Acta Psychologica 24 (1965) 398–408; © *North-Holland Publishing Co*

SCORING THE STROOP TEST [1]

ARTHUR R. JENSEN

University of California, Berkeley, Calif., USA

The Stroop Test, also known as the color-word naming test, has been used in over sixty published psychological studies since the test was originally described by Stroop (2). The test has been used to attempt to measure a variety of processes and traits in the cognitive and personality domains, which it is not the purpose of this article to review. The psychometric characteristics of the Stroop Test, however, have never been adequately determined. The purpose of this article is to provide some normative data on the Stroop Test in the kind of population in which it is most typically used, to determine the reliability of the various measurements derived from the Stroop, and to arrive at some conclusion concerning the meaning of the variety of "scores" that have been derived from the test in the literature. There would surely seem to be considerable redundancy among the various scoring methods suggested by different investigators and the method of scoring the Stroop Test is in need of clarification and symplification.

General description of the Stroop Test

There is no standardized version of the Stroop, but all the variations on Stroop's original procedure have this in common: there are three cards—the "color card" (card A) on which there are 100 patches of from three to five different colors, the "word card" (card B) on which are printed (in black and white) the names of the colors, and the "color-word card" (card C) on which are printed the names of the colors, but printed in an ink of a conflicting color (e.g. the word RED might be printed in green, yellow, or blue, but never in red). Each card has 100 items to be named. The subject's (*S*'s) task on card A

[1] The research reported herein was supported through the Cooperative Research Program of the Office of Education, U.S. Department of Health, Education and Welfare, and by a National Science Foundation grant to the Institute of Human Learning.

is simply to utter the names of the colored patches as rapidly as possible, scanning the rows from left to right. On card B the *S* reads aloud the color names as rapidly as possible. On card C the *S* is required to name the colors of the inks while ignoring the conflicting printed color names. The *S*'s basic score on each card is the total time (in seconds) he takes to utter the 100 names.

Stroop (2) used five colors, but there has been no consistency in the number of colors used by other investigators; the number ranges from three to five, and the same colors are not always used. The size and dimensions of the cards, the print, the spacing of items, etc. also are not standardized. Furthermore, in some versions the color patches and color-words are printed on a white background and the words are printed black on white, while other versions use a black background on all three cards, with card B having the words in white on a black background.

Method

Because of his lack of standardization it seemed advisable to begin by making a version of the Stroop Test which would combine what seem to be some of the best features of the various forms described in the literature.

Materials

The cards were made large enough to be used as wall charts, so that a *S* standing four feet from the chart would have no visual difficulty in discriminating the colors or printed words. It had been found previously that when small cards were used, which the *S* would hold in his hands, it was impossible to control the *S*'s pointing with the finger or other variations in behavior which interfered with standardized administration.

The plates, which were made photographically, were each 17¾ in. × 25¼ in. and were mounted on heavy (⅛ in. thick) cardboards 19 in. × 26 in.

Card A consisted of ten rows and ten columns of evenly-spaced colored dots against a flat-back background. The dots were ⅝ in. in diameter and their centers were separated by 1½ in. All the colors were vivid and easily discriminable; red, green, orange, blue, and yellow. The order of the colors was random except for the following restrictions: (*a*) all five colors appeared an equal number of times, (*b*) all colors appeared in each row of ten dots, and (*c*) adjacent dots (in the order in which they were read, from left to right) were never of the same color.

Card B consisted of twenty rows and five columns of the printed names of the colors. The words were in white against a flat-black background. All letters were block caps $^5/_{16}$ in. high. Their line width was $^1/_{16}$ in. The words were distinctly separated, with the columns exactly in line. The names were in a random order except for the same restrictions that applied to card A (above). However, the order of the color names on card B was never congruous with the order of the colored dots on card A.

Card C consisted of the same plate as was used for card B, but with the words vividly tinted with the five colors, the actual color, of course, always conflicting with the color name. The order of the colors was the same as the order on card A.

