Task Format Modulates the Relationship between Reading Ability and Stroop Interference

Laoura Ziaka[1,2], Dimitra Skoteinou[1], and Athanassios Protopapas[2]

[1]Department of Philosophy and History of Science, University of Athens, Greece,
[2]Department of Special Needs Education, University of Oslo, Norway

Author note

Laoura Ziaka https://orcid.org/0000-0002-0559-7331

Athanassios Protopapas https://orcid.org/0000-0002-7285-8845

Data and analysis scripts are available at https://osf.io/pjvrc/

Correspondence should be addressed to Laoura Ziaka, Department of Special Needs Education, University of Oslo, P.O. box 1140 Blindern, 0318 Oslo, Norway. E-mail: laoura.ziaka@isp.uio.no

Word count: 8,670

Abstract

Previous research has shown that Stroop interference and reading ability are negatively related, with higher reading skills associated with less interference. A direct link between interference and the speed of inhibition of the task-irrelevant dimension (i.e., word) has been proposed to explain this relationship. If that were the case then it should apply regardless of the format of the Stroop task, that is, whether stimuli are presented simultaneously (multi-item version) or individually (single-item version). Here we examine data from six experiments using single-item and multi-item Stroop tasks and their relationship to reading measures. Our results indicate that reading performance is primarily related to the multi-item version of the Stroop task and not to the single-item version. These findings question the direct link between inhibition and interference as an interpretation of the reading–interference relationship. We argue that cascaded processing of successive items, and the ability to monitor and control this process, is the cognitive mechanism regulating the relationship between reading and interference. Therefore, we propose that the link between Stroop interference and reading is indirect and their relationship is determined by the efficiency in temporally overlapping processing of adjacent items.

Public Significance Statement— When compared to poor readers, skilled readers have been previously found to be less prone to word interference when they have to name the color in which a different color-word is printed. This has been interpreted as a result of speedy blocking of the irrelevant word, which is available earlier in better readers. Our experiments show that skilled reading is associated with higher color-naming performance only when items are presented simultaneously in an array, as in previous studies, but not when they are displayed one by one. This format-specific association suggests that the advantage of skilled readers is not due to faster word recognition but perhaps due to efficient simultaneous processing of adjacent items, thereby weakening the theoretical claim

of a direct connection between individual word recognition speed and interference from words to color naming.

Task Format Modulates the Relationship between Reading Ability and Stroop Interference

The concept of automaticity emerged early in the domain of experimental psychology along with the distinction between automatic and controlled processes. Automatic processes are thought to be fast, unintentional, obligatory, and effortless, whereas controlled processes are slow, intentional, effortful, and capturing attentional resources (Cohen et al., 1990; Moors & De Houwer, 2006). Cattell (1886) first introduced the notion of automaticity arguing that word reading becomes highly automatic due to extensive practice. To this day, skilled word reading is considered to be fast, unavoidable, and effortless; that is, an automatic process.

One of the most extensively used tasks in the study of automatic and controlled processes is the Stroop task. In the incongruent condition of the task color words are printed in a different color (e.g., the word "red" in green colored letters) and the participants' task is to name the color and ignore the word, whereas in the control condition participants have to name the color of neutral stimuli (such as color patches or arrays of Xs). It is a robust finding that it takes longer to name colors in the incongruent condition compared to the control condition; this difference in response times between conditions is termed Stroop interference. The automaticity account proposed by MacLeod (1991) assumes that Stroop interference stems from the highly automatized process of reading. That is, word reading is unavoidable and interferes with the less practiced task of color naming, causing response delays due to conflict. A connectionist model of this account (Cohen et al., 1990) has implemented the crucial insight that the strength of interference mirrors the degree of reading automaticity. If this were the case, then skilled readers should suffer from greater interference, compared to developing readers or readers with difficulties.

However, a number of studies have reported the opposite result, namely that children with reading difficulties exhibit greater interference than typically developing readers. Everatt et al. (1997) showed that children with dyslexia exhibited greater interference than age–

matched controls. This finding has been confirmed across languages and ages (Di Filippo & Zoccolotti, 2011; Faccioli et al., 2008; Helland & Asbjørnsen, 2000; Kelly et al., 1989; Reiter et al., 2005; Kapoula et al. 2010; Protopapas et al., 2007; Wang & Gathercole, 2015). In some studies, the group differences in interference have been attributed to inhibition impairments (van der Schoot et al., 2000; Reiter et al., 2005; but cf. Wang & Gathercole 2015) or to a general slowness of readers with dyslexia (Di Filippo & Zoccolotti, 2011). In contrast, Protopapas et al. (2007) proposed that there is a direct link between reading skill and Stroop interference, not involving executive functions, causing reading speed to be negatively correlated with interference. In their first study, children with dyslexia were compared to age-matched controls and were found to exhibit greater interference. In the second study, the continuous relationship between reading skill and Stroop interference was examined in the general school population and poorer reading skills were found to be associated with greater interference.

Based on a production model of interference (Roelofs, 2003), Protopapas et al. (2007) posited that reading skill affects the time course of suppression of the task-irrelevant response. Specifically, poor readers are slower in reading the word, which therefore takes longer to emerge. Because a task-irrelevant response can only be suppressed after it has been activated and detected as inappropriate, the delayed activation of the word leads to its delayed suppression, before the task-relevant color naming response can prevail. These consecutive delays amount to greater interference, compared to skilled readers, whose speedy word reading permits rapid suppression of the task-irrelevant dimension and therefore faster task-relevant response, that is, less interference. This proposal was subsequently supported by a practice study. Specifically, Protopapas et al. (2014) found that word reading practice (i.e., reading the color words) in children attending Grades 4 and 5 reduced Stroop interference whereas color naming practice did not. This is consistent with a direct link between reading ability and Stroop interference hinging primarily on the speed of word reading.

One aspect of the Stroop task that has not been sufficiently scrutinized in relation to reading ability is the format of the task. Specifically, two different formats of the Stroop task are commonly used in the literature, that is, the *multi-item* format, in which all stimuli are presented simultaneously on a sheet or computer screen, and the *single-item* format, in which each stimulus appears individually on the screen. The single-item format can be administered blocked, with each condition presented in a separate block of trials, or mixed, with all conditions presented within the same block. It is notable that, with the sole exception of Faccioli et al. (2008), who adopted the single-item mixed version, all other studies reporting a disadvantage of poor readers compared to skilled readers in the Stroop task have implemented the multi-item version. The question may therefore be raised as to whether the reported relationship between Stroop interference and reading ability may be specific to or dependent on the multi-item format of the Stroop task.

Evidence from the Rapid Automatized Naming (RAN) literature seems to favor the idea that the interference-reading relationship may be modulated by task format. RAN tasks require rapid naming of repeated familiar items, and may be composed of alphanumeric (i.e., letters and numbers) or non-alphanumeric stimuli (i.e., colors and objects; for a review see Wolf & Bowers, 1999, and Norton & Wolf, 2012). Like the Stroop task, rapid naming tasks can be either serial (i.e., multi-item) or discrete (i.e., single-item). It has long been known that reading fluency correlates more strongly with serial naming tasks than with discrete naming tasks (e.g., Bowers & Swanson, 1991; Chiappe et al., 2002; de Jong, 2011; Protopapas et al., 2013, 2018; Stanovich et al., 1983). This robust finding has been termed serial superiority effect (Altani et al., 2017) and suggests that fluent reading and serial naming of familiar symbols share common cognitive processes which are absent in discrete naming tasks.

Serial Stroop and RAN are generally thought to resemble each other (Norton & Wolf, 2012), to the extent that some studies treat the control condition of the Stroop task as a color RAN task (e.g., Helland & Morken, 2016; Stringer et al., 2004), while it is found that RAN

and Stroop tasks load on the same factor as similar naming tasks (Di Filippo & Zoccolotti, 2011). Moreover, it has been argued that not only RAN, but also word list reading and, presumably, passage reading can been seen as serial naming tasks, because they also require recognition of individual items (i.e., words) to be named as well as sequential processing of multiple items (Altani et al., 2019, 2020; Protopapas et al., 2018; Zoccolotti et al., 2014). Shared demands across these tasks are crucial when considering the Stroop task and its relationship to reading. In particular, they suggest that reading will have a stronger relationship to the multi-item than the single-item Stroop task because the multi-item format (a) is essentially a serial naming task and (b) encompasses word list reading, given that reading emerges to produce the "irrelevant response" in the incongruent condition of the task.

