Interpreting Stroop Interference:
An Analysis of Differences Between Task Versions

Ruth Salo
University of California, Davis

Avishai Henik
Ben Gurion University of the Negev

Lynn C. Robertson
University of California, Berkeley and Veterans Administration

The present study investigated methodological differences between the clinical version of the Stroop Color and Word Test and the computerized single-trial version. Three experiments show that different presentations of the Stroop task can produce different levels of interference. The 1st experiment examined the effect of blocking; the 2nd experiment examined different control conditions. Greater interference in the blocked clinical version appears to result from lower response times (RTs) in the neutral condition, not from greater RTs in the incongruent condition. Experiment 3 examined the impact of shifting attention across locations while responding to Stroop stimuli. The present set of findings sheds light on the inconsistency in the clinical literature and demonstrates that the method and selection of neutral stimuli (that provide the baseline by which interference is measured) are critical because they clearly can change performance.

The study of cognitive abilities in clinical populations has yielded a wealth of information over the years (Farah & Feinberg, 2000; Grant & Adams, 1996). Both clinicians and researchers have used the results from a wide range of test batteries to enrich their understanding of a variety of neurological and psychiatric disorders. One approach that has contributed greatly to our understanding of cognitive deficits in clinical populations is neuropsychological testing (Lezak, 1995). Full neuropsychological batteries provide a rich clinical assessment of relatively large categories of problem solving, memory, spatial organization, verbal skills, and selective attention. Performance deficits on neuropsychological indexes contribute valuable information about the cognitive sequelae associated with neurological damage. However, precision is often lacking and often would be increased by additional exploration within a particular cognitive domain.

Attention is one fundamental cognitive domain that is often assessed with clinical neuropsychological batteries. Attention is multifaceted, encompassing divided, sustained, and selective attention factors. Many clinical subtests are used to assess attentional capabilities, but one of the best known for testing attentional selection is the Stroop Color and Word Test (Stroop, 1935). In the original Stroop task, subjects were presented with sheets of paper on which were written multiple columns of words or nonwords. On the color-word sheet, color names were printed in noncorresponding colors of ink (e.g., blue printed in red); on the neutral sheet, solid squares were printed in the same colors as the color-word sheet. Subjects were instructed to name the colors as quickly as possible and to self-correct any errors. They were then presented with two sheets of each condition (100 stimuli per sheet = 400 stimuli total). Although in the original version administered by Stroop (1935) subjects were instructed to name the colors across the rows, later versions have subjects read down each column. In normal subjects, the direction of naming appears to make no difference, whereas some brain-damaged patients can lose their place on the page when naming the colors across rows (Silverstein & Franken, 1965).

The Stroop task has been widely used for assessing attentional deficits in neurological and psychiatric patient groups (Abramczyk, Jordan, & Hegel, 1983; Bliener, 1993; Buchanon et al., 1994; Drake, Schwartz, Turner, & Rosenthal, 1996; Golden, 1976; Kimberg & Farah, 1993; Mehl & Cromwell, 1969; Schwartz & Shagass, 1960; Vakil, Weisz, Jedwab, Groszwan, & Aberbach, 1995). A consistent finding is that people have difficulty ignoring the word while naming the color, and their responses reflect this problem. Response time (RT) is slower when the word and color are incongruent. Stroop interference is typically calculated as the difference in RT between the word-color sheet (i.e., the sheet in which the Stroop words are printed in a noncorresponding color of ink) and the neutral sheet (i.e., the sheet with color patches or XXXs). Although variations exist in how interference is calculated, most of the reported Stroop studies use this convention (Henik, 1996).
In recent years, computerized experiments have become a mainstay in the laboratories of cognitive psychologists. With the emergence of computers in the study of cognition, many software packages of traditional paper-and-pencil tasks have been developed, resulting in increased flexibility in design and more precise timing measures. The Stroop task is no exception. Teice and Dimartino (1965) were the first to develop a computerized version of the Stroop task, in which single Stroop stimuli (i.e., words or nonwords) appeared at the center of the screen in a series of trials. The task is exactly the same as in the traditional clinical version (i.e., name the ink color as quickly as possible), except this version has no flanking stimuli. RT is typically measured with a voice-onset operated relay (although manual response is sometimes used; see MacLeod, 1991, for a review).

The use of a single-trial computerized version of the Stroop task allows for a measure of per-item RTs as opposed to a summation of RTs across a large stimulus set. Moreover, errors for single words can then be omitted when calculating RT measures. It also allows for the presentation of congruent words (i.e., color names printed in their corresponding color ink, such as red printed in red ink) randomly presented with neutral or incongruent trials, a condition first used by Dalrymple-Alford and Budayr (1966). Comparison of RTs on congruent trials and neutral trials is assumed to yield a measure of facilitation, whereas the difference in RTs between incongruent and neutral trials gives a measure of interference.

