general intelligence. Paper presented at the 29th Annual Conference of the Military Testing Association, Ottawa, Ontario, Canada.

Tabler, R. E., Turnage, J. J., & Kennedy, R. S. (1987, September). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations (DAMD 17-85-C-5095). U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL.

Turnage, J. J., Kennedy, R. S., & Osteen, M. K. (1987, September). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations (DAMD 17-85-C-5095). U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL.

Kennedy, R. S., Lane, N. E., & Kuntz, L. A. (1987, August). Surrogate measures: A proposed alternative in human factors assessment of operational measures of performance. Proceedings of the 1st Annual Workshop on Space Operations, Automation, & Robotics (pp. 551-558). Houston, TX: Lyndon B. Johnson Space Center.

Fowlkes, J. E., Kennedy, R. S., & Hennessy, R. T. (1987, August). Relevance of visual accommodation for performance in spacecraft (NASA Contract NAS9-17745, Final Report). Houston, TX: NASA Lyndon B. Johnson Space Center.

Jones, M. B., Kennedy, R. S., & Kuntz, L. A. (1987, August). Isoperformance: Integrating personnel and training factors into equipment design. Paper presented at the 2nd International Conference on Human-Computer Interaction, Honolulu, HI.

Hettinger, L. J., Berbaum, K. S., Kennedy, R. S., & Westra, D. P. (1987, June 23-26). Human performance issues in the evaluation of a helmet-mounted area-of-interest projector. Paper presented at the IMAGE IV Conference, Phoenix, AZ.

Kennedy, R.S, Baltzley, D.R., Lilienthal, M.G., Allgood, G.O., & Gower, D.W. (1987). Consistency across measures of simulator sickness: Implications for a biocybernetic safety reporting device. Paper presented at the 25th Annual SAFE Symposium.

Kennedy, R. S., Berbaum, K. S., Williams, M. C., Brannan, J., & Welch, R. B. (1987). Transfer of perceptual-motor training and the space adaptation syndrome. Aviation, Space, and Environmental Medicine, 58(9,Suppl.), A29-A33.

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., & Kuntz, L. A. (1987). Development of an Automated Performance Test System for environmental and behavioral toxicology studies. Perceptual and Motor Skills, 65, 947-962.

Dunlap, W. P., Bittner, A. C., Jr., Jones, M. B., & Kennedy, R. S. (1986, November). Factor analysis of composite scores from the Armed Services Vocational Aptitude Battery (ASVAB). Proceedings of the 28th Annual Military Testing Association Conference (pp. 218-224). Mystic, CT: U.S. Coast Guard Academy.

Kennedy, R. S., Lane, N. E., Wilkes, R. L., & Banderet, L. E. (1986, November). Development of behavioral assessment protocols for varied repeated-measures testing paradigms. Proceedings of the 28th Annual Military Testing Association Conference (pp. 568-573). Mystic, CT: U.S. Coast Guard Academy.

Dunlap, W. P., Jones, M. B., Kemery, E. R., & Kennedy, R. S. (1986, November). Optimizing a test battery by varying subtest times. Proceedings of the 28th Annual Conference of the Military Testing Association (pp. 225-230). Mystic, CT: U.S. Coast Guard Academy.

Kennedy, R. S. (1986, October). A menu of performance tests implemented on a portable microcomputer. Proceedings of the 1st Joint IUTOX-IST Symposium on Behavioral Toxicology (to appear in the Journal of Neurobehavioral Toxicology and Teratology), Bari, ITALY.

Jones, M. B., Kennedy, R. S., & Turnage, J. J. (1986, September). Isoperformance: A methodology for human factors engineering design. Proceedings of the 30th Annual Meeting of the Human Factors Society, Dayton, OH.

Kennedy, R. S., & Kuntz, L. A. (1986). Self-monitoring of subjective status during extended operations using an Automated Performance Test Battery. Paper presented at the 37th International Astronautical Congress, Innsbruck, Austria.

Kennedy, R. S., Dunlap, W. P., & Kuntz, L. A. (1986). Application of a portable automated performance test battery for the study of drugs and driving performance. Paper presented at the 2nd International Symposium on Medicinal Drugs and Driving Performance.

*Kennedy, R. S., Wilkes, R. L., & Kuntz, L. A. (1986). Sensitivity of a notebook-sized Portable Automated Performance Test System. Paper presented at the Annual Behavioral Toxicology Society Meeting, Atlanta, GA.

