Naval Biodynamics Laboratory
NBDL-82R014

PROCEEDINGS OF THE 26th ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY SEATTLE, WASHINGTON, 25 - 29 OCTOBER 1982

REPEATED MEASURES OF INFORMATION PROCESSING

Mary M. Harbeson, Michele Krause and Alvah Bittner, Jr.
Naval Biodynamics Laboratory

R.S. Kennedy
Canyon Research Group

Naval Biodynamics Laboratory
P.O. Box 29407
New Orleans, LA 70189-0407

Prepared for

Naval Medical Research and Development Command
Bethesda, MD 20889-5044

PROCEEDINGS OF THE 26TH ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY
SEATTLE, WASHINGTON, 25 - 29 OCTOBER 1982

REPEATED MEASURES OF INFORMATION PROCESSING

Mary M. Harbeson
Naval Biodynamics Laboratory
New Orleans, LA

Robert S. Kennedy
Canyon Research Group
Orlando, FL

Michele Krause and Alvah C. Bittner Jr.
Naval Biodynamics Laboratory
New Orleans, LA

ABSTRACT

Two information processing tasks were considered for inclusion in a test battery which was being developed for repeated measures investigations of adverse environmental effects. The tests, adapted from the Rose Battery were: Baron's Graphemic and Phonemic Analysis, and Posner's Letter Classification. Alternate forms of the the tests were administered on 15 consecutive workdays (Monday to Friday) to 21 Navy enlisted men. A total of 20 measures were taken. These scores were examined to determine when in practice they obtained unchanging or linearly changing means, homogeneous variances, and constant (differentially stable) intertrial correlations of an acceptably high level (task definition). In general, correlations of the basic measures tended to become stable with sufficient practice, but derived measures such as difference, slope, and ratio scores did not attain stability. The Baron and Posner tasks had high reliabilities and were highly correlated with each other. Preliminary analysis indicated that both tests may be measuring the same thing. Either the Baron Sense/Nonsense or the Posner Name score may be recommended for repeated measures experimentation.

INTRODUCTION

The integrity of mental functioning, including the ability to process information, is of prime interest in the assessment of human performance. Therefore, information processing tests were considered for inclusion in the Performance Evaluation Tests for Environmental Research (PETER) battery which was being designed to assess the effects of environmental stress (Kennedy, Bittner, Harbeson, & Jones, 1981). In a series of studies, Rose and cohorts, examined information processing tasks for the purpose of assembling them into a battery to assess individual differences. (Rose 1974; 1978; Rose & Fernandes, 1977; Fernandes & Rose, 1978). In developing the PETER battery, these tasks were borrowed freely with the only intended modification being to extend the number of replications. Rose provided guidance in the extension of his battery for possible use in PETER.

Because many administrations are routine in environmental stress studies, a repeated measures paradigm was adopted. Subjects were tested for approximately 15 minutes per day on each test for 15 consecutive work days. This paradigm entailed a sufficient number of trials to permit means to asymptote or become linearly regular, and provided sufficient data for provocative tests of the stability of variances and correlations (Bittner & Carter, 1981). Although these requirement are generally recognized for analysis and interpretation of repeated measures experimentation (Campbell & Stanley, 1963; Winer, 1971) the authors know of no other performance battery which has been standardized according to these criteria.

Previous reports have documented our experiences with various other tasks from Rose's battery: Stroop (Harbeson, Krause, Kennedy & Bittner, 1982); Grammatical Reasoning (Carter, Kennedy, & Bittner, 1981); Letter Search, and Critical Tracking (Kennedy et al., Nov. 1981). The present report is concerned with two additional tasks from the eight in Rose and Fernandes (1977): Graphemic and Phonemic Analysis (Baron 1973; Baron & McKillop, 1975); and Letter Classification (Posner, & Mitchell, 1967).

These tests were selected because they were purported to measure different information processing constructs. Three other tests were administered at the same time. Lexical Decision Making and Semantic Memory Retrieval will be included in a future report (Harbeson, & Kennedy, in preparation). Short-Term Memory Scanning was reported earlier (Kennedy, Bittner, Carter, Krause, Harbeson, McCafferty, Pepper, & Wiker, 1981). The remaining tests in Rose and Fernandes were not used because they either had low test-retest reliability or they resembled tests which had already been studied (Kennedy et al., July 1981). The purpose of this investigation was to evaluate the suitability of information processing tasks for inclusion in a battery of performance tasks. Evaluations were aimed at statistical suitability of individual measures, and their uniqueness and economy of use. Altogether, the purpose was to provide a basis for including tasks in the Performance Evaluation Tests for Environmental Research (PETER) Battery.

