Problems With Individual Difference Measures Based on Some Componential Cognitive Paradigms

William P. Dunlap, Tulane University

Robert S. Kennedy, Essex Corporation

Mary M. Harbeson, Tulane University

Jennifer E. Fowlkes, Essex Corporation

This article demonstrates that slope and ratio scores may have the same psychometric difficulties—low reliability—as difference scores. Empirically, direct measures and derived scores from Baron's, Collins', Meyer's, and Posner's cognitive paradigms were examined in terms of their reliabilities and cross-correlations. Reliabilities of the direct measures and their intercorrelations were high. The derived measures, which were slope, ratio, and difference scores, had reliabilities near zero and, therefore, their cross-correlations were also low. It is concluded that derived scores, although intuitively appealing as measures of mental operations, may have inherent psychometric difficulties that render them of little value for differential prediction. *Index terms: cognitive paradigms, difference scores, individual differences, prediction, ratio scores, reliability, slope scores.*

In recent years measures of individual differences based on componential cognitive theory have been developed to supplement or replace traditional psychometric measures frequently used for selecting and training individuals (Carroll, 1976; Farr, 1984; Hunt, 1978, 1983, 1984; Office of Naval Research, 1987; Sternberg, 1986). Most componential theories of cognition are variations of an information-processing model involving sequences of mental operations (Carroll, 1988; Farr, 1984; Wickens, 1984). Researchers often attempt to isolate these mental operations by computing derived measures such as slope, difference, or ratio scores. This approach appears promising, as aspects of human cognitive performance exist that are not adequately measured by traditional IQ and ability tests (Carroll, 1976).

This research stemmed from three basic observations concerning the constructs developed in some cognitive paradigms:

1. Although existing cognitive paradigms have intuitive appeal because they are factorially rich (i.e., they measure a number of different mental constructs), the variance shared across various paradigms may be high, because a thread common to many of them involves contrasting latency-based performance on simple versions of a task to performance on increasingly complex forms of the same task;
2. Many derived scores in cognitive paradigms are based on slopes, which may have the same inherent statistical difficulties as difference and percent scores (Carter, Krause, & Harbeson, 1986); and
3. Ratio scores may have the same psychometric difficulties as difference and slope scores, which may cause their reliabilities to be low and render them impotent as predictors.

Clearly, measures denoting structure in intelligence function do not need to be good differential predictors to provide useful descriptions of cognitive processes. The extent to which most persons show the characteristic in question may be impor-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 1, March 1989, pp. 9–17

0146-6216/89/010009-09$1.70

tant, correct, and descriptive; however, if there is not a large and reliable range of individual differences in the characteristic, it will suffer as a predictor. The present study was concerned with the use of the derived scores for predictive purposes.

The use of derived cognitive scores as individual difference measures is typified by the work of Rose and colleagues (Rose, 1978; Rose & Fernandes, 1977), who developed an information-processing performance battery to be used as a selection tool. When developing the battery, they attempted to include information-processing tests with high reliability, statistical independence, and construct validity. Each information-processing task was supposed to involve at least one of the following operations in addition to encoding and responding: constructing, transforming, storing, retrieving, searching, and comparing. Different information-processing tests were assumed to tap different mental operations, and individual differences on different cognitive tasks were assumed to indicate relative skill in each operation.

From a differential prediction standpoint, the question is not whether mental operations or information-processing components exist, but whether the individual differences are sufficient in these constructs to make them useful predictors. Mental operations intuitively appear to take place in stages that take a measurable amount of time. The latency of certain processes can be demonstrated (Brown, 1958; Peterson & Peterson, 1959; Sperling, 1960). The fact that mental events take a finite amount of time, however, does not necessarily mean that there are reliable indicators of individual differences in those events.