Previous studies comparing the single-item (i.e., discrete) and multi-item (i.e., serial) formats of the Stroop task show that the multi-item format is faster than the single-item format, at least in the neutral condition (MacLeod, 2005; Salo et al., 2001). That is, a serial advantage is observed, consistent with findings in other naming tasks (Altani et al., 2017, 2019; Zoccolotti et al., 2013). This is thought to be made possible in the multi-item format because of temporal overlap in processing of successive items (Altani et al., 2020), termed "cascading" (Altani et al., 2019; Protopapas et al., 2013, 2018). It is thus reasonable to assume that temporally overlapped sequential processing will also be present in the multi-item Stroop task and, consequently, will affect the Stroop-reading relationship as an (additional) cognitive mechanism shared between them. Specifically, we hypothesize that cascaded processing efficiency boosts the relationship between Stroop interference and reading in the serial format of the Stroop task.

To sum up, the aim of the present study was to examine whether the format of the Stroop task affects the Stroop–reading relationship. If the speed of inhibition hypothesis is the only reason underlying their association, then the format should have little effect, because inhibition of the irrelevant dimension in the incongruent condition of the Stroop task is

mandatory irrespective of task format. In contrast, if cognitive mechanisms related to serial processing play a crucial role over and above the speed of inhibition, then task format will modulate the Stroop-reading relationship such that it will be strongest when the serial format of the Stroop task is used. To test this hypothesis, we revisit six experiments conducted using single-item and multi-item Stroop tasks and analyze their relationship to reading in adults and children.

Method

Table 1 summarizes the characteristics and measures used in all experiments (Skoteinou, 2008; Protopapas et al., 2014; Ziaka, 2014) included in the analysis.

Studies involving children were approved by the Institute of Educational Policy of the Greek Ministry of Education, Research and Religious Affairs as per applicable regulations and requirements. Written informed consent was obtained from all adult participants and from the parents/legal guardians of non-adult participants. All participants were native speakers of Greek.

For all discrete word reading and single-item Stroop tasks, individual items were presented at the center of a computer screen controlled by DMDX scripts (Forster & Forster, 2003). Individual responses were automatically recorded as wav files for offline determination of response times.

Experiment 1

Participants. The sample consisted of 114 adults (76 women) 18–30 years old.

Tasks and Material

Reading Measures

Discrete word reading. One hundred and ten words were presented individually on a laptop screen. Participants were asked to read the words as fast as possible and try to avoid errors. The words ranged from three-syllable to seven-syllable long (31 three-syllable, 48

four-syllable, 26 five-syllable, 4 six-syllable, and 1 seven-syllable). Words were selected from the word fluency list of Protopapas, Sideridis, et al. (2007) used in reading research.

Stroop Task

Single-item Stroop task. One hundred and forty-four stimuli were presented individually on a laptop screen. Red (κόκκινο /kocino/), green (πράσινο /prasino/), and yellow (κίτρινο /citrino/) were used. The items were 72 color-words appearing in a nonmatching color (e.g, "green" in red ink; incongruent condition), 36 non-word items (e.g., XXXX in red ink; control condition) and 36 stimuli where word and color matched (e.g, "red" in red ink; congruent condition). The presentation of all conditions was mixed and the presentation of items was randomized. Here, only the incongruent and control condition will be analyzed. Each color appeared an equal number of times within each condition, that is, 12 times in the control condition, 12 times in the congruent condition, and 24 times in the incongruent condition. Participants were asked to name the color of the ink as fast as possible and try to avoid errors.

Experiments 2a and 2b

Participants. The study included adults (Experiment 2a) and children (Experiment 2b) from the general school population. The adult sample consisted of 51 individuals 18–30 years old. The school sample consisted of 52 children attending Grade 7 (i.e., 13–14 years old).

Tasks and Material

Reading Measures

Discrete word reading. As in Experiment 1.

Serial Word Reading. For the word reading task, a sheet of paper was given, on which the 84 items to be read were printed in three columns of 28 words each, and the participant was asked to read aloud the words "quickly, but not rushing, to avoid mistakes." Individual responses were noted when incorrect, and the duration of reading was timed using a stopwatch. Test results include the number of incorrectly read items (regardless of number of errors per item; unread or incomplete items count as one error each) and the total reading

time, in seconds. The words ranged from one-syllable to seven-syllable (13 one-syllable, 19 two-syllable, 33 three-syllable, 13 four-syllable, 4 five-syllable, 1 six-syllable, and 1 seven-syllable). This list was taken from Protopapas and Skaloumbakas (2007).

Stroop Tasks

Single-item Stroop task. As in Experiment 1.

Multi-item Stroop task. Two sheets were prepared, one with color stimuli made up of 5 repetitions of the letter X with no spaces (control condition), and the second with color stimuli being the Greek words for red (κόκκινο /kocino/), green (πράσινο /prasino/), blue (μπλε /blé/), yellow (κίτρινο /citrino/), and brown (καφέ /kafé/; incongruent condition). There were 60 stimuli on each sheet, arranged in 3 columns of 20. The stimuli were printed in colored ink, with 12 items on each sheet printed in each of the five named colors, randomly arranged throughout the sheet but counterbalanced in their combinations and equally distributed over columns. In the incongruent condition sheet the correct color-response was identical to the incorrect word-response of the preceding item for 18.3% of the items (i.e., 11 out of 60; see Figure S.1 in Supplemental Material). Errors and total time were noted on the scoring sheet.

Experiment 3

Participants. The sample consisted of 39 children attending Grades 4–5 (i.e., 9–11 years old).

Tasks and Material

Reading Measures

Discrete word reading. Eighty words were presented individually on a laptop screen. Participants were asked to read the words as fast as possible and try to avoid errors. There were 56 three-syllable and 24 four-syllable words. Words appeared on the screen in white 40-pt Arial font on a black background for 2 s.

Stroop task

Single-item Stroop task. The Greek words for red, green, and yellow were used. Stimuli for the neutral color condition were made up of 7 repetitions of the letter X (no spaces) in red,

green, and yellow color. For the incongruent condition the Greek words for red, green and yellow appeared in a non-matching color. All stimuli were presented on a black background. Presentation of conditions was blocked.

Each condition consisted of 24 stimuli, which were presented individually in random order on a 15.5″ laptop screen in white 40-pt Arial font on black background. Four practice trials for the neutral condition and 10 for the incongruent one preceded data collection.

Experiment 4

Participants. The sample consisted of 41 adults (32 women) 21–38 years old, mainly undergraduate and graduate students.

Tasks and Material

Reading measures and Stroop tasks were presented on a 15.5″ laptop screen (resolution 1280×1024 pixels) on a fullscreen Miscrosoft Powerpoint presentation with black background.

Reading Measures

Discrete word reading. One hundred and ten words were presented individually on the laptop screen in white 40-pt Arial font. Participants were asked to read the words as fast as possible and try to avoid errors. The words ranged from three-syllable to six-syllable (31 three-syllable, 49 four-syllable, 26 five-syllable, and 4 six-syllable).

Serial Word Reading. Two slides including words were used. Words were presented in white 20-pt Arial font. Each slide included 50 items of increasing difficulty in five columns of ten rows, including bisyllabic and trisyllabic words. Participants were instructed to read the words as quickly as possible, avoiding errors, from left to right (i.e., across rows) for one slide, and from top to bottom (i.e., down columns) for the other one.

Stroop Tasks

Single-Item Stroop task. As in Experiment 3.

Multi-item Stroop task. Same material as in Experiment 3.

Each condition was presented in a single-screen array of three columns of 20 stimuli, for a total of 60 stimuli per condition, displayed in 20-pt Arial font. Each item extended 5×20 mm on the screen. The vertical distance between adjacent items (nearest edges) was 6 mm and the horizontal distance between columns was 40 mm. For the incongruent condition there were 20 repetitions of each word and 20 repetitions of each color with colors being counterbalanced over columns. The correct color-response was identical to the incorrect word-response of the preceding item for 46.6% of the items (i.e., 28 out of 60; see Figure S.2 in Supplemental Material). For the neutral condition 20 repetitions of each color were used. Colors and color words were randomly ordered with the constraint that adjacent items were not the same. Prior to the tasks the production of the intended colors and understanding of the instructions were verified using sample cards.

Experiment 5

Participants. The total sample consisted of 43 adults (31 women) 21–36 years old, mainly undergraduate and graduate students.

Reading Measures

Discrete word reading. Fifty words of increasing difficulty were presented individually on a laptop screen. Specifically, 25 three-syllable and 25 four-syllable words of increasing difficulty were used. Participants were asked to read the words as fast as possible and try to avoid errors.

