Two distinct patterns of interference in between-attribute Stroop matching tasks

Kerstin Dittrich¹ · Christoph Stahl²

Published online: 15 December 2016
© The Psychonomic Society, Inc. 2016

Abstract In between-attribute Stroop matching tasks, participants compare the meaning (or the color) of a Stroop stimulus with a probe color (or meaning) while attempting to ignore the Stroop stimulus's task-irrelevant attribute. Interference in this task has been explained by two competing theories: A semantic competition account and a response competition account. Recent results favor the response competition account, which assumes that interference is caused by a task-irrelevant comparison. However, the comparison of studies is complicated by the lack of a consensus on how trial types should be classified and analyzed. In this work, we review existing findings and theories and provide a new classification of trial types. We report two experiments that demonstrate the superiority of the response competition account in explaining the basic pattern of performance while also revealing its limitations. Two qualitatively distinct interference patterns are identified, resulting from different types of task-irrelevant comparisons. By finding the same interference pattern across task versions, we were additionally able to demonstrate the comparability of processes across two task versions frequently used in neurophysiological and cognitive studies. An integrated account of both types of interference is presented and discussed.

Keywords Stroop matching task · Response interference · Semantic competition · Selective attention

✉ Kerstin Dittrich
dittrich@psychologie.uni-freiburg.de

¹ Institut für Psychologie, University of Freiburg,
D-79085 Freiburg, Germany

² Department Psychologie, University of Cologne, Köln, Germany

Perhaps the most prominent interference paradigm in psychology is the Stroop color-word task (Stroop, 1935), in which participants name the ink color of a color word while attempting to ignore its meaning. Typically, participants fail to fully ignore the word's meaning, as demonstrated by prolonged RT and higher error rates when word meaning and ink color mismatch (incongruent trial), relative to cases in which they match (congruent trial). To date, numerous Stroop variants have been developed (see MacLeod, 1991, for an overview), one being the Stroop matching task (Dyer, 1973; Treisman & Fearnley, 1969). In this task, a stimulus display consists of two stimuli: a Stroop stimulus and a probe. The Stroop stimulus, a colored color word, has two attributes: word meaning and ink color. The probe is either a color word printed in a neutral ink (e.g., black or white), or a colored nonword stimulus (e.g., a bar, a series of Xs). Thus, the probe has only one attribute: either a word meaning or an ink color. Participants are instructed to compare one of the Stroop stimulus' attributes (the target attribute) with the probe while ignoring the other (the distracter attribute), and to respond “same” if the former two are equivalent, and “different” if they are not.

Treisman and Fearnley (1969) realized four different variants of the Stroop matching task that result from different choices of the target and probe attributes (color or word meaning; examples of all four variants are given in Table 1). They distinguished two types of matching: *within-attribute* matching versus *between-attribute* matching. In within-attribute matching tasks (i.e., the word-word and color-color matching tasks), two attributes of the same type (i.e., two words, two colors) had to be compared. To illustrate, in the word-word matching task, the probe consisted of a color word printed in black; its meaning had to be compared with the Stroop word's meaning. In the color-color matching task, the probe was a series of colored Xs that had to be compared with

Table 1 Overview of trial types of all four Stroop matching task (SMT) versions, the proposed notation, and relevant SMT studies

SMT version	Matching type (Task-relevant attributes)	Correct Response (i.e., P – TA Relation)				
		Same		Different		
Word-Word (W-W)	Within-attributes (P: WM; TA: WM)	RED RED	RED RED	RED BLUE	RED BLUE	RED BLUE
Word-Color (W-C)	Between-attributes (P: WM; TA: WC)	RED RED	RED BLUE	RED BLUE	RED BLUE	RED RED
Color-Color (C-C)	Within-attributes (P: BC; TA: WC)	RED RED	RED BLUE	RED BLUE	RED BLUE	RED RED
Color-Word (C-W)	Between-attributes (P: BC; TA: WM)	RED RED	RED RED	RED BLUE	RED BLUE	RED BLUE
Proposed trial type notation		SSS S: P = TA S: TA = DA S: P = DA	SDD S: P = TA D: TA ≠ DA D: P ≠ DA	DSD D: P ≠ TA S: TA = DA D: P ≠ DA	DDD D: P ≠ TA D: TA ≠ DA D: P ≠ DA	DDS D: P ≠ TA D: TA ≠ DA S: P = DA
Study		Notation				
Treisman & Fearnley (1969): all four SMT versions; first study on SMTs		No differentiation between trial types				
Dyer (1973): C-W and W-C versions First systematic examination of trial types		Congruent Same TA = DA	Incongruent Same TA ≠ DA	Congruent Different TA = DA	Incongruent Different (A) TA ≠ DA	Incongruent Different (B) TA ≠ DA
Simon & Baker (1995); Simon & Berbaum (1988): all four SMT versions		Stroop congruent TA = DA	Stroop incongruent TA ≠ DA	Stroop congruent TA = DA	Stroop Incongruent TA ≠ DA	
Neurophysiological studies using the W-C version ^a , the C-W version ^b , or both versions ^c		Stroop congruent TA = DA	Stroop incongruent/ TA ≠ DA	Stroop congruent TA = DA	Stroop Incongruent/ TA ≠ DA	
Luo (1999): C-W and C-C versions Semantic competition account		Match P = DA	Mismatch P ≠ DA	Mismatch P ≠ DA		Match P = DA
Goldfarb and Henik (2006): C-W SMT Evidence against a semantic competition account; response competition account		WC-BC-WM congruent	Incongruent	WC-WM congruent/ WC-BC incongruent	All incongruent	WC-BC congruent/ WC-WM incongruent
Caldas et al. (2012): C-W SMT Evidence for both a semantic competition account and a response competition account		Congruent/ related TA = DA P = DA	Incongruent/ unrelated TA ≠ DA P ≠ DA	Congruent/ unrelated TA = DA P ≠ DA	Incongruent unrelated TA ≠ DA P ≠ DA	Incongruent related TA ≠ DA P = DA

Note. For each SMT version, examples of the five trial types are displayed. For each study, SMT versions and classification of trial types are listed. SMT = Stroop matching task; W = word; C = color; P = probe; TA = target attribute; DA = distracter attribute; WM = word meaning; WC = word's ink color; BC = bar color. The displayed classification is based on how trial types were aggregated in a study's main analysis (some studies descriptively displayed results in more detail, e.g., Simon & Baker, 1995; Simon & Berbaum, 1988). Only studies using SMTs as described by Treisman and Fearnley (1969) are shown (for other versions, see, e.g., Durgin, 2003; Flowers, 1975; Mascolo & Hirtle, 1990). ^a David, Volchan, Alfradique et al. (2011); David, Volchan, Vila et al. (2011); Menz et al. (2006); Mitchell (2006); Norris et al. (2002); Schroeter et al. (2002); Schroeter et al. (2003); Schroeter et al. (2004); Yanagisawa et al. (2010); Zysset et al. (2001); Zysset et al. (2007). ^b Kim et al. (2005). ^c Machado-Pinheiro et al. (2010)

the Stroop word's ink color. In between-attribute matching tasks (i.e., the word-color and color-word matching tasks), two attributes of different types (i.e., ink color and word meaning) had to be compared. In the word-color matching task, the

probe consisted of a color word printed in black; its meaning had to be compared with the Stroop word's ink color. In the color-word matching task, the probe was a series of colored Xs that had to be compared with the Stroop word's meaning.

In their initial study, Treisman and Fearnley (1969) found negligible interference in both within-attribute matching tasks, but considerable (and comparable) interference in both versions of the between-attribute matching task. The absence of interference in within-attribute matching tasks was explained in terms of different cognitive analyzing modules for ink color and word meaning: In the color-color matching task, when comparing ink colors, the word analyzer can be switched off, such that the distracter attribute of the Stroop stimulus (i.e., word meaning) is not processed and therefore does not cause interference. Similarly, when comparing word meanings in the word-word matching task, the color analyzer can be switched off, implying that the distracter attribute of the Stroop stimulus (i.e., ink color) is not processed and therefore does not cause interference. In contrast, interference arises in between-attribute matching tasks because both types of attributes (i.e., words and colors) need to be processed in order to perform the task-relevant comparison; therefore, both analyzers have to be “switched on,” and, as a consequence, the distracter attribute of the Stroop stimulus (i.e., ink color in the color-word matching task; word meaning in the word-color matching task) can cause interference.

This study focuses on the latter between-attribute Stroop matching tasks and proposes an account of the basic pattern of performance in the different trial types of these tasks. Before introducing the study, we review previous conflicting findings and introduce a notation of trial types that will help in interpreting and integrating these findings.

Trial types in the Stroop matching task

Because Treisman and Fearnley (1969) measured performance for entire blocks, but not for single stimulus displays, they were not able to differentiate between trial types. In fact, target, distracter, and probe attributes can be combined in five different ways, resulting in five different trial types that are best described by the pattern of pairwise relations between attributes. Three relations between pairs of attributes can be distinguished: (1) the task-relevant relation between probe and target attribute, (2) the task-irrelevant relation between target and distracter attribute, and (3) the task-irrelevant relation between probe and distracter attribute. Each of these pairs of attributes can be either same or different. Table 1 illustrates how these pairwise relations result in five different trial types; examples for each trial type are given for each version of the Stroop matching task.

Perhaps in contrast to the classical Stroop task it is not immediately obvious which of the types of Stroop matching task trials would be associated with interference. This is reflected in the literature as an ongoing debate about how trial types should be classified and analyzed (e.g., Dyer, 1973; Goldfarb & Henik, 2006; Luo, 1999). Table 1 lists previous

Stroop matching task studies and gives an overview of the different approaches.

