Acta Psychologica 25 (1966) 36–93; © *North-Holland Publishing Co., Amsterdam*

THE STROOP COLOR-WORD TEST: A REVIEW [1]

ARTHUR R. JENSEN and WILLIAM D. ROHWER JR.
University of California, Berkeley, USA

The Stroop Color-Word Test, which has now been in existence for 30 years, is of considerable psychological interest for several reasons: (a) it yields highly reliable and stable measures of individual differences on what seem to be three quite simple and basic aspects of human performance; (b) though there are reliable individual differences on each of the three time scores obtained from the Stroop test, the three scores maintain the same rank order of magnitude for all subjects (there was not a single exception among over 400 *Ss* tested by the writers); (c) the test has been used in a large variety of studies and has shown significant correlations with a host of other, often more complex, psychological measurements. Indeed, one difficulty in reviewing the literature on the Stroop test is that it cuts across so many diverse types of research and schools of thought in psychology. The variety of interpretations of the Stroop phenomena, couched in many different terminologies and unrelated theoretical orientations, testifies to the fact that psychology still has a long way to go in becoming a unified science.

The origins of the Stroop test go back almost to the beginning of experimental psychology. In the first psychological laboratory, in Leipzig, Wilhelm Wundt, in 1883, suggested to one of his students, James McKeen Cattell, that he do his doctoral research on the time it takes to name objects and colors and to read the corresponding words. Cattell reported the first experimental study of the relative speeds of color-naming and color-word reading in *Mind* in *1886*. The fact that color-naming requires more time than word-naming was also noted in William James' *Principles of Psychology* (1908, vol. I, p. 559), and was later

[1] This review 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.

the subject of researches by DESCOEUDRES (1914), W. BROWN (1915), and then, within a few years, by several other investigators whose work is cited in later sections of this review.

The most interesting feature of the Stroop test—the conflict or interference situation in which the subject must name the color of the ink of color-words when the color and the word are incongruous—actually originiated in the laboratory of Erich Rudolf Jaensch in Marburg, Germany, in connection with Jaensch's research on perceptual types (JAENSCH, 1929). The color-word interference test, however, was first introduced into American psychology by John Ridley Stroop (March 21, 1897b.) Stroop was working as a graduate student in the Jessup Psychological Laboratory of the George Peabody College for Teachers, where, under the directorship of Professor Joseph Peterson (1878–1935) there had already been established an interest in individual differences in speed of color naming and word reading. Peterson's theories on this topic (see PETERSON and DAVID, 1918) and his interest in racial psychology had stimulated an earlier student, TELFORD (1930), to use color naming and word reading in the study of racial differences, and also stimulated Stroop's choice of a topic for his doctoral thesis, concerned with interference in serial verbal reactions, in which he used the color-word interference test now generally referred to by his name. The study was published in the *Journal of Experimental Psychology* (STROOP, 1935), and was followed by only one other study using the color-word test (STROOP, 1938). From 1936 to 1964 Dr. Stroop was Head of the Department of Psychology at David Lipscomb College in Nashville, Tennessee, and since 1964 has been Professor of Bible.[2]

FORMS OF THE TEST

There is no standard version of the Stroop test with respect either to the materials of the test, the administration, or the scoring. However, the number of versions seems to be fewer than the number of investigators who have used the test, and each investigator seems to stick to only one version throughout various studies.

The original Stroop

STROOP'S (1935b) test consisted of three cards: a word card (*W*), a color card (*C*), and the incongruous color-word card (*CW*). Five colors

[2] Personal communication from Dr. Stroop, December, 1964.

were used: red, blue, green, brown, and purple. Stroop had first used yellow instead of brown, but found that yellow did not have sufficient contrast with the white card background. The words were printed on a white card in black ink in 14-point Franklin lower-case type. The words and the colors were arranged in a 10×10 matrix of evenly spaced rows and columns. An attempt was made to avoid any regularity in the sequence (horizontally and vertically) of colors. Each of the five colors (or words) occurred twice in each column and each row, and no color was immediately adjacent to itself in either column or row. On the *CW* card each color name appeared an equal number of times in each of the four other colors. The exact size of the cards, the size or shapes of the color patches, or their spacing were not specified in Stroop's writings. A version of the test very similar to Stroop's original version was produced by the C. H. Stoelting Company, manufacturers of psychological laboratory equipment, but it is now out of print and is no longer commercially available.

Variations

Thurstone. If any version of the test comes near being the standard form, it is Thurstone's modification of the original Stroop test. Thurstone made up his own form of the test to be used in his factorial studies of perception (1944), and the test was later described in detail by THURSTONE and MELLINGER (1953). In some ways the Thurstone version seems to be an improvement over the original Stroop, while in other respects the Thurstone would appear to have defects that Stroop took care to avoid. Thurstone used cards *W*, *C*, and *CW*, administered in that order, but the background of all three cards is black. The cards consist of photostatic negatives, with the *C* and *CW* cards tinted with photographic watercolors. The negatives are glued to heavy cardboards and coated with a clear plastic spray. The color patches on card *C* are circular dots $^5/_{16}$″ in diameter. The words and the colored dots are arranged in a 10×10 matrix. On the back of each card is a practice test consisting of a single row of ten of the same kinds of items as on the face of the card. The subject's successful performance on this brief practice task helps to insure that he properly understands the requirements of the test which immediately follows.

All these points seem to be improvements over the original Stroop. The black background helps to accentuate the colors and the cards

have an aesthetically more striking appearance than the versions using a white background.

Thurstone used only four colors: red, green, yellow, and blue. The use of only four colors has the possible disadvantage that each color cannot occur an equal number of times in each row or column, with the consequence that practice effects are not apt to be as evenly distributed throughout the test as in Stroop's version. This could possibly make a difference when time scores based on every 20 items are used —a method known as "serial scoring", which has been the basis of several studies to be described in a later section of this review. But probably the worst feature of the Thurstone version in this respect is that the words and colors are arranged simply in random order, with the one restriction that on card *W* the positions of the words in the 10×10 matrix be different from the positions of the corresponding colors on card *C*. The incongruous *CW* card has the words in the same positions as on card *W* and the colors are in the same positions as on card *C*. This degree of randomness of sequence unfortunately results in card W differing from cards *C* or *CW* in certain characteristics which might affect performance in ways that are irrelevant to differences in color naming and color-word reading. For example, on card *W* the response "yellow" is called for six times in the first two rows while "green" is called for only three times. Also there are many doublets and triplets of the same color, which speeds up responding at these points. With a little practice *S*s begin to perceive these doublets and triplets as single units. The first writer conducted an informal experiment which consisted of having several *S*s read Thurstone's card *W* and then having them read another card with an identical format on which were printed the words corresponding to the sequence of the colors on card *C* of the Thurstone version. The results suggest that the cards differ in difficulty due to the differences in *sequence* of responses. The sequence of responses in card *W* is easier than on card *C*, probably due to the great number of doublets and triplets on card *W*. The difference would probably be even greater if it weren't for the sequence *"green, yellow"* repeated three times in immediate succession on card C. Thus the repetition of certain sequences, which Stroop tried to avoid, is not avoided in the Thurstone version. Unrestricted randomness seems inadvisable as a means of determining the sequence of stimuli.

Smith. Gudmund J. W. Smith, a psychologist in the University of

Lund, Sweden, has carried out a number of studies (see references) using a further modification of Thurstone's version (SMITH and NYMAN, 1959). Smith does not use the *W* card and uses the *C* card mainly as a practice test preliminary to the *CW* card. The *C* card in the Smith version has groups of three colored *X*es rather than colored dots. There is an additional line at the bottom of the *C* card which consists of the color names printed in incongruous colors; if the *S* successfully names the colors rather than the words, the card is turned over to expose the *CW* card, which is essentially the same as Thurstone's.

Clark University. A number of investigations with the Stroop have emanated from the Psychology Department of Clark University, largely in connection with the late Heinz Werner's theories of cognitive development. These studies (see BROVERMAN, 1960; LAZARUS, 1958; PODELL and PHILLIPS, 1959; H. PODELL, 1963; and WAPNER, 1963) all seem to have used the same version of the Stroop, which differs from the previously described versions. Its most complete description is given by COMALLI, WAPNER and WERNER (1962). The cards are 9¼ × 9¼″ and the items are in a 10 × 10 matrix. The *C* card is made up of rectangular patches $^{5}/_{16}'' \times {}^{2}/_{16}''$. Only three colors are used: red, blue, and green. They are printed on a white background. The order of the items is random. At the top of each card is one row of practice items.

Jensen. This version (JENSEN, 1965) was made to overcome some of the deficiencies in other versions, but it probably has certain deficiencies of its own. The cards were made large enough to be used as wall charts. They are placed on an easel at the *S*'s eye level when standing and can be easily read at a distance of four feet. This modification in size resulted from the observation that with small cards which the *S* had to view from a relatively close distance, it was difficult to control such behaviour as card turning, viewing the card at a tilted angle, finger pointing, and other variations in the *S*'s behaviour which interfered with standardized administration. The cards were therefore 18″ × 25″, with colored dots ⅝″ in diameter and letters $^{5}/_{16}''$ high. Five colors (red, green, orange, blue, and yellow) were used. The colors and words were in random order except that there were no doublets or triplets of the same color, every color occurred an equal number of times, and every color occurred in every row. Card *C* was a 10 × 10 matrix, but cards *W* and *CW* had five columns and twenty

rows. With large charts, a row of ten words seemed so long as to make it difficult for *S*s quickly to locate the beginning of each successive row. The rows of ten colored dots occupied approximately the same visual span as the row of five color names. Further details of this version are given elsewhere (JENSEN, 1965).

Other modifications. GARDNER, HOLZMAN, KLEIN, LINTON, and SPENCE (1959) describe a version essentially like Thurstone's except that the background is white and the *C* card is made up of groups of colored asterisks which match the length of the words on card *W*.

The most radical departure from the original Stroop is that of FRASER (1963). Each of the three cards involves words and colors. Card 1 is a 10 × 10 matrix of the words: red, brown, orange and blue printed in incongruous colors of red, brown, orange and blue. Thus it is similar to the usual *CW* card. Card 2 is a 10 × 10 matrix of the words black, yellow, pink, and green printed in incongruous but not directly conflicting colors: red, brown, orange, and blue. Card 3 is like card 1 except for the sequence. Here the color of the *ink* in the stimulus word N is the same as the *color* word in stimulus N + 1. Since all three cards involve some degree of interference, evaluation of the *S*'s performance cannot be compared with the non-conflict performance usually obtained on cards *W* and *C* (this was not Fraser's purpose). There is no evidence that Fraser's Cards 2 and 3 yield any additional information that is not obtained from the usual *CW* card.

Parallel forms

Most investigators who have wanted to obtain repeated measures have simply retested with the same set of cards. STROOP (1935b) made an "equivalent" form of the test by printing each card in reverse order. Truly equivalent forms would seem to be essential if one were to study the effects of practice on the color-word phenomenon per se. It is not now known how much of *S*s' improvement with practice is due to improvement in ability to read words, to name colors, and to overcome the interference on the *CW* card and how much is due to the learning of the specific sequence of responses.

SMITH and BORG (1964) tried to make an equivalent form which would minimize practice effects. This "parallel" version consisted of the words *white, gray,* and *black* printed in incongruous shades of white, gray, and black. Smith and Borg considered this attempt unsuccessful.

Interference effects on the shaded version were small as compared with the colored version, and the scores derived from the two forms had low and statistically insignificant intercorrelations.

Conclusion

No one form of the Stroop test has been generally adopted by investigators, and none of the forms in existence seems rationally or psychologically ideal in all respects. If investigators are going to continue using the Stroop in their researches, it would seem worth while to construct and to generally adopt an improved and standardized form of the test. Since it requires a good deal of time, skill, and expense for the individual investigator to produce a satisfactory set of cards, they should be produced and distributed by one of the commercial psychological supply firms. Several points should be taken into consideration in preparing a more satisfactory form of the test than we now possess: (a) the cards, color patches, and printed words should be large enough to minimize variance due to individual differences in visual acuity over a wide range; (b) the length of the rows of items should not be so great that there is a perceptible gap in reading speed at the end of each row due to difficulty in locating the beginning of the next row; (c) the colors should be sufficiently vivid and dissimilar to minimize variance due to color discrimination per se; (d) it should be adequately demonstrated that the differences in performance times between cards *W*, *C*, and *CW* are negligible when the particular sequence on each of the cards is presented in identical form, i.e. either as words in black and white or as color patches; (e) doublets of the same color or word in immediate succession should be avoided; (f) every color (and every word) should follow every other color an equal number of times; (g) sequential practice effects from one card to another would best be avoided by requiring a different order of responses on each of the cards (a feature that is lacking in all present forms); (h) there should be from five to ten authentically parallel forms, consisting simply of different arrangements of the items on each card. It is known that certain scores derived from the Stroop have very low reliability when based on a single administration; obtaining repeated measures from parallel forms would minimize individual differences in improvement with practice due to the learning of specific sequences of responses. If research should show that repeated practice on the same cards results in no greater improvement in performance than

does practice on parallel forms in which the sequence of items is varied, then, of course, parallel forms could be dispensed with.

METHODS OF ADMINISTRATION

The test generally takes five to eight minutes to administer; retesting is considerably faster, since detailed instructions are unnecessary. The cards are usually presented in the order *W*, *C*, and *CW*. Some investigators use only cards *C* and *CW*. Most forms of the test include a row or two of practice items before the test proper; this brief warm-up helps to insure that the *S* has understood the instructions before beginning the test itself. Since color blindness is usually detectable on Card *C*, it would seem sensible to give this card first. Out of 436 *S*s tested by the authors no *S* seemed to have difficulty in naming the colors correctly. It is advisable for the experimenter (*E*) to tell the *S* the color names before the test begins; a few overly conscientious *S*s are prone to search their color vocabularies for more precise descriptions. The instructions themselves are obvious, but they should include the point that this is a test of speed and accuracy. The *S* begins with the top row and reads from left to right. Overt errors are very infrequent and are usually spontaneously corrected by the *S*.

THURSTONE and MELLINGER (1953) suggest that *E* follow *S*'s performance on a typed or mimeographed key, checking the items on which overt errors are made. Some *E*s tap the table with a pencil whenever an overt error occurs; this seems to boost the *S*'s vigilance. *E* records the total time taken for each card, starting the stopwatch with *S*'s first response and stopping with the last. These time measurements should be made as carefully and accurately as possible, since *S*s' performance on the Stroop is remarkably reliable. A considerable proportion of the error variance in Stroop scores is probably true measurement error, so pains should be taken to minimize this source of unreliability.