A number of potentially important differences exist between the single-trial and the clinical form of the Stroop task, differences that may obscure the underlying mechanisms contributing to attentional pathology. The standard clinical version presents a “cluttered field” to the subject, who must not only respond to and name the ink color of the attended stimulus but must also ignore the adjacent irrelevant flanker words and colors. Another difference between the two forms is that in the clinical version the conditions are blocked. The subject is presented with either an entire sheet of incongruent stimuli or an entire sheet of control stimuli (strings of Xs or color patches). In the computerized single-trial version, the conditions typically are randomly presented throughout the experiment rather than being blocked. In addition, the single-trial version typically presents each Stroop stimulus at central fixation. This stimulus display does not require the subject to generate eye movements or to move attention down columns of stimuli.

Given these fairly substantial differences in the presentation of the Stroop task, it is quite remarkable that the results from both the standard and the computerized versions are interpreted in much the same way within the clinical literature. Interference effects on the Stroop task (i.e., increased RT on the incongruent minus the neutral trials) are reported in both versions, and these effects are usually interpreted as reflecting the inability to ignore the task-irrelevant information (i.e., to avoid reading the word). In addition, these task differences may interact with the presence or absence of particular clinical disorders. It is quite possible that the presence, or size, of differences in Stroop task performance between patients and control participants may depend on which version of the task is used as well as the disorder in question. Studies that used the clinical version of the Stroop task reported increased interference effects in both brain-damaged individuals (Blumen, 1993; Vakil et al., 1995) and psychiatric patients (Abrams-Guy et al., 1983; Buchanan et al., 1994; Mehl & Cromwell, 1969). However, a number of studies using computerized single-trial versions of the Stroop task failed to replicate the traditional findings of increased Stroop interference in some psychiatric patient groups (Carter, Robertson, & Nordahl, 1992; Henik, Tzelgov, Cohen, & Henik, 1999; Salo, Robertson, & Nordahl, 1996; Taylor, Kornblum, & Tandon, 1996). Comparisons of medicated and unmedicated schizophrenia patients and matched controls also failed to yield significant differences in Stroop interference when the single-trial version was used (Salo, Robertson, Nordahl, & Kraft, 1997).

In addition, differences on different versions of the Stroop task have been described among groups of children and older adults (Henik et al., 1999). Henik et al. (1999) found that both children and older adults exhibited increased interference on the standard clinical version of the Stroop task compared with young adults (see Comalli, Wapner, & Werner, 1962, for a similar finding), but there were no corresponding differences on the computerized single-trial version. Brown, Engle, and Jones (1992) also reported differences between the clinical and single-trial version of the Stroop task, with subjects exhibiting greater Stroop interference on the clinical version compared with the single-trial version. Boucart, Mobarek, Cuervo, and Danion (1999) compared Stroop interference between schizophrenic patients and control subjects and found equivalent Stroop interference when stimuli were presented individually on the screen but increased interference in the patients when distractors appeared in the spatial surround of the Stroop target item. In contrast, Perlstein, Carter, Barch, and Baird (1998) reported no differences between schizophrenia patients and controls on either the clinical or the single-trial version of the Stroop task. Although differences between the two versions of the Stroop task have been reported, a systematic investigation into the source of these differences has not been carried out.

Rationale

The goal of the present study was to investigate the individual task components that contribute to the methodological differences between the standard clinical version of the Stroop task and the computerized single-trial version. The understanding of how these task components contribute to Stroop effects may resolve inconsistencies in the clinical literature. If attentional tasks are to be used by researchers and clinicians alike to assess cognitive deficits in patient populations, it is important to understand what the tasks are measuring. Tasks that require multiple cognitive operations, such as the clinical version of the Stroop task, may not be the best tools for understanding cognitive dysfunction in patients. Also, if one is interested in linking cognitive dys-
function to brain regions or in comparing one patient group with another, tests that measure specific cognitive operations would be more informative (Posner & Snyder, 1975). Behavioral interventions, clinical maintenance, and prognosis may well be more effective when focusing on specific areas of cognition.

Grouping of physically similar stimuli has been shown in other paradigms to modulate attentional processes (Lubow & Kaplan, 1997; Tzelgov, Henik, & Berger, 1992). Findings from both a visual search task (Lubow & Kaplan, 1997) and a Stroop task with different expectancy conditions (Tzelgov et al., 1992) provide evidence that individual subjects can take advantage of grouped stimuli to modulate attentional selection. Because of these findings, it is a reasonable hypothesis that the grouping of conditions in the clinical version of the Stroop may be one factor contributing to differences between the clinical and single-trial versions of the Stroop. Selection of a single stimulus versus selection in a cluttered field may also impact selective attention processes. An increased number of flankers in a visual display has been shown to increase RT in numerous visual attention tasks, particularly when those flankers share common attributes, such as form and color, with the target (Treisman & Gelade, 1980). In addition, moving attention from one spatial location to another takes time, so it is also reasonable to hypothesize that the requirement of shifting attention across a large array of Stroop stimuli in the clinical version may create additional RT costs, relative to maintaining attention at one spatial location, and that such shifting might take more time for more difficult items. All of these methodological differences may be important factors contributing to the differences in interference effects between the two versions in clinical populations. By isolating the contribution of these factors, the underlying mechanisms of the interference effects may be better understood.