Dyer (1973) first recorded RTs for the five trial types (plus two neutral control trial types) of both between-attribute matching tasks. To differentiate between the trial types, he proposed a classification primarily based on the relation between target and distracter, albeit with inconclusive results. To illustrate, a trial was termed congruent-different when target attribute and probe did not correspond (i.e., called for a different response), but the target attribute corresponded to the distracter attribute (i.e., they were congruent). Because the relation between distracter and probe is the same as between target and probe, performance in trials of this type was predicted to be superior to trials in which this relation is different: The redundancy was expected to speed up the decision process. However, contrary to this prediction, the congruent-different type turned out to be associated with interference (this was already noted by Dyer, at least for the word-color matching task, and confirmed later by Goldfarb & Henik, 2006, for the color-word matching task). Nevertheless, several subsequent studies have used this empirically inadequate classification; results obtained in these studies should be interpreted with caution (see Table 1).

Luo (1999) subsequently proposed a classification of trial types that focuses instead on the relation between probe and distracter. Specifically, he distinguished between match and mismatch trials (separately for “same” trial types, i.e., those requiring a “same” response, and “different” trial types, i.e., those requiring a “different” response), depending on whether the distracter attribute matched or mismatched the probe (see Table 1). This classification, however, distinguished only four categories of trial types; as a result, one of the categories (i.e., the mismatch-different category; see Table 1) conflated two trial types that differ with regard to the target-distracter relation. Goldfarb and Henik (2006; see also Caldas, Machado-Pinheiro, Souza, Motta-Ribeiro, & David, 2012) showed empirically that performance in these two conflated cases differs substantially. As we will discuss, this finding is difficult to reconcile with Luo’s semantic competition account (but see the [General Discussion](#) for an explanation of this finding in terms of semantic competition).

Finally, Goldfarb and Henik’s (2006) response competition account of the Stroop matching task focused on the relation between the target and distracter attribute as critical in explaining performance across trial types. Importantly, in contrast to Dyer’s (1973) approach, it does so in an empirically adequate manner; for instance, it predicts interference for Dyer’s congruent-different trials. The response competition account is described in more detail below.

It is clear from this brief review that there exists no consensus on the classification of trial types, and equivalently,

about the basic pattern of performance in Stroop matching tasks. This fact complicates or even precludes the comparisons of findings obtained in different studies, and it results in failures to fully and adequately characterize basic patterns of performance across trial types. Still, two basic findings can be secured: First, performance in “same” trial types is typically superior to performance in “different” trial types (e.g., Dyer, 1973; Goldfarb & Henik, 2006; Luo, 1999). As a consequence, these two groups of trial types are typically analyzed separately (e.g., Goldfarb & Henik, 2006). Second, with regard to “same” trials, the findings are clear: Previous studies have consistently shown better performance for the trial type in which all three attributes match, as compared to the trial type in which the distracter attribute mismatches target and probe (e.g., Dyer, 1973; Goldfarb & Henik, 2006; Luo, 1999). As can be seen by the variety of classifications in Table 1, the picture is much less clear for the “different” trial types.

To clarify this picture, and to help establish the basic pattern of performance in Stroop matching tasks, we suggest a notation of trial types that (1) considers all relevant stimulus features, (2) is descriptive in the sense that it avoids a priori assumptions about interference associated with a trial type, and (3) is independent of the specific choice of relevant and irrelevant attributes and thus applicable to all four Stroop matching task variants. It is illustrated in Table 1. Representing the three possible pairwise relations between attributes, a trial type is characterized by three binary features, each denoted by one of two symbols, S for “same” and D for “different.” The first S or D denotes the relation between the task-relevant attributes, and thus, the required response. Based on their first feature, trial types can be subsumed into “same” trial types and “different” trial types, depending on whether they require a “same” or a “different” response. The second symbol (S or D) denotes the task-irrelevant relation between target attribute and distracter attribute of the Stroop stimulus; this is always a relation between different attributes of the same object (i.e., ink color and meaning of the Stroop word). The third symbol (again, S or D) denotes the task-irrelevant relation between the distracter attribute of the Stroop stimulus and the probe; in the between-attribute matching tasks investigated here, this is always a relation within the same attribute dimension across two different objects. In within-attribute matching tasks, it is a relation between different attributes across objects. Thus, the nature of this third relation (within-attributes or between-attributes) always differs from the nature of the task-relevant relation.

With the help of this descriptive classification of trial types, we attempt to fully characterize the interference pattern of both versions of the between-matching task and to identify the processes underlying these patterns. As starting point, we focus on two studies that have investigated the color-word matching task (i.e., a between-attribute matching task using

a colored bar as probe), and have accounted for their findings in terms of competing theoretical accounts (Goldfarb & Henik, 2006; Luo, 1999). We will then extend our research to the word-color matching task. Thereby, we will provide empirical evidence concerning the comparability of these two between-attribute matching tasks that have been interpreted as equivalent in the literature (e.g., David, Volchan, Vila, et al., 2011; Goldfarb & Henik, 2006; Luo, 1999; Norris, Zysset, Mildner, & Wiggins, 2002; Zysset, Müller, Lohmann, & von Cramon, 2001).

Current accounts of the Stroop matching task

The semantic competition account (Luo, 1999) assumes that for the color-word matching task, both ink colors—word color and bar color—activate their respective semantic unit within a semantic system. Further, it is assumed that information from the semantic unit is mapped or translated into its verbal representation in a verbal-lexical unit, in which it is then compared with word meaning. Impaired identification of the relevant ink color (i.e., bar color) is predicted when the two colors differ; for instance, a word-color/bar-color mismatch should impair color identification because two different semantic units are activated and might interfere with each other. Importantly, it is assumed that this interference occurs at a translational stage located before the response selection stage. Based on this assumption, Luo (1999) divided trial types into match stimuli (i.e., trials in which word color and bar color match; e.g., a red bar presented along a word in red ink) and mismatch stimuli (i.e., trials in which word color and bar color mismatch; e.g., a red bar presented along a word in blue ink). As is evident from Table 1, this considers only four different types of trials; Luo’s “different/mismatch” type was in fact a mixture of two different types (i.e., DDD and DSD). Luo (1999) reported two experiments to test this account. In a first study, as predicted, robust interference was obtained for “same” trial types (i.e., increased RTs for the SDD trial over the SSS trial), and smaller but significant interference for “different” trial types (i.e., increased RTs for the mixture of DDD and DSD trials over DDS trials). Furthermore, when the colored bar preceded the colored word by one second or more (Luo, 1999, Experiment 2), interference was completely eliminated and even reversed for “different” trial types: Responses for the DDS trial were now significantly slower than those for the mixture of DDD and DSD trials. Luo’s interpretation was that, by delaying the presentation of the Stroop word—and thereby, the comparison process assumed to occur in the postulated verbal-lexical unit—the relevant color of the bar could be encoded semantically before the task-irrelevant color of the Stroop word could interfere.

However, this interpretation must be treated with caution, as the above conflation of trial types yielded misleading

results. Goldfarb and Henik (2006) presented evidence showing that Luo's (1999) classification of stimuli is not adequate. They compared RTs for all five trial types described above (see also Table 1). Focusing here on the "different" trial types, results showed that RTs were significantly larger for the DSD type than for the DDD type, suggesting that pooling both trial types, as in Luo's studies, is not warranted. Also, contrary to Luo's prediction, RTs for DDS and DDD trials did not differ significantly. Both of these findings cannot easily be reconciled with Luo's semantic competition account (but see the [General Discussion](#) for an explanation of these findings in terms of semantic competition). Goldfarb and Henik's (2006) response competition account argues that interference in the Stroop matching task arises at the response selection stage and is a consequence of task conflict (MacLeod & MacDonald, 2000), resulting in irrelevant comparisons being made. As pointed out previously, in addition to the task-relevant comparison between probe and target, two possible task-irrelevant comparisons can be made: a comparison between target and distracter (i.e., word meaning and word color) and a comparison between distracter and probe (i.e., word color and bar color in the color-word matching task). The authors argued that the former comparison is made, but not the latter, based on the assumption that attention is primarily object based, not feature based (Kahneman & Henik, 1981; Wühr & Waszak, 2003; for a review, see Chen, 2012). In other words, Goldfarb and Henik (2006) assumed that when focusing on the meaning of the Stroop word, its ink color cannot be ignored. As a result, when target and distracter attribute of the Stroop stimulus correspond (which is the case for the DSD type), a tendency toward a "same" response is elicited, competing with the required "different" response. This accounts well for the increased RTs for DSD as compared to DDD trials. In contrast, the second possible irrelevant comparison between the two ink colors is assumed not to be made; this assumption was supported by similar RTs for the DDD and DDS trials. From the perspective of Goldfarb and Henik's response competition account, therefore, the target–distracter relation is critical for interference, implying that the SDD and DSD types would be associated with interference. In contrast, the probe–distracter relation is assumed to be irrelevant, implying that the DDS type would not be associated with interference.