A *Pretest Card* was used at the beginning of the test to aid in explaining the task to the *S*. It consisted of the same plate as card A, but without any colors on the dots, which were simply white on a black background.

Instructions

The cards are placed on an easel at about the *S*'s eye level. The *S* stands facing the easel at a distance of approximately four feet. The Pretest Card is shown first and *S* is told he will next be shown a similar card with colored dots and that he is to name the colors, going from left to right, as rapidly as possible and without stopping until the end. The five colors are named by the experimenter (*E*). Then card A is presented, *E* says "Go!" and simultaneously starts a stopwatch. *E* taps the card with a pencil whenever *S* makes an overt error. This procedure seems to militate against careless performance and overt errors are very rare. The procedure is similar for cards B and C. Prior to the presentation of each card, *S* is told what is expected of him. On card B *S* is told to read the color names as rapidly as possible, and on card C he is told to name the colors and ignore the printed words.

In the present study one group of *S*s ($N = 386$) was tested on two occasions and another group ($N = 50$) was tested on ten occasions, separated by at least one day. Instructions were slightly abbreviated after the first administration and the Pretest Card was discarded.

The *S*s in this study were 436 undergraduate students in an introductory course in educational psychology at the University of California. Approximately two-thirds to three-fourths of the *S*s were women.

Results

Scoring Formulas

Eleven formulas for deriving scores from the three basic Stroop scores (A. B, and C) are found in the Stroop literature. The data were scored by all these methods. The various scores are claimed to represent somewhat different psychological functions and each formula is associated in the literature with some kind of rationale connected with the particular investigator's use of the Stroop Test.

The scoring formulas are presented in table 1. Formulas M and N were originally used by Thurstone in his factorial study of perception (3); formulas D, F, G, H, I, J, K, and L were devised by Thurstone in a more recent study of the Stroop Test (Thurstone and Mellinger, (4)). Many of these formulas have since been used by other investigators. Formula E was originally proposed by Callaway (1).

TABLE 1
Basic Stroop scores and scoring formulas.

Basic scores	Time measures (seconds)
A	Color card
B	Word card
C	Color-word card
Derived scores	**Scoring formulas**
D	$A/(A + B)$
E	$C - A$
F	A/B
G	A/C
H	$A - B$
I	$(A - B)/(A + B)$
J	B/A
K	$(C - A)/B$
L	$C_z - 2A_z + 10$*
M	$B(C - A)/A$
N	$B(C - A)/AC$

* The C and A raw scores are converted to *Z* scores in this formula.

Normative data and reliability of scores

One group of *S*s ($N = 50$) had a test-retest interval of only two or three minutes; another group ($N = 50$) had an interval of one day: and the remaining *S*s ($N = 336$) had an interval of one week. The length of the test-retest interval, within these limits, made no appreciable or statistically significant differences on any of the scores. All scores showed some slight practice effect, regardless of the time intervening between test and retest. Therefore all the data were combined. The means and standard deviations for all the scores on each of the two administrations are presented in table 2, along with the reliability coefficients estimated for a single administration. The reliability was determined by the intraclass correlation (R_1) between the first and second administrations. The reliability of the two administrations combined can be obtained, of course, by means of the Spearman-Brown formula, boosting the length of the test by a factor of two.

Two points should be especially noted in table 2: (*a*) some scores show greater "practice effects" from the first to the second administration than do others; (*b*) the reliabilities of the various scores differ

TABLE 2

Means and standard deviations for first and second administrations of Stroop Test, the reliability (R_1). ($N = 436$).

