Psychological Reports, 1971, 29, 523-530.

RELIABILITY AND LEARNING FACTORS ASSOCIATED WITH COGNITIVE TESTS

JOSEPH J. FLEISHMAN AND E. RALPH DUSEK[1]

U. S. Army Research Institute of Environmental Medicine, Natick, Mass.

Summary.—Reliability, practice effects, and factor loadings were investigated for 21 paper-and-pencil tests selected from the French, *et al.* Kit of Reference Tests for Cognitive Factors. Criteria of brevity and ease of administration were employed for selecting tests to be studied. The results are helpful in selecting those tests of cognitive abilities which may be most useful in studies of the effects of climatic and environmental variables on behavior. Such studies frequently require repeated measurements on a small group of *S*s over several days, thus making it extremely important that practice and environmental effects not be confounded. Test-retest reliabilities of the tests were generally quite good; 16 of the tests had a Pearson *r* of .8 or greater. Eight of the tests showed enough stability in mean performances over six trials to suggest that they may be used repetitively under environmental extremes without serious confounding by practice. The authors caution that factor definitions may not be the same under extreme conditions and under normal environmental conditions.

In assessing the effects of environmental stress on psychological performance, it is important that behavior be evaluated across a variety of abilities. In this regard, tests presumed to measure relatively independent behavioral factors, such as those in physical fitness, psychomotor performance, or cognition, have been identified through factor analysis techniques. For example, French, Ekstrom, and Price (1963) have summarized recommendations for measuring 24 cognitive abilities most consistently identified in previous factor analytic studies.

The purpose of this investigation was to determine the effects of practice on tasks which might be selected for use in investigating the effects of environmental stress on psychological performance. In such studies, repeated daily measures of psychological functioning may be required on each *S*; furthermore, only limited numbers of *S*s may be available. In addition, the use of test chambers in studying environmental effects on physiological and psychological functioning make counterbalancing or randomizing presentations of experimental conditions to subjects an unwieldy procedure. As a result, repeated measures designs are frequently used by necessity and effects attributable to order, sequence, boredom, practice, etc. may be confounded with the effects of the environmental variables. Consequently, it is desirable to determine whether under

[1]The authors wish to express their appreciation to Specialists James Sullivan, Mark Cohan, Morton DeMott, Lawrence Obrist, and Dominic Mammola for their assistance in scoring and collating the data. Special thanks are due to Mrs. Mary Ann Wall, who programmed the factor analyses for the GE 225 computer and Miss Ella Munro, who did most of the statistical analyses, for their invaluable assistance, without which this study would never have been completed.

repeated use there are appreciable practice effects. If there are, then the experimental design must be such that the effects attributable to practice and environmental stress are not confounded.

METHOD

French, *et al.* (1963) summarized recommendations for measuring 24 cognitive ability factors most consistently identified in the factor analytic literature, and made recommendations for three tests to measure each factor in which each test was relatively pure factorially and easy to administer as a paper-and-pencil test. Table 1 lists the Battery of Cognitive Tests (BCT) used in the present study with names of the tests, time to administer, and factor definitions. The tests were selected from those recommended by French, *et al.* (1963). Among the criteria for selecting tests were brevity and ease of administration.

Subjects

Ss were 90 Army enlisted men. Using 10 *Ss* a week for 9 wk. (90 *Ss*), reliability measures were obtained with 20 of the 25 tests and practice effects were studied with the same 20 tests using a limited sample (10 *Ss*). Regarding the remaining five tests, 10 *Ss* were used for obtaining reliability scores for four tests (Ms-1), (V-2), (N-1), and (Ma-2); the same limited sample was used to study practice effects on these tests. For the last test (Rs-1), reliability was obtained using the entire group of 90 *Ss*; however, no measures of practice effects were taken for this test.

Procedure

Scores for computing test-retest reliability were obtained by administering one-third of the tests in a randomized order to each *S* during the morning and repeating the same tests in the afternoon session. The testing lasted approximately 6 hr. on each of the three days required to administer all of the tests.[2] Two 15-min. rests were given during each of the 3-hr. sessions. An hour for lunch separated the morning and afternoon test periods. The tests were administered to the group by a single test administrator in a room large enough to comfortably seat 10 or more *Ss*. The standard test instructions, also printed on the first page of the test booklet, along with practice problems, were read aloud to *Ss* by the tester.