For example, it has been known since the time of Donders (1868/1969; Smith, 1968) that reaction time for four choices takes longer than simple reaction time. However, the reaction time slope, which purportedly measures speed of complex information processing, has long been known to be unreliable. Another example is the Stroop (1935) phenomenon. It virtually always takes longer to name the color of words printed in conflicting colors than it does to read the names of the colors printed in black and white. Harbeson, Krause, Kennedy, and Bittner (1982), however, found that the difference score, which supposedly measures interference, is not reliable. In addition, the difference score does not appear to measure a construct separable from the scores from which it was derived.

In both examples, although differences exist in group performance between a simple and a more complex condition, individual difference scores are not reliable. In each case, the basic scores were reliable and were as highly correlated with each other as was statistically possible.

Slope, ratio, and difference scores do not always have reliabilities so low that they would incapacitate derived scores as predictors. For examples of derived scores with reasonably useful reliabilities, see Jensen (1965) regarding derived scores from the Stroop (1935) phenomenon, or Rose (1974) and Rose and Fernandes (1977) for derived scores from other paradigms. However, given the following derivation and demonstration, it behooves researchers to address the reliability of derived scores before recommending them as predictors.

Reliability

Practice effects. Evaluating performance based on tests that are not stable with repeated testing creates at least two related dangers. First, constructs that change over time are unstable, and predicting from measures of that construct will be compromised. Second, measurements of nonstable constructs can be misconstrued as unreliable. An adequate assessment of reliability can only be performed on a stabilized variable. Because reliability defines the upper limit to validity, it is essential to establish reliabilities of putative measures of individual differences of cognitive ability prior to employing such scores in prediction or selection.

Difference scores. It has long been recognized that difference scores frequently have low reliability (Cronbach & Furby, 1970). Whenever the correlation between different tasks is near the reliabilities of those tasks, difference scores will be severely limited in reliability. This can be seen from the following equation for reliability of difference scores (Cohen & Cohen, 1975, p. 64):

$$r_{(a-b)(a-b)} = \frac{\frac{r_{aa}+r_{bb}}{2} - r_{ab}}{1-r_{ab}} , \quad (1)$$

where r_{aa} and r_{bb} are the reliabilities of Task A and Task B, and r_{ab} is the correlation.

Slope scores. Measures based on slope scores, which appear with increasing frequency in cognitive research paradigms, have demonstrable inherent statistical weaknesses. Empirical studies have shown that a number of slope measures have low reliabilities (Carter et al., 1986), which indicates that slope scores may have lower reliabilities—often near zero—than the scores from which they are derived. This occurs because, mathematically, slope scores can be recognized as either difference scores or as weighted averages of difference scores. The same statistical difficulties arise under common conditions (Cronbach & Furby, 1970). If the base line (usually task difficulty) is fixed, then the slope can be defined in terms of linear trend coefficients. For three points along the base line, these coefficients are -1, 0, $+1$, representing nothing more than the difference between the most and least difficult task. For four points, the coefficients are -3, -1, $+1$, $+3$, amounting to three times the difference score derived from the most extreme tasks, plus one times the difference score on intermediate tasks.

Ratio scores. When the base line is not fixed, but a variable in its own right, the slope becomes a ratio of two variables. The reliability of ratio scores, as shown below, under quite reasonable conditions is identical to the reliability of difference scores or slope scores, and thus suffers from the same statistical deficiencies. The reliability of a ratio can be derived by substituting a and b appropriately into Cohen and Cohen's (1975, p. 68) equation, where a is the numerator and b is the denominator of the ratio variable, which results in

$$r_{(a/b)(a/b)} = \frac{r_{aa}S_a^2 + r_{bb}S_b^2 - 2r_{ab}S_aS_b}{S_a^2 + S_b^2 - 2r_{ab}S_aS_b} . \quad (2)$$

Under the condition that S_a (the standard deviation of scores on Task A) equals S_b (the standard deviation of scores on Task B), Equation 2 obviously simplifies to Equation 1; therefore, difference scores and ratio scores share the common statistical difficulty of low reliability whenever r_{ab} approaches the average of r_{aa} and r_{bb}.