Score	Administration				R_1
	First		Second		
	Mean (sec)	SD	Mean (sec)	SD	
A	58.24	10.17	56.11	10.57	.79
B	38.09	5.84	37.19	5.62	.88
C	100.36	19.50	88.34	15.93	.71
D	.60	.04	.60	.04	.72
E	42.12	14.96	33.23	10.50	.48
F	1.54	.25	1.52	.26	.65
G	.59	.09	.64	.08	.31
H	20.15	8.73	18.92	8.93	.65
I	.21	.07	.20	.08	.71
J	.66	.10	.68	.11	.71
K	1.12	.39	.88	.28	.44
L	10.00	1.55	10.00	1.39	.43
M	27.99	11.13	21.81	7.98	.48
N	.28	.08	.25	.08	.46

TABLE 3

Means and standard deviations for administrations 1–10 of Stroop Test, with *F* ratio for differences between administrations and the reliability (R_1). ($N = 50$)

score	1		2		3		4		5		6		7		8		9		10		*F*(9,441)*	R_1
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD		
A	60.30	9.80	55.80	8.60	51.70	7.60	50.40	8.40	50.10	8.30	49.00	8.20	47.50	7.70	48.00	7.90	46.90	7.50	45.90	7.20	101.64	.86
B	39.90	5.70	38.60	5.30	36.90	5.10	36.50	5.70	36.30	5.80	35.90	5.50	35.80	5.60	35.20	5.50	35.20	5.80	34.40	5.10	32.43	.86
C	102.30	16.50	88.70	14.00	80.00	11.70	75.90	12.50	73.50	12.40	71.40	11.90	69.50	13.40	66.60	11.10	66.90	11.60	64.60	11.50	261.01	.84
D	.60	.03	.59	.03	.58	.03	.58	.03	.58	.04	.58	.04	.57	.03	.58	.04	.57	.03	.57	.04	17.54	.77
E	42.00	13.00	32.90	9.30	28.30	7.60	25.50	7.70	23.40	6.80	22.30	7.40	22.00	8.50	18.60	6.80	20.00	6.90	18.70	7.30	85.86	.56
F	1.52	.21	1.45	.17	1.42	.20	1.39	.18	1.39	.22	1.38	.23	1.34	.19	1.38	.22	1.34	.19	1.35	.21	16.40	.78
G	.60	.08	.63	.07	.65	.06	.67	.07	.68	.06	.69	.07	.69	.07	.72	.08	.70	.07	.72	.08	29.12	.47
H	20.40	7.70	17.30	6.10	14.90	6.70	13.90	6.30	13.80	7.10	13.10	7.20	11.70	6.10	12.80	6.70	11.70	6.10	11.50	6.30	38.65	.77
I	.20	.06	.18	.06	.17	.07	.16	.06	.16	.07	.15	.08	.14	.07	.15	.07	.14	.07	.14	.07	17.53	.77
J	.67	.08	170	.08	.72	.10	.73	.09	.73	.10	.74	.11	.76	.10	.74	.11	.76	.11	.76	.11	17.46	.76
K	1.05	.30	.86	.22	.77	.19	.71	.21	.65	.17	.62	.18	.62	.22	.53	.19	.58	.19	.55	.20	53.11	.46
L	10.00	1.59	10.00	1.40	10.00	1.38	9.99	1.35	10.00	1.26	10.00	1.36	10.00	1.33	10.00	1.35	10.00	1.30	10.00	1.35	.00	.68
M	28.47	10.42	23.07	7.35	20.53	6.66	18.65	6.09	17.26	5.84	16.75	6.58	16.74	6.46	13.97	5.81	15.16	5.43	14.23	5.64	48.57	.56
N	.27	.08	.26	.07	.26	.07	.24	.07	.23	.06	.23	.07	.24	.07	.21	.07	.23	.07	.22	.07	9.15	.56

* For 9 and 400 df an *F* of 2.46 is required for significance at the .01 level.

markedly. Not surprisingly, the derived scores, consisting of differences and ratios among the basic scores, are less reliable than the basic scores. As further evidence will show, all these reliabilities can be boosted to a highly satisfactory level by obtaining repeated measurements on each *S*.