For one group of 10 *Ss* the same BCT tests (except Rs-1) were given for four additional days, each day consisting of two 3-hr. sessions.

Material

The Kit of Reference Tests for Cognitive Factors as well as a manual and

[2]In addition to the paper-and-pencil tests listed in Table 1, during this time part of the equivalent forms paper-and-pencil tests developed by Moran and Mefferd (1959) were administered. The norms obtained with the latter group of tests are being prepared for separate publication by Fleishman and Dusek.

TABLE 1

BATTERY OF COGNITIVE ABILITIES TESTS (BCT)

Ability or Factor	Test Name	Time for Admin.	Factor Code	Definition
I. Linguistic				
1. Verbal Comprehension	1. Vocabulary	8 min.	V	The ability to understand the English language
	2. Vocabulary	8 min.		
	3. Extended range	6 min.		
	4. Advanced	4 min.		
2. Word Fluency	1. Word endings	6 min.	Fw	Facility in producing isolated words that contain one or more structural, essentially phonetic, restrictions, without reference to word meaning.
	2. Word beginnings	6 min.		
	3. Word beginnings and endings	6 min.		
II. Numerical				
3. Number Facility	1. Addition	4 min.	N	The ability to manipulate numbers in arithmetical operations, rapidly.
	2. Division	4 min.		
III. Memory				
4. Associative Memory	1. Picture-number	14 min.	Ma	The ability to remember bits of unrelated material.
	2. Object-number	10 min.		
	3. First and last names	10 min.		
5. Memory Span	1. Auditory number span	15 min.	Ms	The ability to recall perfectly for immediate reproduction a series of items after only one presentation of the series.
IV. Reasoning				
6. Induction	3. Figure classification	16 min.	I	Associated abilities involved in the finding of general concepts that will fit sets of data; the forming and trying out of hypotheses.
7. General Reasoning	1. Math aptitude	20 min.	R	The ability to solve a broad range of reasoning problems including those of a mathematical nature.
	2. Math aptitude	20 min.		
8. Syllogistic Reasoning	1. Nonsense syllogisms	8 min.	Rs	The ability to reason from stated premises to their necessary conclusions.
V. Conceptual				
9. Semantic Redefinition	1. Gestalt transformation	10 min.	Re	The ability to shift the function of an object or part of an object and use it in a new way.
10. Spatial Orientation	2. Cube comparisons	6 min.	S	The ability to perceive spatial patterns or maintain orientation with respect to objects in space.

(Continued on next page)

TABLE 1 (Cont'd)
BATTERY OF COGNITIVE ABILITIES TESTS (BCT)

Ability or Factor	Test Name	Time for Admin.	Factor Code	Definition
11. Visualization	2. Paper folding 3. Surface development	6 min. 12 min.	Vz	The ability to manipulate or transform the image of spatial patterns into other visual arrangements.
VI. Discrimination				
12. Flexibility of Closure	1. Hidden figure	20 min.	Cf	The ability to keep one or more definite configurations in mind so as to make identification in spite of perceptual distractions.
13. Perceptual Speed	1. Finding A's 3. Identical pictures	4 min. 3 min.	P	Speed in finding figures, making comparisons and carrying out other very simple tasks involving visual perception.
14. Spatial Scanning	1. Maze Tracing	6 min.	Ss	Speed in visually exploring a wide or complicated spatial field.

sample tests, designed by French, *et al.* (1963), were purchased from Educational Testing Service, Princeton, N. J.

RESULTS

The Pearson product-moment test-retest correlations for the BCT are given in Table 2. All correlations were significant at the .01 level of significance. An $r \geq .90$ was accepted as very high reliability and the range, $.80 \leq r \leq .89$, as reasonably high reliability. Table 2 shows that four tests of the BCT achieve the highest criterion, and 12 more fit the latter criterion. Of the remaining five tests in the BCT, three achieved an r of .70 or greater and two fell below .70.