Three experiments were conducted to investigate differences between the standard neuropsychological version of the Stroop task and the computerized single-trial version. Experiment 1, consisting of three parts, compared the computerized version of the clinical card Stroop task with blocked and random presentations of the single-trial Stroop task. The goal of this experiment was to determine how interference effects differed between a computerized version of the clinical test and two versions of the single-trial Stroop task (blocked and unblocked). Experiment 2 examined the effect of different neutral conditions on interference, given that interference is a difference score from a neutral baseline. Because the clinical version of the Stroop task requires the shifting of attention across spatial locations, Experiment 3 presented Stroop stimuli one at a time in different locations on the computer screen. The examination of these types of variables is important in understanding the contributions of the different task requirements in the clinical Stroop tasks and thus in interpreting the findings in different clinical populations.

Experiment 1

Method

Subjects

For Experiments 1 and 3, 30 normal volunteers (17 women and 13 men) were recruited through advertisements targeting students at the University of California, Davis and in nearby communities. The mean age was 29.8 years (SD = 9.3 years). Mean level of education was 15.9 years (SD = 2.3 years). All subjects reported both normal color vision and normal or corrected-to-normal visual acuity. Experiments 1 and 3 were counterbalanced across subjects for order; half performed the tests in one order (Experiments 1A, 1B, 1C, 3A, and 3B) and half performed the tests in the reverse order (Experiments 3B, 3A, 1C, 1B, and 1A). Thus, half of the subjects performed the clinical Stroop task first and half performed the clinical Stroop task last. Experiment 2 was a follow-up experiment that tested 15 new subjects. Written consent was obtained, and each participant was paid $15.

Apparatus

For all experiments, IBM 386 PC computers were used to control stimulus timing and data collection. Response timing was to 1-msec resolution and was controlled by the 8253 chip. Stimulus timing was tied to the vertical synchronization pulse. Stroop stimuli were presented on a VGA color monitor (View Sonic Corp., Walnut, CA) in four colors (red, yellow, blue, and green) and were matched for luminance using a Minolta chromatic luminance meter (Minolta Company, Ltd., Ramsey, NJ). RTs for the multiple-item Stroop task were recorded using a Gravis buttonpress (Gravis, San Mateo, CA) interfaced with the computer and monitored through the game port. In the single-trial Stroop task, responses on each trial were recorded with a Gerbrands voice-operated relay interfaced with the computer and monitored through the game port.

Experiment 1A: Computerized Clinical Version

Stimuli. In the clinical version, 36 Stroop stimuli were presented simultaneously on the VGA color monitor in four columns of nine items. The incongruent stimuli were created by printing each of the four color names in the three other ink colors three times each. The congruent stimuli were created by printing each of the four color names in its own color nine times each. The neutral stimuli consisted of strings of XXX printed in one of the four colors of ink nine times each. Each letter within the stimulus items was uppercased and subtended 1 degree vertically. The width of each item display varied as a function of the item presented (range = 3–6 letters). Two screens of each condition (congruent, incongruent, neutral) were presented for a total of six screens and 216 responses. The conditions were counterbalanced for order across all subjects.

Procedure. The subjects were instructed to name the ink color of each word beginning in the upper left-hand corner, to read down each column, and to finish with the stimulus in the lower right-hand corner. Each experimental task was explained by the experimenter, and each subject received 20 practice trials before each experimental block of trials. Total RT per screen was divided by the number of items presented to calculate an estimated per-item RT measure. Subjects were given instructions that discouraged a speed-accuracy trade-off in that they were told to respond as rapidly as possible without making too many errors. The experimenter initiated and terminated the presentation of the stimulus display with a buttonpress. In the majority of standard clinical
versions, a summation of correct responses is recorded within a given time period, and error responses are not always reported. In this computerized clinical version, total time to respond to all stimuli within a block was recorded with a button press, interfaced with the game port, and recorded by the computer to the nearest millisecond. Errors were recorded by the experimenter on prepared forms. The recording (or nonrecording) of errors is, in fact, another difference between the clinical and single-trial procedures.

**Experiment 1B: Random Single-Trial Version**

**Stimuli.** Three blocks of 72 Stroop items were presented, and each individual Stroop item appeared in the center of the screen. Parameters for each single word display were the same as in Experiment 1A. Trial type was randomized within each block: One third of the trials were incongruent, one third were neutral, and one third were congruent. Although equivalent, the order of the blocks was counterbalanced across subjects.

**Procedure.** Subjects were instructed to say aloud as rapidly as possible the color of ink in which each item was printed while ignoring the item itself. They were given the same speed-accuracy trade-off instructions as in Experiment 1A. Voice onset triggered the relay switch (recorded by the computer to the nearest millisecond) and terminated the stimulus display on the screen. The experimenter then typed in the first letter of the response, which recorded accuracy, and the keypress initiated the subsequent trial.

**Experiment 1C: Blocked Single-Trial Version**

**Stimuli.** Parameters for each word were the same as in Experiment 1B.

**Procedure.** Three blocks of 216 trials (72 congruent, 72 incongruent, and 72 neutral) were presented. Each block contained only one condition. Except for the blocking manipulation, the stimuli and procedure were identical to those described for Experiment 1B. Block order was counterbalanced across subjects.