The goal of the present research was to study the reliability of derived measures from four cognitive paradigms. These paradigms were: graphemic and phonemic analysis (Baron, 1973; Baron & McKillop, 1975); semantic memory retrieval (Collins & Quillian, 1969); lexical decision making (Meyer, Schvaneveldt, & Ruddy, 1974); and letter classification (Posner & Mitchell, 1967).

The paradigms studied used a common method, reaction time to letters or words; therefore, the raw scores should share common variance. The derived scores were either differences, slopes, or ratios; the question addressed was whether such scores were sufficiently reliable to support their use in selection or prediction. The usefulness of these derived scores in understanding or describing cognitive processing was not questioned; rather, these paradigms were examined in terms of the psychometric usefulness of the emergent scores for differential prediction.

Method

Examinees

The examinees were 19 Navy enlisted men between the ages of 18 and 24 who volunteered for duty at the Naval Biodynamics Laboratory in New Orleans. All examinees were recruited, evaluated, and employed in accordance with procedures specified in Secretary of the Navy Instruction 2900.30 Series and Bureau of Medicine and Surgery Instruction 3900.6. These instructions are based on voluntary consent and meet or exceed the provisions of prevailing national and international guidelines.

Task Descriptions

Graphemic and phonemic analysis. This task was developed by Baron (1973; Baron & McKillop, 1975) to study visual versus auditory or articulatory reading strategies. Examinees were re-

quired to judge whether phrases made sense or not under three conditions: Sense (e.g., our new car), Homophone (e.g., it's knot so), or Nonsense (e.g., a drop of ran). These were combined in pairs to form three basic conditions, Sense vs. Nonsense, Sense vs. Homophone, and Homophone vs. Nonsense. Theoretically, graphemic encoders would do better on Sense phrases and acoustic encoders would do better on Homophone phrases. There were 20 phrases in each condition, and the stimuli were displayed for approximately 3.5 seconds each. Four variables were recorded: response time for each of the three conditions, and the ratio of Sense vs. Homophone time to Homophone vs. Nonsense time.

Semantic memory retrieval. Collins and Quillian (1969) designed this task to study the hierarchical organization of semantic information in memory. Examinees were required to judge whether sentences describing three complexities of superset and property relationships were true; for example, A peach is a peach, A peach is a fruit, or A peach is food. Sentences describing more complex relationships were expected to require more processing time than those describing simple relationships. There were 24 sentences per day including two sentences in each category. Each sentence was displayed for approximately 3.5 seconds. The variables were response times for the correct positive sentences for superset and property at three levels of difficulty (accounting for six variables), and the slopes of superset and property sentences, for a total of eight scores.

Lexical decision making. Meyer et al. (1974) used this task to study the effects of graphemic and phonemic factors on word recognition. Letter strings were paired according to graphemic and phonemic similarity or dissimilarity: for example, clash, flash; cheap, deep; rough, dough; brake, note. Each item was displayed for approximately 3.5 seconds. Examinees were asked to judge whether they were words or nonwords. Each day the examinees were shown a list of 40 letter strings, of which half were words and half were nonwords. Nine scores were recorded, including the response times to the second stimulus of each pair that were either phonemically or graphemically the same or different (four basic scores); three difference scores derived from the basic scores representing Phonemic Facilitation, Graphemic Interference, and Phonemic Similarity; and response times for words and for nonwords correctly identified (Rose & Fernandes, 1977).

Letter classification. Posner and Mitchell (1967) used this task to study matching or recognition of stimuli of various levels of complexity. Examinees made 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, or not matched, such as AB). There were 36 trials per day in each of the first two conditions and 32 in the third. The stimulus items were each displayed for 2 seconds. Five scores were calculated, including response times for each of three conditions for same judgments, and two difference scores (search time for name, and search time for category).

Apparatus and Procedure

The stimulus material was presented by means of black and white slides shown on a Kodak Ektograph 240 Audio Viewer™. The rate of presentation was controlled by preprogrammed tape cassettes. Examinees responded by pushing one of two buttons (yes or no) on boxes that were fastened to their desk tops. The answer and the response times were displayed on an automatic timing device and recorded on an answer sheet by the experimenter.