METHOD

Task Descriptions

Graphemic and Phonemic Analysis. This task was developed by Baron to study visual (graphemic encoding) versus articulatory (auditory encoding) reading strategies. Subjects were required to judge whether phrases made sense or not under three conditions: Sense (our new car), Homophone (its knot so), or Nonsense (a drop of ran). These were combined in pairs to form three basic conditions, Sense/Non-

sense (SN), Sense/Homophone (SH), and Homophone /Nonsense (HN). Theoretically, graphemic encoders would do better on S phrases and acoustic encoders would do better on H phrases. But, since graphemic encoding is faster with normal readers, and is more common, it would be expected that response times would be least for SN, and greatest for HN. There were 20 phrases in each condition, and the interstimulus interval was approximately 4 seconds. Following Rose and Fernandes (1977), twelve variables were recorded: response times for each of the phrases as a function of condition (6); ratio of SH time to HN; response time for each of the three conditions; percent of errors; and mean error time across conditions.

Letter Classification. Posner and Mitchell (1967) used this task to study matching or recognition of stimuli of various levels of complexity. Subjects were to make same or different judgments on pairs of letters based on three criteria. Letters were classified by physical appearance (AA vs. AB), name identity (Aa vs. Ab), or category (both vowels or consonants such as AE or BC , vs. not matched, such as AB). There were 36 trials per day in each of the first two conditions and 32 in the third. The interstimulus interval was approximately 4 seconds. Eight scores were calculated including response times for each condition for same judgments, response times for all different judgments, two difference scores, percent errors and mean error time.

Subjects

The subjects were 21 Navy enlisted men between the ages of 18 and 24 who had volunteered for duty at the Naval Biodynamics Laboratory. One subject was dropped from the analysis in Graphemic and Phonemic analysis because his daily score sheet was lost. All subjects were recruited, evaluated and employed in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 Series and Bureau of Medicine and Surgery Instruction 3900.6. These instructions are based upon voluntary consent, and meet the provisions of prevailing national and international guidelines. For a detailed description of the subject selection procedure, signal see Thomas, Majewski, Ewing, and Gilbert (1978).

Apparatus and Procedure

The stimulus material was presented by means of black and white slides shown on a Kodak Ektograph 450 Audio Viewer ®. The rate of presentation was controlled by preprogrammed tape cassettes. Each trial was preceeded by a cueing signal of two clicks. Subjects responded by pushing one of two buttons (yes or no) on boxes which were fastened to their desk tops. The response time was measured from the onset of the stimuli to the time the subject pushed his answer button. The answers and the response times were displayed on an automatic timing device and recorded on an answer sheet by the experimenter. The subjects were tested in groups of four beginning at 8:00 AM for 15 consecutive work-days. The five tests were administered in the same order to each group of subjects, but the order was varied across groups and days. There was a break of 2 or 3 minutes between tests while the experimenter changed carousels and cassette tapes, and a five minute break between tests halfway through testing. Total testing time was approximately an hour and a half including breaks.

RESULTS

Analysis

Means, standard deviations, and cross-session correlations were calculated for each measure. In an intitial sensitivity screening, measures with correlations which averaged below .50, or which were obviously unstable, were dropped from further analysis. The reliability and the stability of the correlations of the remaining measures were determined by a general computer program developed by Steiger (1980) To avoid problems with last day effects, the 15th day was dropped from the stability analysis (cf., Carter et al., 1981). Fmax (Winer, 1971) was used to test for the homogeneity of the variances. Graphical Analysis was employed to examine the stability of the means. Within each task those measures which achieved stability were compared. Correlations were calculated between average scores for each subject over stable days on each measure.[1] These values were adjusted using the Spearman-Brown Prophecy Formula and the correction for attenuation to estimate the between measure correlations. A similar procedure was followed to compare measures across tasks.