**Data Analysis**

Mean RTs were calculated for each subject for each condition. Analysis of variance (ANOVA) procedures for repeated measures were used to analyze the data in a 3 x 3 within-subjects ANOVA with task version (clinical vs. random single trial vs. blocked single trial) and condition (congruent vs. incongruent vs. neutral) as variables. Incorrect responses were not included with the ANOVA for RT in the single-trial versions. However, because a total time to respond to all of the stimuli within a block is collected in the clinical version, error responses were included in the RT analysis for that version. Further analyses examined the effect of error responses on interference and facilitation. Planned comparisons of differences between task version for facilitation (neutral minus congruent RT) and interference (incongruent minus neutral) were performed across all experiments.

**Results**

**RT Analysis**

Table 1 summarizes the data for the three versions of the Stroop task. Overall, RT analyses revealed a main effect of task version, $F(2, 58) = 23.30$, $MSE = 4.499$, $p < .001$, and condition, $F(2, 58) = 294.35$, $MSE = 2.774$, $p < .001$, as well as an interaction between task version and condition, $F(4, 116) = 29.13$, $MSE = 1.340$, $p < .001$. Differences

<table>
<thead>
<tr>
<th>Condition</th>
<th>RT</th>
<th>SD</th>
<th>Error SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1A (clinical)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent</td>
<td>473</td>
<td>80</td>
<td>.003</td>
</tr>
<tr>
<td>Neutral</td>
<td>491</td>
<td>62</td>
<td>.020</td>
</tr>
<tr>
<td>Incongruent</td>
<td>680</td>
<td>116</td>
<td>.030</td>
</tr>
<tr>
<td>Experiment 1B (random single trial)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent</td>
<td>580</td>
<td>78</td>
<td>.020</td>
</tr>
<tr>
<td>Neutral</td>
<td>587</td>
<td>69</td>
<td>.030</td>
</tr>
<tr>
<td>Incongruent</td>
<td>681</td>
<td>77</td>
<td>.090</td>
</tr>
<tr>
<td>Experiment 1C (blocked single trial)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent</td>
<td>488</td>
<td>81</td>
<td>.020</td>
</tr>
<tr>
<td>Neutral</td>
<td>553</td>
<td>74</td>
<td>.030</td>
</tr>
<tr>
<td>Incongruent</td>
<td>713</td>
<td>102</td>
<td>.050</td>
</tr>
</tbody>
</table>

Note. Response times (RTs) are in milliseconds.

Score analyses also revealed a significant interaction between interference and version, $F(2, 58) = 30.93$, $MSE = 1.150$, $p < .001$, as well as between facilitation and version, $F(2, 58) = 11.74$, $MSE = 1.238$, $p < .001$. Further analyses of these interactions revealed significant RT differences to neutral conditions as a function of task version, $F(2, 58) = 49.09$, $MSE = 1.462$, $p < .001$. RTs to neutral conditions were significantly faster in the clinical version than in both the random single-trial version, $t(29) = 9.93$, $MSE = 1.414$, $p < .001$, and the blocked single-trial version, $t(29) = 5.14$, $MSE = 2.216$, $p < .001$. RTs to the neutral conditions also differed significantly between the random and the blocked single-trial version, $t(29) = 4.78$, $MSE = 7.56$, $p < .001$. RTs to congruent conditions were significantly longer in the random single-trial version than in both the blocked single-trial version and the clinical version, $F(2, 58) = 30.47$, $MSE = 3.318$, $p < .001$. The incongruent trials in the blocked single-trial version were significantly longer than those in both the random single-trial version, $t(29) = 3.16$, $MSE = 1.463$, $p < .01$, and the clinical version, $t(29) = 2.38$, $MSE = 2.766$, $p < .05$, which did not differ from each other ($F < 1$).

Because interference and facilitation are of interest in calculating Stroop effects, additional planned comparisons revealed that subjects showed significantly greater interference on the clinical version (189 ms) than on both the random single-trial (94 ms), $F(1, 29) = 50.35$, $MSE = 1.353$, $p < .001$, and the blocked single-trial (160 ms), $F(1, 29) = 5.57$, $MSE = 1.225$, $p < .05$, versions. Interference also differed between the two single-trial versions, $F(1, 29) = 36.45$, $MSE = 873$, $p < .001$. Analyses further revealed a significant interaction between facilitation and version, $F(2, 58) = 11.74$, $MSE = 1.238$, $p < .001$. Facilitation was significantly greater in the blocked single-trial version (65 ms) than in both the clinical (18 ms), $F(1, 29) = 9.77$, $MSE = 1.755$, $p < .01$, and the random single-trial (7 ms), $F(1, 29) = 25.76$, $MSE = 994$.

1 Means were used as the unit of analysis because it was not possible to trim outliers from the clinical version.
physically dissimilar stimuli. In contrast, the RT to the neutral trials differed significantly among the three versions (standard clinical = 491 ms, random single trial = 587 ms, blocked single trial = 553 ms). This is an important finding because the magnitude of interference (incongruent minus neutral) is presumed to reflect the efficiency of selective attention (i.e., the ability to ignore the incongruent word). The control condition is characterized by repetitive nonword stimuli, which may not require suppression; thus, differences in RT to the control condition might not be a very convincing measure of selective attention (i.e., the ability to ignore the irrelevant word). The present data show that this subtraction (incongruent minus neutral) might not adequately capture the complexity of the effects and questions some common assumptions (e.g., the meaning of neutral and its similarity and stability in various versions; see also Jonides & Mack, 1984; Lindsay & Jacoby, 1994).