Serial Word Reading. As in Experiment 4 with the difference that three-syllable and four-syllable words were used.

Stroop tasks

Single-item Stroop task. The single-item task consisted of 60 stimuli in 20-pt Arial font displayed in random order on a black background. Each stimulus remained on the screen for 2 s. The interstimulus interval was 167.7 ms. Presentation of conditions was blocked; each condition lasted approximately 3 minutes. Three representative stimuli from each condition

were presented to the participants prior to the single-item version of the task to verify understanding of the instructions.

Multi-item Stroop task. As in Experiment 4.

Experiment 6

Participants. The sample consisted of 45 children attending Grades 4–5.

Reading Measures

Discrete word reading. Fifty words of increasing difficulty were presented individually on a laptop screen. Twenty-five three-syllable and 25 four-syllable words of increasing difficulty were used. Participants were asked to read the words as fast as possible and try to avoid errors.

Serial Word Reading. As in Experiment 4.

Stroop tasks

Single-item Stroop task. As in Experiment 5.

Multi-item Stroop task. As in Experiment 5.

Dependent Measure

For all experiments and tasks the dependent measure was mean response time per item in milliseconds. Following standard practice in the field, for the serial reading and multi-item Stroop tasks, this was calculated by dividing the total time spent reading/naming over the number of displayed stimuli, ignoring errors. For the discrete reading and single-item Stroop tasks, time for each item was measured from the moment the item appeared on the screen until the onset of the spoken response, determined offline on the recorded responses using CheckVocal (Protopapas, 2007). Raw times for correct responses were then averaged by participant and condition. Thus, response times included articulation in serial and multi-item tasks but not in discrete and single-item tasks. (See also Salo et al., 2001, for a similar approach to single-item and multi-item tasks).

Table 1

Summary of Sample and Measure Information for all Experiments

Experiment	Population	N	Single-item Stroop task	Multi-item Stroop task	Discrete Reading	Serial Reading (TB)	Serial Reading (LR)
Experiment 1	Adults	114	✓[a,b]		✓		
Experiment 2a	Adults	51	✓[a,b]	✓	✓	✓	
Experiment 2b	Children (G7)	52	✓[a,b]	✓	✓	✓	
Experiment 3	Children (G4-5)	39	✓		✓		
Experiment 4	Adults	41	✓	✓	✓	✓	✓
Experiment 5	Adults	43	✓	✓	✓	✓	✓
Experiment 6	Children (G4-5)	45	✓	✓	✓	✓	✓

Note. TB: top-to-bottom (list) scanning; LR: left-to-right scanning; G=Grade(s). [a]mixed presentation of conditions; [b]congruent condition included.

Results

Results are first presented using the traditional calculation of interference by subtraction of condition means, in comparison to individual conditions. Because of important issues that have been raised concerning this kind of calculation, especially when used for individual differences analyses, analyses of reliability are presented next followed by regressions using residualized scores.

Spearman's Rank Correlations

For this part of the analysis, Stroop interference was calculated for each participant as the difference in mean time to name the color in the incongruent condition minus the mean time to name the color in the neutral condition (across all items). Spearman's rank correlations between reading measures and the individual Stroop conditions as well as with interference were calculated in R 4.0.3 (R Core Team, 2020). We used Spearman's coefficients to ensure that results are not affected by deviations of variables from the normal distribution. For the forest plots we used functions metacor (of the metacor package; Laliberté, 2019) and forest (of the meta package; Balduzzi et al., 2019). Descriptive statistics for each experiment and task are presented in the Supplemental Material. Figure 1 displays correlations between reading measures and Stroop interference and individual conditions, and their meta-analytic overall effect estimate; correlations with *serial* reading measures are shown on the top row and with *discrete* reading measures on the bottom row. Performance for all tasks is measured in terms of time, so that greater values are associated with slower (i.e., poorer) performance in the reading measures and individual Stroop conditions. Thus the predicted negative relationship, namely greater Stroop interference associated with poorer reading, would manifest itself in *positive* correlation coefficients in these analyses.

Correlations With Serial Reading

Multi-item Format. For serial reading and Stroop interference in the multi-item format the meta-analytic effect was close to zero (Figure 1, top left). However, Spearman's rank correlation varied from low to moderate for the individual Experiments. Presumably, the negligible overall effect emerges because the combination of negative and positive correlations cancels out.

Part of this variation seems to stem from correlations between serial reading and individual Stroop conditions of the multi-item version. The meta-analytic effect for the neutral condition was 0.44 [95% CI 0.35, 0.53] (Figure 1, top middle) with all individual correlations being somewhat low to high, whereas for the incongruent condition it was 0.33 [95% CI 0.23, 0.43] (Figure 1, top right) and individual correlations varied from low to moderate, with the exception of Experiment 2b (Grade 7 children) where the correlation was high ($r_s = 0.57$). Inspection of correlations in individual experiments suggests that when the neutral condition captured more reading variance than the incongruent condition, then the correlations of reading with interference (i.e., the difference between the two conditions) were negative or close to zero (Experiments 2a, 4, and 5).

Single-item Format. For serial reading and Stroop interference in the single-item version, the meta-analytic overall effect estimate was essentially zero and correlations in all experiments were very low ($|r_s| \leq 0.14$), with little variation (Figure 1, top left).

For the individual Stroop conditions (i.e., neutral and incongruent; Figure 1, bottom middle) the meta-analytic effect estimate was close to that for multi-item format (if slightly lower), with no substantial variation between conditions in individual experiments.

Correlations With Discrete Reading

Multi-item Format. For discrete reading and Stroop interference in the multi-item version the meta-analytic effect was low to moderate (Figure 1, bottom left). For individual

experiments in children, correlations varied from moderate to high, whereas for adults from low to moderate.

More importantly, the meta-analytic effect estimate for the individual conditions manifested itself differently than in serial reading tasks. Specifically, for the neutral condition (Figure 1, bottom middle) the total effect was moderate (0.35) with the individual experiments varying from somewhat low (0.28) to high (0.54) with the exception of Experiment 2a in adults where correlations were close to zero. For the incongruent condition (Figure 1, bottom right) the meta-analytic effect was moderate (0.43) with correlations for individual experiments between 0.22 (Experiment 2a) to 0.59 (Experiment 2b) and the rest lying in between. This combination allowed to the meta-analytic overall effect estimate for Stroop interference as a difference index to appear enhanced when compared to serial reading and not to cancel out.

Single-item Format. For the single-item version of the Stroop task and its relation to discrete reading the meta-analytic effect estimate was low (Figure 1, bottom left). All correlations for individual Experiments were low ($|r_s| \leq 0.18$) and positive.

The meta-analytic effect estimate for the individual Stroop conditions was high for both the neutral and the incongruent condition. For the neutral condition (Figure 1, bottom middle) all individual correlations were moderate to high and a similar pattern of results was observed for the incongruent condition (Figure 1, bottom right).