More refined measurement techniques have been used for special purposes. In order to make a detailed analysis of various types of errors in Stroop performance, RAND, WAPNER, and WERNER (1963) obtained complete tape recordings of the *S*'s performance and converted these to visual form by means of a Grass Polygraph. SMITH and KLEIN (1953) recorded the time for every 20 responses (two rows of items) and had *S*s do the *CW* card five times in succession, with one-minute

rest intervals, in order to obtain special kinds of scores, which are described in the following section on methods of scoring.

The cards are best placed before the *S* on a table or an easel, so that *S* cannot handle the cards. The *CW* card is made easier when viewed in such a way that the words are not perceived as clearly as they would be under optimal conditions. *E* must be alert to *S*s who try to take advantage of this fact by viewing the cards at an angle, squinting their eyes, "defocusing", or deaccommodating. *S*s rarely have time to discover these "tricks" on the first administration, but they tend to crop up when repeated measurements are sought. Subtle forms of these techniques may contribute in part to the clear-cut practice effects observed in repeated testing.

Group administration

Kipnis and Glickman (1959, 1962) made the first attempt to produce a form of the Stroop for group administration. There were only two cards: *C* and *CW*. Card *C* contained 150 rectangles colored red, yellow, and blue. Card *CW* contained 150 colored-words; some were incongruous and some congruous. On a separate answer sheet the *S* would identify the color of each item by writing the initial letter of each color. Three minutes were allowed for each card; the *S*'s score was the number of items correct. Kipnis and Glickman (1959) doubt that this group form gets at as much of the same variance as the individual form of the test. Kipnis has concluded that they have not been successful in building a group form of the Stroop and that all the valid variance in the group form is based mainly on the factors of speed and accuracy that are measured by clerical aptitude tests.[3]

Uhlmann's (1962a) adaptation of the Stroop for group administration consisted of four subtests, each printed on separate pages. The first two subtests are used as a warm-up exercise and are not scored. The second two subtests each consists of a 10×10 matrix of incompatible color-words (red, yellow, blue, and green). On one subtest the *S* has to print the first initial of each word, ignoring the color, on a line directly below the word. On the other subtest *S* has to print the initial of each color, ignoring the word. The time limit is one minute for each subtest and the *S*'s score is his number of correct responses. An interference measure is derived from the difference between the scores on the two subtests.

[3] Personal communication from Dr. Kipnis, February, 1963.

H. PODELL (1963) made up a group form of the test by using the Clark University cards and having *S*s enter a stroke in one of three columns labeled red, blue, and green as they scanned cards *W*, *C*, and *CW*. The score is the number of strokes made in 45 seconds for each card.

As Kipnis has pointed out, in all of these group forms one wonders to what extent the true variance one wishes to measure by means of the Stroop technique is swamped by the variance due to a clerical speed factor. No one has yet performed the obvious experiment—intercorrelating the group and individual forms. Until their equivalence has been demonstrated the use of group forms of the Stroop would seem to be risky.

SCORING

Probably no other psychological test, with the exception of the Rorschach, has yielded so many different scores as the Stroop test. These scores fall into one of three classes: (a) the basic time scores and all the derived scores to which their algebraic manipulation gives rise, (b) refinements of scoring the *S*'s performance to yield various kinds of error scores, and (c) temporal patterns of responses, which are concerned, not with comparisons of performance on each of the cards, but with changes in performance during the course of responding to only one card (usually *CW*).

Basic and derived scores

The literature reveals no fewer than sixteen scores derived from the three basic time scores on cards *W*, *C*, and *CW*. A large variety of psychological interpretations has been given to each of these derived scores. The various scoring formulas are shown in table 1.

Stroop used only the basic scores in his two investigations with the test (1935b, 1938). The derived scores have been contributed by later investigators. Scores H, N and O (table 1) were originally used by THURSTONE (1944) in his factorial study of perception; scores A, B, C, D, K, and M were later proposed by THURSTONE and MELLINGER (1953). Score J was first used by CALLAWAY (1959). Scores E and I were attributable to BROVERMAN (1963), as well as score G (BROVERMAN, 1960a). Score L was first used by Klein (see KLEIN, 1954). Score F was added by JENSEN (1965). Score P is due to CALLAWAY and STONE (1960).

It is readily apparent that there is a great deal of redundancy among the derived scores in table 1. Scores A, B, C, D, and E, as a little algebraic manipulation will show, are all linear functions of one another and thus have 100 per cent redundancy. H and I are similarly redundant. Scores G and L, being derived from the difference between an obtained basic score and a predicted score based on the regression of one score on another, have the disadvantage of being computationally complicated and of involving parameters in the regression equation which are specific to the particular sample of *S*s under investigation. A similar objection may be made to score M, which, however, is the score Thurstone thought best, on the grounds that it showed more significant correlations with personality inventory items than several of the other scores (THURSTONE and MELLINGER, 1953). This score, how-

TABLE 1

Basic Stroop scores and scoring formulas

Basic scores	Time measures (seconds)
W	Word card
C	Color card
CW	Color-Word card
Derived scores	**Scoring formulas**
A	C/W
B	W/C
C	$(C+W)$
D	$(C-W)/(C+W)$
E	$(C-W)/W$
F	$C-W$
G	$C-C_p$ *
H	C/CW
I	$(mW-C)/C$
J	$CW-C$
K	$(CW-C)/W$
L	$CW-CW_p$ **
M	CW_z-2C_z+10 ***
N	$W\times(CW-C)/C$
O	$W\times(CW-C)/(C\times CW)$
P	$C+CW$

*) C_p = Predicted value of C based on the regression of C on W.

**) CW_p = Predicted value of CW based on the regression of CW on C.

***) The CW and C raw scores are converted to z scores in this formula.

ever, turns out to be factorially more complex than most of the other scores and is therefore less desirable for theoretical purposes than some of the others. Similarly, scores N, O, and P each confound all three Stroop factors.

Error scores

Highly refined scoring of the *S*'s performance has been used by WAPNER (1963) and by RAND, WAPNER, WERNER, and MCFARLAND (1963). The *S*'s performance was tape-recorded and then converted to a visual record by means of a Grass polygraph. Some of the scorable features of performance were: (a) reading the word rather than naming the color, (b) inappropriate color responses other than word reading, (c) contaminated responses (e.g. "breen" for *green*), (d) inarticulate utterances, (e) inserted color words in addition to naming color, (f) omissions, (g) inserted linguistic words or phrases (e.g. "that's *green*"), (h) inserted nonliguistic utterances, (i) part-wrong responses, i.e. *S* begins to make wrong response, then corrects it, (j) whole-wrong corrected response, (k) jumbled order of response (*S* loses place, repeats, etc.), and (l) the duration of silent intervals.

GARDNER et al. (1959) handled overt reading errors in a unique way. The number of errors was multiplied by the reading time per unit (i.e. total time/100) and this value was added to the reading time for the particular card. Though this method penalizes the *S* for making errors by giving him a poorer time score, there is the danger that it might involve the admixture of "impurities" into the basic Stroop scores. It would thus seem advisable to record the time and the overt errors separately. In the writers' experience overt errors are so infrequent as to make it questionable whether they should be scored at all, especially since most *S*s spontaneously correct their errors, thereby causing error tendencies to be reflected in the basic time scores. SMITH and NYMAN (1962) have reported that 95 per cent of their *S*s make fewer than two overt errors on the *CW* card and that errors are virtually absent on cards *C* and *W*. Smith and Nyman adopted the practice of eliminating *S*s from their studies who made more than 10 errors in five administrations of the test.

Serial scoring

A method known as "serial scoring" was first proposed by SMITH and KLEIN (1953). This method was adopted, according to SMITH and

NYMAN (1962), because the usual Stroop scores, particularly the interference score *CW-C*, "had resisted all reasonable attempts at predicting behavior in a series of new cognitive tasks" (p. 2). The method has been described in at least four articles (SMITH and KLEIN, 1963; SMITH, 1959ab; SMITH and NYMAN, 1962), and, in addition to these studies, has been used in all of Smith's extensive research with the Stroop and in a study by GARDNER, HOLZMAN, KLEIN, LINTON, and SPENCE (1959).

The method is based on the time the *S* takes for every two rows (20 responses) on the *CW* card. Thus, each *S* has five time scores; the point of interest is the pattern of these scores. The total variability for a given S can be analyzed into the variability due to linear regression (i.e. improvement in speed from the first set of 20 responses to the fifth set of 20 responses) and the residual variability. *S*s are then classified as one of four types in terms of the amounts of these two sources of variability in their performances. The four types are:

(a) *Cumulatives:* high on regression and low on residual variability.
(b) *Dissociatives:* low on regression and high on residual.
(c) *Stabilized:* low on regression and low on residual.
(d) *Cumulative-dissociatives:* high on regression and high on residual.

The cut-off scores for determining a *S*'s classification are determined by the degree of differentiation the investigator desires to achieve among the four classes of *S*s. If the medians of the distributions of regression and residual scores are used, no *S*s are lost in the assignment to classes. But in one study in which sharper distinctions were desired (SMITH and KLEIN, 1953) and more stringent criteria were used (*S* had to fall into the same classification on at least three out of five administrations of the test), 40 per cent of the *S*s had to be discarded as not being sufficiently clear-cut examples of any one of the four types.

Obviously the serial scoring method appears to get at quite different aspects of performance than do the traditional methods of scoring. The serial scoring method seems to tap some combination of practice effects, the cumultative effects of fatigue or response inhibition, and possibly fluctuations in attention.

PSYCHOMETRIC PROPERTIES

Normative data

Most investigators have reported the means and standard deviations of the particular Stroop scores they have used in their own studies, but there has been no real attempt by anyone to develop Stroop norms.

About the only purpose such norms could serve at the present time would be to make it possible for investigators to compare the scores derived from their particular samples with those of some well-defined population and some clearly described form of the test, method of administration, and method of scoring. The possibility of making such comparisons would perhaps enhance conclusions concerning the generality of any particular study.

The closest thing we have to such norms is provided in a study by JENSEN (1965), which presents the means and *SD*s of the basic scores and derived scores (table 1) obtained on 436 university undergraduates. Means and *SD*s of these scores are also presented for each of ten repeated administrations of the test. The mean basic time scores are quite typical of those reported by other investigators: $W = 38.09$ sec, $C = 58.24$ sec, $CW = 100.36$ sec. THURSTONE and MELLINGER (1953) also present "exploratory norms" in the form of frequency distributions on the basic Stroop scores and four derived scores (B, H, K, and M in table 1) based on 99 students at the University of Chicago.

The intercorrelations of all of the basic and derived scores are also presented by JENSEN (1965). The product-moment correlation between W and C (first administration) was .52, r between W and $CW = .43$, r between C and $CW = .66$. Corresponding correlations reported by BROVERMAN (1960b) for 92 male college freshmen were .74, .57, and .76; and for a group of 35 male volunteers from a church organisation who were very heterogeneous with respect to educational and occupational criteria the corresponding correlations were .80, .63, and .81, respectively.

Serial scoring norms

SMITH and NYMAN (1962) have presented quite elaborate tables of norms for their serial method of scoring based on various psychiatric groups, and on various age groups from 12 to 60 years of age. The *N*s in their groups range from 10 to 109.

Reliability

Test-retest reliabilities of the basic and derived scores, with an average test-retest interval of one week, based on 436 *S*s are presented by JENSEN (1965). It was found that the length of the test-retest interval made no appreciable difference within the range of from a few minutes to one week. The reliabilities of the basic scores for a single

administration were $W = .88$, $C = .79$, and $CW = .71$. The derived scores, because they consist of differences and ratios, have somewhat lower reliabilities for a single administration, ranging from .31 to .72. Repeated testing has the effect of improving the reliabilities considerably more than would be predicted by the Spearman–Brown prophecy formula. The composite of ten administrations of the test, for example, raises the reliabilities of all of the derived scores above .90. The derived scores which measure interference (formulas involving C and CW) do not have satisfactory reliability for a single administration, and it is suggested that a composite of at least three administrations be used if derived scores are to be used. The table of reliabilities given by JENSEN (1965) can be used in connection with the Spearman–Brown formula to determine how many administrations are needed to obtain a given level of reliability on any particular score; the Spearman–Brown prophecy will almost certainly insure at least the desired level of reliability which the investigator has inserted in the formula for determining the necessary number of retestings. Ten administrations of the test yield reliabilities that are probably higher than those of any other psychometric tests (JENSEN, 1965).

The reliability of the serial scoring method is quite another matter. Since serial scoring is based on the pattern of change in a S's performance throughout the course of the test, and since practice effects strongly interact with these patterns of change, the determination of reliability by retesting is almost certainly bound to produce unsatisfactory results. Such, in fact, was found to be the case when SMITH and BORG (1964) attempted to determine the test-retest reliabilities of their serial scores. The reliability coefficients of the various scores ranged from —.16 to +.51, with a mean of .24. Smith and Borg therefore tried a "parallel" form of the Stroop consisting of shades of gray rather than of colors in hopes of obtaining more satisfactory retest results. The attempt was wholely unsuccessful in producing anything resembling a truly parallel form.

Factor analysis of Stroop scores

In order to reduce the redundancy among derived scores (table 1), JENSEN (1965) intercorrelated them, along with the basic scores, and performed a principal axes analysis and a rotation of the principal axes to simple structure by the varimax method. The results, based on an N of 436, were very clear-cut, and rotation of the principal axes

made no difference in the conclusions. Only three factors can be extracted from all the Stroop scores; these three factors account for at least 99 per cent of the variance contained in all of the scores. The percentages of variance accounted for by factors I, II and III before rotation were 51, 31, and 17 and after rotation 46, 33, and 20, respectively. Two of the basic scores and most of the derived scores are an admixture of all three factors, but a few of the scores emerged as almost "pure" (i.e. independent) measures of a particular factor.

Factor I is best called a *color difficulty* factor and is equally well represented by scores A, B, C, D, and E (see table 1), which are all equivalent in the analysis. Each of them correlates approximately .99 with factor I and has correlations less than .05 with the two other factors. The only basis for choosing among these scores is computational simplicity.

Factor II is best identified as the *interference* factor and score F (*CW–C*) is clearly the purest measure of this factor; it correlates .97 with factor II, .07 with factor I and .24 with factor III. The ratio of *CW/C* (or *C/CW*) for some reason is less pure, having a greater loading on factor I than does *CW–C*. Score M, which was favored by Thurstone as an interference measure turned out to be a mixture in almost equal parts of all three factors.

Factor III is best called a *speed* factor; Thurstone referred to it as "personal tempo" (THURSTONE and MELLINGER, 1953). Only one score is a clear-cut measure of this factor—the basic time score on card *W*. It correlates .97 with factor III and —.34 and .06 with factors I and II, respectively.