Most of the supporting literature on expectancy and blocking comes from display presentations of less than 100% pure blocks (Lindsay & Jacoby, 1994; Tzelgov et al., 1992). Tzelgov et al. (1992) reported that in Stroop displays with a low percentage of neutral trials (25%) relative to a high percentage of neutral trials (75%), subjects were able to reduce interference effects. Tzelgov et al. interpreted this finding as evidence that subjects were able to adjust to the increased probability that an upcoming stimulus would be incongruent and could thus better ignore the color name that was in conflict with the ink color. It is quite possible, however, that attentional expectancies operate differently when stimuli appear in pure blocks versus varying proportions. Research on blocking (Lund, 1927; Seifert & Johnson, 1994) showed that blocking by conceptual categories (i.e., incongruent conditions; red printed in blue and green printed in yellow) may differ from the blocking of physically identical stimuli (i.e., neutral XXXs printed in red, blue, yellow, and green ink). If this is true, it is possible that participants are not able to reduce RTs to incongruent trials but are able to reduce their RTs to the neutral conditions on the basis of physical similarity. This could be partially responsible for the apparently increased interference observed in the blocked condition. This possibility is explored in Experiment 2.

Experiment 2

Method

Experiment 2 examined the impact of grouping on different types of neutrals. This manipulation was motivated by the results of Experiment 1, which suggested that grouping creates increased interference effects as a function of reduced RTs to the physically identical condition (XXXs). Because Stroop stimuli in the clinical version are always blocked and neutrals can vary across Stroop task versions, it was important to determine whether grouping had the same effect when the neutrals were blocked but were physi-
INTERPRETING STROOP INTERFERENCE

467

cally dissimilar. Fifteen new subjects were tested in this experiment. We used three types of neutral: nonword strings of XXXs as in Experiment 1, nonword but physically dissimilar neutrals (XXX, WWWW, SSXXS, MMMMM/MM); see MacLeod, 1998), and animal names (dog, bear, tiger, monkey). The stimuli and the method of presentation were identical to those in Experiment 1C (blocked single-trial version). Three blocks of each neutral type were presented for a total of 540 trials (180 congruent trials, 180 incongruent trials, and 60 trials each of the three neutral types). Each neutral type was presented in a triad of blocks; the incongruent block was presented first in each trial. The trials were counterbalanced for order across the 15 subjects.

Results

RT Analysis

Analyses revealed a main effect of neutral type, $F(2, 28) = 5.29$, $MSE = 863$, $p = .01$. RTs to animal names were significantly slower than those to the physically identical strings of XXXs, $F(1, 14) = 10.25$, $MSE = 872$, $p < .01$, and the physically dissimilar letter strings, $F(1, 14) = 5.41$, $MSE = 661$, $p < .05$. Although it did not reach statistical significance, the RTs to the physically identical neutral type were faster than to the physically dissimilar letter strings, $F(1, 14) = 1.14$, $MSE = 1.058$, $p > .05$. Planned comparisons revealed that RTs to the incongruent trials (mean = 749 ms) and congruent trials (mean = 548 ms) did not differ between the triads containing different neutral types ($F < 1$). Facilitation did not differ significantly between the trials, $F(2, 28) = 1.45$, $p < .05$.

Error Analysis

Errors were analyzed in the same design. There was a main effect of condition, $F(2, 28) = 9.6$, $MSE = 0.001$, $p < .001$, with subjects making more incongruent errors than neutral errors, $t(14) = 3.3$, $p < .01$, and congruent errors, $t(14) = 3.5$, $p < .01$. There were no differences between neutral and congruent errors ($F < 1$; see Table 2).

Discussion

The results show that RTs to neutral types differ as a function of both physical similarity and lexical properties. RTs to the physically identical neutrals (XXXs) were the fastest, whereas RTs to the lexical neutrals (animal names) were the slowest. The RTs to the nonlexically physically dissimilar neutrals (XXX, WWWW, SSXXS, MMMMM/MM) fell midway between. This continuum of RTs to different neutral types suggests that physical similarity speeds RT compared with lexical properties of the stimulus. We also carried out additional analyses examining change in RTs to the neutral types across the block. A post hoc comparison revealed no significant change in RTs to any of the three neutral conditions across the block, $F(2, 28) = 1.37$, $p > .05$. Thus, at least in this experimental situation, the data do not suggest that the difference in RT to the neutral types results from differential habituation.

Although it was not possible to vary the properties of the incongruent stimuli, it was possible to group the incongruent and congruent stimuli with different neutral types. Analysis of the RTs to the different triads (XXX triad, XXSSMMWWW triad, and animal neutral triad) revealed no difference in RT to the incongruent and congruent stimuli as a function of which neutral was included. Because interference scores were calculated as a difference score between incongruent and neutral trials, the data suggest that it is the type of neutral used that influences the interference score rather than the incongruent trial. Because RTs are reduced to the physically similar neutral stimuli, interference effects are larger than when physically dissimilar neutrals are used.