The present study

We propose that the target–distracter relation may not be sufficient to fully account for interference in the Stroop matching task. Treisman and Fearnley's (1969) finding of fast and efficient within-attribute comparisons suggests that the lack of an RT effect does not rule out that a task-irrelevant probe–distracter comparison (e.g., between

two ink colors in the color-word matching task) is made. Such an irrelevant comparison could, if it is indeed made, result in accuracy costs in trials in which the result of this comparison is incongruent with that of the relevant comparison (i.e., as in the DDS type). Presumably, a physical match (of either two ink colors or two words) would be detected quickly and efficiently, and it would elicit a "same" response. A respective tendency would lead to incorrect responses based on the wrong within-attribute matching in the absence of RT costs. Initial support for this hypothesis can be found in the work of Goldfarb and Henik (2006), who observed descriptively increased error rates for DDS than for DDD trials (see also Caldas et al., 2012; Dyer, 1973). On the basis of these considerations, we expect that two types of task-irrelevant comparisons are inadvertently made when performing the Stroop matching task that differ in their relative speed of processing: (1) the task-irrelevant between-attribute comparison is assumed to be a slow comparison process (e.g., a process in which two different modules are involved) and should therefore result in both RT and accuracy costs, whereas (2) the task-irrelevant within-attribute comparison is assumed to be fast and should therefore produce accuracy costs in the absence of RT costs.

This study investigates the possibility of two distinct sources of interference in the Stroop matching task, not only at the level of mean RT and accuracy but also by analyzing the RT and error distributions. The RT distribution will be analyzed by means of delta plots (De Jong, Liang, & Lauber, 1994; Schwarz & Miller, 2012; Speckman, Rouder, Morey, & Pratte, 2008), which depict the magnitude of the interference effect at different RT percentiles. Specifically, the delta plot maps (on the *y*-axis) the difference between two RT distributions (e.g., the difference of the RT distributions of incongruent and congruent trials) across percentiles, against (on the *x*-axis) the average of the two RT distributions across percentiles. Thus, the slope of the delta plot captures the RT differences between incongruent and congruent trials across response speed, with a positive slope indicating that the RT difference (i.e., the interference effect) is larger for slower responses, and a negative slope indicating that the RT difference is smaller for slower responses. Delta plots of classical Stroop data typically reveal positive slopes (Dittrich, Kellen, & Stahl, 2014; Pratte, Rouder, Morey, & Feng, 2010), reflecting the finding that Stroop interference is larger for slow responses than for fast responses. Our assumption concerning the relative speed of processing of the two irrelevant comparisons results in the prediction of a steeper delta plot slope for trials involving the slower task-irrelevant between-attribute comparison (i.e., DSD trials) in comparison to trials involving the faster task-irrelevant within-attribute comparison (i.e., DDS trials). Assuming that the within-attribute comparison is fast and

efficient (Treisman & Fearnley, 1969), and that it would result in the direct activation of an incorrect response (Gratton, Coles, & Donchin, 1992), we expect to see a greater proportion of fast errors on DDS trials when compared to trials involving task-irrelevant between-attribute comparisons (i.e., DSD trials). This assumption will be tested with the help of conditional accuracy functions (CAFs; Ridderinkhof, 2002), in which accuracy is displayed as a function of RT. We tested these predictions in both between-attribute matching tasks in Experiment 1a and 2a. Experiment 1b and 2b replicate these results in a procedure with slightly modified trial type proportions.

Experiment 1a

In Experiment 1a, the color-word Stroop matching task was applied. We expected to replicate Goldfarb and Henik's (2006) results at the level of mean RT: RTs of the "different" trial types were expected to be faster when word color and word meaning mismatch (DDD = DDS < DSD), indicating that a slow task-irrelevant between-attribute matching is made. For RT, it should make no difference whether bar color matches or mismatches word color (DDD = DDS), whereas we expected to find accuracy costs for DDS trials relative to DDD trials. The latter result would be in line with our assumption that a task-irrelevant within-attribute comparison is also made. Delta plot slopes and CAFs will help us to characterize both task-irrelevant comparisons: We expected to find a more negative delta plot slope for trials that involve the task-irrelevant within-attribute comparison (i.e., the DDS trial type) in comparison to the delta plot slope for trials involving the task-irrelevant between-attribute comparison (i.e., the DSD trial type). Moreover, we expected to find a greater proportion of fast errors for DDS than DSD trials. Both assumptions are derived from the hypothesis that the task-irrelevant within-attribute comparison occurs fast and relatively effortlessly, resulting in specifically fast errors along with negligible RT interference.

Method

Participants Participants were mostly University of Freiburg students with different majors. They were recruited in lectures of different subjects and via flyers distributed in the city of Freiburg. Thirty-two persons participated (18 women, 14 men; mean age was 22 years, ranging from 18 to 29 years). All participants were native speakers of German with normal or corrected-to-normal vision and participated for course credit or as paid volunteers. Two participants were extreme outliers (with mean error rates of 0.26 and 0.35, respectively, in the total sample's distribution of error rates, $M = 0.05$, $SD = 0.22$). These participants were excluded from analyses.

Stimuli and apparatus The German words *rot* [red], *grün* [green], *gelb* [yellow], and *blau* [blue] were presented in capital letters, in one of four colors (red, green, yellow, and blue). A colored bar was presented above the colored word in either one of these four colors. The length and height of the colored bar matched those of the longest colored word. Stimuli subtended a visual angle of approximately $1.5^\circ \times 1.0^\circ$.

Trial types were generated as described by Goldfarb and Henik (2006): Word color, bar color, and word meaning were factorial combined, resulting in 64 possible combinations; 16 were "same" stimuli and 48 were "different" stimuli. To elicit the same number of "same" and "different" responses, the "same" stimuli were presented three times. One block consisted of 96 trials (i.e., $3 \times 16 = 48$ "same" and 48 "different" trials). Participants responded by pressing the keys *j* for "same" and *k* for "different" on a standard computer keyboard. All instructions and colored words were presented in 20-point Lucida Sans Regular font. Stimuli were presented on a light gray background on a 48.3 cm CRT screen at an approximate viewing distance of 60 cm.

Procedure and design Each trial started with a fixation cross presented for 500 ms. A 300-ms blank screen followed, and afterwards the stimulus was presented. Participants' task was to decide whether bar color and word meaning were same or different. The response window was 1,500 ms. In case of an erroneous key press or a missing response, participants saw the word "*Fehler*" [Error] printed in red for 500 ms; in case of no error, a blank screen was presented for 500 ms to keep the trial length identical. The intertrial interval was 500 ms. Participants first practiced the task during one block of 96 trials. Subsequently, participants performed two experimental blocks of 96 trials each. The experiment lasted approximately 20 minutes.

Results

Errors (5.28%) were excluded from all analyses of RTs. Outliers in RTs (0.72%) were removed from each individual's distribution of latencies according to Tukey's criterion (i.e., latencies more than 3 interquartile ranges below the first or above the third quartile). Mean RT was 696 ms ($SD = 106$), mean error rate was 0.05 ($SD = 0.22$).

Mean RTs and mean error rates Figure 1 (upper row) presents RTs and error rates for the five trial types. RTs and error rates were analyzed separately for the "same" trial types and "different" trial types (i.e., ANOVAs with trial type as a within-subject factor were conducted). For the "same" trial types, a main effect of trial type was found both in RTs, $F(1, 28) = 195.27$, $MSE = 545.77$, $p < .001$, $\eta_G^2 = .145$, and error rates, $F(1, 28) = 28.51$, $MSE = 0.00077$, $p < .001$, $\eta_G^2 = .162$,

Fig. 1 Experiment 1a (color-word matching task). The upper row displays mean RTs (left panel) and mean error rates (right panel) for all trial types. Error bars show the 95% within-subject confidence interval

(see Morey, 2008). The lower row displays delta plots for DSD and DDS interference (left panel) and conditional accuracy functions (right panel) for the DSD and DDS trials

indicating that RTs and error rates were larger for SDD trials than for SSS trials (see Fig. 1, upper row). The “different” trial types also differed in their RTs, $F(1.76, 49.41) = 5.87$, $MSE = 889.23$, $p = .007$, $\eta^2_G = .009$, and error rates, $F(1.9, 53.2) = 5.95$, $MSE = 0.002$, $p = .005$, $\eta^2_G = .075$. Evidence for a task-irrelevant between-attribute comparison was obtained when comparing DSD with DDD trials: RT was higher, and more errors were made, for DSD trials than for DDD trials, $M_d = 23.77$, 95% CI [7.93, 39.61], $t(28) = 3.07$, $p = .005$ and $M_d = 0.04$, 95% CI [0.01, 0.06], $t(28) = 3.42$, $p = .002$, respectively. RT was also higher for DSD than for DDS trials, $M_d = 19.14$, 95% CI [2.29, 35.99], $t(28) = 2.33$, $p = .027$, while in the error rates, this difference was not significant, $t(28) = 0.23$, $p = .817$ (see Fig. 1, upper row). Importantly, evidence for a relatively fast and efficient task-irrelevant within-attribute comparison was obtained when comparing DDS with DDD trials: RT did not differ between DDD and DDS trials, $M_d = 4.63$, 95% CI [-7.46, 16.73], $t(28) = 0.78$, $p = .439$; yet, error rates were higher for DDS trials compared to DDD trials as can be seen in Fig. 1 (upper row), $M_d = 0.03$, 95% CI [0.01, 0.06], $t(28) =$

2.92, $p = .007$. Note that this pattern is similar to the one observed by Goldfarb and Henik (2006), but no statistical tests of accuracy were reported in that study.

Delta plot slopes To compute delta plots, we mapped (on the y-axis) the difference between the DSD and DDD RT distributions across five RT bins, against (on the x-axis) the average of the two RT distributions across five RT bins; the respective delta plot slope captures interference from a task-irrelevant between-attribute comparison. The same procedure applies for the difference between the DDS and DDD RT distribution, capturing interference from a task-irrelevant within-attribute comparison. Figure 1 (left panel, lower row) depicts respective delta plot slopes. As expected, delta plot slope was positive for the task-irrelevant between-attribute comparison, $M = 0.04$, 95% CI [-0.07, 0.16] and negative for the task-irrelevant within-attribute comparison, $M = -0.08$, 95% CI [-0.17, 0.01]. Moreover, delta plot slopes of DSD and DDS interference differed significantly, $M_d = 0.13$, 95% CI [0.00, 0.25], $t(28) = 2.11$, $p = .044$.