There would seem to be little justification for using any other of the known derived scores than those mentioned above as having the greatest factorial independence. It is also apparent that the basic *CW* score, which has been used in this raw form by so many investigators, is not just an interference measure but is an amalgam of all three factors, being loaded .38 on factor I, .66 on factor II, and .64 on factor III.

THE EFFECTS OF PRACTICE

Interest in the effects of practice on color naming has a much longer history than the Stroop test. In 1915 Hollingworth gave a color naming test to each of 19 *S*s 100 times over a period of 10 to 40 days. There was 30 per cent improvement in speed of color naming, and yet after

the 100 trials of practice the speed of color naming was still 37 per cent slower than the speed of reading color words. At about the same time WARNER BROWN (1915) reported that the initial speed of word reading was almost twice as fast as color naming but that both improved to about the same extent with practice when improvement was expressed as the percentage of decrease in time scores. Subsequent studies have not agreed with Brown's conclusion on this point. Color naming generally benefits more from practice than does word reading, regardless of whether improvement is measured on an absolute or on a relative basis. JENSEN (1965), for example, found 23 per cent improvement for color naming and only 15 per cent improvement for word reading over the course of 10 administrations.

Despite the significant improvement in color naming with practice, individual differences in color naming speed show remarkably little interaction with practice; *S*s maintain pretty much the same rank order at every stage. The first systematic investigation of this point was carried out by GATES (1922), who administered a color naming task (200 color patches) 25 times to each of 23 women students. The mean time on the first three tests correlated .72 with the mean of the last three, while the mean of the second set of three tests (i.e. administrations 4, 5, and 6) correlated .90 with the mean of the last three. The largest part of the practice effect occurred between the first trial and the median of the next three trials. Beyond the first three trials the intercorrelations among the subsequent trials are all over .90. The remarkable stability of performance on this test is shown by the correlations between the first trial, which is the least reliable, and the median performance on each of the subsequent 8 sets of three trials each; the correlations were .78, .74, .85, .75, .77, .79, and .72, respectively.

STROOP (1935b) was the first to investigate the effects of practice on all three of the cards simultaneously. Performance on the *CW* card improved most with practice over eight trials; card *W* showed the least effect of practice. STROOP's (1935b) results are in very close agreement with JENSEN's (1965), shown in fig. 1.

On every card the practice effects shown in fig. 1 are significant well beyond the .001 level, though their absolute magnitude is quite small, particularly for word reading. Most of the practice effect occurs within the first few trials. SMITH and NYMAN (1959) also found that performance became more or less asymptotic after five trials. Stroop found

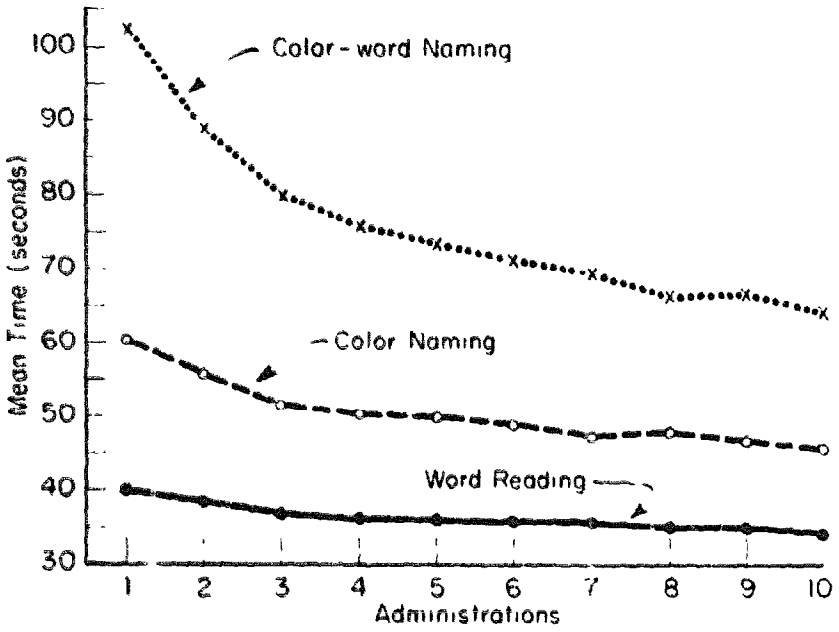

Fig. 1. Mean time for each Stroop card as a function of the number of administrations (Jensen, 1965).

a slight increase in the coefficient of variability on *CW* performance and a similar trend is apparent in Jensen's data, although the absolute amount of intersubject variability decreases slightly with practice. The mean intercorrelation among the ten trials is .86 for *W*, .86 for *C*, and .84 for *CW* (JENSEN, 1965).

Since in all the studies we have reviewed practice effects were measured by administering the form of the test repeatedly, it is not known exactly how much of the improvement is due to increasing familiarity with the particular sequence of responses on each card and how much is due to actual improvement of the abilities we are mainly interested in measuring with the Stroop. As was previously suggested, the construction of truly equivalent forms of the test is a necessary condition for the rigorous assessment of practice effects on the Stroop. The boosting of reliability by means of repeated measurements could conceivably be even greater by the use of equivalent forms rather than repetitions of the same form, since in the former case individual differences in the learning of specific sequences, which is only error variance as far as Stroop performance is concerned, would be ruled out.

One point of interest concerning the effects of practice has been noted by SMITH and NYMAN (1962), who have measured performance at five stages in the course of a single trial and have done this over

several trials. This is the fact that a systematic practice effect does not show up during the course of a single trial but only from one trial to the next. It is as if a reminiscence effect occurred between trials. If this is indeed a true reminiscence phenomenon it should be possible to increase the effect by having *S*s respond continuously to, say, 200 or 300 items, rather than merely 100, on each trial.

The closest any study has come to these conditions is that of BILLS (1931), who was the first investigator to study silent intervals or "blocks" during the course of color naming. A block was defined as a pause in responses equivalent to the time taken for two or more average responses. Bills administered a color naming task consisting of six colors; *S*s had to name the color patches as rapidly as possible for 10 minutes. Responses were recorded by *E*'s tapping a Morse key —a far from satisfactory method for measuring the durations of blocks. It was found that the speed of color-naming correlated —.33 with frequency of blocks and —.70 with the duration of blocks. Over the 10-minute period of color-naming the mean number of responses per minute decreased from 96 to 82 and the number of blocks per minute increased from 3 to 4; the length of the blocks however showed no systematic change over the 10-minute period.

Just what improvement with practice actually consists of is not known, but it is known that *S*s can adopt various strategies which can enhance performance. We have already noted that squinting or deaccommodating can make performance easier on card *CW*. And KLEIN (1964) found that the *CW* card is made easier for *S*s if they overtly read the word before naming the color of the ink. *S*s were required to say aloud first the word then the color. When the time for words alone (card *W*) was subtracted from the double response time (i.e. the time for saying the word then the color), the time for naming the colors on card *CW* was found to be significantly less under the double response condition than under the usual, single response condition. Also, *S*s reported that the task seemed subjectively easier when they could "release" the printed word overtly before naming the color. Interestingly enough, the reverse condition (i.e. being required to name the color and then say the word) made for greater difficulty on card *CW*.

A similar phenomenon was discovered by ROUSE and MAYER (1961), who increased the speed of performance on the *CW* card by having *Ss* use separate response channels for color naming and word reading, such as pressing one of three keys labeled *yellow, red,* and *blue* in

response to the words while naming the incongruous colors of the inks. When *Ss* are permitted to make both of these responses overtly, response competition and interference delay are markedly lessened. Practice under these conditions, Rouse and Mayer found, results in positive transfer when the Stroop is administered in the conventional manner.

FRASER (1963) was able to improve performance on *CW* by means of "verbal satiation" of the color words, which presumably reduced their response strength through inhibition and consequently their competitive interference with the colors. *Ss* simply repeated each of the four color words as fast as possible for 15 seconds before taking the *CW* test, on which they then performed significantly ($p < .02$) better than a control group which had "satiated" on the words "dog", "hand", "bridge", and "doctor."

THEORIES OF THE PHENOMENA

Colors versus words

Why does one take more time to name colors than to read color names? Though interest in this question dates back at least as far as JAMES MCKEEN CATTELL's (1886) investigations of the problem, no generally accepted theory of the phenomenon has been proposed. CATTELL (1886) and, later, STROOP (1938) demonstrated that the phenomenon extends beyond color naming. It also takes more time to name common objects than to read the names of the objects. In fact, one study showed that the differences between naming and reading is slightly greater for objects than for colors (STROOP, 1938).

Several rather similar theories have been proposed to account for this general phenomenon. The most common explanation is that of differences in amount of practice in color naming and word reading. CATTELL (1886), GARRETT and LEMMON (1924), LUND (1927), PETERSON (1918), PETERSON, LANIER, and WALKER (1925), and STROOP (1935a, b, 1938) all offered this explanation with slight variations. Adults do not spontaneously react to every object or color they see by giving its name, while the mere act of recognition of printed words implies a covert, if not overt, verbal response. Consequently the habit strength for responding verbally to printed words is presumed to be greater than the habit strength for verbally responding to objects and

colors. PETERSON, LANIER, and WALTER (1925) and STROOP (1938) suggested further that only one particular and dominant response habit is associated with each word, while objects and colors are associated with a variety of response tendencies only one of which is the specific naming response. Thus the naming response supposedly suffers greater response competition than the reading response.

WOODWORTH and WELLS (1911) offered an explanation which does not emphasize the long-term practice effects of color naming and word reading but rather the response competition generated in the task itself. "The real mechanism here may well be the mutual interference of the five names, all of which, from immediately preceding use, are "on the tip of the tongue", all are equally ready and likely to get in one another's way" (1911, p. 52). Since it is not stated why this mutual interference should be greater for color naming than for word reading, this explanation seems to miss the crux of the matter.

WARNER BROWN (1915) was the only radical dissenter from the differential practice theory. While he did not ignore practice as a contributing factor, he believed it was not the essential basis of the phenomenon, which he regarded as a more profound aspect of brain functioning than could be explained in experiential terms. "... the association process in naming simple objects like colors is radically different from the association process in reading printed words" (1915, p. 34). This conclusion seems to be supported by the fact that *S*s who are given a great amount of practice (up to 100 administrations of the *C* and *W* cards) seem to attain quite different asymptotes on color naming and word reading. It is possible that much more extensive practice than anyone has yet attempted could bring about a gradual convergence of the performance levels on naming and reading. Such an experiment would seem necessary for further evaluation of the differential practice hypothesis.

Experimental investigations of the phenomenon have yielded other interesting facts. LUND (1927), for example, presented plates containing either colors, color names, or geometric forms. One set of tasks consisted of scanning the plates as rapidly as possible in order to find and name all the items in a single class. This latter procedure minimizes the difficulty on the response side but apparently increases the difficulty on the perceptual side. Here are Lund's results in terms of mean seconds per item:

	sec.
Reading a word (color)	.36
Naming a color	.56
Naming a form	.80
Finding a word (color)	1.00
Finding a color	.56
Finding a form	.73

The interesting point here is that it takes almost twice as long to find a word as to find a color. (Essentially the same finding was earlier reported by CATTELL (1886).) So why should color naming be more difficult than word reading? These results suggest that the answer is probably not to be found in terms of the relative speeds of the perceptual or recognitive processes for words and colors, which seem to be just the reverse of what would be needed to explain the difference between the speeds of color naming and word reading.

GARRETT and LEMMON (1924) used a similar procedure to sort out the relevant factors. *S*s were presented a plate of randomly ordered color patches and had either to name them in succession or scan the rows and name only all the patches of one color then of another color, and so on. Performance on this color-finding task was 16 per cent faster than color naming. The ratio of the time for color-naming/color-finding was called an "index of interference". In addition, Garrett and Lemmon used letter and word cancellation tasks to measure speed of recognition, and they also had *S*s read a series of two-digit numbers to measure speed of speech.

Three factors were hypothesized to account for individual differences in speed of color naming. (a) Speed of recognition (measured by the cancellation tests), (b) speed of speech (measured by the 2-digit task), and an interference or inhibition factor (measured by the color-naming/color-finding ratio). The criterion—color naming—correlated significantly with all three measures and the multiple R was .88. The relative weights of the three factors in the regression equation were .58 (interference), .36 (speed of recognition), and .06 (speed of speech). Unfortunately this study does not tell us to what extent the "interference" factor would be found in a parallel word-reading test. JENSEN's (1965) factor analysis of the Stroop indicates that the "interference" hypothesized by Garrett and Lemmon as accounting for difficulty in color naming is decidedly not the same kind of interference that is generated on the incongruous *CW* card. In the principal components

analysis the *C* card had a loading of only .37 on the interference factor defined by the CW test.

LIGNON (1932) proposed a 3-factor theory of the phenomenon which is both obvious and unenlightening, the three factors being color-naming, word-reading, and a common factor.

Color-word interference

Relatively little doubt or disagreement exist concerning the nature of *S*s' behavior on card *CW*. Furthermore, the *CW* phenomenon throws some light on the nature of the difference between color naming and word reading.

When the color-word is incongruous with the color of the ink in which it is printed it is almost twice as difficult to name the color of the ink as when the ink is presented merely as a color patch. Though there are reliable individual differences in the magnitude of this phenomenon, apparently all literate persons are subject to it. Not a single one of the more than 400 *S*s in JENSEN's (1965) study, for example, was able to name colors on card *CW* as rapidly as on card *C*, even after 10 days of practice. The difference between *C* and *CW* shows up not only in the large difference in time scores but in various behavioral manifestations as well. It was largely for this reason that THURSTONE and MELLINGER (1953) regarded the Stroop as an alternative to the "stress interview", a situation in which the *S* is embarrassed, annoyed, or frustrated by the examiner or by other outside forms of disturbance, the object being to see how the *S* reacts to various forms of stress. The gross behavioral effects of card *CW* as compared with cards *W* and *C* were inadvertently impressed upon the first writer when he occasionally entered a room adjoining the sound-proof laboratory in which *S*s were tested to observe the procedure through a one-way-vision window. Since the Stroop cards were out of view through the one-way window, it was rarely possible to detect from the *S*'s behavior whether he was reading card *W* or card *C*; except for the difference in speed, there is little difference in *S*s' behavior on these two cards. In marked contrast, there is seldom any difficulty in telling when *S*s are responding to card *CW*. They become more tense, they strain forward, they take on the expression of eyestrain, they gesture with the arms and hands, and occassionally they stamp their feet. Exaggerated vocal emphasis is also characteristic. A few *S*s even break down with laughter and the test has to be given again to obtain

a fair score. On retesting *S*s did not hide their displeasure at the prospects of having to face *CW* again, and they usually heaved a sigh of relief when the test was over. Repeated retesting decreases these overt signs of stress, though *S*s never come to regard the *CW* task with the same bored equanimity that they finally show toward cards *C* and *W*.