Table 2

<table>
<thead>
<tr>
<th>Condition</th>
<th>XXX</th>
<th>XXSSMMWWW</th>
<th>Animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency (ms)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent</td>
<td>547</td>
<td>64</td>
<td>544</td>
</tr>
<tr>
<td>Neutral</td>
<td>624</td>
<td>60</td>
<td>637</td>
</tr>
<tr>
<td>Incongruent</td>
<td>745</td>
<td>75</td>
<td>752</td>
</tr>
<tr>
<td>Error rate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent</td>
<td>.01</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>Neutral</td>
<td>.01</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>Incongruent</td>
<td>.02</td>
<td>.02</td>
<td>.02</td>
</tr>
</tbody>
</table>

Note. XXX = physically identical neutrals; XXSSMMWWW = physically dissimilar neutrals.

Experiments 3A and 3B

Method

Experiments 3A and B were designed to examine the impact of shifting attention across spatial locations on interference and facilitation. This shifting is a requirement in the clinical version of the Stroop task but not in the single-trial version. In the standard single-trial version, the Stroop stimulus appears at central fixation such that the subject responds to sequentially presented stimuli at the same spatial location. In contrast, the clinical version of the Stroop task requires subjects to move attention down the screen as they respond to each item.

Subjects

The same 30 subjects who participated in Experiment 1 took part in Experiments 3A and B.

Stimuli

Six blocks of 36 stimuli were presented (216 stimuli). Stimuli appeared one at a time on the screen in different locations. The locations at which they appeared were matched to the same locations occupied by the stimuli in Experiment 1A (i.e., the computerized version of the standard clinical task). The stimuli appeared sequentially beginning in the upper left-hand corner and terminating with a stimulus in the lower right-hand corner. Each stimulus

2 We thank Colin MacLeod for pointing out the necessity of this manipulation.
remained on the screen until the subject responded. Voice onset triggered the relay switch (recorded by the computer to the nearest millisecond) and terminated the stimulus display on the screen. After the subject responded, the experimenter keyed in the response, which then triggered the appearance of the next Stroop word on the screen. In Experiment 3A the conditions were randomly presented within the blocks, whereas in Experiment 3B the Stroop stimuli were blocked by condition, corresponding to Experiments 1B and 1C.

**Procedure**

The procedure was identical to that of Experiments 1B and 1C, except that the words changed position on the screen. Subjects were informed of this location shift before the experiment began.

**Data Analysis**

Mean RTs were calculated for each subject for each condition. ANOVA procedures for repeated measures were used in a 2 × 3 factorial design with two within-subject variables: blocking (blocked vs. random) and condition (congruent, incongruent, and neutral). Incorrect responses were not included in the ANOVA for RT. In addition, the random version of the location shift experiment (Experiment 3A) was compared with that of Experiment 1B to determine the effects of shifting location independent of blocking manipulations. The blocked versions of the single trial (Experiment 1C) and the blocked location shift trial (Experiment 3B) were also compared.

### Results

**RT Analysis**

Analyses revealed a main effect of blocking, F(1, 29) = 61.48, MSE = 2.227, p < .001, and condition, F(2, 58) = 222.70, MSE = 1,998, p < .001, as well as a significant interaction between blocking and condition, F(2, 58) = 26.77, MSE = 773, p < .001. Planned comparisons revealed that subjects showed significantly greater interference on the blocked location shift version (143 ms) than on the random location shift version (114 ms), F(1, 29) = 10.36, MSE = 592, p < .01. Facilitation also differed significantly as a result of blocking (random location shift = 11 ms; blocked location shift = 58 ms), F(1, 29) = 33.14, MSE = 506, p < .001. Simple planned paired t tests revealed significant RT differences to neutral conditions as a function of blocking (random location shift = 612 ms and blocked location shift = 563 ms), t(29) = 6.73, MSE = 792, p < .001. RTs were 49 ms faster in the random version than in the blocked version. RTs to congruent trials also differed between the two blocking conditions (blocked location shift = 505 ms and random location shift = 601 ms), t(29) = 9.60, MSE = 1,507, p < .001. Finally, RTs to incongruent trials differed significantly as a function of blocking (random location shift = 726 ms and blocked location shift = 706 ms), t(29) = 2.08, MSE = 1,451, p < .05.

**Error Analysis**

Errors were analyzed in the same design. On the random location shift version, subjects made 4% errors; on the blocked location shift presentation, they made 3% errors. There was a main effect of version, F(1, 29) = 21.06, MSE = 0.0008, p < .001, and condition, F(2, 58) = 32.43, MSE = 0.0007, p < .001, as well as a Blocking × Condition interaction, F(2, 58) = 16.18, MSE = 0.0005, p < .001. Planned comparisons revealed that subjects made significantly more errors on incongruent trials (6%) than on either neutral trials (2%), t(29) = 6.06, MSE = 0.0009, p < .001, or congruent trials (2%), t(29) = 7.16, p < .001. Subjects made more incongruent errors in the random location shift version than in the blocked location shift version, t(1, 29) = 5.87, MSE = 0.0009, p < .001. Neutral and congruent errors did not differ between the two versions, nor did error rates differ significantly between congruent and neutral conditions (F < 1). For both the blocked location shift version (Experiment 3B) and the random location shift version (Experiment 3A), there was no evidence of a speed-accuracy trade-off (blocked, r = .14; random, r = .17; see Table 3).