Conditional accuracy function To compute CAFs, we plotted (on the y -axis) accuracy against RT (x -axis) across five RT bins, separately for DSD and DDS trials; Fig. 1 (right panel, lower row) depicts CAFs of DSD and DDS trials. Note that DDS trials showed a relatively high proportion of fast errors (i.e., in the first RT quintile), whereas there were fewer errors in slower responses. Contrasting DDS and DSD trials, DDS accuracy was lower than DSD accuracy in the first, and it was higher in the second quintile. We compared the slopes of the CAFs over the fastest two quintiles as an index for a predominance of fast errors. In line with the notion that the efficient task-irrelevant within-attribute comparison directly activates the incorrect response, DDS trials elicited significantly more fast errors than DSD trials, $t(28) = 2.73$, $p = .011$, $d = 0.64$.

Discussion

The results of mean RTs of Experiment 1a suggest that the DSD trial causes the greatest amount of interference, while RTs for the DDD and DDS trials were comparable. The DDS trial also caused interference, as shown in the mean error rates (i.e., reduced accuracy for the DDS trial as compared to the DDD trial). This suggests that not only a task-irrelevant between-attribute comparison is made, but perhaps also a task-irrelevant within-attribute comparison between the two ink colors. The assumption that both task-irrelevant comparisons are made when performing the Stroop matching task was also supported by different delta plot slopes and CAF properties. First, a positive delta plot slope (as typically observed for Stroop interference) was found for DSD trials, whereas a negative slope was found for DDS trials. Second, in contrast to DSD trials, the accuracy distribution of DDS trials was characterized by a high proportion of especially fast errors.

Experiment 1b

In Experiment 1a, we replicated the trial type proportion used by Goldfarb and Henik (2006): we factorial combined word color, bar color, and word meaning. However, this results in an unequal proportion of trial types (see the Appendix). Importantly, the three “different” trial types occur with unequal frequency, with DDD being presented twice as often as DSD and DDS. As it is debated which of these three trials is to be classified as “incongruent” or “congruent,” and given the well-established finding that the proportion of congruent and incongruent stimuli affects the amount of interference in a task (e.g., Kane & Engle, 2003; Lindsay & Jacoby, 1994; Logan & Zbrodoff, 1979), the unequal distribution of trial types might have affected the result pattern. In the extreme, evidence for the second task-irrelevant within-attribute comparison might be an artifact of the relatively small proportion

of DDS trials and might not be found with equal proportions of trial types. To rule out this possibility, we replicated Experiment 1a with a constant trial type proportion within the “same” and “different” trial types (i.e., the proportions were one fourth each for the SSS and SDD trials and one sixth each for the DSD, DDD, and DDS trials).

Method

The method was identical to Experiment 1a except for the differences indicated below. Participants were $N = 40$ University of Freiburg students with different majors (27 women, 13 men; mean age was 22 years, ranging from 18 to 46 years). All participants were native speakers of German with normal or corrected-to-normal vision and participated for course credit or as paid volunteers. One participant was an extreme outlier (with mean error rate of 0.19 in the sample’s distribution of error rates, $M = 0.05$, $SD = 0.22$) and was excluded from analyses.

Stimuli and apparatus were identical to Experiment 1a, except that the proportion of trial type changed: proportions were one fourth each for the SSS and SDD trials and one sixth each for the DSD, DDD, and DDS trials. Participants practiced the task within one block of 72 trials and then performed three experimental blocks of 72 trials each. In case participants responded slower than 1,200 ms, a message (“Zu langsam” [too slow]) printed in red was shown for 500 ms; the response window was 2,000 ms. The experiment lasted approximately 20 minutes.

Results

For RT analyses, data were preprocessed as in Experiment 1a (i.e., excluding error trials, 4.97%, and outliers, 0.57%). Mean RT was 687 ms ($SD = 103$); mean error rate was 0.05 ($SD = 0.22$).

Mean RTs and mean error rates The result pattern for mean RT and mean error rate of Experiment 1a was replicated in Experiment 1b (see Fig. 2, upper row): First, for the “same” trial types, a main effect of trial type emerged both for RTs, $F(1, 38) = 231.44$, $MSE = 958.01$, $p < .001$, $\eta_G^2 = .242$, and error rates, $F(1, 38) = 31.15$, $MSE = 0.0016$, $p < .001$, $\eta_G^2 = .258$; it reflects the finding that RTs and error rates were larger for SDD than for SSS trials. Second, the “different” trial types also differed in their RTs, $F(1.82, 69.05) = 11.80$, $MSE = 1, 133.76$, $p < .001$, $\eta_G^2 = .014$, and error rates, $F(1.95, 74.07) = 12.28$, $MSE = 0.001$, $p < .001$, $\eta_G^2 = .133$. Specifically, as in Experiment 1a, RT was greater, and more errors were made, for DSD trials than for DDD trials, RT: $M_d = 31.77$, 95% CI [16.00, 47.55], $t(38) = 4.08$, $p < .001$; error rate: $M_d = 0.03$, 95% CI [0.02, 0.05], $t(38) = 5.00$, $p < .001$. RT was again

Fig. 2 Experiment 1b (color-word matching task). The upper row displays mean RTs (left panel) and mean error rates (right panel) for all trial types. Error bars show the 95% within-subject confidence interval

greater for DSD trials than for DDS trials, $M_d = 29.23$, 95% CI [13.34, 45.13], $t(38) = 3.72$, $p = .001$, whereas both trial types did not differ in their error rates, $M_d = 0.01$, 95% CI [-0.01, 0.02], $t(38) = 0.75$, $p = .457$. Importantly, the pattern was reversed for DDD and DDS trials: RT did not differ between DDD and DDS trials, $M_d = 2.54$, 95% CI [-9.62, 14.70], $t(38) = 0.42$, $p = .675$, but error rates were significantly higher for DDS trials compared to DDD trials, $M_d = 0.03$, 95% CI [0.01, 0.04], $t(38) = 3.84$, $p < .001$.

Delta plots and conditional accuracy functions Delta plot slopes and CAFs are displayed in Fig. 2 (lower row). Replicating results of Experiment 1a, delta plot slopes were positive for the task-irrelevant between-attribute comparison ($M = 0.11$, 95% CI [0.01, 0.20]) and negative for the task-irrelevant within-attribute comparison ($M = -0.08$, 95% CI [-0.17, 0.02]). This difference in slopes was again statistically significant, $M_d = 0.18$, 95% CI [0.07, 0.30], $t(38) = 3.34$, $p = .002$. Regarding the CAFs, the DSD–DDS crossover in slopes was again obtained, descriptively, $d = 0.29$, but this time it was not statistically significant, $t(38) = 1.29$, $p = .203$.

(see Morey, 2008). The lower row displays delta plots for DSD and DDS interference (left panel) and conditional accuracy functions (right panel) for the DSD and DDS trials

Discussion

Results of Experiment 1a were largely replicated. Thus, the specific trial type proportion realized in Experiment 1a (following the trial type proportion used by Goldfarb & Henik, 2006) was not responsible for the pattern of significant and nonsignificant comparisons across trial types. In particular, the finding of DDS interference was not an artifact of this methodological choice.

Experiment 2a and 2b

In the literature of the Stroop matching task, both the color-word and the word-color versions of the between-attribute matching task have been repeatedly used (David, Volchan, Vila, et al., 2011; Goldfarb & Henik, 2006; Luo, 1999; Zysset et al., 2001), and results gathered from both task versions have been directly compared, assuming that processes underlying both matching tasks are identical. However, an empirical investigation of the comparability

of processes has been missing so far. In Experiment 2, our prediction of two sources of interference was tested in the word-color matching task, allowing us to draw an empirical conclusion on the comparability of processes underlying both between-attribute matching tasks. A replication of the result pattern of Experiment 1 would not only support our notion of two sources of interference in the Stroop matching task but also would demonstrate that the task-irrelevant comparisons are made irrespective of their specific properties (word or color). Specifically, it would demonstrate that a task-irrelevant within-attribute comparison is not only made for ink colors but also for words. Finally, if the result pattern replicates the findings of Experiment 1, this would also lend credence that processes underlying both between-attribute Stroop matching tasks are comparable, justifying the comparison among studies using different versions of the task.

In Experiment 2a, we used the same trial type proportion as in Experiment 1a (word color, bar color, and word meaning were factorial combined). In Experiment 2b, we realized the word-color matching task using the same constant trial type proportion within the “same” and “different” trial types that we realized in Experiment 1b.

Method

The method was identical to Experiment 1a, except for the differences indicated below.

Participants Participants in Experiment 2a were $N = 32$ University of Cologne students with different majors (26 women, six men; mean age was 24 years, ranging from 19 to 39 years). Participants in Experiment 2b were $N = 24$ University of Freiburg students with different majors (19 women, five men; mean age was 24 years, ranging from 20 to 42 years). All participants were native speakers of German with normal or corrected-to-normal vision and participated for course credit or as paid volunteers.

Stimuli and apparatus The German words *rot* [red], *grün* [green], *gelb* [yellow], and *blau* [blue] were used, presented in capitalized letters. Two words were presented one above another. The (upper) probe word was presented in white on a black background (Exp. 2a), or in black on a light-gray background (Exp. 2b); the (lower) Stroop word was presented in one of the four colors (red, green, yellow, or blue). During the experiment, participants were required to compare the word meaning of the probe with the ink color of the Stroop word. A “same” or “different” response was required equally often. Experiment 2a applied the trial type proportion of Experiment 1a (word color, probe word, and Stroop word were factorial combined); Experiment 2b applied the trial type proportion of Experiment 1b (proportions were one fourth

each for the SSS and SDD trials and one sixth each for the DSD, DDD, and DDS trials).