Intercorrelations (rs) were obtained between each of the 21 tests. Intercorrelations were larger between tests that represented the same factor. For example, r for R1 and R2 was .79, between V1 and V3 and V1 and V4 .83 and .81, respectively. The centroid method of factor analysis (Thurstone, 1947) was applied to the intercorrelations between the 21 tests. The analysis generated five factors, the factor loading distribution being presented in Table 3.[3] A factor loading of +.5 and greater or —.5 and less was used as a criterion for fac-

[3]A copy of the table of intercorrelations can be obtained from the authors or from ASIS as Document NAPS-01534. Order from ASIS National Auxiliary Publications Service, c/o CCM Information Corp., 909 Third Ave., 21st Floor, New York, N. Y. 10022. Remit $2.00 for microfiche or $7.40 for photocopy. On the mathematical side, one can reproduce all the original correlations from the factor loadings (see Fleishman, 1964, for procedure).

TABLE 2

TEST-RETEST RELIABILITY COEFFICIENTS, SAMPLE SIZE, SIGNIFICANCE LEVELS

Variable No.	Factor Code	r	N*	P
1	V-1	.94	85	.01
2	V-3	.89	82	.01
3	V-4	.56	83	.01
4	Fw-1	.82	82	.01
5	Fw-2	.86	86	.01
6	Fw-3	.81	83	.01
7	N-2	.96	83	.01
8	Ma-1	.71	82	.01
9	Ma-3	.83	86	.01
10	I-3	.86	82	.01
11	R-1	.87	83	.01
12	R-2	.81	83	.01
13	Rs-1	.33	73	.01
14	Re-1	.90	82	.01
15	S-2	.76	83	.01
16	Vz-2	.84	82	.01
17	Vz-3	.92	86	.01
18	Cf-1	.72	85	.01
19	P-1	.87	79	.01
20	P-3	.82	83	.01
21	Ss-1	.81	83	.01

*Variation in number of *Ss* taking different tests was due to an *S* occasionally being unavailable to complete the study.

TABLE 3

FACTOR LOADING FOR FIVE FACTORS CONTRIBUTING TO PERFORMANCE VARIANCE ON 21 BCT TESTS

Tests	1	2	3	4	5
Fw-1	.733	.067	.354	.034	−.206
Fw-2	.785	−.086	.150	.206	−.315
Fw-3	.666	−.211	.119	−.110	−.467
Re-1	.635	−.410	.302	.208	.216
V-1	.805	−.168	−.043	.359	−.040
V-3	.852	−.127	−.175	.269	−.051
V-4	.749	−.243	−.029	.490	−.093
P-1	.083	−.096	.822	.185	.013
N-2	.361	.113	.270	.767	−.060
R-1	.423	−.318	.121	.594	−.385
R-2	.328	−.210	.000	.688	−.345
Cf-1	.015	−.673	.081	−.093	.002
Ss-1	.131	−.663	.487	−.088	−.198
Vz-3	.287	−.649	−.078	.238	−.372
I-3	−.016	−.636	−.052	.453	−.082
P-3	.200	−.704	.417	.079	.085
S-2	.250	−.728	−.066	.189	−.336
Ma-1	−.055	−.193	.393	.363	−.548
Rs-1	.254	−.055	−.016	.207	−.700
Vz-2	.140	−.566	−.126	.161	−.597
Ma-3	.305	−.055	.216	.381	−.195

tor communality. It can be seen that for the pattern of 21 variables that were factor analyzed, Fw-1, Fw-2, Fw-3, Re-1, V-1, V-3 and V-4 formed Factor 1; Cf-1, I-3, P-3, S-2, Ss-1, Vz-2 and Vz-3 formed Factor 2. On Factor 3, P-1 had a positive loading of greater than .8. Tests N-2, R-1, and R-2 produced Factor 4; Ma-1, Rs-1, and Vz-2 formed Factor 5.

Practice Effects

A treatments by subjects design (Lindquist, 1953) assessed whether a significant practice effect had occurred on a test. When the *F* ratio for analysis of variance was significant at the .05 level, critical differences were calculated to determine whether differences between the means were significant at the .05 level (Li, 1964).