Lane, N. E., Kennedy, R. S., & Jones, M. B. (1986). Overcoming unreliability in operational measures: The use of surrogate measure systems. Proceedings of the 30th Annual Meeting of the Human Factors Society (pp. 1398-1402). Dayton, OH: Human Factors Society.

Thomley, K.E., Kennedy, R.S., & Bittner, A.C. (1986). Development of postural equilbrium tests for examining environmental effects. Perceptual and Motor Skills, 63, 555-564.

Wilkes, R. L., Kennedy, R. S., Dunlap, W. P., & Jones, M. B. (1986). Development of a portable human assessment battery for environmental and behavioral toxicology studies. Unpublished manuscript.

Wilkes, R. L., Kennedy, R. S., Dunlap, W. P., & Lane, N. E. (1986). Stability, reliability, and cross-mode correlations of tests in a recommended 8-minute performance assessment battery (TR No. EOTR 86-4 for NASA Contract No. NAS9-17326). Essex Corporation, Orlando, FL.

Bittner, A. C., Harbeson, M. M., Kennedy, R. S., & Lundy, N. C. (1985, October). Assessing the human performance envelope: A brief guide. Paper presented at the Aerospace Technology Conference & Exposition, Long Beach, CA.

Kennedy, R.S., Dunlap, W.P., Jones, M.B., Wilkes, R.L., & Bittner, A.C. (1985, October). A automated portable test system (APTS): A performance envelope assessment tool (SAE Technical Paper Series). Aerospace Technology Conference & Exposition, Long Beach, CA.

Bittner, A. C., Jr., Smith, M. G., Kennedy, R. S., Staley, C. F., & Harbeson, M. M. (1985). Automated portable test system (APTS): Overview and prospects. Behavior Research Methods, Instruments and Computers, 17, 217-221.

Johnson, J.H., & Kennedy, R.S. (1985, August). Literature review and critique of methods to assess human performance in dynamic vehicle/operator settings (Task 1) (Final Report). Orlando, FL: Essex Corporation.

Johnson, J. H., Kennedy, R. S., Smith, M. G., & Dutton, B. (1985). On the use of portable microprocessors as field data collection units. Proceedings of the Annual Scientific Meeting of the Aerospace Medical Association, San Antonio, TX.

Kennedy, R. S. (1985). A portable battery for objective, nonobtrusive measures of human performance. Proceedings of the Workshop on Advances in NASA-relevant, minimally invasive instrumentation (pp. 4.17-4.30). Pacific Grove, CA.

Kennedy, R. S. (1985). What are the advantages of self monitoring by the Automated Performance Test System (APTS)? Paper presented at the 38th International Air Safety Seminar, Boston, MA.

Kennedy, R. S., Dunlap, W. P., Jones, M. B., Lane, N. E., & Wilkes, R. L. (1985). Portable human assessment battery: Stability, reliability, factor structure and correlation with tests of intelligence (Final Report NSF/BNS 85001; also EOTR 85-3). Washington, DC: National Science Foundation. (NTIS PB88-116645/A03)

Kennedy, R. S., Dunlap, W. P., Wilkes, R. L., & Lane, N. E. (1985). Development of a portable computerized performance test system. Proceedings of the 27th Annual Conference of the Military Testing Association (pp. 107-112). San Diego, CA: Navy Personnel Research & Development Center.

Kennedy, R. S., Jones, M. B., Dunlap, W. P., Wilkes, R. L., & Bittner, A. C., Jr. (1985). Automated Portable Test System (APTS): A performance assessment tool. SAE Technical Paper Series (Report No. 81775). Warrendale, PA: Society of Automotive Engineers.

Kennedy, R.S., Wilkes, R.L., Lane, N.E. (1985). Preliminary evaluation of a microbased repeated measures testing system. (EOTR-85-1), Orlando, FL: Essex Corporation.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L. (1985). Preliminary evaluation of a microbased repeated measures testing system (EOTR 85-1; NASA CR-172038). Washington, DC: National Aeronautics and Space Administration.

Merkle, P. J., Kennedy, R. S., Smith, M. G., & Johnson, J. H. (1985). Microprocessor based field testing for human performance assessment. Proceedings of the 27th Annual Military Testing Association Conference (pp. 398-403). San Diego, CA: Navy Personnel Research & Development Center.

Banderet, L. E., Benson, K. P., McDougall, D. M., Kennedy, R. S., & Smith, M. (1984). Development of cognitive tests for repeated performance assessment. Proceedings of the 26th MTA Conference (pp. 375-380), Munich, Germany.