The examinees were tested in groups of four beginning at 8 a.m. for 15 consecutive workdays. Each group was tested at the same time each day. The four tests were administered in the same order to each group of examinees, but the order was varied across groups and days. There was a break of two or three minutes between tests while the experimenter changed carousels and cassette tapes, and a five-minute break between tests halfway through testing. The duration of each test was approximately 15 minutes per day. Total testing time was approximately 1.5 hours per day per group, including breaks.

Results

The 26 variables analyzed are presented in Table 1. Scores 5, 6, and 7 from Meyer's paradigm and

Table 1
Descriptions and Means and Standard Deviations (in msec) for the 26 Variables Analyzed

Paradigm and Score	Description	Definition of Derived Scores	Mean	Standard Deviation
Meyer				
1	Phonemic and Graphemic Similar		163.01	59.38
2	Phonemic Similar/Graphemic Dissimilar		166.69	57.92
3	Phonemic Dissimilar/Graphemic Similar		162.12	52.82
4	Phonemic and Graphemic Dissimilar		162.90	60.78
5	Phonemic Facilitation	5 = 4 − 1	2.65	57.83
6	Graphemic Interference	6 = 4 − 3	21.57	81.03
7	Phonemic Similarity	7 = 4 − 2	3.32	65.34
8	Mean Reaction Time to Words		166.13	44.08
9	Mean Reaction Time to Nonwords		206.76	56.40
Posner				
10	Physical Match		119.93	27.11
11	Name		142.34	33.01
12	Category		172.55	47.02
13	Name Retrieval	13 = 11 − 10	22.41	17.75
14	Category Retrieval	14 = 12 − 11	30.95	28.49
Collins				
15	Superset 0		231.05	78.34
16	Superset 1		245.97	78.04
17	Superset 2		272.96	73.98
18	Property 0		269.71	72.62
19	Property 1		281.64	82.43
20	Property 2		300.47	87.13
21	Slope for Superset	21 = 17 − 15	14.13	48.06
22	Slope for Property	22 = 20 − 18	8.92	46.79
Baron				
23	Sense vs. Nonsense		294.27	65.52
24	Sense vs. Homophone		308.13	72.65
25	Homophone vs. Nonsense		340.70	72.20
26	SH/HN	26 = 24/25	.90	.10

Scores 13 and 14 from Posner's paradigm are difference scores and, therefore, might be anticipated to have psychometric problems (Carter et al., 1986). Slope Scores 21 and 22 from Collins' paradigm should show psychometric problems similar to difference scores. Finally, the ratio score (Score 26) from Baron's paradigm should be vulnerable to the same psychometric problems as the difference and slope scores above.

Means and standard deviations for each variable (shown in Table 1) and average reliabilities and cross-correlations were computed in the following manner. Preliminary examination revealed that most variables had stabilized by day 5, and that performance on the final two days, 14 and 15, tended to be somewhat erratic on some tests. Therefore, the data were averaged over days 6 to 13, eight data points per examinee per measure. For each measure, the 28 possible reliabilities [$(8 \times 7)/2$] were computed, then these correlation coefficients were averaged. For cross-correlations, the 64 possible correlation coefficients were computed and averaged. These averaged reliabilities (diagonal elements) and cross-correlations (off-diagonal elements) are presented in Table 2.

As the correlation matrix in Table 2 shows, the rows and columns corresponding to the difference, slope, and ratio scores have consistently low correlations. On the diagonal, it can be seen that these variables also have very low reliabilities. Therefore, the failure of these variables to predict other variables is not surprising: They do not even predict themselves. The average *r*s shown in Table 2 demonstrate that the derived scores are poorly related

Table 2
Intercorrelations and Reliabilities (On the Diagonal) of 26 Measures From the Cognitive Paradigms; Difference, Slope, and Ratio Coefficients Are In Rows 5, 6, 7, 13, 14, 21, 22, and 26 (Decimal Points Omitted)