Figure 1

Correlations of Stroop Interference and Individual Stroop Conditions with Serial and Discrete Reading Measures
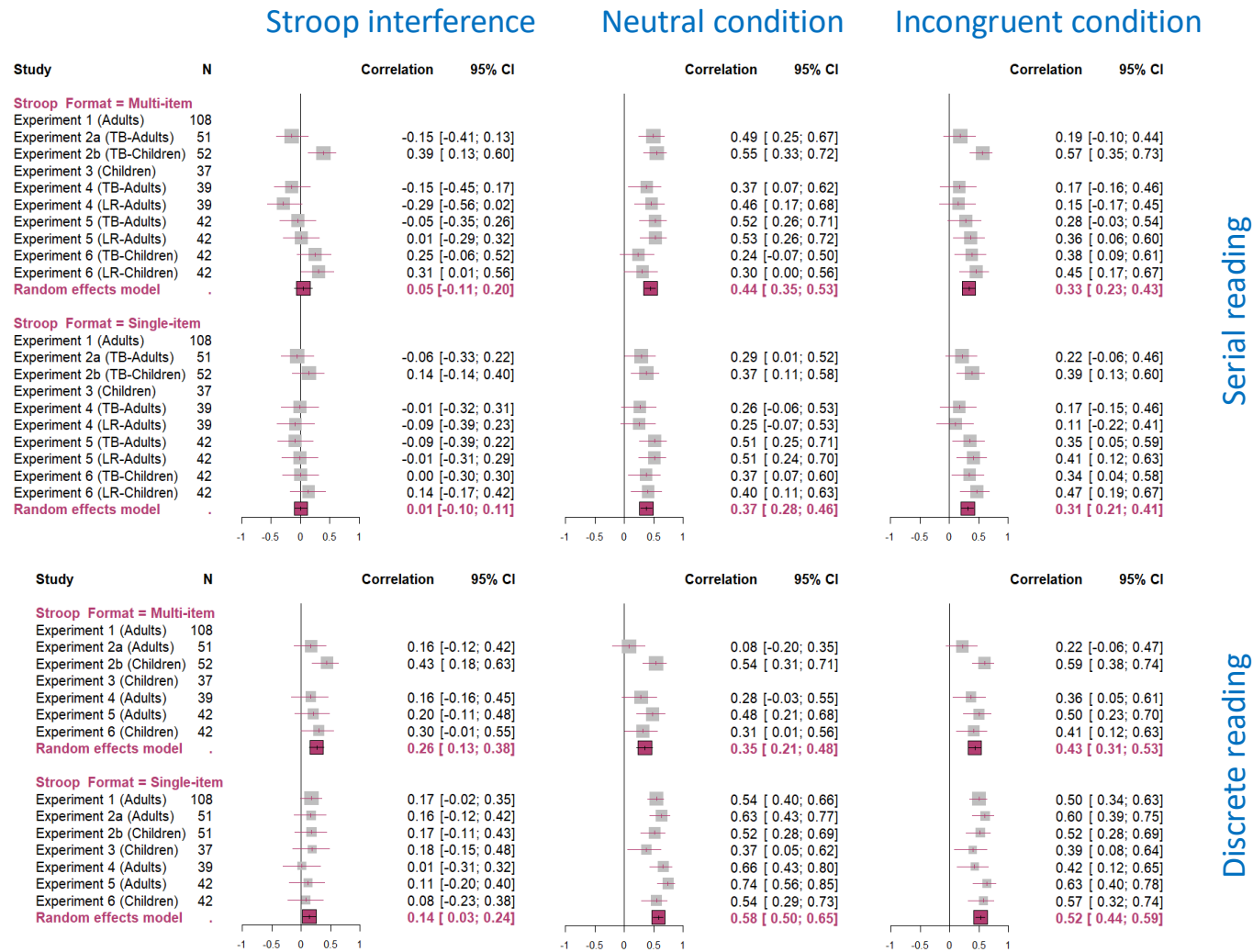
Note. Spearman's rank correlations and their overall effect estimate of interference (calculated as the difference between the incongruent and the neutral condition) and individual Stroop conditions with reading measures. Error bars show 95% confidence intervals.

Internal Consistency: Split-half Correlations

Following the recommendation of Parsons et al. (2019), we computed split-half reliability estimates for the Stroop tasks of Experiments 3–6[1], taking into account that difference scores (as in Stroop interference) are less reliable than their component scores if (a) the component scores are not perfectly reliable and (b) the two component scores are not completely independent (Draheim et al., 2019). In the first step of the analysis, we calculated Spearman's rank correlation for overall performance between individual conditions for each Stroop task. Split-half reliability for individual conditions and interference was computed using the Spearman-Brown formula (Brown, 1910; Spearman, 1910).

Specifically, for Experiment 3 and the single-item version of Experiment 4 we divided items based on order of presentation, that is, first half and second half. For the multi-item task of Experiment 4 and all Stroop tasks of Experiment 5 and 6, we excluded the first 20 items from the analysis[2] and computed reliability for the second and third part (column) of the tasks in a fixed order[3]. In the second step, we calculated split-half reliability measures for interference, by first computing interference from the corresponding parts of the control and incongruent condition.

Table 2 presents Spearman's rank correlation among the individual conditions (neutral and incongruent) of the Stroop task, and the corresponding interference, and split-half reliability estimates for each experiment and format. As shown in Table 2, correlations between the neutral and incongruent condition were high in all experiments and in both task versions. Split-half reliability for the neutral condition ranged between 0.80 and 0.93 for the single-item format and between 0.78 and 0.90 for the multi-item format. Somewhat greater

---

[1] Due to lack of trial-level data, split-half reliability estimation was not possible for Experiments 1 and 2.
[2] Based on a previous study in our laboratory, the first column (20 items) in the multi-item version of the Stroop task behaves differently compared to subsequent items (Ziaka & Protopapas, under review)
[3] Because it is not possible to perform permutation-based split-half analysis for the multi-item version of the Stroop task, a fixed order was adopted for all tasks. However, the Spearman-Brown formula was preferred to Cronbach's alpha in order to account for underestimation when splitting the number of observations into halves (Parsons et al., 2019).

variation was obtained for the incongruent condition in both formats, with split-half reliability ranging between 0.60 and 0.94.

It is therefore evident that component scores are neither completely reliable nor independent. We thus computed split-half reliability for the interference scores for each experiment and format. As shown in Table 2, there was great variability in the reliability of Stroop interference, which ranged between 0.27 and 0.85 in the multi-item version and between 0.43 and 0.71 in the single-item version.

To sum up, split-half reliability measures of interference in Experiments 3–6 showed that reliability fluctuated from low to high, indicating that interference as a result of subtraction may exhibit weak associations with other variables (here, reading measures) and may underestimate individual differences (Hedge et al.,, 2018; Draheim et al., 2019; Parsons et al., 2019). To mitigate this problem our final analysis captured interference as residualized scores.

Table 2

Intercorrelations and Reliability of Stroop Task Conditions and Interference

Measure	Single-item			Multi-item		
	N	I	S	N	I	S
Experiment 3 (children)						
Neutral Condition	**0.80**			–		
Incongruent Condition	0.65	**0.80**		–	–	
Stroop Interference	0.02	0.72	**0.43**	–	–	–
Experiment 4 (adults)						
Neutral Condition	**0.90**			**0.90**		
Incongruent Condition	0.53	**0.76**		0.56	**0.60**	
Stroop Interference	–0.04	0.79	**0.68**	–0.32	0.50	**0.27**
Experiment 5 (adults)						
Neutral Condition	**0.93**			**0.93**		
Incongruent Condition	0.83	. **0.94**		0.69	**0.92**	
Stroop Interference	0.11	0.60	**0.71**	0.02	0.66	**0.85**
Experiment 6 (children)						
Neutral Condition	**0.92**			**0.78**		
Incongruent Condition	0.63	**0.85**		0.67	**0.83**	
Stroop Interference	–0.18	0.57	**0.61**	0.08	0.72	**0.63**

Note. Spearman's rank correlations among the individual conditions (neutral and incongruent) of the Stroop task in the single-item and multi-item format, and the corresponding interference (calculated as the difference between the two conditions), below the diagonal. Split-half reliability for both conditions and Stroop interference, calculated using the Spearman-Brown formula, on the diagonal in bold. N=neutral condition; I=incongruent condition; S=Stroop interference.

Difference/Interference Scores as Residualized Scores

In our final analysis Stroop interference was expressed as the residual of the incongruent condition regressed on the baseline (i.e., neutral condition; using function lm), rather than as difference scores. This approach allowed us to focus on the relationship between reading and interference (specifically affecting the incongruent condition) by excluding variability accounted for by the neutral-condition baseline (Draheim et al., 2019). Residualization is preferable to difference scores because residualized scores provide best within-construct correlations (Kane et al., 2016). Specifically, in the first step we regressed the incongruent condition on the corresponding neutral condition and in a second step we regressed the reading measure on the residuals of the first step.

An initial round of regression analyses was performed with the raw times (and differences), resulting in residuals diverging substantially from the normal distribution in most cases. Therefore, all data were subsequently log-transformed and all analyses reported below concern log-transformed data. In addition, one participant in Experiment 6 was excluded as a severe outlier, leaving 41 participants for analysis. There was no difference in the pattern of results between the analyses with transformed and untransformed data. See Tables S.8 and S.9 in the Supplemental Material for Shapiro-Wilk normality tests of the residuals using raw and log-transformed data, and Table S.10 for results of analyses using untransformed data.

Table 3 presents the results of these regression analyses (i.e., R^2, calculated using function lm of base R) for the single-item and multi-item Stroop tasks for each experiment and each reading measure. For the single-item Stroop task, the proportion of variance of the incongruent condition that was accounted for by the neutral condition ranged greatly (between .29 and .84) across experiments. However, regressing a reading measure on the residuals in the second step did not account for any significant additional variance. In most

cases, R^2 was less than .01; it was highest (.05) in Experiment 6, with children in Grades 4–5, but this was not significant.