The effect of repeated measurements

One group of *S*s ($N = 50$) was tested every day for 10 days; the interval between sessions was generally one day, but every *S* had one interval of two days because of the weekend. The means and standard deviations for each administration are shown in table 3. In the last column are given the reliabilities estimated for a single administration so as to make possible direct comparison with the reliabilities in table 2. The reliabilities in table 3, when determined for the composite of all ten administrations, range from .89 to .98. The Spearman-Brown formula can be applied to the reliabilities in table 3 to determine the number of administrations required to obtain any desired level of reliability for any given score. It should be noted that the reliabilities are generally improved by repeated measurements, even when they are estimated for a single administration. The reason seems to be that unwanted variance due to practice effects tends to vanish after the first few administrations. Also shown in table 3 is the *F* ratio for the differences between administrations, which was obtained from the analysis of variance of each score. Though all the *F*s but one (score L) are highly significant, the actual magnitudes of the differences between administrations, especially after the third, are practically negligible. As indicated by the *F* ratio, the conflict card (card C) showed the greatest improvement with practice, while the word card (card B) showed the least effect of practice.

Table 4 shows the complete analysis of variance of the basic scores summarized in table 3; the last column gives the percentage of the total variance attributable to each of the main effects and their interactions. The excellent reliability of the basic Stroop scores with repeated administrations probably is largely due to the nonsignificant interaction between *S*s and administrations.

Intercorrelations and factor analysis

To attempt to reduce the redundancy in having so many derived scores, all the scores were intercorrelated and subjected to multivariate

Table 4

Analysis of variance of basic Stroop scores (A, B, C) for 10 administrations.

Source of variation	MS	df	*F*	% Variance
Scores (S)	213 230.38	2	874.14**	43.92
Administrations (A)	5 228.50	9	20.69**	4.85
Subjects (*S*s)	1 947.44	49	7.98**	9.83
S × A	1 441.80	18	5.91**	2.67
S × *S*s	481.95	98	1.98*	4.86
A × *S*s	252.76	441	< 1	11.48
Residual	243.93	882		22.16

* $p < .01$
** $p < .001$

analysis. The intercorrelations among the scores for the first and second administrations are presented in table 5. The composite scores of all ten administrations for the one group of 50 *S*s were also intercorrelated; the correlations did not differ appreciably from those in table 5, nor did the subsequent factor analyses.

Since the intercorrelations among scores are highly similar for the first and second administrations, the intercorrelations were obtained for the composite scores of both administrations. These correlations

Table 5

Intercorrelations of Stroop scores (first administration above diagonal, second administration below diagonal). ($N = 436$)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A		52	66	58	18	61	31	82	58	–56	–05	–85	–06	–50
B	54		43	–38	21	–35	06	–07	–38	39	–21	–38	35	16
C	76	51		29	86	31	–49	48	29	–28	67	–20	66	22
D	63	–30	39		–02	99	28	93	100	–100	14	–56	–40	–71
E	14	24	75	–05		00	–85	06	–02	02	90	32	90	63
F	64	–29	39	99	–05		28	94	99	–97	14	–57	–38	–69
G	39	08	–29	37	–84	37		32	28	–28	–88	–71	–87	–87
H	85	01	58	93	02	94	41		93	–92	08	–74	–31	–69
I	63	–30	39	100	–05	99	37	93		–100	14	–56	–40	–71
J	–62	30	–38	–100	05	–97	–37	–92	–100		–14	54	41	71
K	–11	–23	51	09	88	09	–88	02	09	–09		49	74	56
L	–82	–45	–38	–52	26	–52	–68	–69	–52	51	47		50	78
M	–13	34	50	–46	90	–45	–89	–37	–46	46	72	45		86
N	–56	09	02	–74	61	–72	–88	–73	–75	75	56	72	86	

were subjected to a principal components analysis, which was rotated to approximate simple structure by the varimax method. The results are shown in table 6.