	Meyer									Posner					Collins								Baron				Mean
Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	r
1	**62**	55	55	60	−06	13	07	75	72	68	74	70	33	31	65	53	54	58	58	55	−17	10	62	62	56	27	.46
2	55	**45**	49	53	−03	10	02	65	60	59	63	63	28	31	54	45	46	49	48	46	−11	08	53	52	48	21	.40
3	55	49	**48**	54	−05	04	05	65	60	60	64	61	28	26	53	46	46	49	51	49	−11	11	54	53	49	23	.40
4	60	53	54	**53**	−02	12	08	70	66	63	69	65	33	26	60	52	52	56	55	53	−12	08	61	60	55	25	.44
5	−06	−03	−05	−02	**−04**	05	04	−06	−06	−06	−05	−06	00	−05	−04	−01	−01	−02	−06	−05	04	−04	−01	−01	−01	00	−.02
6	13	10	04	12	05	**00**	06	12	13	13	14	12	07	02	13	10	11	13	07	07	−03	−03	12	11	09	08	.09
7	07	02	05	08	04	06	**−02**	06	06	07	08	03	04	−05	07	05	06	07	04	03	−02	−01	06	07	05	07	.05
8	75	65	65	70	−06	12	06	**85**	82	77	84	82	38	38	73	61	62	66	67	65	−17	13	73	71	66	29	.53
9	72	60	60	66	−06	13	06	82	**79**	71	78	76	37	36	72	60	59	64	66	64	−17	13	73	71	65	30	.51
10	68	59	60	63	−06	13	07	77	71	**77**	80	73	27	27	62	49	53	56	55	52	−15	07	58	56	48	33	.45
11	74	63	64	69	−05	14	08	84	78	80	**84**	78	38	29	70	58	61	65	62	59	−15	06	68	67	59	32	.51
12	70	63	61	65	−06	12	03	82	76	73	78	**87**	34	56	67	57	56	59	61	58	−17	11	65	64	60	24	.49
13	33	28	28	33	00	07	04	38	37	27	38	34	**21**	11	37	35	33	36	32	31	−04	01	38	38	37	11	.26
14	31	31	26	26	−05	02	−05	38	36	27	29	56	11	**55**	30	27	22	22	29	29	−11	12	29	30	31	03	.22
15	65	54	53	60	−04	13	07	73	72	62	70	67	37	30	**68**	61	61	65	63	62	−15	13	73	72	66	30	.48
16	53	45	46	52	−01	10	05	61	60	49	58	57	35	27	61	**54**	57	59	57	58	−09	12	70	68	66	23	.43
17	54	46	46	52	−01	11	06	62	59	53	61	56	33	22	61	57	**53**	61	56	55	−07	07	68	68	65	24	.43
18	58	49	49	56	−02	13	07	66	64	56	65	59	36	22	65	59	61	**64**	60	59	−09	07	72	72	67	28	.46
19	58	48	51	55	−06	07	04	67	66	55	62	61	32	29	63	57	56	60	**58**	62	−12	16	70	69	65	25	.45
20	55	46	49	53	−05	07	03	65	64	52	59	58	31	29	62	58	55	59	62	**60**	−13	20	71	69	67	22	.44
21	−17	−11	−11	−12	04	−03	−02	−17	−17	−15	−15	−17	−04	−11	−15	−09	−07	−09	−12	−13	**00**	−07	−11	−11	−09	−06	−.10
22	10	08	11	08	−04	−03	−01	13	13	07	06	11	01	12	13	12	07	07	16	20	−07	**08**	17	16	19	00	.09
23	62	53	54	61	−01	12	06	73	73	58	68	65	38	29	73	70	68	72	70	71	−11	17	**85**	85	83	27	.51
24	62	52	53	60	−01	11	07	71	71	56	67	64	38	30	72	68	68	72	69	69	−11	16	85	**84**	83	28	.50
25	56	48	49	55	−01	09	05	66	65	48	59	60	37	31	66	66	65	67	65	67	−09	19	83	83	**83**	18	.47
26	27	21	23	25	00	08	07	29	30	33	32	24	11	03	30	23	24	28	25	22	−06	00	27	28	18	**22**	.20

to other variables, a fact easily predictable from their very low reliabilities and from the literature (Carter et al., 1986). Although this lack of cross-correlation may indicate that different domains of cognitive functioning are being sampled, examination of the reliabilities reveals a more parsimonious explanation: Because reliability places an upper limit on validity, such measures were psychometrically deficient at the outset.