CW obviously has all the essential characteristics of a conflict situation, and all investigators since JAENSCH (1929) and STROOP (1935b) have interpreted the *CW* phenomenon strictly as an interference effect due to response competition between habits of unequal strengths, the stronger habit (word reading) having to be inhibited in in favor of the weaker (color naming). The fact that the habit strength of word reading is dominant over that of color naming is indicated, not only by the *C–W* difference, but also by STROOP's (1935b) finding that reading the *words* on the *CW* card suffers no appreciable interference from the incongruous colors and is practically as easy as reading the words on card *W*. Stroop found an average increase of 5.6 percent in time for reading the words on *CW* as compared with *W*. On the other hand, there was a 74.3 percent increase in color naming time on *CW* as compared with *C*. This fact is a convincing demonstration of differential response strengths for colors and words. Practice or repetition is, of course, known to increase response strength, and so it is the favored explanation for the dominance of word reading over color naming. This hypothesis also results in predictions concerning the interaction of color-word conflict with age and reading ability, which is examined in the following section.

STROOP (1938) tried to test the differential practice theory by having a group of 20 *S*s practice giving nonsense syllable names to five unfamiliar symbols; each *S* had to make 1200 such naming responses. A control group made only 200 naming responses but continued increasing their familiarity with the symbols by making a total of 1000 other responses involving the symbols, such as sorting, checking, etc. Both groups were then tested for naming speed for 200 items. As one would expect, groups with more practice in naming performed much faster than the control group (.57 vs .81 sec per item).

While apparently all investigators regard *CW* as an interference phenomenon, they have not all couched their interpretations in terms of S–R theory, and some have attributed broader psychological meaning to the phenomenon than we have indicated in our discussion thus far. COMALLI, WAPNER, and WERNER (1962), for example, regard *CW*

performance as reflecting a general capacity to maintain a course of action in the face of intrusion by other stimuli. This ability to resist interference is related to a basic principle of cognitive development which has been of central interest to Werner and the Clark University group, viz. the interpretation of cognitive development as increasing differentiation and hierarchic integration. This "organismic-developmental" approach has given rise to a great deal of Stroop research, which is reviewed in later sections. Another member of the Clark group, Broverman, refers in several of his publications to "verbalness" and "ego strength" as measured by certain Stroop indices. A highly verbal *S* should gain relatively more practice in word reading than in color naming and should therefore show a relatively high score on $(C–W)/W$. A high degree of such verbal specialization, as indicated by this Stroop index, is considered an indicator of obsessional or anal personality organization. Ego strength, referring to the control and regulation of response, is reflected in the ratio $(CW–C)/C$.

LANGER and ROSENBERG (1964) have discovered an interesting phenomenon closely related to the Stroop which is probably best regarded as a type of semantic generalization, though its discoverers believe it does not easily lend itself to interpretation within an S–R framework. Langer and Rosenberg presented *S*s with 25 "sonic symbols" (e.g. mumle, skat, zab, oom, tut, verd, sool, and klak), which had no obvious semantic or structural resemblance to color names, and asked the *S*s to classify them subjectively as either red, blue, green, or yellow. *S*s who were in high agreement with the modal classifications were called "concensualizers", those in low agreement were "non-concensualizers". *S*s in each of these groups, in addition to a control group which did not take part in the classification, were then given a test analogous to the *CW* Stroop card, but in which the sonic symbols were printed in the concensually "incongruous" colors. There was also a "congruent" card in which the symbols were matched with their concensual colors. Color naming on the incongruous card was significantly ($p < .01$) slower than for the congruous card for all three groups. Consensualizers showed the greatest interference and non-consensualizers the least, with the controls intermediate.

KLEIN (1964), in one of the most interesting sets of experiments ever performed with the Stroop test, demonstrated a kind of semantic gradient of the capacity of words to interfere with the color-naming response. Six analogues of the *CW* card were made up in which the

printed verbal units consisted of (a) nonsense syllables, (b) rare words, (c) common words, (d) color-related words, (e) distant color names, i.e. color names which were different from any of the colored inks on the card, (f) close color-names, i.e. the usual *CW* condition. The control card (card *C*) consisted of sets of asterisks printed in different colored inks. The amount of interference, as measured by the increase in total response time over that required on the control card (card *C*), was significant even for the nonsense syllables, and it increased as a positively accelerated function of the semantic gradient represented by the order of the conditions listed above. The closer the semantic relationship between the required response (i.e. color-naming) and the competing response which has to be held in check (i.e. the verbal units), the greater is the interference and the slower is the response time

Relationship to other variables

Age

Because of the prominence of differential amounts of practice in color naming and word reading as an explanation of the Stroop phenomena, age has been one of the most extensively studied independent variables in relation to Stroop performance. The results of the various studies show a high degree of agreement.

Before the Stroop test was invented there was an interest in the relative dominance of color and form as a function of age. Descoeudres (1914) had groups of *S*s from 3-year olds to adults perform card-sorting tasks involving either colors or simple geometric forms as well as the printed names of colors. Speed of sorting increased with age for all types of materials, but there was a strong age *X* materials interaction: young children sort faster on the basis of color than of form, while the reverse is true for older children and adults. The same is true for color vs. color words. The change from color to form or word dominance comes on the average between six and seven years of age. Brian and Goodenough (1929) investigated color-form matching as a function of age from two years of age to adulthood, with total $N = 474$. The task allowed *S*s to match cards on the basis either of color or of form, but not both. Strangely enough, below the age of three matching was based predominantly on form; from three to six years of age color matching was predominant; and from six years to adulthood form increasingly predominated over color.

LIGON (1932) made the first investigation of color naming and word reading as a function of age. His *S*s were 6 to 18 years of age. He found a progressive improvement in speed of both color naming and color word reading, but noted that the absolute difference in time scores (i.e. number of seconds per 100 items) between colors and words was approximately constant across all age groups. He therefore concluded that differential practice was not an adequate explanation of the difference between color naming and word reading. STROOP (1935a) criticized Ligon's interpretation on the basis that he should have compared the age groups in terms of the *relative* rates of responding to colors and words rather than in terms of the absolute difference. When performance on *C* and *W* are compared on the basis of number of reactions per 100 seconds, rather than the total time for 100 reactions, the difference between *C* and *W* increases in the ratio of 1 to 4 from the age of 6 to 18. Thus the increase in relative superiority of word reading over color naming would seem to support the differential practice hypothesis.

The most comprehensive investigation of age involving all three basic Stroop scores is that of COMALLI, WAPNER, and WERNER (1962), which spanned the age range from 7 to 80. Five- and six-year olds were excluded, since preliminary investigation of this age group led the investigators to conclude that below seven years of age reading ability was not sufficiently established to serve as a potent factor of interference on the *CW* card; presumably there is less interference on *CW* below the age of seven. (Unfortunately there are no published results on *CW* for *S*s under seven years of age.) The results of the study by COMALLI et al. are summarized in fig. 2. These curves are in close agreement with the results of other investigations which have studied only various segments of the age range represented in fig 2 (LIGON, 1932; LAZARUS, 1955; GARDNER, et al., 1959; RAND, WAPNER, and WERNER, 1962; UHLMANN, 1962b; LEEDY, 1963).

COMALLI et al. explain the age changes on the Stroop, particularly on *CW* and the relative difference between *CW* and *C*, in terms of Werner's organismic-developmental theory, with its emphasis on the increase of perceptual and cognitive differentiation and hierarchic integration with increasing maturity. Older *S*s (over 50 or so) show some regression in differentiation and integration of functions. Though the time scores of children and of old adults are similar, COMALLI et al. note that the young and the old achieve their scores by somewhat

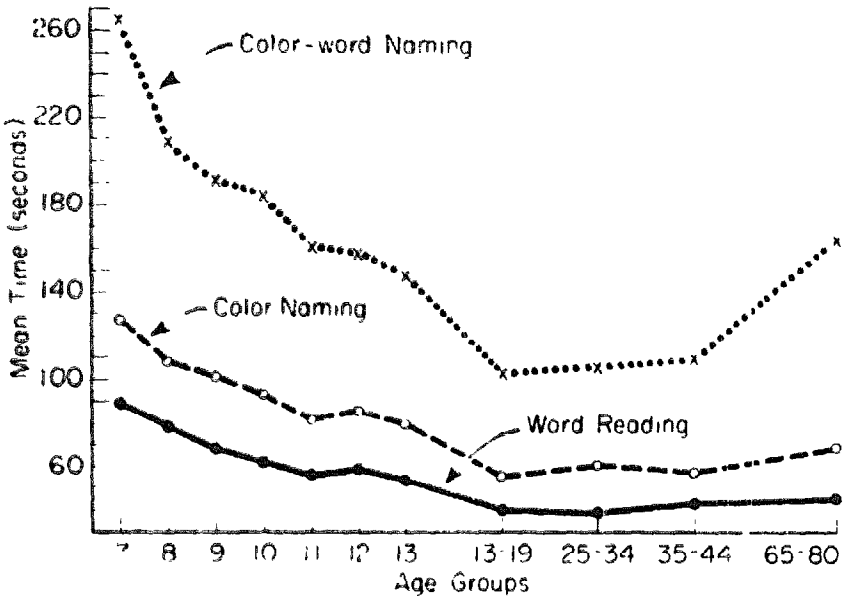

Fig. 2. Changes in Stroop performance from 7 to 80 years of age. (From table 2 in Comalli, Wapner, and Werner, 1962).

different means. Children try to overcome interference by pointing the finger at the items and by rhythmically accenting their verbal responses. In contrast to this concrete pointing, the aged *S*s sometimes used "verbal pointing" by preceding each response with "that's a . . ."

In addition to time scores, RAND, WAPNER, WERNER, and MCFARLAND (1963) obtained a variety of error scores (see "error scores" in section on Scoring Methods) in four age groups: 6, 9, 12 and 16 years of age, with 10 *S*s in each group. The total of these deviations from an ideal performance was least for card *W* and greatest for card *CW*, with card *C* intermediate. Most of the deviant behavior scores showed a significant interaction with age although the form of the interaction is not the same for all scores. Five of the error categories (inappropriate color responses, contaminated responses, inserted color words, and omissions) showed a general decrease in frequency with age; one category (inserted nonlinguistic utterances) showed an increase with age; and none category (inarticulate utterances) decreased in frequency with age and then increased in the oldest group. The error categories were interpreted in terms of two sets of processes hypothesized as underlying Stroop performance: the "Process of Identification" of the appropriate aspect of the stimulus item and the "Process of Serial Organization" of the responses. The interaction of these processes with age was interpreted in terms of organismic-developmental theory. Since this

very fine-grained analysis was made possible by tape recording *Ss*' performance, the same type of analysis could be refined still further by examining these variables in terms of the temporal progress of performance within each card. RAND et al. state that such an analysis is being undertaken and will be reported in a forthcoming study. Another finding from this experiment not reported in the article by RAND et al. is presented by WAPNER (1963). The duration of silent intervals or blocks during the *S*'s performance was measured; it was found that these intervals become shorter with increasing age over the range of 6 to 16 years, and this was true for all three Stroop cards.

SMITH and NYMAN (1962) have presented evidence on the serial scoring of the Stroop *CW* performance in relation to age over a range from 12 to 60. In brief, total variability of within-subject performance and the linear component reflecting within-card improvement in the course of a single trial both decrease with age.

Conclusion. The trend shown in fig. 2 is not consistent with the simple theory which attributes differences in interference to differential response strengths of color naming and word reading as a function of previous practice. The dominance of words over colors, which is presumably the cause of the interference on card *CW*, should be expected to result in increasing interference with increasing age. But just the reverse occurs. Performance on *CW* improves with age, and indices of interference such as *CW–C* and *CW/C* show a decrease with age up to about age 60, when they begin to increase. There is, of course, some evidence from experiments on verbal rote learning that interference effects, such as associative interference and retroactive and proactive inhibition, first increase and then decrease as a function of the degree of learning of the competing tasks (e.g. POSTMAN and RILEY, 1959). Overlearned verbal habits do not seem to result in as much interference in proactive and retroactive paradigms as do habits learned to only a moderate degree. But the degrees of learning or overlearning involved in these experiments are of quite a different order than the degree of learning of color names, etc. involved in the Stroop test, and there are many other obvious differences in the two sets of conditions as well. This explanation of the decline in Stroop interference with age therefore must be regarded as quite gratuitous. The only serious attempts to explain this phenomenon are those of Comalli et al. and Rand et al., which invoke developmental rather than learning principles. A critique of this particular approach, which involves a

whole theory of cognitive development, is beyond the purpose of this review.

Sex

The study of sex differences on the Stroop has always been incidental to other variables. Only one fact regarding sex differences is quite certain; girls and women are better than boys and men at color naming (i.e. card *C*); this has been found in every study in which the sexes were compared (WOODWORTH and WELLS, 1911; BROWN, 1915; LIGON, 1932; STROOP, 1935; JENSEN, 1965). JENSEN (1965) found a biserial correlation between sex and color factor score derived from the Stroop of .23, $p < .01$. There was no significant correlation between sex and either the speed (card *W*) or interference (*CW–C*) factors. Significant sex differences on *W* and *CW* have not been found in any study.

STROOP (1935) attributed female superiority in speed of color naming to differential practice in color naming, interest in and responsiveness to colors presumably being a feminine trait. In support of the differential practice notion Stroop notes that the sex difference diminishes with practice in color naming, but too few relevant facts are known to permit evaluation of this hypothesis.

In a factor analytic study involving the Stroop test, GARDNER, HOLZMAN, KLEIN, LINTON, and SPENCE (1959) thought it advisable to do separate factor analyses for men and women, on the grounds that the two sexes produced somewhat different intercorrelations among Stroop scores and some 30-odd other variables. But it is not unlikely that these differences are unreliable, since there were only 30 *Ss* in each group and few of the correlations involving the Stroop scores were significantly greater than zero; even fewer of the correlations showed significant sex differences. The picture is further obscured by the fact that in this sample sex was significantly correlated with age, a variable which among men correlated significantly (.37 $p < .05$) with amount of time taken on card *C*.

Race

PETERSON, LANIER, and WALKER (1925) compared groups of 10 and 12 year old white and Negro children on *C* and *W* tasks involving five colors and 250 responses per card. At ten years of age the Negro children were significantly ($p < .001$) slower than white children on

both *C* (by 20 percent) and *W* (by 19 percent). At 12 years of age, however, the Negro-white differences, though in the same direction, were minute (6 to 8 percent) and insignificant. This change was explained in terms of assumed differences in the learning histories of Negro and white children with respect to reading and color naming which could cause Negro and white children to attain more or less the same asymptotic level of performance at different ages. TELFORD (1930), one of Peterson's students, compared groups of white and Negro college students on 10 administrations of the *C* and *W* cards. There was no significant race difference overall or on any trial for either *C* or *W*. The interpretation of these findings is obviously limited by the well-known difficulties and shortcomings of sampling and methodology that vitiate so many studies of racial differences.