The *F* ratios, mean square for the trials effect, and the significance levels of trials for each of the tests are presented in Table 4. Trials were not statistically significant for the cognitive tests V-1, V-3, V-4, Fw-2, and Ms-1. The effects attributable to practice were statistically significant for the remaining tests. The trial means for each of the 24 cognitive variables are given in Table 5.

Discussion

Reliability and practice effects were studied using cognitive tests. From a practical point of view, each of the tests is given with time limitations and is

TABLE 4

Summary of Statistical Analysis of Effects of Trials on 24 BCT Tests

Factor Code	*MS*	*F*	$p_{.05}$	Critical Diff.*	Range, Trial *Ms*
V-1	6.05	1.59	no		23.16— 24.90
V-2	32.11	4.14	yes	3.72	16.08— 21.12
V-3	8.01	1.62	no		16.82— 19.10
V-4	.56	1.00	no		13.32— 13.95
Fw-1	418.46	17.65	yes	6.50	25.70— 43.50
Fw-2	77.31	2.65	no		22.30— 29.20
Fw-3	187.46	11.73	yes	5.33	13.60— 24.80
N-1	129.07	5.18	yes	6.70	51.80— 59.30
N-2	593.96	35.24	yes	5.50	42.70— 59.30
Ma-1	104.53	3.11	yes	7.70	25.20— 34.10
Ma-2	83.46	4.43	yes	5.80	12.90— 20.60
Ma-3	110.90	11.55	yes	4.10	18.90— 27.10
Ms-1	3.09	1.00	no		9.50— 11.20
I-3	5833.14	14.95	yes	26.31	91.13—153.70
R-1	47.61	5.40	yes	3.96	17.10— 23.00
R-2	25.82	3 81	yes	3.47	13.60— 17.38
Re-1	8.63	6 26	yes	1.57	11.20— 13.80
S-2	196.86	6.80	yes	7.17	16.40— 28.40
Vz-2	40.45	10 01	yes	2.68	10.42— 15.68
Vz-3	275.81	11.28	yes	6.59	39.90— 50.48
Cf-1	245.77	12.03	yes	6.02	10 50— 23.45
P-1	817.66	17.56	yes	9.10	66 20— 91.10
P-3	1332.11	29.15	yes	9.01	68 60— 94.68
Ss-1	191.27	19.56	yes	4.20	24 10— 36.20

*All significant at .05 level (Li, 1964).

brief and easy to administer. Furthermore, the stimuli are easily adapted for presentation on slides with responses being recorded on tape recorders. Thus, in studying the effects of environmental variables, e.g., cold, heat, hypoxia, on cognitive abilities, it should be possible to minimize effects of environmental variables on sensory input or motor response, e.g., writing.

The data provide a basis for selecting measures of cognitive abilities as dependent variables in studies where it is essential to obtain brief repeated measures of cognitive performance. As stated earlier, in evaluating effects of environmental variables on performance, it may be methodologically beneficial to employ tasks that remain relatively stable with repeated use under normal environmental conditions. Thus, stable measures can facilitate interpretation of the experimental effects since it is difficult to present the environmental conditions to *S* in a randomized manner. In addition, small numbers of available subjects almost always prevent the use of simple randomized groups designs. For the time intervals used in this study the learning curves obtained on these tests of cognitive abilities indicated that tests V-1, V-3, V-4, Fw-2, and Ms-1 are stable with repeated use. In examining the critical differences of Table 4 and juxtaposing them on the order of trial means (Table 5) one finds suggestions for other tests that are relatively insensitive to the effects of practice, e.g., Ma-1