Graphemic and Phonemic Analysis

Preliminary analysis was done on 12 scores. Three of these were dropped from further analysis, including the SH/HN Ratio which had a reliability of essentially zero. Table 1 shows the results of the differential stability analysis for the remaining 9 scores. It can be seen that all but H(S) attained stability, H(N) and HN did not become stable until rather late in practice. Fmax tests were nonsignificant for all measures. Figure 1 shows the means of SN, SH, and HN. All appear to become stable by about Day 6. As was predicted in the literature the HN phrases required the most time and the SN phrases the least. However, SN, and SH times became more alike with practice. Correlations between all of the measures were quite high. Those for SN, SH, and HN are shown in Table 2.

TABLE 1. Graphemic and Phonemic Analysis: Differential Stability Analysis

Score	Stable Days	r	χ^2	df	p
S(N)	3-14	.76	78.90	65	.12
N(S)	3-14	.79	61.33	65	.61
SN	3-14	.84	78.46	65	.12
S(H)	4-14	.78	61.58	54	.22
H(S)	----	NOT	STABLE	--------	
SH	6-14	.84	45.13	35	.12
H(N)	12-14	.90	1.57	2	.46
N(H)	4-14	.73	67.09	54	.11
HN	10-14	.88	14.62	9	.10

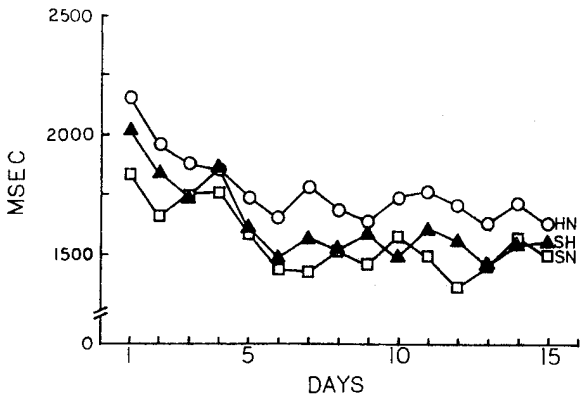

FIGURE 1: Baron, Task; Mean Response Times on SN, SH, and HN over 15 days (N=20).

TABLE 2. Graphemic and Phonemic Analysis: Differentially Stabilized Correlations*

Score	SN	SH	HN
SN	(.85)	.84	.85
SH	1.00	(.84)	.86
HN	.99	1.00	(.88)

Letter Classification

Of the 8 Posner scores, only the three basic measures qualified for further analysis. The results of the differential stability analysis are shown in Table 3. Considerable practice was required, but all conditions eventually became stable and reliabilities were very respectable. Again, Fmax tests were non-significant for all measures. Examining Figure 2, it can be seen that the means appear almost level for the Name and Physical conditions from about Day 2, and certainly for all conditions by at least Day 6. The relationship between the means is exactly as predicted in the literature. As can be seen in Table 4, the three measures were highly correlated.

TABLE 3. Posner Task: Differential Stability Analysis

Score	Stable Days	$\bar{r}$	χ^2	df	p
Physical	10-14	.81	10.21	9	.33
Name	8-14	.83	25.11	20	.20
Category	12-14	.89	3.61	2	.16

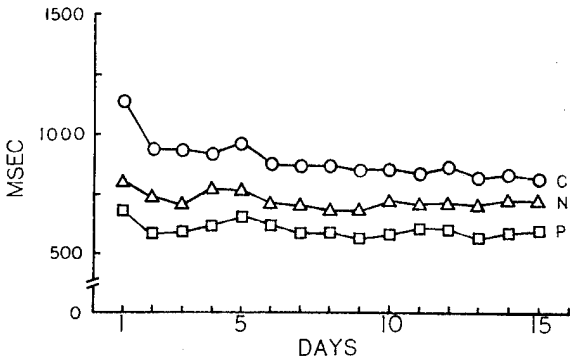

FIGURE 2: Posner Task mean response times on Physical, Name, and Category over 15 days (N=21).

TABLE 4: Letter Classification: Differentially Stabilized Correlations*

Score	Physical	Name	Category
Physical	(.81)	.82	.76
Name	.99	(.83)	.80
Category	.90	.94	(.89)

*Correlations above, reliabilities along, and corrected-for-attenuation estimates below the diagonal.