For the multi-item Stroop task, the proportion of variance of the incongruent condition that was accounted for by the neutral condition varied less than in the single-item format, ranging between .29 and .53 across experiments. When regressing discrete reading on the residuals in a second step, residuals accounted for a significant additional proportion of variance (up to 9% in Experiment 5) in three out of five experiments (with both children and adult populations). The additional variance was not significant (R^2 of .03) in the other two experiments. In contrast, in serial reading measures residuals accounted for a marginally significant additional proportion of variance (5%, $p = .087$) in only one out of five experiments, ranging between less than 1% and 4% in the other four (not significant), across adult and children samples. Thus, overall, when compared to the single-item format, the multi-item format exhibited more shared variance between reading measures and the incongruent condition after controlling for the neutral condition.

Table 3

Results of Residualized Scores Regression Analyses on Log-transformed Data

Step	Independent variable	Dependent variable	Population	Single-item		Multi-item	
				B	*R²*	*B*	*R²*
	Experiment 1		Adults				
1	Neutral	Incongruent		.98	.79	–	–
2	Residuals	Discrete Reading		−.04	< .01	–	–
	Experiment 2a		Adults				
1	Neutral	Incongruent		.92	.84	.64	.29
2	Residuals	Discrete Reading		.37	.01	.38	.08*
2	Residuals	Serial Reading (TB)		−.15	< .01	.02	< .01
	Experiment 2b		Children (G7)				
1	Neutral	Incongruent		.85	.73	.91	.53
2	Residuals	Discrete Reading		.23	< .01	.41	.08*
2	Residuals	Serial Reading (TB)		.54	.01	.55	.05
	Experiment 3		Children (G4–5)				
1	Neutral	Incongruent		.80	.47	–	–
2	Residuals	Discrete Reading		.32	.03	–	–
	Experiment 4		Adults				
1	Neutral	Incongruent		.71	.29	.49	.29
2	Residuals	Discrete Reading		.04	< .01	.28	.03
2	Residuals	Serial Reading (TB)		−.12	< .01	−.15	< .01
2	Residuals	Serial Reading (LR)		−.17	< .01	−.34	.02
	Experiment 5		Adults				
1	Neutral	Incongruent		.94	.65	.83	.45
2	Residuals	Discrete Reading		.15	.01	.26	.09*
2	Residuals	Serial Reading (TB)		−.15	< .01	−.07	< .01
2	Residuals	Serial Reading (LR)		−.09	< .01	−.03	< .01
	Experiment 6		Children (G4–5)				
1	Neutral	Incongruent		.65	.38	.77	.41
2	Residuals	Discrete Reading		.25	.03	.19	.03
2	Residuals	Serial Reading (TB)		.23	.01	.25	.02
2	Residuals	Serial Reading (LR)		.52	.05	.35	.04

Note. For Step 1, the dependent variable was the incongruent condition of the Stroop task in

single-item and multi-item format and independent variable was the corresponding neutral

condition. Step 2 refers to a follow-up analysis in which the dependent variable was the indicated reading measure and the independent variable was the residuals from Step 1. TB: top-to-bottom (list) scanning; LR: left-to-right scanning; G=Grade(s); *p < .05

In sum, examining the prediction of performance on the incongruent condition after controlling for performance on the neutral condition (instead of calculating interference as a difference score) indicated that word reading (both discrete and serial) was not significantly related to the single-item version of Stroop interference. In contrast, reading was significantly related to the multi-item version of Stroop interference in some experiments. The amount of shared variance varied, but was always higher for discrete than for serial reading.

Discussion

The findings of the reported study question the direct link between Stroop interference and reading due to speed of inhibition and show that the relationship under examination is format dependent. Stroop interference and reading are consistently related only in the multi-item Stroop task format. If present, the hypothesized direct link between interference and reading should have manifested itself in a significant association between discrete word reading and interference as measured in the single-item format. This would be in accordance with the theoretical proposal that the reading-interference link is based on the timing of inhibition during the processing of individual incongruent stimuli (Protopapas et al., 2007). The lack of a significant association in the single-item format across experiments and analytical approaches indicates that the proposed link is either very weak or altogether absent.

Specifically, the meta-analytic overall effect estimate for the correlation between Stroop interference and discrete reading was 0.24 for the multi-item format, with individual correlations ranging from low to moderate, whereas it was only 0.15 for the single-item version, with individual correlations being close to zero or very low ($|r_s| \leq 0.18$)[4]. Interference in the single-item format of the Stroop task did not correlate significantly with

---

[4] Note that for serial reading and Stroop interference in the multi-item format the negligible overall effect emerges because the combination of negative and positive correlations cancels out; in contrast to the observed pattern in the single-item format, where all correlations were positive.

any reading measure in either adults or children and irrespective of whether conditions were presented mixed or blocked[5]. Crucially, serial and discrete reading correlated with Stroop interference only in the multi-item format (especially in children, i.e., Experiments 2b and 6), indicating that the format of the Stroop task plays a crucial role in whether such an association will be observed.

Our results are in accordance with the empirical pattern reported in the study of Protopapas et al. (2007), where a negative relationship was observed in children between reading and Stroop interference in the multi-item format. In that study, the data were modeled using the speech production model WEAVER++ (Roelofs, 2003) to implement the speed of inhibition hypothesis by manipulating relevant model parameters to match the observed relationship. However, Protopapas et al. (2007) acknowledged that "the value of this particular optimization is limited, because the human tasks (taking several seconds to read lists of words and name lists of colored items) do not precisely match the modeled task (responding to individual stimuli within a few hundred milliseconds)" (pp. 274–275). Our findings, from multiple experiments, corroborate the importance of this limitation and emphasize that the format in which tasks are administered should be taken seriously into consideration when examining the relationship between Stroop interference and reading.

Processing complexity in single- vs. multi-item Stroop tasks

Based on findings related to RAN tasks and their known similarity to Stroop tasks, we postulated the hypothesis that the common requirement for temporally overlapped sequential

---

[5] Correlations between the incongruent condition in mixed presentation and serial reading were .22 and .39 for Experiments 2a and 2b, respectively, whereas for block presentation ranged between .11 and .47 (Figure 1, top right). For discrete reading, mixed presentation resulted in correlations between .50 and .60 and blocked between .39 and .63 (Figure 1, bottom right). Although correlations between Stroop interference—as a result of subtraction—and reading measures depended equally on the neutral condition (Figure 1, middle), mixed presentation did not seem to modulate their relationship (Figure 1, left), which was not significant in any single-item versions (Table 3).

processing (i.e., cascaded processing) in serial tasks might underlie the relationship between Stroop interference and reading in the serial format of the Stroop task. The cascaded processing hypothesis is based on the observation of the serial advantage, that is, that the time to name a set of items in a serial task such as RAN is less than the sum of individual times to name the same items in a discrete task. This serial advantage is attributed to temporally overlapping (i.e., parallel) processing of successive items, such that one item is processed while the previous one is uttered and the next one is viewed (and, possibly, an item further down is previewed). Because the serial advantage is observed in both naming and reading tasks, and because reading fluency (a serial task) is predicted much more strongly by serial than by discrete naming, the cascaded processing hypothesis has been put forward to account for both (i.e., the serial advantage and the serial superiority effect; Protopapas 2013, 2018). One might expect the same to hold for the relationship of serial word reading to all conditions of the multi-item version of the Stroop task.

Our findings are consistent with this hypothesis when the individual conditions are considered: As seen in Figure 1, the correlations between reading and color naming are greater for same-format (serial with multi-item and discrete with single-item) than for different-format comparisons, and this difference is much larger for the neutral condition (akin to a color-RAN task) than for the incongruent condition. However, when interference itself is considered, the situation is less clear-cut. A negative relationship between Stroop interference in the multi-item format and reading was indeed present, consistent with the prediction based on format-specific associations due to temporal overlap in processing of successive items in sequential tasks. Still, it was discrete reading—and not serial reading— that predicted the incongruent condition of the multi-item format to a greater extent, in both populations (i.e., children and adults), after controlling for the neutral condition. Discrete reading times are largely determined by the speed of single word identification, whereas serial

reading times (for nonbeginner readers) are increasingly dominated by cascaded processing efficiency as indexed by RAN (Altani et al., 2020; Protopapas, 2018). Therefore, our present findings reveal a relationship between single word recognition and the multi-item version of Stroop interference (but not the single-item version).