Procedure and design Each experiment consisted of four experimental blocks with 96 trials each that were preceded by an adaptive training phase: Participants performed at least two practice blocks with 10 trials each. In the second practice block, participants were only allowed to make three errors; more than three errors resulted in the instruction to make fewer errors, and a new training block, up to a maximum of five rounds of practice. The response window was identical to Experiment 1b. Each experiment lasted approximately 20 minutes.

Results

For RT analyses, data were preprocessed as in Experiment 1 (i.e., excluding error trials, 5.50% and 6.74%, as well as outliers, 0.67% and 0.76%, respectively, for Experiments 2a and 2b). In Experiment 2a, mean RT was 670 ms ($SD = 97$); mean error rate was 0.05 ($SD = 0.23$); in Experiment 2b, mean RT was 719 ms ($SD = 58$); mean error rate was 0.07 ($SD = 0.25$).

Mean RTs and mean error rates Figures 3 and 4 (upper rows) depict mean RT and mean error rates for the trial types. The result pattern for mean RT and accuracy of Experiments 1a and 1b was replicated. First, for the “same” trial types, a main effect of trial type emerged both for RTs and error rates, Exp. 2a, RTs: $F(1, 31) = 200.98$, $MSE = 582.49$, $p < .001$, $\eta_G^2 = .194$; Exp. 2a, error rates: $F(1, 31) = 34.40$, $MSE = 0.0013$, $p < .001$, $\eta_G^2 = .241$; Exp. 2b, RTs: $F(1, 23) = 181.30$, $MSE = 715.72$, $p < .001$, $\eta_G^2 = .431$, Exp. 2b, error rates: $F(1, 23) = 94.63$, $MSE = 0.00087$, $p < .001$, $\eta_G^2 = .436$. It reflects the finding that RTs and error rates were larger for SDD than for SSS trials. The “different” trial types also differed in their RTs and error rates, Exp. 2a, RTs: $F(1.55, 48.08) = 46.55$, $MSE = 716.99$, $p < .001$, $\eta_G^2 = .047$; Exp. 2a, error rates: $F(1.75, 54.26) = 14.95$, $MSE = 0.0017$, $p < .001$, $\eta_G^2 = .133$; Exp. 2b, RTs: $F(1.76, 40.39) = 26.78$, $MSE = 454.67$, $p < .001$, $\eta_G^2 = .071$; Exp. 2b, error rates: $F(1.46, 33.6) = 13.08$, $MSE = 0.0018$, $p < .001$, $\eta_G^2 = .198$. As in Experiments 1a and 1b, RT was greater, and more errors were made, for DSD trials than for DDD trials, Exp. 2a, RTs: $M_d = 43.28$, 95% CI [33.71, 52.86], $t(31) = 9.22$, $p < .001$; Exp. 2a, error rates: $M_d = 0.05$, 95% CI [0.03, 0.07], $t(31) = 6.19$, $p < .001$; Exp. 2b, RTs: $M_d = 37.27$, 95% CI [25.84, 48.69], $t(23) = 6.75$, $p < .001$; Exp. 2b, error rates: $M_d = 0.05$, 95% CI [0.03, 0.07], $t(23) = 4.73$, $p < .001$. Importantly, RT did not differ between DDD and DDS trials, Exp. 2a: $M_d = -10.31$, 95% CI [-21.34, 0.71], $t(31) = -1.91$, $p = .066$; Exp. 2b: $M_d = 1.46$, 95% CI [-8.71, 11.63], $t(23) = 0.30$, $p = .769$, but error rates were significantly higher for DDS trials compared to DDD trials, Exp. 2a: $M_d = 0.03$, 95% CI [0.01, 0.05],

Fig. 3 Experiment 2a (word-color matching task). The upper row displays mean RTs (left panel) and mean error rates (right panel) for all trial types. Error bars show the 95% within-subject confidence interval

$t(31) = 3.67$, $p = .001$; Exp. 2b: $M_d = 0.04$, 95% CI [0.03, 0.05], $t(23) = 5.65$, $p < .001$.

Delta plots Figures 3 and 4 (lower rows, left panels) depict delta plots of DSD and DDS trials obtained in Experiments 2a and 2b, respectively. Replicating results of Experiments 1a and 1b, delta plot slopes were positive for the task-irrelevant between-attribute comparison, Exp. 2a: $M = 0.16$, 95% CI [0.06, 0.25]; Exp. 2b: $M = 0.20$, 95% CI [0.13, 0.27], but negative for the task-irrelevant within-attribute comparison, Exp. 2a: $M = -0.15$, 95% CI [-0.24, -0.06]; Exp. 2b: $M = -0.04$, 95% CI [-0.13, 0.05]. The difference in slopes was again statistically significant: Delta plot slopes of DSD and DDS differed for Experiment 2a, $M_d = 0.31$, 95% CI [0.19, 0.43], $t(31) = 5.27$, $p < .001$, and Experiment 2b, $M_d = 0.24$, 95% CI [0.17, 0.31], $t(23) = 7.05$, $p < .001$.

Conditional accuracy functions Figures 3 and 4 (lower rows, right panels) depict CAFs of DSD and DDS trials separately for Experiments 2a and 2b, respectively. In both experiments, the CAF slope was steeper for DDS trials than for DSD trials (Exp. 2a: $d = 0.32$, Exp. 2b: $d = 0.54$) suggesting a greater proportion

(see Morey, 2008). The lower row displays delta plots for DSD and DDS interference (left panel) and conditional accuracy functions (right panel) for the DSD and DDS trials

of fast errors for the former. The slope difference did not, however, reach statistical significance in Experiment 2a, $t(31) = 1.22$, $p = .233$, and was only marginally significant in Experiment 2b, $t(23) = 1.91$, $p = 0.68$.

Discussion

Results of the word-color matching task in Experiments 2a and 2b replicated the findings obtained for the color-word matching task in Experiments 1a and 1b. First, overall performance was comparable across Stroop matching task versions, with overall mean RTs in the range of 670–720 ms and error rates below 0.08. More importantly, the word-color matching task revealed a similar pattern of performance across trial types as the one observed for the color-word matching task: A clear RT cost was found for the DSD trial whereas accuracy costs were found in DDS trials. Moreover, delta plot slopes were positive for the task-irrelevant between-attribute comparison, but negative for the task-irrelevant within-attribute comparison. Additionally, CAFs descriptively revealed a steeper slope for DDS than DSD trials between the first and second

Fig. 4 Experiment 2b (word-color matching task). The upper row displays mean RTs (left panel) and mean error rates (right panel) for all trial types. Error bars show the 95% within-subject confidence interval

(see Morey, 2008). The lower row displays delta plots for DSD and DDS interference (left panel) and conditional accuracy functions (right panel) for the DSD and DDS trials

quintile, reflecting a greater proportion of fast errors in DDS than DSD trials.

We interpreted the finding of a steeper CAF slope for DDS than DSD trials in the fastest quintiles as evidence that, when compared to DSD trials, the task-irrelevant within-attribute comparison in DDS trials leads to especially fast errors. However, although we have consistently observed this descriptive crossover pattern across all of the Stroop matching task studies in our lab, the finding was weak in individual studies and repeatedly failed to reach statistical significance. To probe whether it reflected reliable evidence, we conducted joint analyses across all four studies. First, we computed a meta-analytic estimate of the slope difference, which was clearly positive, $d = 0.44$ ($p < .001$, 95% CI [0.20, 0.69]). Second, we computed a meta-analytic Bayesian t test (Rouder & Morey, 2011) that compared the alternative hypothesis of a greater slope for DDS than DSD trials (we used a Cauchy prior with scale parameter $r = .707$) to the null hypothesis of equal slopes. The Bayes factor was $BF_{10} = 57.49$; that is, the data obtained across the four studies are about 57 times more likely under the alternative hypothesis of more fast

errors for DDS than DSD trials, as compared to the null hypothesis of comparable levels of fast errors. This shows that, over all of the present studies, there was reliable evidence for a greater proportion of fast errors in DDS than in DSD trials.

General discussion

We examined the basic performance pattern across trials of the Stroop matching task. First, Experiments 1a and 1b used a color-word task and replicated the results reported by Goldfarb and Henik (2006). We found evidence for the response competition account in the form of reliable word-color/word-meaning match costs in RT and accuracy for DSD trials. No evidence was found for semantic competition, as a word-color/bar-color match did not facilitate responses in DDS trials. In contrast, a robust pattern of increased error rates in DDS trials was found that replicated across all four experiments. This finding suggests the existence of two sources of interference in Stroop matching tasks. Specifically, we assume that accuracy costs for DDS trials result from a second type of

task-irrelevant comparison between word color and bar color. In line with Treisman and Fearnley's (1969) finding of fast and efficient within-attribute comparisons, we assume that the task-irrelevant comparison in DDS trials is quickly and efficiently made and therefore does not result in RT costs; however, it directly activates a "same" response and thereby produces a high proportion of fast errors that are also reflected in overall accuracy costs for DDS trials. Distributional analyses confirmed the prediction of two sources of interference: As expected for a slow process, RT costs increased with increasing mean RT for DSD trials, but they decreased for DDS trials. In addition, as predicted by the notion of a fast and direct response activation underlying DDS accuracy costs, a greater proportion of fast errors was found for DDS trials.