TABLE 6

Principal components and varimax rotation of Stroop scores (first + second Administration). ($N = 436$)

	Principal-Components			Varimax-Solution		
Score	I	II	III	I	II	III
A	73	37	57	57	01	82
B	—08	19	97	—34	06	93
C	29	88	37	38	66	64
D	93	23	—29	100	—05	03
E	—20	97	05	07	97	24
F	92	24	—28	99	—04	04
G	59	—75	28	24	—94	23
H	94	32	03	93	—02	36
I	93	24	—29	100	—05	03
J	—92	—23	30	—99	05	—02
K	—16	88	—42	23	94	—22
L	—79	09	—56	—50	43	—71
M	—58	79	18	—37	90	21
N	—90	42	—04	—68	69	—17
% Var.	51	31	17	46	33	20

DISCUSSION

Only three factors account for all the variance in the various Stroop scores. The varimax solution is particularly clear-cut and makes for easy identification of the factors.

Factor I, which is most unambiguously represented by score D (score I is almost completely redundant), can best be regarded as a color-naming factor. Score D is a measure of individual differences in the degree of difficulty *S*s have in naming colors, with the "speed" factor (i.e. speed of reading words) partialed out.

Factor II, represented most clearly by score E, is the interference factor. It is a measure of the increment in difficulty of color naming brought about by the interference of the conflicting printed words. Thus, it is clear that whatever it is that causes color naming per se to be more difficult than word reading, it is not at all the same kind of difficulty that makes for slowness on the color-word interference card.

Factor III is the only factor which is most clearly represented by one of the basic scores—score B. It is best regarded as a reading speed factor. Thurstone (4) has referred to it as a "personal tempo" factor, but the generality implied by this designation should be based on some empirical demonstration of a general factor of "personal tempo" in a variety of tasks in addition to the Stroop.

This analysis clearly indicates that little, if anything, is to be gained from the proliferation of scoring formulas for the Stroop Test. It is suggested that scores D (color-factor), E (interference factor) and B (speed factor) are the most satisfactory scores and contain all the essential information that can be derived from the Stroop Test. These scores have the highest loadings on each of their respective orthogonal factors and are only slightly intercorrelated with each other. The three together account for practically all the variance in all of the derived scores that have been proposed in the literature. These particular scores, furthermore, are recommended by their reliabilities. The reliabilities of scores B and D are quite high even for a single administration of the test (.88 and .72, respectively), while the reliability of score E (.48) cannot be considered satisfactory for a single administration. But the interference factor cannot be reliably assessed by a single administration of the test on any of the scoring formulas. It is thus necessary to obtain repeated measures if one is to adequately assess individual differences in the interference factor. The reliabilities of the composite of ten administrations are highly satisfactory, however, being .98, .97, and .93 for scores B, D, and E, respectively. It should also be noted that scores B and D are relatively insensitive to practice effects, as indicated in table 3, while score E is affected to a greater degree by practice. Most of the practice effect on score E, however, occures in the first two or three administrations, after which individual differences remain quite stable.

Summary

A modified version of the Stroop Test was administered to over 400 university students. All *S*s received at least one retest; 50 *S*s were tested ten times at approximately one-day intervals. The basic Stroop scores (time taken for each of the three cards) were entered into eleven different scoring formulas reported

in the literature. The scores were factor analyzed and it was found that only three factors emerged: (*a*) a color-naming factor, (*b*) an interference factor, and (*c*) a speed factor. Thus it is possible to reduce the redundancy in Stroop scoring formulas to only three scores. Normative statistics and test-retest reliability estimates were given for all 14 Stroop scores, and recommendations concerning the use of particular scores were made.

References

1. Callaway, E., The Influence of Amobarbital (Amylobarbitone) and Methamphetamine on the Focus of Attention, *J. Ment. Sc.* 1959, *105,* 382—392.
2. Stroop, J. R., Studies of Interference in Serial Verbal Reactions, *J. Exp. Psych.* 1935, *18,* 643—661.
3. Thurstone, L. L., A Factorial Study of Perception, *Psychometric Monograph No. 4.* Chicago: University of Chicago Press, 1944.
4. Thurstone, L. L. and Mellinger, J. J., The Stroop Test, *Psychometric Laboratory Report No. 3,* May, 1953.