We propose that these findings can be understood if we consider the substantial differences between Stroop and RAN tasks beyond their superficial similarities as naming tasks. In particular, in comparison to standard RAN tasks (or the neutral condition), the incongruent condition of the multi-item Stroop task stands out due to its double nature: It includes both alphanumeric content (the words) and non-alphanumeric content (the colors), suggesting that successful performance requires both automatic and controlled processes, respectively (Roelofs, 2006). Moreover, the multi-item Stroop task is not a simple, serial single-task but a multi-task (Ziaka & Protopapas, under review). Successful performance in the incongruent condition of the multi-item task requires additional parallel processes, absent in serial RAN tasks, namely: (a) parafoveal identification of the automatic dimension (i.e., word) of the upcoming item to be flagged as irrelevant, and (b) processing of the two dimensions of the current item, namely the integrated, controlled, and slower color dimension and the flagged-as-irrelevant word dimension. These requirements lead to at least three simultaneous active responses, including two to be identified[6] and flagged as irrelevant, and one to be articulated. Pre-processing of the upcoming color may also be simultaneously underway, although the relatively slower nature of color naming may not consistently produce a corresponding response in time to compete with the others. (For simplification, preview of items further down is not taken into account, although it is plausible when considering the spatial proximity of items; Rayner, 1998; Rayner et al., 2005). And all this happens during execution of the

---

[6] Identification encompasses perception, conceptualization, lemma retrieval, and word-form encoding (Roelofs, 2003; 2007).

vocal response of the preceding item, creating an eye-voice span (Gordon & Hoedemaker, 2016; Huang, 2018; Pan et al., 2013). Consequently, successful performance in the incongruent condition of the multiple-item Stroop task requires that each correct color naming response is protected against interference not only from the integrated word of the current item but also from adjacent items. In other words, there is both within-item and between-item interference.

To protect against multiple sources of interference and secure successful performance, a serial strategy can be adopted in the incongruent condition of the multi-item Stroop task whereby perception of the next item is delayed until planning of the response to the current item has been completed (Roelofs, 2007). This strategy, termed "lockout scheduling" (Meyer & Kieras, 1997, p. 20), effectively blocks the high degree of parallelism that is proposed by the cascaded processing hypothesis for serial RAN and reading. As a result, the format-specific associations that are observed between RAN and reading (Altani et al. 2019) are greatly diminished when the incongruent condition of the multi-item Stroop task takes the place of RAN. Thus, although serial processing of each item through successive stages and serial processing of successive items are preserved, parallel processing of multiple items in the various stages is impaired in this condition, leading to stronger correlations with discrete reading than with serial reading. Indeed, participants' performance in the incongruent condition of the multi-item Stroop task resembles the performance of beginner readers, for whom serial RAN correlates with both serial and discrete reading (de Jong, 2011; Protopapas et al., 2013, 2018).

Support for this interpretation is provided by evidence related to picture-word Stroop tasks and gaze-to-speech lag under dual-task conditions. Specifically, Roelofs (2007) modified his earlier WEAVER++ production model (Roelofs, 2003) to account for dual-task performance and showed that in dual-tasks "participants strategically set and adjust a criterion

for shifting gaze to the secondary task, dependent on variables such as the primary task and the relationship between picture and word" (Roelofs, 2007, p. 248; henceforth we refer to this as "shift criterion" adjustment). He proposed that in dual-task situations gaze-shift behavior is modulated based on the demands of the task. If task demands are high, the shift criterion increases so that disengagement from the current task to the secondary task through gaze shifts is postponed. Interestingly, if both tasks require vocal responses, the level of attentional demands—implemented by the shift criterion in the model—is raised even further to protect the current task from between-items interference. "By adopting a serial strategy, the planning of the first vocal response is protected against interference from planning the second vocal response" (Roelofs, 2007, p. 234). Because the multi-item Stroop task requires vocal responses for all items, current and nearby, in the face of multiple sources of within-item and between-item interference, adoption of such a serial strategy seems well justified. In other words, the incongruent condition of the multi-item Stroop task is resistant to cascaded processing and therefore shares fewer cognitive processes with serial word reading than the neutral condition does.

To summarize, in the incongruent condition of the multi-item Stroop task, processing of the current color-word requires attentional engagement. The automatic (but task-irrelevant) word dimension of the current item activates its word form in parallel with the controlled (but task-relevant) planning of the color-word response. To protect the color-word response from interference by the printed word, attentional engagement is maintained until word-form encoding of the color-word has been completed. Once accomplished, gaze shifts may occur.

Reading skills alter the shift criterion. Because word-form encoding of the irrelevant word is more automatic for skilled readers, compared to poor readers, the level of complexity in processing the current item becomes lower. Consequently, the shift criterion for unlocking attention is positioned earlier in time, allowing skilled readers to move faster to the next item

and increase the degree of parallel processing. Efficient, automatic reading captures fewer attentional resources, thus reducing the demand for protection of the color response against interference, resulting in effective attentional decoupling. In contrast, word-form encoding is less automatic for poor readers, therefore it lies closer to color naming along the continuum of processing automaticity. This makes attentional decoupling more demanding (Moors & De Houwer, 2006) and increases the potential of interference, thereby causing the shift criterion to be placed at a later time point (Roelofs, 2007). Consequently the association between reading skill and interference is indirect, because it results from partially temporally overlapping processing of adjacent items, rather than from faster total processing time of individual items.

This interpretation is consistent with our findings for the neutral condition of the multi-item format, which correlated more strongly with serial reading than with discrete reading (Figure 1, top middle). We submit that this result emerged because of the lower attentional demands in the neutral condition, compared to the incongruent condition, due to the absence of the conflicting word dimension. Consequently, the shift criterion decreases and gaze shifts occur more rapidly, elevating the degree of parallelism and shifting processing of the neutral condition closer to the sequential processing of serial reading.

A remaining open question concerns the incongruent condition of the single-item format, where word-form encoding of the (within-item) word should be less demanding, at least for skilled readers, than in the multi-item format, due to the lack of between-item interference. Yet this does not seem to result in less interference compared to poor readers. To address this question we again turn to Roelofs (2007), who found that distractor effects (e.g., interference in naming pictures or reading words in picture-word stimuli) are evident only in gaze shifts but not in vocal responding. Roelofs noted that "the signal to move the eye is not simply the completion of a predetermined aspect of the programming of the vocal response" and

suggested that "the gaze shifts indexed shifts in the focus of attention in performing the task" (p. 244). The crucial distinction is between gaze shifts (modulated by attentional disengagement) and vocal responses (issued upon completion of processing). In the multi-item format, greater automatization of word-form encoding by skilled readers (compared to beginner or disabled readers) permits earlier attentional disengagement (and, presumably, increased gaze-to-speech lag), leading to improved performance not primarily due to faster individual item processing but to parallel processing of successive items. In contrast, in the single-item format there is no upcoming item to turn the attention to, therefore the aspect of processing that is associated with reading skill, namely word-form encoding and inhibition of the irrelevant response, does not have a chance to affect response latency; hence the diminished association with reading. This is entirely consistent with the idea that the relationship between interference and reading skill is indirect, attributable to shifting latency rather than to item processing duration, contradicting the proposal of a direct relationship put forth by Protopapas et al. (2007).

Our interpretation on the basis of between-items interference is in accordance with previous research examining negative priming in the Stroop task. The negative priming effect refers to the increase of response times in the incongruent condition when the incorrect word-response of the preceding item corresponds to the correct color-response of the currently named item (e.g., the word "green" printed in red followed by an item printed in green color). Negative priming is stronger in the multi-item Stroop task than in the single-item (Dalrymple-Alford & Budayr, 1966; Neill, 1977). To explain these findings, Neill (1977) proposed that "color naming may be slowed by a tendency to look ahead to the subsequent item while trying to respond to the current one ... the data may reflect backward rather than forward interference" (p. 445), similar to our between-items interference proposal.

"Negative-priming" items were present in our multi-items tasks, especially in Experiments 4–6 (i.e., 28 out of 60 items). If negative priming were the main reason for our finding of differential association of multi-item and single-item interference to reading, this difference should be most evident in these experiments. However, the relationship between Stroop interference and reading was evident even when negative priming was presumably minimal (Experiments 2a and 2b; 11 out of 60 items). This suggests that the relationship between interference and reading in the multi-item format is not critically dependent on the negative priming effect, in accordance with previous findings showing that negative priming did not differentiate between readers of different levels (Di Filippo & Zoccolotti, 2011). In addition, the relationship in question was also evident in the multi-item neutral condition, in which negative priming is absent. We acknowledge that these observations do not preclude a role of negative priming. This can be investigated in future research by directly manipulating the proportion of negative-priming items in the composition of multi-item lists.