Experiments 2a and 2b provided evidence that the two types of interference observed in Experiments 1a and 1b are not limited to the color-word matching task. In these experiments, we used a word-color matching task in which participants compared the meaning of a probe word printed in a neutral color with the ink color of a Stroop word (Dyer, 1973; Treisman & Fearnley, 1969). Results revealed the same basic interference pattern across trial types as in the color-word matching task at the level of mean RT and error rates, as well as at the level of distributional analyses. Thus, similar interference effects were associated with the DSD and DDS trial types across task versions, suggesting that the response-competition account proposed by Goldfarb and Henik (2006) for the color-word matching task, and extended herein, can also explain interference in the word-color matching task. Additionally, similar interference effects across experiments suggest that the notation we put forward is appropriate for both between-attribute matching task versions. Notably, the interference pattern appear to be independent of whether the target attribute of the Stroop stimulus is word meaning (as in the color-word matching task) or word color (as in the word-color matching task); put differently, interference appears to be independent of the nature of the to-be-ignored attribute (color or word). This finding is in stark contrast with the classical Stroop naming task, in which interference is typically found to vary greatly with the to-be-named and to-be-ignored attributes.

Finally, whereas Experiments 1a and 2a applied the trial type structure used by Goldfarb and Henik (2006) with unequal trial type proportions, we replicated the results in Experiments 1b and 2b in which the proportion of trial types was held equal within "same" and "different" trials. Thus, the specific trial type proportion did not affect the pattern of performance in both types of the Stroop matching task.

Additional findings

In all experiments, RT was faster for the "same" responses than for the "different" responses; a finding that is consistently found in the literature (e.g., Dyer, 1973; Goldfarb & Henik, 2006; Luo, 1999; Simon & Baker, 1995). One explanation

given for this robust finding is that different processes might be involved in "same" and "different" judgments (e.g., Bindra, Donderi, & Nishisato, 1968; Dyer, 1973); another assumption is that a "different" response requires an additional rechecking process (Krueger, 1978). Although we could replicate this basic pattern, results of our experiments do not directly address the processes responsible for the difference in "same" and "different" judgments.

With respect to interference in "same" trials, we may speculate that SDD trials may be associated with both types of interference, due to a combination of both a between-attribute mismatch as well as a within-attribute mismatch. Across studies, performance costs in accuracy were often greatest in SDD trials, suggesting that both types of irrelevant comparisons indeed co-occur in that condition.

An extended response-competition account: Two types of irrelevant comparisons

In the Stroop matching task, participants are instructed to compare two of the three attributes in the stimulus display (i.e., probe and target), and to ignore the third (i.e., distracter) attribute. Interference comes about via task-irrelevant comparison processes involving the distracter attribute.

Our results support the assumption that two types of irrelevant comparisons are made in between-attribute matching tasks. First, as suggested by Goldfarb and Henik (2006), target and distracter (i.e., the Stroop word's meaning and its word color) may erroneously be compared, perhaps driven by stimulus confusion, such that a between-attribute (meaning-color) comparison is performed as instructed, but the wrong attribute (i.e., the distracter instead of the probe) is subjected to comparison. Object-based attention can explain why the correct task rule is incorrectly applied: Goldfarb and Henik argued that attention might be specifically drawn to attributes of the same stimulus (Kahneman & Henik, 1981); an assumption that is also confirmed by Wühr and Waszak (2003), who reported object-based attention effects favoring the processing of task-irrelevant features of the same object over task-irrelevant features of different objects with constant spatial distance (for a review of object-based attention, see Chen, 2012). As already suggested by Goldfarb and Henik, this first type of task-irrelevant comparison might thus reflect a top-down process of task-rule application misled by object-based attention, presumably causing interference in DSD trials.

Results of our experiments suggest that a second type of task-irrelevant comparison is made between distracter attribute and probe (i.e., between word color and bar color in the color-word matching task and between the Stroop word and the probe word in the word-color matching task). We assume that this within-attribute comparison occurs involuntarily, and that it is fast and relatively effortless as suggested by Treisman

and Fearnley (1969). The assumption of a fast comparison of colors is supported by studies on feature search, demonstrating speeded visual search for colors (e.g., Nothdurft, 1993; Treisman & Gormican, 1988), although it is still under debate why visual search of colors is quickly executed; it had been speculated, for example, that color features are processed preattentively and in parallel (Nothdurft, 1993). Similar processes (based on line orientation features) might be involved for the perceptual comparison of two letter strings. Thus, this second task-irrelevant comparison is perhaps best described as an involuntary bottom-up effect that automatically operates when different objects in a stimulus display carry attributes of the same type. As an involuntary process, it directly activates the associated response, thereby interfering with the task-relevant response and producing accuracy costs (Gratton et al., 1992). And to the degree that the within-attribute comparison is indeed fast and efficient, this explains the absence of RT costs obtained for DDS trials. Evidence for these assumptions was obtained via distributional analyses of DDS performance, showing negative delta-plot slopes as well as a high proportion of fast errors. This pattern of fast errors can also be accounted for by task-set errors (Bub, Masson, & Lalonde, 2006), as erroneous responses based on this task-irrelevant within-attribute comparison would have been given before the correct task set is instantiated.

Still, the discrepancy between RT and error data with evidence for within-attribute interference only in the error data might appear surprising, given that effects in Stroop tasks are most often found in RT data (see, e.g., MacLeod, 1991). Similar discussions can be found in studies applying the process dissociation procedure for analyzing Stroop task data (e.g., Klauer, Dittrich, Scholtes, & Voss, 2015; Lindsay & Jacoby, 1994). Process dissociation models (Jacoby, 1991) rely on error rates instead on RT data and it is debated whether it is reasonable to draw conclusions from studies that apply a process dissociation model of the Stroop task that do not take RT results into account (Hillstrom & Logan, 1997) or that do not even find effects in RT data due to the specific procedure needed to estimate parameters in the process dissociation model (for a discussion of this issue, see Lindsay & Jacoby, 1994). Given this ambiguity, it is desirable to further investigate the nature of the task-irrelevant within-attribute matching. One way of gaining further insights into the nature of both sources of interference would be to apply a stimulus-onset asynchrony (SOA) manipulation: If the task-irrelevant within-attribute matching reflects a fast visual-search-like comparison, then a sequential presentation might reduce this involuntary matching. On the other hand, a sequential presentation might also increase within-attribute interference. This possibility is supported by a study in which the Stroop stimulus was presented before the probe (Simon & Berbaum, 1988; see also Flowers, 1975). Results showed descriptively larger RTs for DDS trials in comparison to DSD trials, a pattern

consistent with an involuntary task-irrelevant within-attribute comparison that was intensified by the sequential stimulus presentation. SOA manipulation studies may help to gain additional insights into the processes underlying both sources of interference. If SOA manipulation would indeed dissociate the two sources of interference experimentally, this would yield more conclusive evidence for the present claim that DSD and DDS interference result from distinct underlying processes.

The nature of task-irrelevant comparisons in the Stroop matching task might be comparable to task conflicts reported for the classical Stroop task: Goldfarb and Henik (2007) assumed that in the classical Stroop task, not only an “informational conflict” occurs between the relevant (surface color) and irrelevant information (word meaning), but also a “task conflict” (a conflict between the instructed task of naming the surface color and an involuntarily activated word reading “task”). This assumption is in line with MacLeod and MacDonald (2000), who reported a larger activation of the anterior cingulate cortex (ACC) in incongruent and congruent trials than in neutral trials, which was interpreted to indicate greater conflict when congruent or incongruent stimuli were presented as compared to when neutral stimuli were presented. However, behavioral data typically do not confirm this assumption: Descriptively at least, a facilitation effect in terms of a performance benefit in congruent compared to neutral trials is often reported (see MacLeod, 1991). Goldfarb and Henik (2007) argued that in classical Stroop tasks, a task-conflict control is activated that eliminates the evoked task conflict. By using a specific procedure (proportion manipulation and cueing procedure) they were able to reduce the task-conflict control and found a behavioral expression of the task conflict: Slower responses for congruent compared to neutral trials. In the Stroop matching task, the two task-irrelevant comparisons can also be described as task conflicts, and here as well, task-conflict control mechanisms might explain why the behavioral expression of these conflicts are weak, at least for the task-irrelevant within-attribute matching. Additional research is needed to directly investigate the role of task conflict in the Stroop matching task.

Interference at the response-selection stage

Goldfarb and Henik (2006) suggested that interference in the (color-word) Stroop matching task occurs at the response selection stage. However, in contrast to the classical Stroop task, in which interference can be explained by dimensional overlap (Kornblum, Hasbroucq, & Osman, 1990), Stroop matching tasks are characterized by the lack of such overlap between stimuli (i.e., colors) and responses (i.e., the judgments “same” or “different”) that could explain the interference pattern. Instead, response interference in the Stroop matching task arises because of the competition between the different judgments or response tendencies (“same” or

“different”) that are the result of relevant and irrelevant comparisons (Goldfarb & Henik, 2006). Interestingly, Zysset et al. (2001) demonstrated in an fMRI study that performing a Stroop matching task does not induce an activation of the ACC, although it has been shown that the ACC is activated when participants perform the classical Stroop task (e.g., Pardo, Pardo, Janer, & Raichle, 1990). Zysset et al. (2001) argued that the ACC might indicate motor preparation processes. Thus, the ACC might be activated when stimuli and responses overlap, leading to “late” response inhibition processes as in the classical Stroop task, whereas in Stroop matching tasks, “early” response selection processes might be involved in which two response tendencies compete prior to motor preparation processes. Recent theoretical and empirical developments support this distinction between two types of response-related interference processes (e.g., Aron, 2011; Sebastian et al., 2013; Stahl et al., 2014).