Effects of drugs

The Stroop test seems to be a favorite device of investigators studying the effects of various drugs on behavior. In some cases the use of the Stroop is clearly indicated by the nature of the hypotheses under investigation, while in other cases the Stroop seems to have been included quite arbitrarily among a collection of other psychological measurements, probably in the hope that one test or another might show the drug effect. The results of these studies do not appear very enlightening. In general, stimulant drugs improve performance on all Stroop cards and decrease interference measures, while depressants and psychotomimetics (viz. LSD) have the opposite effect.

First, those studies which at least have the virtue of testing hypotheses to which both the Stroop and the particular drugs used seem relevant and appropriate:

Callaway and various co-workers have hypothesized a psychological continuum called "narrowed attention", which is a response to certain psychophysiological changes (CALLAWAY and DEMBO, 1958). Stimulant and sympathomimetic drugs are hypothesized to induce greater focusing and "narrowness" of attention, with a reduction in the *S*'s sensitivity and responsiveness to peripheral or extraneous, distracting stimuli. Depressant drugs, on the other hand, have the opposite effect, that of "broadening" attention and increasing the *S*'s responsiveness to a broad influx of environmental stimuli. Since "narrowed attention" implies enhanced ability to select relevant and to ignore peripheral or irrelevant stimuli, the Stroop interference score (*CW*–*C*) should be

expected to decrease under a stimulant drug and to increase under a depressant drug. The appropriate experiment was performed on three groups of college students: a depressant drug (amobarbital), a stimulant drug (methamphetamine), and a placebo. *S*s were given the Stroop before taking the drug and again one hour after taking the drug. As compared with the placebo group, the amobarbital group showed an increase in interference while the methamphetamine group showed a decrease in interference (CALLAWAY, 1959). CALLAWAY and STONE (1960) reanalyzed these data using a derived score $CW + C$ rather than $CW - C$, and found a somewhat larger improvement effect of the stimulant drug. Rather complicated theoretical reasons were given for this finding; a simpler explanation lies in the fact that the stimulant drug improves performance on both *C* and *CW* and the sum of these two scores is more reliable than their difference. The addition of *W* to the score would probably have shown still higher statistical significance for the drug effect, but we wouldn't know much more. Since it would seem more important to know how these drugs affect each of the Stroop factors separately, $CW - C$ is a more defensible score than $CW + C$.

In two other studies atropine was used instead of methamphetamine with similar effects, although the results were more ambiguous, since atropine has both depressant and stimulant properties depending on the dosage and individual differences in reaction to a particular dose; furthermore, atropine produces visual side effects which probably affected Stroop performance (CALLAWAY and BAND, 1958; CALLAWAY and DEMBO, 1958). QUARTON and TAILAND (1962) repeated CALLAWAY's (1959) experiment with slight modifications, using pentobarbital and methamphetamine. These drugs in the dosages used produced no significant effects on Stroop interference, although the amphetamine increased *S*s' memory span for digits and pentobarbital had the opposite effect.

The one other theoretically oriented drug study is by WAPNER and KRUS (1960), whose hypothesis was derived from a comparative-developmental theory which analyzes psychological phenomena and processes in terms of the concepts of differentiation and hierarchic integration. On the basis of previous studies, performance on *CW* was taken as a measure of capacity to differentiate and organize responses in accordance with the instructions on the *CW* task, which requires the suppression of stronger response tendencies (word reading) in favor of weaker tendencies (color naming). It was hypothesized that children,

schizophrenics, and normal adults when under the influence of a psychotomimetic drug (LSD-25) would all show developmentally "less mature" behavior on *CW* performance than normal adult *S*s. This study compared normals and schizophrenics under LSD and placebo conditions. On cards *W*, *C*, and *CW* normals performed significantly better than schizophrenics under both placebo and LSD condition LSD significantly impaired performance on cards *W*, *C*, and *CW* in both normals and schizophrenics.

Stimulants. Caffeine improved speed of color naming (*W* and *CW* not used) (HOLLINGWORTH, 1912). Adrenalin had no significant effect on Stroop scores, though the dosage was sufficient to affect performance on certain motor tasks (BASOWITZ, KORCHIN, and OKEN, 1955). The performance of drug addicts and of normal controls were compared under placebo and under d-amphetamine: this drug improved Stroop *CW* performance only among the addicts (LEHMAN and KNIGHT, 1961). The Stroop was included among a battery of tests in an investigation of imipramine hydrochloride, a stimulant used in the treatment of depressive states, but no Stroop results were reported, which probably means nothing significant was found (GERSHON, HOLMBERG, MATTSON, MATTSON, and MARSHALL, 1962). Phenylephrine, an amphetamine with effects similar to those of benzedrine, was given to *S*s after they had practiced the Stroop to a point where a plateau was reached; there was no significant effect of the drug on the interference score *CW* — *C* (OSTFELD and ARUGETE, 1962).

Depressants. Alcohol impaired speed and accuracy of color naming in direct proportion to the dose (HOLLINGWORTH, 1923; MCFARLAND and BARACH, 1936). Scopolamine (hyoscine) significantly increased interference as measured by *CW* — *C* (OSTFELD and ARUGUETE, 1962). Secobarbital impaired Stroop performance (exact measures not given) in both drug addicts and normal controls (LEHMAN and KNIGHT, 1961).

Cognitive styles

The Stroop has been used to select criterion groups representing various dimensions of "cognitive styles". "Cognitive style" is a generic term for the distinctive ways in which individuals come to grips with reality. BROVERMAN (1960b) conceives of cognitive styles as representing manifestations of different response probabilities or response strengths in certain classes of behavior. Certain cognitive styles are operationally defined by scores derived from the Stroop. Criterion

groups are obtained by selecting a given percentage of *S*s from the top and bottom of the distribution on a particular score, and these groups are then compared on various cognitive, motor, or perceptual tasks as a means of testing the investigator's hypotheses concerning the behavioral manifestations of a particular cognitive style. There have been three more or less systematic programs of research along these lines. The dimensions of cognitive style, as defined by particular Stroop scores, afford a framework for summarizing some of the findings regarding the relationship of the Stroop to other variables.

Sensorimotor versus conceptual dominance. This dimension of cognitive style has been defined by BROVERMAN (1960b) and is measured by comparing the *S*'s speed on *W* to his speed on *C*. Different highly intercorrelated indices have been used in various studies: $C - W$, $(C - W)/W$, C/W, and $C - Cp$, where Cp is the predicted value of C from its regression on W. This dimension has also been referred to in some studies as "verbalness". A high C/W ratio indicates conceptual dominance; a low C/W ratio indicates sensorimotor dominance. Variations in this index are asumed to be due to differences in the amount of learning the individual has accumulated over his lifetime in responding to conceptual and to sensorimotor stimuli. Conceptually dominant *S*s appeared to have little interest in sensorimotor tasks, such as painting or judging distances, as compared with sensorimotor dominant *S*s (BROVERMAN and LAZARUS, 1957, 1958). Conceptually dominant *S*s were also found to be faster and more careless in painting geometric designs than were sensorimotor dominant *S*s, and when conceptual *S*s were asked to listen to a recorded passage of prose and then write down what they heard, they tended to paraphrase the content of the passage, while sensorimotor *S*s made a more literal transcription (LAZARUS, BAKER, BROVERMAN, and MAYER, 1957). The two types did not differ in arithmetic ability except under distraction (a voice reading figures through earphones), where conceptually dominant *S*s did beter. When the distractor was a sensorimotor task (disjunctive reaction time), however, the sensorimotor dominant *S*s did relatively better (BROVERMAN and LAZARUS, 1958). On more difficult arithmetic problems, a conceptual group was superior to a sensorimotor group both with and without simultaneous distraction, while on a perceptual-motor task (tracing a difficult pattern) sensorimotor *S*s showed less performance decrement due to distraction (holding a rotating handle with the other hand) than did conceptual *S*s (BROVERMAN,

1960a). AMSTER (1965) gave *Ss* a concept attainment task under instructional sets for either incidental or intentional learning of the concept. Conceptual *Ss* (i.e. high *C/W* ratio) showed significant superiority of intentional over incidental learning, while sensorimotor *Ss* (low *C/W* ratio) were not significantly affected by this particular instructional set.

Automatization and constructed versus flexible cognitive control. Automatization, BROVERMAN's (1960a) term for the tendency of certain acts to become automatic and to require little conscious effort or attention, is operationally defined by the same Stroop scores as KLEIN's (1954) concept of cognitive control. Strong automatization and flexible control, as contrasted with weak automatization and constricted control, are represented by low scores on the interference factor as measured by $CW - C$ or one of its equivalents (derived scores H, J, K, and L in table 1). To avoid confusion this dimension will henceforth be referred to simply as interference proneness.

Low interference (LI) *Ss*, as compared with high interference (HI) *Ss*, have shown a faster tapping rate, smaller handwriting, and were faster at making up three-letter words (BROVERMAN and LAZARUS, 1958). LIs did beter at mental arithmetic (simple addition) under distracting conditions (listening to a voice reading numbers) than HIs; and LIs also did better at tracing a line under motor distraction in which the other hand had to hold on to a rotating handle (BROVERMAN, 1960a). LIs were "over-achievers"; they attained a higher occupational level than HIs and had occupations further above those of their fathers, when rated on the education and occupation scales of Hollingshead, and this holds true when LIs and HIs are equated for intelligence and educational background (BROVERMAN, 1962). LIs also obtained higher college grade point averages than did HIs (BROVERMAN, 1963). HIs were relatively better at incidental learning than intentional learning, as measured by a recall test, while LIs showed the reverse (AMSTER, 1965). This finding seems to be consistent with CALLAWAY's (1958b) concept of narrowed and broadened attention which is also reflected in Stroop interference; narrowed attention would favor intentional learning at the expense of incidental learning, while broadened attention should have the opposite effect. Groups of HIs and LIs which were dichotomized into high and low "need for independence" as measured by the Edwards Personal Preference Scale, were compared on the Wechsler–Bellevue block design test while performing under

one of three conditions: nonstressful, threat of electric shock, and negative verbal reinforcement. Differences between the groups showed up only under the stress conditions, in which the high independence LI group had the best performance and the low independence HI group had the worst (HARDISON and PURCELL, 1959).

In the perceptual realm, comparisons of extreme groups on the interference dimension have yielded more significant results than correlational approaches involving the entire range of interference scores (e.g. THURSTONE, 1944). In a size estimation task in which the *S* had to adjust a spot of light to match a standard disc, LIs tended to overestimate and HIs to underestimate. LIs were also beter than HIs at tachistoscopic picture recognition; the LIs made wider visual scannings of the stimulus; but these perceptual differences between HIs and LIs showed up significantly only under a condition of high drive (thirst) (KLEIN, 1954). LOOMIS and MOSKOWITZ (1958) compared HIs and LIs in dealing with ambiguous stimulus situations, such as judging the point at which a subjective change occurs in a series of gradually changing pictures, and summarizing an ambiguously worded character sketch. It was concluded that LIs more than HIs tended to integrate the competing, overlapping, and contradictory elements in a stimulus situation, while HIs were more likely to keep apart intrusive or contradictory ambiguities. Cognitive rigidity as reflected in perseverative tendencies on the gradually changing picture series test, however, was not related to interference proneness.

Cumultatives, Dissosiatives, and Stabilizers. This typology is based on the method of serial scoring originally proposed by SMITH and KLEIN (1953) and described earlier in this article in the section on scoring. In brief, performance time on the *CW* card is measured five times—after every 20 responses—and these five time scores form some kind of pattern for each *S*. Three main types of pattern are discernible and form the basis of this typology. The *Dissociatives'* curve rises and falls discontinuously; this is claimed to reflect a faltering of the attention required in the process of isolating the relevant stimulus. The *Cumulatives'* curve tends toward increasingly slower reading time over the five time scores; these *S*s show continuously aggregating difficulties throughout the performance. The *Stabilizers'* curve maintains an even course, remaining more or less horizontal over the five time scores; these *S*s are most adequate to the interference task.

The three types of *S*s have been found to differ on other psycho-

logical tests (SMITH and KLEIN, 1953). For example, in judging the size of squares, Cumultatives progressively underestimate throughout the course of the experiment; Dissociatives are inconsistent, sometimes overestimating and sometimes underestimating; and Stabilizers make the most consistent judgements, i.e. each *S* has relatively little fluctuation around his own mean. In a test requiring the detection of camouflaged faces in a larger picture, the Dissociatives fabulized more faces, i.e. saw faces where none actually existed. In the Gottschaldt test Cumultatives were slower than Dissociatives or Stabilizers in discovering the embedded figures, and Dissociatives showed greater variability in performance than Stabilizers. In the serial learning of four lists of pseudowords the Cumulatives and Dissociatives differed significantly on lists 3 and 4; the Cumulatives required more trials to attain criterion. Also, the Cumulatives produced a more irregular serial-position curve than the Dissociatives or Stabilizers; the Cumulatives' serial-position curve tended to break in half, as if the serial list had been learned as two shorter lists.

Perceptual-motor correlates

Though the evidence regarding the relationship of the Stroop to various perceptual abilities is rather inconsistent, it at least affords the conclusion that the relationship is a highly tenuous one. The Stroop is clearly more a cognitive than a perceptual task, and where significant correlations with perceptual tests are found there is usually some cognitive or problem-solving aspect that can be discovered in the perceptual test.

THURSTONE (1944) included a number of Stroop scores in his factor analytic study of perceptual abilities, which involved 58 other perceptual measures. The correlations between the Stroop and the other perceptual tests were all so low that Thurstone excluded the Stroop from his factor analysis. He later stated, "In view of the highly specific and unknown nature of the variance of the (Stroop) test, it should prove to be an interesting test to investigate . . ." (1953, p. 2). Other investigators have been only slightly more successful in finding perceptual-motor correlates of the Stroop. PODELL and PHILLIPS (1959) performed a cluster analysis of a battery of 20 tests of motor, perceptual, and conceptual abilities, including the Stroop. Various Stroop scores had appreciable loadings on some of the factors, but the factor loadings of the various tests were quite inconsistent in two different

samples, so that the results of this study, at least as regards the Stroop, are practically impossible to interpret. It does show, however, that the Stroop shares in some of the common factor variance in a large battery of perceptual and cognitive tests. The one other large-scale correlational study involving perceptual tests is that of GARDNER, HOLZMAN, KLEIN, LINTON, and SPENCE (1959). They correlated Stroop Interference and Color naming with 30-odd perceptual measures, including apparent motion thresholds, the rod and frame test, recognition time, size constancy, size estimation, and embedded figures. The pattern of Stroop correlations was quite different for men and women. The interference score had no significant correlations with the other variables among the men, but correlated significantly (between .37 and .63) with several of the tests (size estimation, rod and frame, embedded figures, and a field-dependence index) among the women. In all cases, "inferior" performance on a particular perceptual test was directly related to "inferior" Stroop performance (i.e. greater interference). Color naming on the Stroop showed one significant correlation for the men (size constancy) and one for the women (size estimation). (There were a few other significant correlations which had nothing to do with perceptual variables and which are mentioned elsewhere in this review.)