We do not wish to suggest that word reading automaticity and, consequently, speed in word-form encoding and inhibition, play no role in the emergence of Stroop interference. Indeed, interference surfaces because of the obligatoriness of reading with attentional decoupling in serial processing of items depending on the level of automaticity. What we propose is that word reading automaticity is not sufficient on its own to distinguish between readers with different levels of reading skill and to explain why the relationship between Stroop interference and reading is only evident in the multi-item Stroop format.

Interference Scores: Is Subtraction Misleading?

The present study confirms that subtraction between conditions as an index of interference should be treated with caution, in agreement with previous admonitions (Hedge et al., 2018; Draheim et al., 2019; Parsons et al., 2019). Indeed, reliability for Stroop

interference was highly variable across experiments in our study, consistent with the two subtracted conditions being intercorrelated and far from perfectly reliable. In a similar vein, Di Filippo and Zoccolotti (2011) found that all naming tasks (including the neutral and incongruent Stroop condition) share much of their variance and attributed the relationship between reading and interference to an overadditivity effect. Overadditivity refers to "the general tendency of 'slower' individuals to respond more to experimental manipulations that affect the general difficulty of a given task over and above the specific influence of the manipulation itself" (Di Filippo & Zoccolotti, 2011, p. 357) and is attributed to factors that are not task-specific but more global (e.g., general slowness; Di Filippo & Zoccolotti, 2011; Faust et al., 1999). In the context of the Stroop task, this means that an increased difference in response times between the neutral and the incongruent condition cannot be safely ascribed to sensitivity to the incongruence per se (i.e., the task-specific factor constituting the experimental manipulation) but may in part be an exaggerated response to the difference in general difficulty between the two conditions. In other words, those individuals who are most challenged by the neutral condition will be disproportionally challenged by the incongruent condition because of its greater difficulty and not because of the qualitative difference in how the two conditions are processed (i.e., presumed effect of interference).

In short, Stroop interference calculated as a difference between two conditions (which are robustly correlated as they constitute minor variations of a naming task) may be unsuitable for individual differences research. Our alternative analysis approach, using regression on residualized scores, shows that the relationship between reading and interference survives in the case of the multi-item format of the Stroop task (but not the single-item format) even when common variance between different naming conditions (i.e., neutral and incongruent) is taken into account. At the same time, log transformation of the data addresses the overadditivity issue noted by Di Filippo and Zoccolotti, as it turns proportional differences in

response times (between groups or individuals) into additive differences (Faust et al., 1999; cf. Madden, 1990) that can be modeled within the general linear framework.

In addition, the "cow-canary paradox" (Capitani et al., 1999) was evident in our data. This term refers to that fact that the difference between two variables (here, incongruent and neutral condition) will always correlate much more strongly with the variable with the greatest variance, running counter to the intuition that both variables will contribute to the difference estimates to a similar degree. In our data, this pattern was especially prevalent in Experiments 2a, 4, and 5, where the correlation between serial reading and Stroop interference in the multi-item format was negative because the neutral condition had greater variance than the incongruent condition. The cow-canary paradox is a statistical observation concerning how the relative variance of two variables is related to their relative correlation with their difference. It is not an expression of cognitive processing, and it is unrelated (hence additional) to the aforementioned issues with low reliability and overadditivity. At any rate, it is also effectively addressed by expressing interference as residualized scores rather than differences.

Although the investigation of using difference scores (such as interference) in individual differences research has recently attracted attention, Lindsay and Jacoby (1994) are to be credited with pointing out the dangers of subtraction for experimental research decades ago. They ended up proposing that no conclusion can be drawn by only examining Stroop interference estimates without considering both conditions separately and that "doing otherwise runs the risk of creating confusing and conflicting results because of the inability to take differences in baseline into account" (p. 233). This was evident in our study as well (Figure 1, top left).

Limitations

Some limitations of the present study should be addressed. First, although results were obtained from six different experiments, the sample of five of them was rather small to draw any definite conclusions and could cause concern with respect to statistical power. However, the meta-analytic approach taken here greatly reduces the impact of sample size on the overall results. Second, we only tested participants from the general population without any screening for the presence of reading difficulties. In contrast, the majority of studies examining the relationship between reading and Stroop interference have included readers with difficulties in their samples. Future studies have to verify the task-format dependence of the interference-reading relationship across reader skill levels. Finally, it should be noted that in three of our six experiments material properties such as font size, equality of set size, and number of items across tasks and conditions, were not fully balanced. Fortunately, this did not seem to have an impact on the pattern of correlations, which was similar across experiments with comparable samples.

Conclusion

The present study showed that the previously established negative relationship between reading and Stroop interference is not universal but depends on the format of the tasks. This finding challenges the hypothesized direct link between reading skill and interference via the speed of inhibition. Instead, we propose that reading skill modulates the degree of cascaded processing based on an attentional disengagement strategy, which allows participants to shift their gaze between successive items earlier or later in time. Future studies can test the limits of this proposal across task variants, populations, and reading materials.

References

Altani, A., Protopapas, A., & Georgiou, G. K. (2017). The contribution of executive functions to naming digits, objects, and words. Reading and Writing, 30, 121–141. https://doi.org/10.1007/s11145-016-9666-4

Altani, A., Protopapas, A., Katopodi, K., & Georgiou, G. K. (2019). Tracking the serial advantage in the naming rate of multiple over isolated stimulus displays. *Reading and Writing*, *33*(2), 349–375. https://doi.org/10.1007/s11145-019-09962-7

Altani, A., Protopapas, A., Katopodi, K., & Georgiou, G. K. (2020). From individual word recognition to word list and text reading fluency. *Journal of Educational Psychology*, *112*(1), 22–39. https://doi.org/10.1037/edu0000359

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence Based Mental Health*, *22*(4), 153–160. https://doi.org/10.1136/ebmental-2019-300117

Bowers, P. G., & Swanson, L. B. (1991). Naming speed deficits in reading disability: Multiple measures of a singular process. *Journal of Experimental Child Psychology*, *51*(2), 195–219. https://doi.org/10.1016/0022-0965(91)90032-n

Brown, W. (1910). Some experimental results in the correlation of mental abilities1. *British Journal of Psychology, 1904-1920*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Capitani, E., Laiacona, M., Barbarotto, R., & Cossa, F. M. (1999). How can we evaluate interference in attentional tests? A study based on Bi-variate non-parametric tolerance limits. *Journal of Clinical and Experimental Neuropsychology*, *21*(2), 216–228. https://doi.org/10.1076/jcen.21.2.216.934

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, *11*(41), 63–65. https://doi.org/10.1093/mind/os-XI.41.63

Chiappe, P., Stringer, R., Siegel, L. S., & Stanovich, K. E. (2002). Why the timing deficit hypothesis does not explain reading disability in adults. *Reading and Writing: An Interdisciplinary Journal, 15*(1-2), 73–107. https://doi.org/10.1023/A:1013868304361

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*(3), 332–361. https://doi.org/10.1037/0033-295x.97.3.332

Dalrymple-Alford, E. C., & Budayr, B. (1966). Examination of some aspects of the Stroop Color-Word Test. *Perceptual and Motor Skills, 23*(3, PT. 2), 1211–1214. https://doi.org/10.2466/pms.1966.23.3f.1211

De Jong, P. F. (2011). What discrete and serial rapid automatized naming can reveal about reading. *Scientific Studies of Reading*, *15*(4), 314–337. https://doi.org/10.1080/10888438.2010.485624

Di Filippo, G., & Zoccolotti, P. (2011). Separating global and specific factors in developmental dyslexia. *Child Neuropsychology*, 396–391. https://doi.org/10.1080/09297049.2011.613809

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508–535. https://doi.org/10.1037/bul0000192

Everatt, J., Warner, J., Miles, T. R., & Thomson, M. E. (1997). The incidence of Stroop interference in dyslexia. *Dyslexia*, *3*(4), 222–228. https://doi.org/10.1002/(sici)1099-0909(199712)3:4<222::aid-dys12>3.0.co;2-p

Faccioli, C., Peru, A., Rubini, E., & Tassinari, G. (2008). Poor readers but compelled to read: Stroop effects in developmental dyslexia. *Child Neuropsychology*, *14*(3), 277–283. https://doi.org/10.1080/09297040701290040

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125,* 777–799. doi:10.1037/0033-2909.125.6.777

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124. https://doi.org/10.3758/BF03195503