Comparison of Tasks

Table 5 shows the intercorrelations of the three basic measures from each task across differentially stabilized days. There appears to be a large overlap between the tasks. There is no notable difference between the correlations of any of the Baron scores. The distinct pattern in the correlations of the Posner measures might be explained by the parallel trend in their internal reliabilities (see Table 4). Since the Posner scores all correlate highly with each other it is most likely that the differences seen in Table 5 are the result of error variance rather than a difference in what is being measured.

TABLE 5. Baron and Posner Task Intercorrelation*

Score	Physical	Name	Category
SN	.58 (.70)	.66 (.78)	.75 (.86)
SH	.57 (.69)	.66 (.79)	.76 (.88)
HN	.53 (.63)	.62 (.73)	.74 (.84)

*Corrected for attenuation values in parenthesis

DISCUSSION

Comparison with Previous Research

The results of the present study are consistant with past research. Some of the means obtained in the present study are higher than those obtained previously (i.e. Rose & Fernandes 1977), but not dramatically so. Subject differences are not surprising considering that the other studies used college students and the present study employed enlisted men. The important point is that the same patterns were obtained in all of the studies.

Comparison of Tasks and Subtasks

Two main findings were revealed in the comparison of measures within tasks. The first was that all of the basic scores in each test appeared to be measuring the same thing. This phenomenon has been noted in other studies in this program (Harbeson, et al., 1982), but cases in which subtasks do not converge, even after considerable practice are also in evidence (Bittner, Lundy, Kennedy, & Harbeson, 1982). The results also indicate that there is a large overlap between the Baron and Posner tasks. Some overlap is to be expected, as many of the same information processing operations are involved (Carroll, 1974). However, it is possible that both tasks are measuring the same underlying ability. When tasks or subtasks are used to measure different constructs or abilities, it is important to examine the differerential relationships. In this study all of the scores may be measuring different levels of the same thing. Further research using factor analysis is needed to reach a more definitive conclusion and will be included in a future report (Harbeson & Kennedy, in preparation).

The second finding was that derived measures such as slope, difference scores, and ratio scores were either unstable or had extremely low internal reliability. This has been observed in previous reports from this laboratory (e.g. Carter & Krause, 1982; Kennedy et al., July 1981). The lack of reliability of derived measures has also been commented on by other authors (Cronbach & Furby, 1970).

Some Practical Considerations

The methodology used in this study was obviously adequate as stable reliable measures were obtained. However, the administration time was rather long as compared to the actual time the subjects were exposed to the stimulus material. An interactive computer implementation would probably be more efficient, eliminating the time taken to record responses and change carousels and cassettes. This would also decrease subject fatigue. It would be helpful to have built in time limits to reduce problems with outliers. It might also be possible to produce numerous alternate forms in a manner similar to that used by Carter and Sbisa (1982). With a few changes in procedure, these tests could be administered in a paper and pencil format to groups or individuals for use in environments where other equipment would be impractical. Further standardization studies would be required with any changes in procedure.

Conclusions

Any of the Baron or Posner basic measures would be suitable for repeated measures testing. Since they appear to be redundant, at least within tests, one from each task would be sufficient. Baron SN and Posner Name have the best psychometric qualities. Alternate forms are easier to construct for the Posner task. Future research is need to compare the two tasks and to determine the qualities of single composite measures. At present, the Baron SN measure and the Posner Name measure may be recommended for inclusion in a repeated measures test battery.

ACKNOWLEDGEMENTS

This work was funded by the Naval Medical Research and Development Command and was performed under Navy Work Unit No. MF58.524-002-5027. The authors would like to acknowledge the contributions of Harvey Sbisa, Susan Jones, Michael Shewmake, Debra Andrews, and Patrina Garrity.

FOOTNOTE

[1]Correlating averages was more convenient to use with this large data set than averaging correlations which is a more statistically efficient technique (see Bittner, Dunlap, & Jones (1982).

REFERENCES

Baron, J. Phonemic stage not necessary for reading. Quarterly Journal of Experimental Psychology, 1973, 25, 241-246.

Baron J., & McKillop, B.J. Individual differences in speed of phonimic analysis, visual analysis, and reading. Acta Psychologica, 1975, 39, 91-96.

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

Bittner, A.C., Jr., Dunlap, W.P., & Jones, M.B. Averaged cross-correlations with differentially-stable variables: Fewer subjects required for repeated measures. Proceedings of the 26th Annual Meeting of the Human Factors Society, Seattle, WA, Oct. 1982.