### Blocked Location Shift and Blocked Single-Trial Versions (Experiment 3B vs. Experiment 1C)

The only difference between these two versions was consistent central presentation versus location shift for the presented items. Analyses revealed a main effect of condition, F(2, 58) = 260.96, MSE = 2,767, p < .001, but no Task Version × Condition interaction, F(2, 58) = 3.27, MSE = 921, p > .05.

### Random Location Shift and Random Single-Trial Versions (Experiment 3A vs. Experiment 1B)

Analyses revealed main effects of task version, F(1, 29) = 9.70, MSE = 4,297, p < .01, and condition, F(2, 58) = 243.67, MSE = 979, p < .001, as well as a significant interaction between task version and condition, F(2, 58) = 4.35, MSE = 573, p < .05. Planned comparisons showed that facilitation did not differ between the random single-trial and random location shift versions (F < 1). Although only marginally significant, interference was greater (20 ms) in the random location shift version (114 ms).

### Table 3

**Experiment 3: Means and Standard Deviations for Response Times and Error Proportions for the Two Versions of Location Shift**

<table>
<thead>
<tr>
<th>Condition</th>
<th>RT</th>
<th>SD</th>
<th>Error SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 3A (location shift, random)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent</td>
<td>601</td>
<td>99</td>
<td>.03</td>
</tr>
<tr>
<td>Neutral</td>
<td>612</td>
<td>95</td>
<td>.03</td>
</tr>
<tr>
<td>Incongruent</td>
<td>726</td>
<td>136</td>
<td>.08</td>
</tr>
<tr>
<td>Experiment 3B (location shift, blocked)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent</td>
<td>505</td>
<td>98</td>
<td>.02</td>
</tr>
<tr>
<td>Neutral</td>
<td>563</td>
<td>80</td>
<td>.02</td>
</tr>
<tr>
<td>Incongruent</td>
<td>706</td>
<td>108</td>
<td>.04</td>
</tr>
</tbody>
</table>

Note. Response times (RTs) are in milliseconds.
ms) than in the random single-trial version (94 ms), F(1, 29) = 3.99, MSE = 729, p = .05.

Discussion

The absence of differences between the blocked location shift and the blocked single-trial versions suggests that requiring an individual to move attention from one location to another does not create increased interference when stimuli are blocked. In contrast, the increased interference observed in the random location shift version is less clear. However, this effect was small and only marginally significant. Given that the clinical version uses a blocked design, the current data demonstrate that switching location does not contribute to differences between single-trial and clinical versions. However, randomizing conditions versus blocking might make a small difference.

General Discussion

The goal of this study was to examine potential sources for differences in Stroop interference that may have influenced differences between populations reported in the literature, particularly those related to psychiatric or neurological patients. Most studies that reported increased interference in clinical populations used the clinical version of the Stroop test. This version presents many Stroop items distributed across a sheet of paper with RTs measured as a summation of multiple responses. In contrast, those studies that failed to report increased Stroop interference in patient populations used the computerized single-trial version, which allows an analysis of RT to individual Stroop items and both the elimination of outliers and the analysis of errors.

The interpretation of Stroop interference is especially important because the amount of interference (computed as the RT difference between neutral and incongruent conditions) is typically used as a measure of the ability to inhibit or ignore irrelevant information. When this difference is larger than normal, inhibitory abilities are often said to be impaired (Abramczyk et al., 1983; Bliemer, 1993; Buchanan et al., 1994; Yaki et al., 1995). This interpretation fits well with the clinical observation that certain patients appear very distractible (e.g., those with frontal lobe syndromes or with schizophrenia). However, in a test such as the Stroop test, increased interference should be observed on both clinical and single-trial versions when comparing patient groups with healthy control subjects. As noted earlier, only the clinical version has reliably reported increased interference.

Implications for Clinical Literature

By comparing different versions of the Stroop task, the present data showed that normal RTs to incongruent trials changed relatively little and in some conditions not at all. The major changes introduced by the variables we examined were in RTs to neutral stimuli, thus increasing or decreasing what is the general measure of Stroop interference across versions. The present set of findings shed some light on why there is inconsistency in the clinical literature and demonstrates that the method and the selection of neutral stimuli (that provide the baseline by which interference is measured) are critical because they clearly can change performance. Healthy control subjects were affected by these changes in systematic ways. Overall, RT was faster for the clinical version of the task (in which all conditions are blocked), somewhat slower for the blocked single-trial version, and slowest for the random single-trial version. The RT changes were largest in the neutral conditions. When the neutral condition contained identical irrelevant letters (e.g., XXX) across trials, RT was 62 ms faster in the clinical version than in the blocked single-trial version and 96 ms faster than in the single-trial version. This made the interference in the clinical version appear larger than in the other two versions. These findings make it very difficult to interpret the meaning of increased interference reported in the clinical literature. For instance, it may mean that the effect of identical neutrals on a page may actually facilitate color-naming RT for patients. We are not advocating this interpretation, but the data demonstrate that conclusions regarding differences in Stroop findings between groups need reconsideration.