*Kennedy, R. S., Bittner, A. C., Jr., Smith, M. G., & Harbeson, M. M. (1984). Development of a portable performance assessment system for behavioral toxicology. Paper presented at the Behavioral Toxicology Society Meeting, Toronto, Canada.

Lintern, G., & Kennedy, R. S. (1984). A video game as a covariate for carrier landing research. Perceptual & Motor Skills, 58, 167-172.

McCormick, B. K., Dunlap, W. P., Kennedy, R. S., & Jones, M. B. (1983). The effects of practice on the Armed Services Vocational Aptitude Battery (Rep. No. TR-602). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (NTIS No. AD A148314)

Smith, M. G., Krause, M., Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. (1983). Performance testing with microprocessors--Mechanization is not implementation. Proceedings of the 27th Annual Meeting of the Human Factors Society (pp. 674-678). Norfolk, VA: Human Factors Society.

Harbeson, M. M., Kennedy, R. S., Krause, M., & Bittner, A. C., Jr. (1982). Repeated measures of information processing. Proceedings of the 26th Annual Meeting of the Human Factors Society (pp. 818-822). Seattle, WA: Human Factors Society.

Harbeson, M. M., Krause, M., Kennedy, R. S., & Bittner, A. C., Jr. (1982). The Stroop as a performance evaluation test for environmental research. Journal of Psychology, 111, 223-233.

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B. (1982). Television computer games: A "new look" in performance testing. Aviation, Space, & Environmental Medicine, 53, 49-53.

Bittner, A. C., Jr., Jones, M. B., Carter, R. C., Shannon, R. H., Chatfield, D. C., & Kennedy, R. S. (1981). Statistical issues in performance testing: Collected papers (NBDL 81R010). New Orleans, LA: Naval Biodynamics Laboratory. (NTIS No. AD A111086)

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. (1981). Grammatical reasoning: A stable performance yardstick. Human Factors, 23, 587-591.

Damos, D. L., Bittner, A. C., Jr., Kennedy, R. S., & Harbeson, M. M. (1981). The effects of extended practice on dual-task training. Human Factors, 23, 627-631.

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

*Bittner, A. C., Jr., Kennedy, R. S., & Harbeson, M. M. (1980). Apparatus testing for aviation performance assessment and selection: A technology ready to come of age. Paper presented at the 28th International Congress of Aviation and Space Medicine, Montreal, Quebec, Canada.

Bittner, A. C., Jr., Kennedy, R. S., & McCauley, M. E. (1980). Time estimation: Repeated-measures testing and drug effects. Proceedings of the 7th Psychology in the DoD Symposium (pp. 445-459). Colorado Springs, CO: USAF Academy.

*Harbeson, M. M., Kennedy, R. S., & Bittner, A. C., Jr. (1980). Application of memory tests for assessment of the effects of exotic environments on humans. Paper presented at the 72nd Annual Meeting of the Southern Society for Philosophy & Psychology, Birmingham, AL.

Harbeson, M. M., Krause, M., & Kennedy, R. S. (1980). The comparison of memory tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 349-353). Los Angeles, CA.

Kennedy, R. S., Jones, M. B., & Harbeson, M. M. (1980). Assessing productivity and well-being in Navy workplaces. Proceedings of the 13th Annual Meeting of the Human Factors Association of Canada (pp. 108-113). Point Ideal, Ontario, Canada.

Harbeson, M. M., Kennedy, R. S., & Bittner, A. C., Jr. (1979). A comparison of the Stroop Test to other tasks for studies of environmental stress. Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada, Bracebridge, Ontario, Canada.

Kennedy, R. S., & Bittner, A. C., Jr. (1978). The stability of complex human performance for extended periods: Application for studies of environmental stress. Proceedings of the 49th Annual Meeting of the Aerospace Medical Association, New Orleans, LA.

MANUSCRIPTS IN PREPARATION

Kennedy, R. S., Dunlap, W. P., & Wilkes, R. L. (1990). Relations between global measures of intelligence and performance tests as functions of practice. Manuscript in preparation.

Kennedy, R. S., Odenheimer, R. C., Baltzley, D. R., Dunlap, W. P., & Wood, C. D. (1989). Differential effects of scopolamine and amphetamine on micro-computer-based performance tests. Submitted for publication to Aviation, Space, and Environmental Medicine.

Kennedy, R. S., Lane, N. E., & Baltzley, D. R. (1989). Assessment of vision measures for repeated-measures applications: The VISTECH contrast sensitivity test. Submitted for publication to Journal of Optometry and Vision Science.

Kennedy, R. S., & Turnage, J. J. (1988). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations. Orlando, FL: Essex Corporation.