Coming down to specific tests, we find that the perceptual task used most frequently in connection with the Stroop is the Gottschaldt embedded figures test. And as most of the investigators had predicted, this test bears some slight but significant relationship to Stroop interference: low interference-prone *S*s do better on the Gottschaldt. CALLAWAY (1959) found a correlation of .38 ($p < .05$) between $CW - C$ and Gottschaldt solution time; UHLMANN (1962b) found a correlation of .36 ($p < .01$); GARDNER, et al. (1959) found a correlation of 54, ($p < .01$) for women only; nonsignificant for men.

Motor ability. This subject has understandably received scant treatment. But one finding is quite interesting: GATES (1922) found correlations ranging from .32 to .67, over 25 administrations of a color naming test, between speed of color naming and speed of tapping a stylus. The correlations increased with practice. This suggests that rate of color naming (or, more specifically, probably the speed factor, which is found in purer form in the word-reading test) and tapping rate both reflect some more general factor of "personal tempo", to use Thurstone's designation. Rate of color naming also correlated significantly with speed of performing simple arithmetic problems and with speed

of word building (making a number of shorter words from a longer word). The intercorrelations of these phenotypically diverse tasks strongly suggests a general "tempo" factor.

SMITH (1959a) found a relationship between rate of adaptation to the Stroop interference task and rate of adaptation to a mirror-drawing test, which is another type of perceptual-motor interference task. Certain scores derived from *Ss*' preformance on these two interference tasks showed low but significant intercorrelations.

Mental abilities and aptitudes

Intelligence. Stroop scores are only tenuously related to intelligence. There is no reported instance of the Stroop ever having been used with *Ss* much below the normal range of intelligence, and most studies have used college students. The obtained correlations, therefore, may be somewhat depressed by the restriction of range of ability. LIGON's (1932) study comes nearest to assessing the entire range of intelligence, but he used only cards *W* and *C*. The correlation between color naming speed and IQ was nonsignificant ($r = .02$), while there was a significant but low correlation (.15, $p < .01$) between IQ and speed of word-reading. In college-level samples, correlations have ranged from zero to .34. The Stroop interference factor ($CW - C$) had near-zero correlations with Raven's Prograssive Matrices in three studies (CALLAWAY, 1959; JENSEN, 1965; LEEDY, 1963). In one study (UHLMANN, 1962b) the interference factor correlated .34 ($p < .01$) with intelligence as measured by the Wonderlic Personnel Test, a general intelligence test used in business and industrial situations. Uhlmann, however, used a group form of the Stroop which may tap additional factors including general intelligence. Speed of *CW* performance, which of course contains the speed and color-difficulty factors as well as the interference factor, showed a significant correlation with Raven's Progressive Matrices (SMITH and NYMAN, 1962). The fact that $CW - C$ has not been found to correlate with intelligence suggests that whatever correlation might exist between *CW* and intelligence is not due to the interference factor. JENSEN (1965) found a correlation of .31 ($p < .05$) between the color-naming factor (W/C) and the Progressive Matrices; the interference ($CW - C$) and speed (W) factors, however did not show significant correlations with the Matrices in this study.

Memory span and short-term retention. JENSEN (1965) found significant correlations between Stroop factors and memory span measured

under various conditions among college students. The speed factor (*W*) correlated between .27 and .40 ($p < .01$) with digit span under various conditions. The interference factor (*CW* — *C*) correlated (—.28, $p < .05$) with digit span only when there was a 10 seconds delay interval between presentation and recall, during which the *S* engaged in a simple discriminative perceptual-motor task. The color factor (*C*/*W*) did not correlate significantly with digit span but correlated —.36 ($p < .01$) with memory span when the items in the series were colored forms. Thus there can be little doubt that Stroop factors are related to memory span, the speed factor being the most general and most potent and the interference factor showing up when some interfering activity is interposed during the interval between presentation and recall.

Learning and retention. In the same study (JENSEN, 1965) the Stroop factors were found to be significantly correlated with ability in serial learning. *S*s learned by the usual anticipation method. The speed factor (*W*) correlated .45 ($p < .01$) with total errors to criterion in the serial learning of words, .38 ($p < .01$) with trigrams, and .31 ($p < .05$) with color-forms. An index of oscillation tendency in serial learning, that is, the tendency for a correct anticipation on one trial to be followed by an incorrect or omitted anticipation on a subsequent trial, correlated with the speed factor (.31 for words, .40 for trigrams). The interference factor correlated significantly (.43) only with the serial learning of trigrams.

UHLMANN (1962) had *S*s listen to a tape-recorded account of a building explosion and then write down whatever details they could recall immediately after hearing the tape; a similar recall was requested after a delay of three hours. The recall was scored in terms of the number of anxiety and non-anxiety words recalled. Stroop interference correlated with immediate recall —.18 ($p < .05$) for anxiety words and —.15 (n.s.) for non-anxiety words, and, with delayed recall, anxiety and non-anxiety words both correlated —.19 ($p < .05$) with Stroop interference, i.e. high interference was associated with poor recall.

Personnel selection. KIPNIS and GLICKMAN (1962) investigated a group form of the Stroop for predicting evaluations of radioman performance among 125 naval aviation machinists mates. The men were categorized as below average (lower one-third) or above average (upper one-third) in performance. Biserial correlations between this dichotomy and Stroop cards *C* and *CW* were .14 and .11, respectively: both *r*s are nonsignificant. For specific criteria of job performance, however,

card *C* had correlations significant at the .01 level (*r*s between .25 and .30) for ratings on "maintaining equipment", "stays clam in an emergency", "gets along with shipmates", and "overall effectiveness". Card *CW* yielded correlations significant at the .05 level with "sound ideas", "military appearance", and "overall effectiveness". In all cases superior Stroop performance indicated superior criterion ratings. This study was replicated with 261 men entering the Nuclear Power School of the U.S. Navy, with essentially similar results (KIPNIS and GLICKMAN, 1961). Kipnis and Glickman point out that cards *C* and *CW* correlated .54 and .55, respectively, with the clerical aptitude test of the Navy Basic Test Battery. This is essentially a test of speed and accuracy, factors which are clearly involved in the group form of the Stroop used in this study. Kipnis believes that all the valid variance in this form of the Stroop is probably due to this clerical skill factor (personal communication, 1964).

THURSTONE (1944) found no significant correlations between any one of several Stroop scores and salaries of public administrators and personnel analysts.

Scholastic abilities. We have already mentioned the correlation of color naming speed with speed in simple arithmetic (addition and multiplication) and with word-building (GATES, 1922). THURSTONE (1944) tried to differentiate fast and slow readers (46 college freshmen) by means of Stroop scores *C*, *CW*, and *C/CW*. Fast readers were significantly ($p < .05$) faster at *C*, while the two other Stroop scores showed no significant relationship to reading speed. The *C* score is, of course, a mixture of the speed factor and the color-difficulty factor. The correlation of C with reading speed might well be due to its saturation on the speed factor, which is represented more purely by W. Unfortunately, Thurstone did not include *W* in this set of correlations. JENSEN (1965) found a correlation of —.31 ($p < .05$) between college grade point average (GPA) and the color factor (*C'/W*); the other Stroop factors showed no significant correlations with GPA.

Personality correlates

Automatization. This dimension, which is the same as interference proneness and which was discussed previously in connection with cognitive styles, also has correlates in the personality, interpersonal, and socioeconomic spheres. Again, we will refer to weak and strong automatization as high and low interference proneness (HI versus LI)

as measured by one of the Stroop interference scores (e.g. $CW - C$). BROVERMAN (1962) has made comparisons of personal history data among groups of *Ss* selected for high or low scores on this dimension. Summing up his findings from this investigation, Broverman states, "A considerable amount of evidence has now accrued all of which suggests that the strong automatizer (i.e. low interference proneness) is an interpersonally dominant, assertive, and effective individual" (1962, p. 35). Without going into the details of Broverman's theory and methodology, here are some of his findings: Married couples in which the husband is more interference prone than the wife report relative dissatisfaction with their marriage, while the reverse dyadic relationship reports relative harmony. The relatively low interference partner tends to assume the bulk of everyday responsibilities in marriage, such as handling correpondence, visiting childrens's teachers, etc. In general, LIs tend to assume dominant relationships to HIs. Among delinquent girls LIs preferred friends who were younger than themselves, while HIs preferred older companions. LIs reported having their first dating experience at an older age than HIs. Among delinquent boys the first truancy from school appeared earlier in LIs than in HIs. Among chronic schizophrenics more LIs were reported to be assaultive than HIs. Broverman concludes that the LI style manifests itself "as an inner push towards interpersonal dominance, personal independence, avoidance of submissive social roles, non-conformity, and rebelliousness to authority". Some other ways in which LIs differ from HIs: female HIs started drinking a year earlier than LIs and had their first sexual intercourse 10 months earlier. Male HIs admitted greater frequency of masturbation than LIs and also married 2.1 years earlier than LIs. HIs come from higher socioeconomic backgrounds and have parents of higher occupational and educational levels than LIs. LIs tend to be first-born. Finally, probably indicating some physiological basis underlying this dimension, it was found that LIs awaken from sleep more easily and do not need as much time as HIs to become fully awake on rising in the morning.

Personality inventories. The first investigators to look for personality correlates of the Stroop by means of personality inventories were THURSTONE and MELLINGER (1953). Their investigation warrants detailed examination. Four personality inventories totalling 475 items and "covering the range of normal personality" were administered, along with the Stroop test, to 99 students. Biserial correlations were

obtained between each of the 475 items and seven different Stroop scores (basic scores *W*, *C*, *CW*, and derived scores B, H, K, and M in table 1). The primary interest in this study was in the personality correlates of the interference aspect of the Stroop. The following prediction was made: "Persons with a high degree of control, who are cool and collected and who have a high frustration tolerance, will calmly override the distraction introduced by the (*CW*) test and proceed to name the color of the tints with relatively little slowing down. These are the stable, unexcitable, deliberate, and determined people. The other possibility is that persons who easily separate different kinds of experience, who tend toward dissociation, will maintain speed on the *CW* card" (p. 2). Of the seven Stroop scores used, Thurstone and Mellinger reported the results for only one, mainly because it yielded a slightly greater number of statistically significant correlations than any of the other scores. The score they settled on was score M in table 1 ($CWz - 2Cz + 10$). A worse choice would have been difficult to make in terms of the interpretability of the results. When this particular score is included in a factor analysis with all the other Stroop scores, it turns out to be the one score with the least factorial purity (JENSEN, 1965). Furthermore, it is more heavily loaded on the color difficulty and speed factors than on the interference factor, with which it correlates only .43. An additional shortcoming of this score is that it has one of the poorest test-retest reliabilities of all the scores (.43). In the enormous correlation matrix obtained in the Thurstone and Mellinger study, the chance number of significant correlations, assuming complete independence among the 457 personality items, would be 35 *r*s at the .01 level and 168 at the .05 level. In fact, there were 50 *r*s at the .01 level and 235 at the .05 level. Since the Stroop score which yielded the greatest number of significant *r*s was used here, and since the personality items are certainly not independent, these results are at best only suggestive that the Stroop has some variance in common with the personality domain. While none of the individual item correlations was high enough to be of predictive value, the specific items showing significant correlations fall into clusters which are fairly homogeneous psychologically.

First of all, the original prediction of low interference being associated with traits of perseverence and stability was not borne out in the least. The results were more or less in the opposite direction, although one must not forget that the Stroop score used here is an

amalgam of all three Stroop factors in almost equal parts, so that the personality constellation described by Thurstone and Mellinger cannot be identified with interference proneness alone. These personality traits are associated with relatively greater speed of color naming (*C*), greater speed of word reading (*W*), and less interference on the *CW* card. The various clusters of personality items are summarized as: a procrastinating, not energetic or intensive workers, unsteady, erratic, changeable, not persevering; (b) pleasure-seeking, easy going, not ambitious, lazy; (c) not systematic, neat, or orderly; (d) not responsible or decisive, big talkers, not practical; (e) active, boisterous, enjoy risks; (f) talkative, sociable. Thurstone and Mellinger conclude, "The picture revealed by this score ($CWz - 2Cz + 10$) is that persons who are relatively fast on card *CW* in spite of the increased stress have irregular habits and seek pleasure in life. They procrastinate, do not work intensively, lack perseverance, are not ambitious, and are not orderly or neat. They seem to dodge hard work and responsibility, they like to talk, and they have little control over their habits. This interpretation is interesting in that it refutes the hypothesis that the controlled and determined person will plow right through the distraction. It seems instead that the deliberate, regular and energetic worker takes pains to do the task carefully and systematically. He finds it difficult to effect the required degree of dissociation to read card *CW* easily" (p. 13).

This personality description bears a striking resemblance to descriptions of the psychopathic personality. EYSENCK (1957) has found that within his system of personality classification psychopaths are quite extraverted and are emotionally unstable or neurotic. Thus, it seems a reasonable hypothesis that the personality correlates of the Stroop might be described in terms of two highly pervasive superfactors that account for most of the variance in any comprehensive battery of personality measurements: Extraversion-Introversion and Neuroticism (EYSENCK, 1960).

JENSEN (1965) attempted to test this hypothesis by administering the Stroop and the Extraversion (*E*) and Neuroticism (*N*) scales of the Maudsley Personality Inventory (MPI) to 436 university students. The correlation between *E* and *N* in this sample was —.087, which is not significant at the .05 level. The personality scales were correlated with each of the Stroop scores listed in table 1. To improve reliability the Stroop scores were a composite of two administrations. The correla-

tions between *E* and total time on *W*, *C*, and *CW* were —.10 ($p < .05$), —.13 ($p < .01$), and —.14 ($p < .01$), respectively. Correlations of *N* with these three basic Stroop scores were .08 (n.s.), .11 ($p < .05$), and .10 ($p < .05$), respectively. More revealing are the correlations between the personality scales and the three Stroop scores which most purely represent the three factors measured by the Stroop (*W*, and derived scores C and J in table 1). Since these scores differ in reliability, comparisons of their correlations with the personality variables should be preceded by correction for attenuation; these corrections also take into account the reliability of the personality measures, which is approximately .80 for both *E* and *N*. The resulting corrected correlations of *E* with the Stroop factors of speed (*W*), color-difficulty [$(C/C + W)$], and interference ($CW - C$) were —.11 ($p < .05$), —.07 (n.s.), and —.12 ($p < .05$), respectively. The corresponding correlations with *N* were .09 (borderline significance at the .05 level); .08 (n.s.), and .06 (n.s.). These are all quite meagre correlations indeed, considering that they are corrected for attenuation of both the Stroop and the personality measures. A study by Callaway (1959) based on only 28 *S*s showed a correlation of —.43 ($p < .05$) between $CW - C$ and the *E* scale of the MPI; the correlation of $CW - C$ with the *N* scale was —.15 (n.s.).