It is undeniable that ignoring the word in the incongruent condition makes reporting the color more difficult. However, the current findings do question the meaning of a larger difference between neutral and incongruent RTs. In Experiment 2, we varied the physical identity of the neutral items in different blocks of trials using the blocked single-trial version only. RT to incongruent stimuli did not change, but RT to the neutral stimuli increased as similarity decreased. When only a string of Xs was used throughout a block, RT was 13 ms faster than when strings of Xs, Vs, Vs, and Ms were intermixed and 35 ms faster than when neutrals composed of animal names were presented. Again, this changed the magnitude of the interference effect as measured by conventional means. Klein’s (1964) evidence of a gradient of interference clearly had merit, even for noncolor items.

Reconsidering Stroop Interference in Schizophrenia: A Case in Point

Many studies that failed to find increased Stroop interference in patients with schizophrenia used the computerized single-trial version (Carter et al., 1992; Henik et al., 2000; Salo et al., 1996, 1997; Taylor et al., 1996). In contrast, those studies that reported increased Stroop interference among schizophrenia patients used the clinical version (Abramczyk et al., 1983; Buchanan et al., 1994; Mehl & Cromwell, 1969). The ability to inhibit interfering information appears to be intact in schizophrenia patients when the stimuli are presented one at a time, as in the computerized single-trial Stroop.

3 The summation of RT across multiple trials includes errors, self-corrections, and outlier responses, all of which may inflate RT. The summation of RT also includes all processing that occurs during the stimulus–response interval.
Facilitation (Neutral Minus Congruent)

The interpretation of increased facilitation in the blocked versions of the Stroop task is more problematic. The reduced RT to the congruent conditions in all blocked versions may be a result of subjects switching strategies and reading the word (MacLeod, 1996). The finding that RTs to congruent conditions were significantly longer in the random single-trial version (580 ms) than in both the blocked single-trial version (488 ms) and the clinical version (473 ms) supports this hypothesis.

When the absolute RTs for congruent conditions are considered in isolation, the largest effect was for blocked versus random presentation. In Experiment 1, single presentation increased RT for the congruent condition only by 15 ms compared with the congruent condition in the clinical version; however, presenting the congruent conditions intermixed increased RT by 107 ms compared with the clinical version and 98 ms compared with the blocked single-trial version. The random--blocked manipulation produced similar effects in Experiments 3A and B, in which RT was 96 ms faster for the blocked than for the random presentation.

These findings are consistent with the suggestion that subjects read the word rather than name the color in congruent conditions. However, it demonstrates that this is far more likely when there is less uncertainty about the type of stimulus condition that will appear, when the conditions are blocked. In both the clinical and blocked single-trial versions of the task, the subject knows that the words will match the colors in all cases, whereas in the random version the strategy of reading the word is not as effective and would have the undesirable effect of producing more errors when an incongruent stimulus appeared.

For this reason, the random version is preferable because it reduces the likelihood that subjects will read the word. In fact, when random presentation was introduced, Stroop facilitation was statistically absent by conventional measures (7 ms in Experiment 1B and 11 ms in Experiment 3A). Randomly presented congruent stimuli may, in fact, be the best neutral type by which to evaluate interference because the words used do not differ from those in the incongruent condition. By randomly presenting the stimuli, concerns about the strategy of word reading appear to be reduced.

Shifts of Location

The clinical version of the Stroop task requires a shift of attention from one word to the next. In healthy, normal individuals, it has been demonstrated that advance knowledge of a target's position can greatly reduce interference from distracting flankers (Kahneman & Henik, 1981). This advance knowledge of target position may allow subjects to focus on the location to be attended, thus narrowing the spotlight of attention to the relevant target stimulus (Eriksen & Yeh, 1985). The finding of increased interference in the clinical version (190 ms), in which conditions are blocked, compared with the blocked single-trial version (159 ms), suggests that blocking of conditions may not be the only factor that contributes to greater interference in the clinical Stroop task. It may be that requiring a subject to disengage attention from a previous target after response and a shift of attention to a subsequent target in the absence of a location cue causes a slowing in RT to the incongruent stimulus. Because the location shift experiments (3A and 3B) involved the abrupt onset of words on each trial, the stimulus presentation might have triggered exogenous or reflexive orienting of attention. The abrupt appearance of each Stroop word on the screen in Experiment 3 differs from the more continuous presentation of the clinical version. More experiments are needed to understand the impact of location shifts on Stroop interference.

Conclusion

Standard neuropsychological assessment tools are valuable for understanding the nature of cognitive functioning in both healthy individuals and patient populations. They provide important data concerning the cognitive sequelae of both neurological and psychiatric disorders. However, failures to replicate at least some of the traditional findings from neuropsychological tests suggest that the interpretation of the results might be somewhat obscured by the global nature of these tasks and the multiple cognitive operations contained within them. In addition, the inconsistency in the literature might be due to methodological issues such as interpretation of baseline measures, mode of stimulus presentation, and differences in how RTs are measured (i.e., summation vs. individual trials), all of which could impact performance and test results. Attention to methodological differences when interpreting studies in patient populations is essential if we are to better understand patterns of cognition in patients and healthy populations.

References


Received March 9, 1998
Revision received February 5, 2001
Accepted March 31, 2001