Bittner, A.C., Jr., Lundy, N.C., Kennedy, R.S., & Harbeson M.M. Performance Evaluation Tests for Environmental Research (PETER): Spoke Tasks. Perceptual and Motor Skills, 1982, 54, 1319-1331.

Campbell, D.T., & Stanley, J.C. Experimental and quasi- experimental designs for research. Chicago: Rand McNally, 1963.

Carroll, J.B. Psychometric tests as cognitive tasks: A new "structure of intellect." Princeton, New Jersey: Educational Testing Service, May 1974.

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

Carter, R.C., Krause, M. Reliability of human information processing: An indictment of the slope score. Manuscript submitted for publication, 1982. (Available from the authors, Naval Biodynamics Laboratory, New Orleans)

Carter, R.C. & Sbisa, H.E. Human Performance tests for repeated measurements: Alternate forms of eight tests by computer. (Research Report No. NBDL-82R003) New Orleans: Naval Biodynamics Laboratory, Jan. 1982. (NTIS No AD A115021).

Cronbach, L. J. & Furby, L. How should we measrure change - or should we? Psychological Bulletin , 1970, 74 , 68 -80.

Fernandes, K. & Rose, A.M. An information processing approach to performance assessment II. An investigation of encoding and retrieval processes in memory. Washington, D.C.: American Institutes for Research, 1978.

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

Harbeson, M.M., & Kennedy, R.S. Repeated measures of information processing: Comparison of four tasks. (Research Report No. 82R014) New Orleans: Naval Biodynamics Laboratory, in preparation.

Kennedy,R.S., Bittner, A.C., Jr., Carter, R.C., Krause, M., Harbeson, M.M., McCafferty, D.B., Pepper, R.L., & Wiker, S.F. Performance Evaluation Tests for Environmental Research (PETER): Collected papers. (Research Report NO. NBDL-80R008) New Orleans, LA: Naval Biodynamics Laboratory, July 1981. (Ntis No. AD A111296)

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

Posner, M.I. & Mitchell, R.F. Chronometric analysis of classification. Psychological Review, 1967, 74, 392-409.

Rose, A.M. Human information processing: An assessment and research battery. (Technical Report No. 46) Ann Arbor: University of Michigan Human Performance Center 1974. (NTIS No. AD 785411)

Rose, A.M. An information processing approach to performance assessment (Report No. AIR 58500-11/78-FR) Washington, D.C.: American Institutes for Research, November 1978.

Rose, A.M., & Fernandes, K. An information processing approacth to performance Assessment: I Experimental investigation of an information processing performance battery. (Report No. AIR-58500-TR) Washington, D.C.: American Institutes for Research, November 1977. (NTIS No. AD A047299)

Steiger, J.H. Tests for comparing elements of a correlation matrix. Psychological Bulletin, 1980, 87, 295-251.

Thomas, D.J., Majewski, P.L., Ewing, C.L., & Gilbert, N.S. Medical qualification procedures for hazardous-duty aeromedical research, AGARD Conference Proceedings No. 231. Nuilly-Sur-Seine, France: AGARD, 1978, A-3: 1-13.

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 82	Interim

4. TITLE AND SUBTITLE

Repeated Measures of Information Processing

5. FUNDING NUMBERS

63216 M0097.001

6. AUTHOR(S)

Mary M. Harbeson, Michele Krause, and Alvah Bittner, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Biodynamics Laboratory
P. O. Box 29407
New Orleans, LA 70189-0407

8. PERFORMING ORGANIZATION REPORT NUMBER

NBDL-82R014

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Naval Medical Research and Development Command
National Naval Medical Center
Building 1, Tower 12
Bethesda, MD 20889-5044

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This publication provides documentation on two information processing tasks were considered for inclusion in a test battery which was being developed for repeated measures investigations of adverse environmental effects. The tests, adapted from the Rose Battery were; Baron's Graphemic and Phonemic Analysis, and Posner's Letter Classification. Alternate forms of the tests were administered on 15 consecutive workdays to 21 Navy enlisted men. In general correlations of the basic measures tended to become stable with sufficient practice, but derived measures such as difference, slope, and ratio scores did not attain stability.

14. SUBJECT TERMS
Graphemic and Phonemic Analysis, Stabilize Correlations

15. NUMBER OF PAGES
5

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	