Gordon, P. C., & Hoedemaker, R. S. (2016). Effective scheduling of looking and talking during rapid automatized naming. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 742–760. https://doi.org/10.1037/xhp0000171

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Helland, T., & Asbjørnsen, A. (2000). Executive functions in dyslexia. *Child Neuropsychology*, *6*(1), 37–48. https://doi.org/10.1076/0929-7049(200003)6:1;1-b;ft037

Helland, T., & Morken, F. (2016). Neurocognitive development and predictors of L1 and L2 literacy skills in dyslexia: A longitudinal study of children 5–11 Years old. *Dyslexia*, *22*(1), 3–26. https://doi.org/10.1002/dys.1515

Huang, Y. T. (2018). Real-time coordination of visual and linguistic processes in novice readers. *Journal of Experimental Child Psychology*, *173*, 388–396. https://doi.org/10.1016/j.jecp.2018.02.010

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention,

memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General*, *145*(8), 1017–1048. https://doi.org/10.1037/xge0000184

Kapoula, Z., Lê, T., Bonnet, A., Bourtoire, P., Demule, E., Fauvel, C., Quilicci, C., & Yang, Q. (2010). Poor Stroop performances in 15-year-old dyslexic teenagers. *Experimental Brain Research*, *203*(2), 419–425. https://doi.org/10.1007/s00221-010-2247-x

Kelly, M. S., Best, C. T., & Kirk, U. (1989). Cognitive processing deficits in reading disabilities: A prefrontal cortical hypothesis. *Brain and Cognition*, *11*(2), 275–293. https://doi.org/10.1016/0278-2626(89)90022-5

Laliberté, E. (2019). metacor: Meta-Analysis of Correlation Coefficients. R package version 1.0-2.1. https://CRAN.R-project.org/package=metacor

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 219–234. https://doi.org/10.1037/0096-1523.20.2.219

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

MacLeod, C.M. (2005). The Stroop task in cognitive research. In A. Wenzel & D. S. Rubin (Eds.) *Cognitive methods and their application to clinical research* (pp. 17–40). Washington, DC: American Psychological Association.

Madden D. J. (1990). Adult age differences in the time course of visual attention. *Journal of Gerontology*, *45*(1), P9–P16. https://doi.org/10.1093/geronj/45.1.p9

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. *Psychological Review*, *104*(1), 3–65. https://doi.org/10.1037/0033-295x.104.1.3

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Neill, W. T. (1977). Inhibitory and facilitatory processes in selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 3*(3), 444–450. https://doi.org/10.1037/0096-1523.3.3.444

Norton, E. S., & Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *Annual Review of Psychology*, *63*(1), 427–452. https://doi.org/10.1146/annurev-psych-120710-100431

Pan, J., Yan, M., Laubrock, J., Shu, H., & Kliegl, R. (2013). Eye-voice span during rapid automatized naming of digits and dice in Chinese normal and dyslexic children. *Developmental Science*, n/a-n/a. https://doi.org/10.1111/desc.12075

Parsons, S., Kruijt, A.-W, & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive behavioural measurements. *Advances in Methods and Practices in Psychological Science*, 378–395. https://doi.org/10.1177/2515245919879695

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, *39*, 859–862. https://doi.org/10.3758/BF03192979

Protopapas, A., Altani, A., & Georgiou, G. K. (2013). undefined. *Journal of Experimental Child Psychology*, *116*(4), 914–929. https://doi.org/10.1016/j.jecp.2013.08.004

Protopapas, A., Archonti, A., & Skaloumbakas, C. (2007). Reading ability is negatively related to Stroop interference. *Cognitive Psychology*, *54*(3), 251–282. https://doi.org/10.1016/j.cogpsych.2006.07.003

Protopapas, A., Katopodi, K., Altani, A., & Georgiou, G. K. (2018). Word reading fluency as a serial naming task. *Scientific Studies of Reading*, *22*(3), 248–263. https://doi.org/10.1080/10888438.2018.1430804

Protopapas, A., Sideridis, G. D., Mouzaki, A., & Simos, P. G. (2007). Development of lexical mediation in the relation between reading comprehension and word reading skills in Greek. *Scientific Studies of Reading*, *11*(3), 165–197. https://doi.org/10.1080/10888430701344322

Protopapas, A., & Skaloumbakas, C. (2007). Traditional and computer-based screening and diagnosis of reading disabilities in Greek. *Journal of Learning Disabilities*, *40*(1), 15–36. https://doi.org/10.1177/00222194070400010201

Protopapas, A., Vlahou, E. L., Moirou, D., & Ziaka, L. (2014). Word reading practice reduces Stroop interference in children. *Acta Psychologica*, *148*, 204–208. https://doi.org/10.1016/j.actpsy.2014.02.006

Skoteinou, D. (2008). *Individual differences in Stroop interference related to reading ability*. [Master's Thesis, Athens University]. Greece. http://cogsci.phs.uoa.gr/fileadmin/cogsci.phs.uoa.gr/uploads/files/diplomatikes/Skoteinou_Dimitra_2008pdf.pdf

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K., Juhasz, B. J., & Pollatsek, A. (2005). Eye movements during reading. In M. J. Snowling and C. Hulme (Eds.) The Science of Reading: A Handbook (pp. 79–97). Blackwell Publishing Ltd.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reiter, A., Tucha, O., & Lange, K. W. (2005). Executive functions in children with dyslexia. *Dyslexia*, *11*(2), 116–131. https://doi.org/10.1002/dys.289

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review*, *110*(1), 88–125. https://doi.org/10.1037/0033-295x.110.1.88

Roelofs, A. (2006). Functional architecture of naming dice, digits, and number words. *Language and Cognitive Processes*, *21*(1-3), 78–111. https://doi.org/10.1080/01690960400001846

Roelofs, A. (2007). Attention and gaze control in picture naming, word reading, and word categorizing. *Journal of Memory and Language*, *57*(2), 232–251. https://doi.org/10.1016/j.jml.2006.10.001

Salo, R., Henik, A., & Robertson, L. C. (2001). Interpreting Stroop interference: An analysis of differences between task versions. *Neuropsychology*, *15*(4), 462–471. https://doi.org/10.1037/0894-4105.15.4.462

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Stanovich, K. E., Feeman, D. J., & Cunningham, A. E. (1983). The development of the relation between letter-naming speed and reading ability. *Bulletin of the Psychonomic Society*, *21*(3), 199–202. https://doi.org/10.3758/bf03334686

Stringer, R. W., Toplak, M. E., & Stanovich, K. E. (2004). Differential relationships between

    RAN performance, behaviour ratings, and executive function measures: Searching for

    a double dissociation. *Reading and Writing*, *17*(9), 891–914.

    https://doi.org/10.1007/s11145-004-2770-x

van der Schoot, M., Licht, R., Horsley, T. M., & de Sergeant, V. R. (2000). Inhibitory deficits

    in reading disability depend on subtype: Guessers but not spellers. *Child*

    Neuropsychology, *6*(4), 297–312. https://doi.org/10.1076/chin.6.4.297.3139

Wang, S., & Gathercole, S. E. (2015). Interference control in children with reading

    difficulties. *Child Neuropsychology*, *21*(4), 418–431.

    https://doi.org/10.1080/09297049.2014.918594

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental

    dyslexias. *Journal of Educational Psychology*, *91*(3), 415–438.

    https://doi.org/10.1037/0022-0663.91.3.415

Ziaka, L. (2014). *Study of intratask components of the Stroop task: Column Time and Pause*

    Time. [Master's Thesis, Athens University]. Pergamos.

    https://pergamos.lib.uoa.gr/uoa/dl/frontend/file/lib/default/data/1318876/theFile

Ziaka, L. & Protopapas, A. (under review). Conflict monitoring or multi-tasking? Tracking

    within-task performance in single-items and multi-items Stroop tasks. Preprint

    available from https://psyarxiv.com/e8bjg/

Zoccolotti, P., De Luca, M., Lami, L., Pizzoli, C., Pontillo, M., & Spinelli, D. (2013).

    Multiple stimulus presentation yields larger deficits in children with developmental

    dyslexia: A study with reading and RAN-type tasks. *Child Neuropsychology*, *19*(6),

    639–647. https://doi.org/10.1080/09297049.2012.718325

Zoccolotti, P., De Luca, M., Marinelli, C. V., & Spinelli, D. (2014). Modeling individual

    differences in text reading fluency: A different pattern of predictors for typically

developing and dyslexic readers. *Frontiers in Psychology*, *5*.

https://doi.org/10.3389/fpsyg.2014.01374