Further support for the response-competition interpretation comes from a within-attribute Stroop matching task in which two ink colors (of a word and a bar) were to be compared and the irrelevant word meanings were “same” or “different” (Egeth, Blecker, & Kamlet, 1969; Kim, Kim, & Chun, 2005). In this within-attribute Stroop matching task, significant RT costs were observed when the irrelevant word (e.g., “same”) mismatched the result of the relevant comparison (e.g., the two colors were different). Thus, whereas irrelevant color words (i.e., those that overlap with relevant color stimuli) do not cause interference in color-color matching tasks (Simon & Baker, 1995; Simon & Berbaum, 1988; Treisman & Fearnley, 1969), irrelevant words that overlap with responses do cause interference in within-attribute matching tasks.

Contrasting with the present account, some behavioral studies conclude that the interference effect in the Stroop matching task is based on processes prior to the response selection stage. Simon and Baker (1995), as well as Simon and Berbaum (1988), argued that interference in Stroop matching tasks can be explained by an impairment of decoding or retrieving the task-relevant information from short-term memory. They argued that interference in Stroop matching tasks cannot be explained by processes at the response selection stage because there is no overlap between the task-irrelevant attribute of the Stroop word and the response. However, as argued above, even in the absence of dimensional overlap, response interference in the Stroop matching task may arise because of competition between response tendencies resulting from the task-relevant and task-irrelevant comparisons. Another prominent account that placed the locus of interference at a stage prior to response selection was put forward by Luo (1999; see the Introduction). Yet the present results show that Luo’s account by semantic competition, as well as the explanation put forward by Simon and Baker (1995) and Simon and Berbaum (1988), fail to fully

explain the effect pattern of Stroop matching tasks when differentiating between all five possible trial types. Note, however, although the present results are interpreted in favor of a response-competition account, alternative explanations are also conceivable. Take for example the result that RT does not differ between DDD and DDS trials; a result pattern that we interpreted as evidence against a semantic competition account (a semantic competition account would expect RT to be faster in DDS trials; see the Introduction). However, comparable RTs in DDD and DDS trials might reflect the sum of different underlying processes that add to the same overall response speed: For example, in DDS trials there might be no semantic competition but a slower (and more error-prone) response selection process, whereas in DDD trials there might be semantic competition but a faster response selection process (given that all comparisons suggest a “different” response). The SOA manipulation suggested above might help to disentangle underlying processes and to more conclusively address the question of whether semantic competition contributes to Stroop matching task interference.

Recently, neurophysiological studies have been conducted to examine the processes underlying interferences in Stroop matching tasks (e.g., Caldas et al., 2012; David, Volchan, Vila, et al., 2011; Norris et al., 2002; Zysset et al., 2001). For instance, David, Volchan, Vila, et al. (2011) analyzed the modulation of early segments of the event-related potential (ERP) in a word-color matching task (while varying the SOA between the Stroop stimulus and the probe). They analyzed only “same” trial types and found that the N1 amplitude (an early ERP component that was assumed to be sensitive to feature-based selective attention) was greater for their “Stroop congruent” trial type (i.e., SSS trials) in comparison to their “Stroop incongruent” trial type (i.e., SDD) in a condition where Stroop stimulus and probe were simultaneously presented (this effect was attenuated and even reversed with increasing SOA). One interpretation of this finding was that the N1 component might be related to the processing of the Stroop word’s distracter attribute, which was assumed to be greatest for the SOA = 0 condition. Perhaps more relevant for the present findings is a study by Caldas et al. (2012), who recorded ERPs (the measured ERP marker N450 was assumed to reflect conflict processing during Stroop tasks) and EMGs (two electrodes were placed on each hand) to examine the contribution of response conflicts, as well as of processes prior to response selection, in their contribution to interference in a color-word Stroop matching task. Evidence for response conflicts was found in EMG activation, which was increased for “Stroop-incongruent same” trials (i.e., the SDD trial) as well as for “Stroop-congruent different” trials (i.e., the DSD type). The results for the ERP marker N450 were interpreted as evidence for processes unrelated to responses: The N450 had a greater amplitude for the DSD and DDD trials in comparison to the DDS trial. The authors argued that the latter

result can be interpreted as evidence for semantic conflicts in the Stroop matching task as proposed by Luo's (1999) semantic competition account (i.e., it presents empirical support for a match/mismatch distinction along the lines proposed by Luo). Alternatively, they acknowledged that the result can also be explained differently: For the DDS trial, a word-color/bar-color comparison might be less effortful (i.e., because both parts of the display share the same color), resulting in a smaller N450 negativity. Interestingly, although it may not reflect the response competition resulting from an irrelevant comparison, the latter interpretation of the N450 marker would be in line with our assumption that a task-irrelevant within-attribute comparison is indeed made in the Stroop matching task. Together with the pattern of EMG results that can be taken to reflect the task-irrelevant between-attribute comparison between target and distracter attributes of the Stroop word, the findings reported by Caldas et al. are consistent with the present account of Stroop matching task performance. In summarizing the behavioral and neurophysiological studies examining processes underlying interference in the Stroop matching task, it appears that evidence for response competition prevails, whereas evidence for interference prior to the response stage is rare and ambiguous (e.g., the results of the N450 amplitude reported by Caldas et al., 2012).

Concluding remarks

This work scrutinizes processes underlying performance in between-attribute Stroop matching tasks. While the procedure closely follows the procedure reported by Goldfarb and Henik (2006), the analyses of error rates as well as the analyses of RT and error distribution allowed us to draw new conclusions: Specifically, as we found evidence that a task-irrelevant within-attribute comparison also plays a role in the Stroop matching task, we argue that a simplified classification of stimuli as “congruent” and “incongruent” suggests inappropriate analyses and might thereby distort results—for instance, when neural substrates of the underlying processes are of interest (Zysset et al., 2001; Zysset, Schroeter, Neumann, & von Cramon, 2007). For instance, some neurophysiological studies have combined SDD and DDD trials into an “incongruent” category, which was then compared with neutral trials (Menz, Neumann, Müller, & Zysset, 2006; Norris et al., 2002; Schroeter, Zysset, Wahl, & von Cramon, 2004; Yanagisawa et al., 2010; Zysset et al., 2007). Findings from these studies should be interpreted with caution because activated regions identified in such an “incongruent” versus neutral contrast may reflect interference, facilitation, or both. Furthermore, some brain regions might have been overlooked because interference (reflected by the SDD trial) and facilitation (possibly reflected by the DDD trial) may have been conflated,

and aggregation across heterogeneous conditions may have masked relevant activations. Even more caution should be applied to findings in which performance in the above “incongruent” class of trials were compared to a “congruent” category of trials, which in fact contained DSD trials that are clearly associated with interference (Schroeter, Zysset, Kruggel, & von Cramon, 2003; Schroeter, Zysset, Kupka, Kruggel, & von Cramon, 2002; Zysset et al., 2001). It is difficult to determine how relative activation of brain regions identified in such an “incongruent” versus “congruent” contrast should be interpreted.

These results have implications not only for research investigating processes underlying the Stroop matching task but also for research applying this task in other fields. For instance, Kim et al. (2005) used the Stroop matching task to investigate material-specific load effects (for other work examining material-specific load effects, see, e.g., Klauer & Zhao, 2004; Park, Kim, & Chun, 2007). Increased interference was obtained with verbal working memory load but not with spatial working memory load. This suggests that verbal, but not spatial, working memory resources are required for the Stroop matching task. However, an interpretation of Kim et al.'s findings should be made with caution, as it is unclear which of the five possible trial types were administered in their study and how they were analyzed. Moreover, given the Stroop matching task's requirement of comparing two different types of attributes, this task might not be well-suited to investigate material-specific load effects, especially when load effects are to be interpreted in terms of resource overlap with target versus distracter processing (Kim et al., 2005). Instead, other tasks might be less ambiguous in this regard (e.g., see Dittrich & Stahl, 2012; Park et al., 2007) and should be preferred in investigations of material-specific load effects.

Although the present work has uncovered some relevant aspects of the Stroop matching task and underlying processes, there are still several open questions. For example, as discussed above, our findings generalize to different types of visual between-matching tasks; yet, it remains to be seen whether the above effects generalize to within-attribute matching tasks, or to other modalities (e.g., auditory or tactile stimuli). Auditory Stroop tasks have been repeatedly examined (e.g., Dittrich & Stahl, 2011; Green & Barber, 1981; Leboe & Mondor, 2007), but to our knowledge, auditory Stroop matching tasks have not. Similarly, we are not aware of reports about Stroop matching tasks in other modalities.

To conclude, although the Stroop matching task has recently received increasing attention, studies vary remarkably in their way of analyzing and, in turn, of interpreting the data. This study provides a notation of trial types (and, thereby, of the different Stroop matching task conditions that should be analyzed separately) that is

descriptive and complete. This approach helped us to fully characterize the effect pattern across trials, to demonstrate comparable results across Stroop matching task versions, and to extend the response-competition account of Stroop matching task performance. It is our hope that this approach will also prove helpful in future research and assist in furthering our understanding of processes underlying the Stroop matching task and related tasks.

Author Note The research reported in this article was supported by a grant of the Landessstiftung Baden-Württemberg foundation to the second author.