Since the *E* and *N* scales together are known to account for a very large proportion of the variance in the personality domain, their small correlations with the Stroop factors suggests that the chief determinants of Stroop variance are not likely to be discovered in the personality domain.

One point should be noted about Jensen's correlations in connection with the Thurstone and Mellinger study. Extraversion is associated with superior performance on each of the Stroop factors, but Neuroticism is associated with poor performance on these factors. Thus it is the stable, non-neurotic extravert, not the unstable extravert or psychopath, who tends to perform relatively well on the Stroop. This conclusion is more in accord with Broverman's description of "strong automatizers", (low interference *S*s) as being aggressive and effective individuals rather than with the somewhat psychopathic personality described by Thurstone and Mellinger.

H. PODELL (1961, 1963) included several Stroop scores in a cluster analysis along with the Minnesota Multiphasic Personality Inventory, the California Personality Inventory, and a number of other cognitive,

affective, and attitude measures. The Stroop measures (W, C, WC, $CW-W$, $(CW-C)/C$) all loaded on one factor on which no other tests had significant loadings. Neuroticism correlated only .05 with this "Stroop factor".

WEISS and SHERMAN (1962) found a correlation of .25 ($p < .05$) between Stroop interference and anxiety as measured by MCREYNOLDS' (1958) scale, a test which assesses anxiety level in terms of the degree of incongruity between the S's moral values and his feelings and desires.

Projective tests. BROVERMAN (1963) has reported some significant relationships between Stroop scores and certain characteristics of response to the Thematic Apperception Test (TAT) and the Rorschach inkblot test. The score $(C-W)/W$ was inversely related to need for achievement (n Ach) as assessed from TAT fantasies, i.e. the high n Ach Ss showed relatively little difference between C and W speeds. N Ach was strongest in Ss with low scores on $(C-W)/W$ who also had high scores on $(CW-C)/W$ (i.e. high interference). The low interference Ss tended to produce TAT fantasies that were characterized by "nice, safe, positive relationships and statements"; they tended not to express hostility or dysphasic feelings.

$(C-W)/W$ was positively correlated with highly integrated whole responses to the Rorschach.

These projective test findings were interpreted in terms of Broverman's hypotheses that high $(C-W)/W$ is an indicator of verbal specialization, which might be an indicator of an obsessional (anal) personality organization, and that low $(CW-C)/W$ is an indicator of ego-strength, which acts as a regulator of both inner drives and outer demands.

Affectivity. Two groups of nine college students at the two ends of a continuum representing range, lability, and threshold of affective responsiveness, as assessed by interview ratings, were compared on the Stroop and showed a significant ($p < .05$) difference, presumably on the interference factor, although the exact Stroop score and the nature of the relationship were not specified (OKEN, GRINKER, HEATH, HERZ, KORCHIN, SABSHIN, and SCHWARTZ, 1962).

Drive. LEEDY (1963) used a Stroop interference score as a measure of drive (D) in an investigation concerning the effects of drive on the range of cue utilization. Since the Stroop showed no significant relationships to any of the other variables in this study, we will bypass a description of the experiment and comment only on Leedy's particular

use of the Stroop as an independent measure of drive strength. The habit strength of word reading (H*w*) is considered to be greater than the habit strength of color naming (H*c*). This situation gives rise to difficulty, due to response competition, on card *CW*. The greater the H*w* — H*c* difference, the greater is the interference of the printed word with the naming of the color of the ink. Now, in the Hull–Spence theory, response strength is a multiplicative function of drive and habit strength (i.e. D × H). Thus D(H*w* — H*c*); and as D increases, the difference in competing response strengths (DH*w* — DH*c*) increases. Consequently, higher drive should result in poorer performance on *CW*. This is very neat and quite in accord with predictions from Hull–Spence theory as it would pertain to the effects of drive on Stroop performance, but no validating evidence exists for this relationship between *D* and *CW*. In fact, there is some evidence in which the outcome is opposite to this prediction (AGNEW and AGNEW, 1963).

One obvious difficulty in using *CW* performance to assess D as an intrinsic drive, that is to say, when D is not manipulated experimentally, is that the assumption must be made that there are no individual differences in H*w* — H*c* independent of individual differences due to D (i.e. DH*w* — DH*c*). Since there is no sound basis for such an assumption we cannot know whether *CW* performance reflects relative differences in habit strength or drive or some combination of the two. Leedy, in fact, found no significant correlations between the Stroop interference score and other purported indicators of intrinsic drive, such as the Taylor Manifest Anxiety Scale, variation in resting skin resistance, and a decision location task—an ambiguous situation in which the *S* is goaded by a noxious stimulus and premature responses are taken as an indication of impulsivity generated by high drive.

KLEIN (1954) has reported an experiment in which high and low *S*s on Stroop interference (*CW* — *C*) were compared on a number of tasks under high and low drive conditions (thirsty vs. satiated for water). Unfortunately, Stroop performance itself was not measured under the drive conditions. There were significant interactions between Stroop interference and drive on tasks such as size estimation and tachistoscopic recognition; the low interference group usually showed superior performance to the high interference group, with accentuated differences under high drive. In a word association test involving thirst-related words the thirsty low interference *S*s gave more remote associations than did the thirsty high interference *S*s. These results

were interpreted in terms of Klein's theory of cognitive control (KLEIN, 1954). Only one study has been reported which compared the Stroop performance of the same *S*s under experimentally induced drive conditions. AGNEW and AGNEW (1963) used electric shock and the threat of shock, plus informing the *S*s that their intelligence was being assessed, to induce high drive. The low drive condition consisted of instructions intended to put the *S* at ease and to reduce ego-involvement. Two groups received both treatments in counter-balanced order and a third group was tested twice under low drive. Simultaneous measurement of heart rate showed highly significant differences under the high and low drive conditions. The Stroop score used was $C + CW$. Analysis of variance showed a significant ($p < .05$) improvement in this score under the high drive condition. Most of the difference was on the first testing. When the low-low and low-high drive group were compared by analysis of covariance, there was no significant drive effect on the second testing. Thus, these results are rather ambiguous, but they clearly do not lend any support to the Hull–Spence type of prediction that drive should impair performance on *CW*. The usual interference score $CW - C$ was also examined and was found to decrease under high drive, though not significantly. The relationship between drive and the $C + CW$ score would not seem very enlightening in terms of the Hull–Spence theory in any case, since this score is an amalgam of all three Stroop factors. The most satisfactory method for studying the effects of drive on Stroop performance would be to assess the effects of drive on each of the three Stroop factors separately.

Psychiatric groups. The only large-scale attempt to relate Stroop variables to psychiatric syndromes has been carried out by Smith and his co-workers; the most comprehensive reference to this work is by SMITH and NYMAN (1962). We have found all of this work exceedingly difficult to read and interpret. Since the usual methods of scoring the Stroop (e.g. table 1) did not yield significant correlations with other measures of cognitive functioning in Smith's earlier investigations, he resorted to the method of serial scoring described in a previous section of this article. This method, which is based on the temporal pattern of time scores on *CW* for successive fifths of the task, has revealed some significant relationships to diagnostic categories. In one study (SMITH and NYMAN, 1962) non-psychiatric orthopaedic patients were compared with psychiatric out-patients (mostly neurotic and psychopaths) and with hospitalized psychotics. In all groups performance

deteriorated from the first to the last fifth of the task; deviations from this linear trend were found to increase in proportion to the severity of the psychiatric disturbance over these three groups of *Ss*. The so-called Stabilized pattern (i.e. relatively consistent performance) was more common among normals and psychopaths than among neurotics or psychotics. The Stabilized pattern also characterized patients who were rated as showing improvement after therapy as compared with patients who failed to improve. The Cumulative pattern (i.e. consistently deteriorating performance throughout the task) showed up most markedly among depressives and patients with hysteroid traits. The Dissociative pattern (i.e. highly variable performance) was associated with hysterical conversion symptoms. Other subcategories of Stroop performance, respresenting various combinations of the three main types above—17 in all--were also correlated with a host of psychiatric symptoms. One wonders how many of these relationships would stand up under cross-validation.

NYMAN and SMITH (1959, 1960) found an increase in the tendency toward the Cumulative pattern over repeated testings among patients in a mixed clinical group who were not responding favorably to therapy in the opinion of the patient's psychiatrist. All types of therapy were lumped together in this study—electro-convulsive therapy,insulin, marsilid, chlorpromazine, ataraxics, sedatives, other pharmacologic treatments, and psychotherapy. Similar Stroop characteristics were found among patients suffering from some form of organic brain disorder.

SMITH and JOHNSON (1964) reported a positive relationship between therapeutic improvement and improvement in Stroop *CW* performance among depressive patients.

It seems rather unfortunate that Smith's extensive psychiatric studies with the Stroop have not also paid more attention to the basic factors measured by simple Stroop scores. A study by WAPNER and KRUS (1960) shows that simple Stroop scores can reflect psychiatric disturbance. They found that schizophrenics were slower than normals on cards *W*, *C*, and *CW* by 46%, 36%, and 54%, respectively (the corresponding significance levels are .10, .05, and .01). On the interference factor ($CW - C$), schizophrenics were on the average 78% higher than normals.

WEISS and SHERMAN (1962) found a complex relationship between Stroop interference, the Taylor Manifest Anxiety Scale, and psychiatric

chronicity. Long term chronic schizophrenics showed a significant correlation of .40 ($p < .05$) between Stroop interference and the MAS, while short-term chronic schizophrenics produced a correlation of only —.05. An important subgroup in producing these different correlations were the paranoid *Ss*. The correlation between interference and the MAS for short-term paranoids was —.41 ($p < .12$), while among the long-term paranoids the correlation was .67 ($p < .01$); the difference between the two correlations is highly significant.

Discussion and conclusions

The high test-retest reliability of the three basic Stroop scores—a reliability which increases with repeated retesting—indicates that the Stroop measures certain highly stable characteristics of individuals. The psychological nature of these characteristics are not well understood. But the fact that various Stroop scores have shown significant, although nearly always quite low, relationships to a diverse host of other psychological variables which are often phenotypically very different from the Stroop task itself, suggests that whatever processes are tapped by the Stroop are of a very basic and broad significance. Since many sources of variance enter into most psychological measurements, those processes having the greatest generality are usually bound to be represented by rather weak correlations with other variables; they are easily swamped by the relatively more task-specific variance contained in any particular measure. The key to the importance of such basic processes lies not in the power to predict a specific criterion with a high degree of precision, but rather in the fact that they have been shown to enter into a broad spectrum of psychological phenomena. It is largely for this reason that the Stroop test, and especially the factors it taps, may be considered worthy of study in their own right.

With few exceptions, however, investigators have accepted their own ad hoc conceptions of the "face validity" of the Stroop and have proceeded to use it for their particular purpose as a reference test in the study of some other, even more obscure, psychological phenomena. Thus, though the Stroop has been used in many studies, the procedures employed have usually not contributed much to our understanding of the processes actually measured by the Stroop, other than to assure us that some basic processes are indeed tapped by this test. Scarcely any other broad or satisfying generalization can be drawn from the body of evidence we have reviewed. Stroop processes seem to enter

most strongly into the cognitive sphere, particularly where learning, tempo, and response competition are involved. Stroop processes seem to enter least into the perceptual realm, and then only in tasks involving some element of discriminative problem solving, such as the Gottschaldt embedded figures test. Stroop processes are manifest rather sporadically and in complex ways in the personality domain; some genuine relationships undoubtedly exist here, but their meaning is hardly decipherable from the present evidence.

The Stroop test probably measures simpler and more fundamental processes than are measured by most of the other tests with which it has been correlated. We know that the Stroop itself contains three dimensions of variance. A simple linear model seems adequate to describe the structure of the Stroop. The three factors can be called Speed (Sp), Color difficulty (Cd), and Interference (Int). Each of the three Stroop tasks taps variance on one or more of these factors. Card *W* taps Sp; card *C* taps Sp + Cd, and card *CW* taps Sp + Cd + Int. The intercorrelations in table 2, obtained from one administration of

TABLE 2

Intercorrelations among Stroop scores ($N = 436$)

Factors	S_p	$S_p + C_d$	$S_p + C_d$ + Int	C_d	Int
Scores	*W*	*C*	*CW*	*C* — *W*	*CW* — *C*
W	—	52	43	—07	21
C			66	82	18
CW				48	86
C — *W*					06

the Stroop to 436 college students (Jensen, 1965) fits this additive model very well. Note that the factors themselves have very low intercorrelations and that large intercorrelations exist only between variables containing common factors. Table 3, based on the same data, shows the average difficulty, in terms of time for 100 responses, for each of the three Stroop factors individually and when they are cumulated in each of the Stroop tasks—*W*, *C*, and *CW*.

The most basic of the Stroop factors is probably the speed factor or "personal tempo", as THURSTONE and MELLINGER (1953) called it.

TABLE 3

Mean time (seconds) for Stroop factors and basic scores ($N = 436$)

Factors	Mean	Variance	Scores	Mean	Variance
Speed	38	34	*W*	38	34
Color difficulty .	20	69	*C*	58	103
Interference . .	42	277	*CW*	100	380

The generality of this factor must, of course, be further established by demonstrating significant correlations among many other measures of "personal tempo", such as normal rate of speaking, tapping rate, preferred metronome rate, the speed factor that exists in many learning and problem-solving tasks, and the like.