To demonstrate the problems inherent in the difference, slope, and ratio scores from a psychometric perspective, Table 3 presents the derived measures (d), the basic scores from which they were derived (i,j), and the relevant cross-correlations and reliabilities (from Table 2), to permit computation of the predicted reliabilities of derived scores from Equation 1. The next to last column in Table 3 presents the predicted reliabilities, and the last column presents the actual reliabilities. It can be seen that only Score 14, Posner's Category Retrieval, has any psychometric basis for reliability and its actual reliability, although a modest .55, was the highest of the derived scores.

Discussion

The results indicate that the theoretically derived scores from these cognitive paradigms have inherent statistical weaknesses that might undermine their usefulness for selection and/or prediction. However, one potential problem in this research is that only 19 examinees were studied. Although a sample size of 19 argues for substantial instability of the correlation coefficients estimated, these are not single estimates but are correlations averaged over a number of trials. Dunlap, Silver, Hunter, and Bittner (1985) and Dunlap, Silver, and Bittner (1986) studied the impact of averaging cross-correlations and reliabilities, respectively, across trials with small sample sizes, and concluded that the standard errors of the resulting estimates were dramatically improved by the averaging process, particularly when the underlying estimated correlation was small. With a population correlation near zero, the precision of a cross-correlation estimated across eight repeated measures, as in the present study with an actual sample of 19, is as accurate as a single estimate of the same correlation from a sample of slightly less than 200 examinees. Reliabilities near zero are estimated with a precision equivalent to a single estimate from somewhat over 100 examinees. Therefore, the low average reliabilities and cross-correlations revealed for difference, slope, and ratio scores in the present study are not an artifact of small sample size.

Another potential problem in interpreting the

Table 3
Predicted and Actual Reliabilities (r_{dd}) of Derived Scores (d) From Basic Scores (i, j) by Equation 1 for Four Cognitive Paradigms

d (i, j)	r_{ij}	r_{ii}	r_{jj}	r_{dd} Eq. 1	r_{dd} Actual
Meyer (Difference Scores)					
5 = 4 − 1	.60	.53	.62	−.06	−.04
6 = 4 − 3	.54	.53	.48	−.07	.00
7 = 4 − 2	.53	.53	.45	−.09	−.02
Posner (Difference Scores)					
13 = 11 − 10	.80	.84	.77	.03	.21
14 = 12 − 11	.78	.87	.84	.34	.55
Collins (Slope Scores)					
21 = 17 − 15	.61	.53	.68	−.01	.00
22 = 20 − 18	.59	.60	.64	.07	.08
Baron (Ratio Score)					
26 = 24/25	.83	.84	.83	.03	.22

correlations may be range restriction, because the examinees were a highly selected group. On the other hand, if range restriction were a substantial problem, all correlations, including the reliabilities and intercorrelations of raw scores, should be low, which is clearly not the case.

Of course, theoretical constructs may be important to description and understanding, although less useful for differential prediction purposes. Cognitive constructs need not be independent or explain new and untapped variance to be useful in structural theory. For differential prediction purposes, however, the latter two qualities are important.

Finally, all difference, slope, or ratio scores are not necessarily unreliable or of low validity (e.g., Mumaw, Pellegrino, Kail, & Carter, 1984); but on statistical grounds, they are quite vulnerable to problems with reliability, and certainly issues of reliability with the target population should be addressed before they are included in a prediction/selection battery.