TABLE 5
TRIAL MEANS FOR EACH OF 24 COGNITIVE TESTS*

Factor Code	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
V-1	23.16	23.16	23.36	24.69	23.66	24.90
V-2	19.25	16.08	21.12	20.32	19.52	20.48
V-3	16.82	17.32	18.08	19.10	18.75	18.70
V-4	13.47	13.32	13.80	13.95	13.78	13.80
Fw-1	25.70	30.90	33.40	38.60	39.30	43.50
Fw-2	22.30	25.00	24.00	25.50	29.10	29.20
Fw-3	13.60	20.90	18.40	23.20	24.80	24.60
N-1	51.80	52.00	49.70	56.70	55.40	59.30
N-2	42.70	49.10	49.70	56.70	55.40	59.30
Ma-1	25.20	29.20	29.60	31.90	34.10	33.10
Ma-2	12.90	19.60	15.50	16.30	18.70	20.60
Ma-3	18.90	21.40	21.50	23.90	27.10	27.10
Ms-1	9.50	10.20	10.10	11.20	10.30	10.00
I-3	91.13	115.45	128.95	150.20	145.05	153.70
R-1	17.10	20.72	20.05	22.68	22.68	23.00
R-2	14.00	16.10	13.60	15.75	17.32	17.38
Re-1	11.20	11.80	12.60	13.10	12.80	13.80
S-2	16.40	22.90	26.30	24.00	28.40	27.90
Vz-2	10.42	12.45	14.20	15.62	15.68	14.02
Vz-3	39.90	47.68	37.77	47.21	50.48	49.30
Cf-1	10.50	13.10	15.08	18.58	23.45	21.20
P-1	66.20	73.40	76.30	83.70	85.20	91.10
P-3	68.60	87.75	72.58	93.28	84.68	93.55
Ss-1	24.10	28.40	28.40	31.20	34.20	36.20

*Scores on each test were calculated according to the formula: $S = R - [W/(K - 1)]$, where '*S*' is the corrected score, '*R*' is the number of items answered correctly, and '*K*' the number of choices (French, *et al.*, 1963).

and R-2. In any case, significant differences between the means can be ascertained from information given in Tables 4 and 5.

The five factors designated in Table 3 may be characterized as follows: Factor 1 is vocabulary ability as is evidenced by the tests that load high in this factor. Re-1, Gestalt Transformation, the only test which in appearance seems not to fit this over-all type of definition, requires that *S* know the use of items. In a sense, this task requires *S*, in part, to define an object by its uses and in this sense it is another indication of vocabulary ability. Factor 2 is composed of tests that require spatial orientation and spatial relations. Factor 3 involves the ability to scan quickly and to make short quick motor movements. Factor 4 clearly loads high in arithmetic ability while Factor 5 seems to require abstract reasoning in conjunction with good memory.

The factor analytic findings seem to substantiate one of the criticisms of such a technique; that is, factor analysis produces loadings that are essentially dependent on the particular pattern of variables that are employed in the factor matrix and it is exceptional to obtain a "pure" task. Thus, the individual tasks which were employed in the present study are not factorially pure when analyzed as part of the pattern of variables employed, and the particular clusters obtained were in all likelihood a function of the specific 21 tests used in the factor analysis.

It is noteworthy that while intercorrelations were generally low (.83 being the largest obtained), they were highest for tests within a category of abilities (listed in Table 1), e.g., V-1, V-3, and V-4.

In considering these tests as dependent variables that might be used under environmental extremes, it may be useful to consider Carver's (1968) warning that factor analytically defined variables which are based on individual difference patterns at normal environmental conditions, may be entirely different in definition if they were obtained under environmental extremes. Thus, interpretation of factors and their meaning may be different when they are used at environmental extremes than when they are defined under normal environmental conditions.

REFERENCES

CARVER, R. P. Brief report; on the danger involved in the use of tests which measure factors. *Multivariate Behavioral Research*, 1968, 3, 509-512.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Manual for Kit of Reference Tests for Cognitive Factors.* (Rev.) Princeton, N. J.: Educational Testing Service, 1963.

LI, C. C. *Introduction to experimental statistics.* New York: McGraw-Hill, 1964.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

MORAN, L. J., & MEFFERD, R. B., JR. Repetitive psychometric measures. *Psychological Reports*, 1959, 5, 269-275.

THURSTONE, L. L. *Multiple factor analysis.* Chicago: Univer. of Chicago Press, 1947.

Accepted June 25, 1971.