The color-difficulty factor is probably least general and may be quite specific to color. Object naming also requires more time than reading the names of objects, but it is not known if there is a correlation between object naming and color naming speeds. In correlating the Stroop with a large battery of learning measures, JENSEN (1965) found that the Cd factor showed significant correlations only with learning tasks which actually involved the *S*'s learning and recall of colored stimuli. Few relationships between card *C* and other variables have been reported in the literature which could not be accounted for in terms of the Sp factor contained in card *C*. The one highly reliable correlation with the Cd factor is sex—in every study women have consistently shown less color difficulty than men. We suggest that the Cd factor is a result of something more basic than differential experience with words and colors. The consistent difference between *W* and *C* beginning at the age the child first learns to read and the almost negligible effect of prolonged practice in diminishing this difference suggests that color-difficulty probably involves some fundamental, physiological difference in the process of reading words and naming colors. The hypothesis of differential degrees of learning the two types of responses seems hardly adequate to account for the facts we have reviewed on this matter. Speed of response is generally accepted as an indicator of response strength, and on this basis it has been assumed that the response strength of the word reading habit, or specifically, of reading the names of colors, is greater than the response strength of naming the colors themselves. Even with prolonged practice these two kinds of habits apparently attain different asymptotes of response

strength, as inferred from speed of response. It might be questioned whether we are dealing here, not with differences in response strength due to different amounts of practice, but with a difference in response latency due to a lesser degree of complexity of the S—R connection for word reading than for color naming. An appropriate analogy might be the difference between simple and complex reaction time.

Thus, on card *CW* the *S* must maintain a "set" (i.e. self-given verbal cues) to inhibit the S—R tendencies with the shorter latencies (words) in favor of those with longer latencies (colors). This results in response competition and interference. The amount of interference that occurs, strangely enough, seems to be scarcely related to the degree of discrepancy between the response strenghts to *W* and to *C* as inferred from speed, for as shown in table 2, there is a correlation of only .06 ($N = 436$) between the color difficulty factor and the interference factor. The large and reliable individual differences on the *CW* factor must be largely attributable to subject variables that are tapped by the *CW* task itself. For the time being, at least, we would label these subject variables "interference proneness". The evidence we have reviewed indicates that this factor has considerable generality and is identifiable to some extent with classical interference effects in learning and retention (JENSEN, 1965).

The Stroop test should continue to be useful in the study of individual differences in basic psychological characteristics such as personal tempo and susceptibility to interference.

SUMMARY

This is a comprehensive review of research on the Stroop Color-Word Test and its predecessors and variants. The Stroop Test is based on the speeds of reading color names, of naming colored patches, and of naming the colors of color-words which are printed in incongruous colors, e.g., the word RED printed in green or blue ink. The test yields highly reliable measures of individual differences on three 'factors': speed or 'personal tempo', color-naming difficulty, and interference proneness. The test has been used in some seventy studies involving such diverse fields as perception, learning, drive, problem-solving, intellectual abilities, cognitive style, personality, psychiatric diagnosis, and psychopharmacology. Significant relationships between Stroop factors and other variables in all these areas have been found. The evidence suggests that the Stroop scores share relatively little variance in common with the perceptual and personality spheres and evince most of their relevance in the cognitive realm. Psychometric and theoretical problems arising from the Stroop test are also discussed.

References

AGNEW, N. and AGNEW, Mary, 1963. Drive level effects on tasks of narrow and broad attention. Quart. J. Exp. Psychol., **15**, 58—62.

AMSTER, H., 1965. The relation between intentional and incidental concept learning as a function of type of multiple stimulation and cognitive style. J. Pers. Soc. Psychol., **1**, 217—223.

BASOWITZ, H., S. J. KORCHIN and D. OKEN, 1955. The evocation of anxiety and performance changes under minimal doses of adrenalin. Amer. Psychol. **10**, 388 (abstract).

BILLS, A. G., 1931. Blocking: A new principle of mental fatigue. Amer. J. Psychol. **43**, 230—245.

BRIAN, C. R. and F. L. GOODENOUGH, 1929. The relative potency of color and form perception at various ages. J. Exp. Psychol., **12**, 197—213.

BROVERMAN, D. M., 960a. Dimensions of cognitive style. J. Pers. **28**, 163—185.

———, 1960b. Cognitive style and intra-individual variation in abilities. J. Pers. **28**, 240—256.

———, 1962. Behavioral correlates of cognitive style. Progress report, U.S.P.H. Grants M-5773 and M-896.

———, 1963. Eight months with the Stroop test. Unpublished manuscript.

——— and R. S. LAZARUS, 1957. Performance under task-induced stress. Paper read to Eastern Psychol. Ass., New York, April.

——— and ———, 1958. Individual differences in task performance under conditions of cognitive interference. J. Pers., **26**, 94—105.

BROWN, W., 1915. Practice in associating color names with colors. Psychol. Rev., **22**, 45—55.

CALLAWAY, E., 1959. The influence of amobarbital (amylobarbitone) and methamphetamine on the focus of attention. J. Ment. Sci. **105**, 382—392.

——— and R. I. BAND, 1958a. Some psychopharmacological effects of atropine: Preliminary investigation of broadened attention. AMA Arch. Neurol. Psychiat., **79**, 91—102.

——— and D. DEMBO, 1958b. Narrowed attention: A psychological phenomenon that accompanies a certain physiological change. AMA Arch. Neurol. Psychiat., **79**, 74—90.

——— and G. STONE, 1960. Re-evaluating focus of attention. In: L. Uhr and J. G. Miller (Eds.). Drugs and behaviour. New York. Pp. 393—398.

CATTELL, J. McK., 1886. The time it takes to see and name objects. Mind, **11**, 63—65.

COMALLI, P. E. Jr., S. WAPNER and H. WERNER, 1962. Interference effects of Stroop color-word test in childhood, adulthood, and aging. J. Genet. Psychol. **100**, 47—53.

DESCOEUDRES, A., 1914. Couleur, forme ou nombre? Arch. de Psychol., **14**, 305—341.

EYSENCK, H. J., 1957. The dynamics of anxiety and hysteria. London: Routledge and Kegan Paul.

————, 1960. The structure of human personality. London: Methuen.

FRASER, Janet H., 1963. Verbal satiation and antagonistic response tendencies. McGill Undergraduate Research Projects in Psychology, 32—38.

GARDNER, R. W., P. S. HOLZMAN, G. S. KLEIN, H. LINTON and D. P. SPENCE, 1959. Cognitive control: A study of individual consistencies in cognitive behavior. Psychol. Issues, 1, 1—185.

GARRETT, H. E. and V. W. LEMMON, 1924. An analysis of several well-known tests. J. Appl. Psychol., 8, 424—438.

GATES, G. S., 1922. Individual differences as affected by practice. Arch. Psychol., 8, no. 58, 1—74.

GERSHON, S., G. HOLMBERG, E. MATTESON, N. MATTESON and A. MARSHALL, 1962. Imipramine hydrochloride. Its effects on clinical, autonomic, and psychological function. Archives of Gen. Psychiat., 6, 96—101.

HARDISON, J. and K. PURCELL, 1959. The effects of psychological stress as a function of need and cognitive control. J. Pers., 27, 250—258.

HOLLINGWORTH, H. L., 1912. Psychological aspects of drug action. Psychol. Bull., 9, 420—423.

————, 1915. Articulation and association. J. Educ. Psychol., 6, 99—105.

————, 1923. The influence of alcohol. J. Abnorm. Soc. Psychol., 18, 204—237.

JAENSCH, E. R., 1929. Grundformen menschlichen Seins. Mit Berücksichtigung ihrer Beziehungen zu Biologie und Medizin, zu Kulturphilosophie und Pädagogik. Berlin: Otto Elsner.

JAMES, W., 1890. The principles of psychology. New York: Holt.

JENSEN, A. R., 1965. Scoring the Stroop Test. Acta Psychol., 24, 398—408.

————, 1965. Individual differences in learning: Interference factor. Final Report, Cooperative Research Project No. 1867, U.S. Office of Education.

KIPNIS, D and A. S. GLICKMAN, 1959. Validity of non-cognitive tests at nuclear power school. U.S.N. Bur. Naval Personnel Tech. Bull., No. 59—6.

———— and ————, 1961. Development of a non-cognitive battery: Prediction of performance aboard nuclear powered submarines. U.S.N. Bur. Naval Personnel Tech. Bull., No. 61—5.

———— and ————, 1962. The prediction of job performance. J. Appl. Psychol., 46, 50—56.

KLEIN, G. S., 1954. Need and regulation. In: M. R. Jones (Ed.), Nebraska symposium on motivation. Lincoln: University of Nebraska Press. Pp. 224—274.

————, 1964. Semantic power measured through the interference of words with color-naming. Amer. J. Psychol., 77, 576—588.

LANGER, J. and B. G. ROSENBERG, 1964. Symbolic meaning and color naming. Unpublished manuscript.

LAZARUS, R. S., 1955. Motivation and personality in psychological stress. Progress Report No. 2, NIMH Grant M-734.

———, R. W. BAKER, D. M. BROVERMAN and J. MAYER, 1957. Personality and psychological stress. J. Pers., **25**, 559—577.

LEEDY, H. B., 1963. An investigation of the range of cue utilization as a function of induced and intrinsic drive. Unpublished doctoral dissertation, Purdue University.

LEHMAN, H. E. and D. A. Knight, 1961. The psychopharmacological profile - A systematic approach to the interaction of drug effects and personality traits. In: Bordeleau, Jean-Marc (Ed.), Systems extra-pyramidal et neuroleptics. Montreal: Editions psychiatriques. Pp. 429—440.

LIGON, E. M., 1932. A genetic study of color naming and word reading. Amer. J. Psychol., **44**, 103—121.

LOOMIS, H. and MOSKOWITZ, S., 1958. Cognitive style and stimulus ambiguity. J. Pers., **26**, 349—364.

LUND, F. H., 1927. The role of practice in the speed of association. J. Exp. Psychol., **10**, 424—433.

MCFARLAND, R. A. and A. L. BARACH, 1936. Relationship between alcoholic intoxication and anoxemia. Amer. J. ment. Sci., **192**, 186—198.

MCREYNOLDS, P. W., 1958. Anxiety as related to incongruencies between values and feelings. Psychol. Rec., **8**, 57—66.

NYMAN, G. E. and G. J. W. SMITH, 1959. A contribution to the definition of psychopathic personality. Lunds Universitets Arsskrift, N.F. Avd. 2, 55, 10. Lund: Gleerup.

——— and ———, 1960. An attempt at describing operationally the effects of psychiatric therapy. Psychiat. et Neur., **140**, 258—280.

OKEN, D., R. R. GRINKER, HELEN A. HEATH, M. HERZ, S. J. KORCHIN, M. SABSHIN and NEENA B. SCHWARTZ, 1962. Relation of physiological response to affect expression. Including studies of autonomic response specificity. Archives of Gen. Psychiat., **6**, 336—351.

OSTFELD, A. M. and ALAYNE ARUGUETE, 1962. Central nervous system effects of hyoscine in man. J. Pharm. Exp. Therapeut., **137**, 133—139.

PETERSON, J. and Q. J. DAVID, 1918. The psychology of handling men in the army. Minneapolis: Perine Book Company.

———, L. H. LANIER and H. M. WALKER, 1925. Comparisons of white and negro children in certain ingenuity and speed tests. J. Comp, Psychol., **5**, 271—283.

PODELL, H. A., 1963. Note on successive dimensional analysis applied to affective, cognitive, and personality traits. Psychol. Rep., **13**, 813—814.

PODELL, J. E. and L. PHILLIPS, 1959. A developmental analysis of cognition as observed in dimensions of Rorschach and objective test performance. J. Pers., **27**, 439—463.

POSTMAN, L. and D. A. RILEY, 1959. Degree of learning and interserial interference in retention. University of California Publications in Psychology, **8**, 271—396.

QUARTON, G. C. and G. A. TALLAND, 1962. The effects of metamphetamine and

pentobarbital on two measures of attention. Psychopharmacologia, **3**, 66—71.

RAND, G. and S. WAPNER, 1962. A behavioral analysis of the color-word test. Paper read at Eastern Psychological Association Meetings, Atlantic City, April.

————, ————, H. WERNER and J. H. MCFARLAND, 1963. Age differences in performance on the Stroop color-word test. J. Pers., **31**, 534—558.

ROUSE, R. O. and F. MAYER, 1961. On the reduction of response competition by practice providing separate response channels. Paper presented at Psychonomic Society, New York, August.

SMITH, G. J. W., 1957. The serial mirror drawing test. Acta Psychol., **13**, 288—298.

————, 1959a. Comparisons between adaptive patterns in two serial experiments. Acta Psychol., Amst., **16**, 302—315.

————, 1959b. Comparisons between adaptive patterns in two serial experiments. Nord. Psykol. 1959, **11**, 250—263.

———— and G. A. V. BORG, 1964. The problem of retesting in the serial color-word test. Psychol. Res. Bull., Lund Univer., Sweden, **4**, no. 6.

———— and G. JOHNSON, 1964. Experimental description of a group of psychiatric patients before and after therapy by means of MCT (the meta-contrast technique) and CWT (the serial color-word test). Psychol. Res. Bull., Lund Univer., Sweden, **4**, no. 7.

———— and G. S. KLEIN, 1953. Cognitive controls in serial behavior patterns. J. Pers., **22**, 188—213.

———— and G. E. NYMAN, 1959. Psychopathologic behavior in a serial experiment. Lunds Universitets Arsskrift, N.F. Avd. 2, **56**, 5. Lund: Gleerup.

———— and ————, 1962. The serial color-word test: a summary of results. Psychol. Res. Bull., Lund Univer., Sweden, **2**, no. 6.

STROOP, J. R., 1935a. The basis of Ligon's theory. Amer. J. Psychol., **47**, 499—504.

————, 1935b. Studies of interference in serial verbal reactions. J. Exp. Psychol., **18**, 643—662.

————, 1938. Factors affecting speed in serial verbal reactions. Psychol. Monogr., **50**, no. 5, 38—48.

TELFORD, C. W., 1930. Differences in responses to colors and their names: Some racial comparisons. J. Genet. Psychol., **37**, 151—159.

THURSTONE, L. L., 1944. A factorial study of perception. Chicago: Univer. of Chicago Press.

———— and J. J. MELLINGER, 1953. The Stroop Test. Chapel Hill, N.C.: The Psychometric Laboratory, Univer. North Carolina, no. 3.

UHLMANN F. W., 1962a. Test of color recognition. Form DE x-27-61. Detroit: The Detroit Edison Co.

————, 1962b. Retention of anxiety material as a function of cognitive style. Unpublished doctoral dissertation. Wayne State Univer.

WAPNER, S., 1963. An organismic-developmental approach to the study of perceptual and other cognitive operations. Paper presented at the Center for Cognitive Studies, Harvard Univer.

——— and D. M. KRUS, 1960. Effects of lysergic acid diethylamide, and differences between normals and schizophrenics on the Stroop color-word test. J. Neuropsychiat., 2, 76—81.

WEISS, R. L. and M. SHERMAN. Anxiety and interfering responses in college students and psychiatric patients. Newsletter Res. Psychol., 1962, 4, 35—40.

WOODWORTH, R. S. and F. L. WELLS, 1911. Association tests. Psychol. Monogr., 13, no. 57, 1—85.