No factor analysis of the present data was attempted because of the small sample size, but the present findings have important implications for factor analyses of the derived scores using these paradigms. In the full-factor model, the diagonal of the correlation matrix factored is replaced with best estimates of the reliabilities of the variables analyzed. These reliabilities establish the upper limit on the communality, h^2, of each variable, where h^2 is the proportion of variance a given variable shares with the resulting factors. In the past, the fact that these derived scores do not load on factors that load heavily on the basic scores has been interpreted as an indication that the derived scores index some "new" or "emergent" factor. However, the present data argue strongly that the reason these derived scores do not load heavily on the more traditional factors is that, because of their low reliability, they have little "true score" variance to share with any factor.

References

Baron, J. (1973). Phonemic stage not necessary for reading. *Quarterly Journal of Experimental Psychology, 25,* 214–246.

Baron, J., & McKillop, B. J. (1975). Individual differences in speed of phonemic analysis, visual analysis, and reading. *Acta Psychologica, 39,* 91–96.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10,* 12–21.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27–56). Hillsdale NJ: Erlbaum.

Carroll, J. B. (1988). Individual differences in cognitive functioning. In R. D. Atkinson, R. J. Herrnstein, G. Lindzey, & R. Duncan Lule (Eds.), *Stevens handbook of experimental psychology* (pp. 813–862). New York: Wiley.

Carter, R. C., Krause, M., & Harbeson, M. M. (1986). Beware the reliability of slope scores for individuals. *Human Factors, 28,* 673–683.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale NJ: Erlbaum.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time for semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 240–247.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin, 74,* 68–80.

Donders, F. C. (1969). On the speed of mental processes. (W. G. Koster, Trans.) *Acta Psychologica, 30,* 412–431. (Original work published 1868)

Dunlap, W. P., Silver, N. C., & Bittner, A. C., Jr. (1986). Estimating reliability with small samples: Increased precision with averaged correlations. *Human Factors, 28,* 685–690.

Dunlap, W. P., Silver, N. C., Hunter, R. E., & Bittner, A. C., Jr. (1985). Averaged cross-correlations: A methodology for validity assessment in small samples. In R. Eberts & C. G. Eberts (Eds.), *Trends in ergonomics/human factors II* (pp. 13–21). Amsterdam: Elsevier.

Farr, M. J. (1984). Cognitive psychology. *Naval Research Reviews, 1,* 33–36.

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.

Hunt, E. (1978). Mechanics of verbal ability. *Psychological Review, 85,* 109–130.

Hunt, E. (1983). On the nature of intelligence. *Science, 219,* 141–146.

Hunt, E. (1984). Intelligence and mental competence. *Naval Research Reviews, 1,* 37–42.

Jensen, A. R. (1965). Scoring the Stroop test. *Acta Psychologica, 24,* 398–408.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1974). Functions of graphemic and phonemic codes

in visual recognition. *Memory and Cognition, 2,* 309–321.

Mumaw, R. J., Pellegrino, J. W., Kail, R. V., & Carter, P. (1984). Different slopes for different folks: Process analysis of spatial aptitude. *Memory and Cognition, 12,* 515–521.

Office of Naval Research. (1987). *Psychological Sciences Division 1985 Programs*. Arlington VA: Office of Naval Research.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193–198.

Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review, 74,* 392–409.

Rose, A. M. (1974). *Human information processing: An assessment and research battery* (Tech. Rep. No. 46). Ann Arbor MI: University of Michigan, Department of Psychology.

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

Rose, A. M., & Fernandes, K. (1977). *An information processing approach to performance assessment: Experimental investigation on an information processing performance battery* (Tech. Rep. No. 1). Arlington VA: American Institutes for Research.

Smith, E. E. (1968). Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin, 69,* 77–110.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74* (Whole No. 498).

Sternberg, R. J. (1986). Inside intelligence. *American Scientist, 74,* 137–143.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Wickens, C. D. (1984). *Engineering psychology and human performance*. Columbus OH: Charles E. Merrill.

Acknowledgments

The authors thank Andrew Rose, who generously loaned the stimulus materials for the basic data collected in these studies.

Author's Address

Send requests for reprints or further information to Robert S. Kennedy, Essex Corporation, 1040 Woodcock Road, Suite 227, Orlando FL 32803, U.S.A.