Appendix

Illustration of the distribution of the trial types used by Goldfarb and Henik (2006)

From their description of materials, it can be inferred that Goldfarb and Henik (2006) presented the trial types with different frequencies: “All the words and all the bars appeared in all four colors in equal proportions. Because the four color words could appear in four different colors and with four different colored bars, 64 different stimuli ($4 \times 4 \times 4$) were created. In 16 of these 64 possible stimuli, BC and WM matched; they did not match in the other 48 stimuli. To produce the same numbers of “same” and “different” responses, we presented each of the 16 matched stimuli three times.” (Goldfarb & Henik, 2006, p. 98). If one takes a close look at how these 96 possible combinations distribute over the trial types, it turns out that trial types are not equally distributed. This is illustrated in Table 2 for bar color red, revealing an unequal distribution of the trial types: For the “same” response, the SDD trial is displayed three times more often than the SSS trial. For the “different” responses, the DDD trial is displayed twice as often as the DSD and DDS trials.

Table 2 Illustration of trial types resulting from factorial combining word color (WC) and word meaning (WM), with four levels each. In this example, the probe is a red bar

	WC			
WM	Red	Green	Yellow	Blue
	– “same” trial types –			
RED	SSS	SDD	SDD	SDD
	SSS	SDD	SDD	SDD
	SSS	SDD	SDD	SDD
	– “different” trial types –			
GREEN	DDS	DSD	DDD	DDD
YELLOW	DDS	DDD	DSD	DDD
BLUE	DDS	DDD	DDD	DSD

References

- Aron, A. R. (2011). From reactive to proactive and selective control: Developing a richer model for stopping inappropriate responses. *Biological Psychiatry*, 69, e55–e68. doi:10.1016/j.biopsych.2010.07.024
- Bindra, D., Donderi, D. C., & Nishisato, S. (1968). Decision latencies of “same” and “different” judgments. *Perception & Psychophysics*, 3, 121–136. doi:10.3758/BF03212780
- Bub, D. N., Masson, M. E. J., & Lalonde, C. E. (2006). Cognitive control in children: Stroop interference and suppression of word reading. *Psychological Science*, 17, 351–357. doi:10.1111/j.1467-9280.2006.01710.x
- Caldas, A. L., Machado-Pinheiro, W., Souza, L. B., Motta-Ribeiro, G. C., & David, I. A. (2012). The Stroop matching task presents conflict at both the response and nonresponse levels: An event-related potential and electromyography study. *Psychophysiology*, 49, 1215–1224. doi:10.1111/j.1469-8986.2012.01407.x
- Chen, Z. (2012). Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, 74, 784–802. doi:10.3758/s13414-012-0322-z
- David, I. A., Volchan, E., Alfradique, I., de Oliveira, L., Pereira, M. G., & Ranvaud, R. (2011). Dynamics of a Stroop matching task: Effect of alcohol and reversal with training. *Psychology & Neuroscience*, 4, 279–283. doi:10.3922/j.psns.2011.2.013
- David, I. A., Volchan, E., Vila, J., Keil, A., Oliveira, L. de, Faria-Junior, A. J. P., ... Machado-Pinheiro, W. (2011). Stroop matching task: Role of feature selection and temporal modulation. *Experimental Brain Research*, 208, 595–605. doi:10.1007/s00221-010-2507-9
- De Jong, R., Liang, C.-C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response correspondence. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 731–750. doi:10.1037/0096-1523.20.4.731
- Dittrich, K., Kellen, D., & Stahl, C. (2014). Analyzing distributional properties of interference effects across modalities: Chances and challenges. *Psychological Research*, 78, 387–399. doi:10.1007/s00426-014-0551-y
- Dittrich, K., & Stahl, C. (2011). Nonconcurrently presented auditory tones reduce distraction. *Attention, Perception, & Psychophysics*, 73, 714–719. doi:10.3758/s13414-010-0064-8
- Dittrich, K., & Stahl, C. (2012). Selective impairment of auditory selective attention under concurrent cognitive load. *Journal of Experimental Psychology: Human Perception and Performance*, 38, 618–627. doi:10.1037/a0024978
- Durgin, F. H. (2003). Translation and competition among internal representations in a reverse Stroop effect. *Perception & Psychophysics*, 65, 367–378. doi:10.3758/BF03194568
- Dyer, F. N. (1973). Same and different judgments for word-color pairs with “irrelevant” words or colors: Evidence for word-code comparisons. *Journal of Experimental Psychology*, 98, 102–108. doi:10.1037/h0034278
- Egeth, H. E., Blecker, D. L., & Kamlet, A. S. (1969). Verbal interference in a perceptual comparison task. *Perception & Psychophysics*, 6, 355–356. doi:10.3758/BF03212790
- Flowers, J. H. (1975). “Sensory” interference in a word-color matching task. *Perception & Psychophysics*, 18, 37–43. doi:10.3758/BF03199364
- Goldfarb, L., & Henik, A. (2006). New data analysis of the Stroop matching task calls for a reevaluation of theory. *Psychological Science*, 17, 96–100. doi:10.1111/j.1467-9280.2006.01670.x
- Goldfarb, L., & Henik, A. (2007). Evidence for task conflict in the Stroop effect. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 1170–1176. doi:10.1037/0096-1523.33.5.1170
- Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of*

- Experimental Psychology: General*, 121, 480–506. doi:10.1037/0096-3445.121.4.480
- Green, E. J., & Barber, P. J. (1981). An auditory Stroop effect with judgments of speaker gender. *Perception & Psychophysics*, 30, 459–466. doi:10.3758/BF03204842
- Hillstrom, A. P., & Logan, G. D. (1997). Process dissociation, cognitive architecture, and response time: Comments on Lindsay and Jacoby (1994). *Journal of Experimental Psychology: Human Perception and Performance*, 23, 1561–1578. doi:10.1037/0096-1523.23.5.1561
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541. doi:10.1016/0749-596X(91)90025-F
- Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 181–211). Hillsdale: Erlbaum.
- Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, 132, 47–70. doi:10.1037/0096-3445.132.1.47
- Kim, S.-Y., Kim, M.-S., & Chun, M. M. (2005). Concurrent working memory load can reduce distraction. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16524–16529. doi:10.1073/pnas.0505454102
- Klauer, K. C., Ditttrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, 144, 198–221. doi:10.1037/xge0000044
- Klauer, K. C., & Zhao, Z. (2004). Double dissociations in visual and spatial short-term memory. *Journal of Experimental Psychology: General*, 133, 355–381. doi:10.1037/0096-3445.133.3.355
- Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility—A model and taxonomy. *Psychological Review*, 97, 253–270. doi:10.1037/0033-295X.97.2.253
- Krueger, L. E. (1978). A theory of perceptual matching. *Psychological Review*, 85, 278–304. doi:10.1037/0033-295X.85.4.278
- Leboe, L. C., & Mondor, T. A. (2007). Item-specific congruency effects in nonverbal auditory Stroop. *Psychological Research*, 71, 568–575. doi:10.1007/s00426-006-0049-3
- 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, 219–234. doi:10.1037/0096-1523.20.2.219
- Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition*, 7, 166–174. doi:10.3758/BF03197535
- Luo, C. R. (1999). Semantic competition as the basis of Stroop interference: Evidence from color-word matching tasks. *Psychological Science*, 10, 35–40. doi:10.1111/1467-9280.00103
- Machado-Pinheiro, W., Volchan, E., Vila, J., Dias, E. C., Alfradique, I., & de Oliveira, L. (2010). Role of attention and translation in conflict resolution: Implications for Stroop matching task interference. *Psychology & Neuroscience*, 3, 141–150.
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163–203. doi:10.1037/0033-2909.109.2.163
- MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences*, 4, 383–391. doi:10.1016/S1364-6613(00)01530-8
- Mascolo, M. F., & Hirtle, S. C. (1990). Verbal coding and the elimination of Stroop interference in a matching task. *The American Journal of Psychology*, 103, 195–215. doi:10.2307/1423142
- Menz, M. M., Neumann, J., Müller, K., & Zysset, S. (2006). Variability of the BOLD response over time: An examination of within-session differences. *NeuroImage*, 32, 1185–1194. doi:10.1016/j.neuroimage.2006.06.003
- Mitchell, R. L. C. (2006). Anterior cingulate activity and level of cognitive conflict: Explicit comparisons. *Behavioral Neuroscience*, 120, 1395–1401. doi:10.1037/0735-7044.120.6.1395
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4, 61–64.
- Norris, D. G., Zysset, S., Mildner, T., & Wiggins, C. J. (2002). An investigation of the value of spin-echo-based fMRI using a Stroop color-word matching task and EPI at 3 T. *NeuroImage*, 15, 719–726. doi:10.1006/nimg.2001.1005
- Nothdurft, H.-C. (1993). The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, 33, 1937–1958. doi:10.1016/0042-6989(93)90020-W
- Pardo, J. V., Pardo, P. J., Janer, K. W., & Raichle, M. E. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences*, 87, 256–259. doi:10.1073/pnas.87.1.256
- Park, S., Kim, M.-S., & Chun, M. M. (2007). Concurrent working memory load can facilitate selective attention: Evidence for specialized load. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 1062–1075. doi:10.1037/0096-1523.33.5.1062
- Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. *Attention, Perception, & Psychophysics*, 72, 2013–2025. doi:10.3758/APP.72.7.2013
- Ridderinkhof, K. R. (2002). Activation and suppression in conflict tasks: Empirical clarification through distributional analyses. In W. Prinz & B. Hommel (Eds.), *Common mechanisms in perception and action: Attention & performance, XIX* (pp. 494–519). Oxford: Oxford University Press.
- Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18, 682–689. doi:10.3758/s13423-011-0088-7
- Schroeter, M. L., Zysset, S., Kruggel, F., & von Cramon, D. Y. (2003). Age dependency of the hemodynamic response as measured by functional near-infrared spectroscopy. *NeuroImage*, 19, 555–564. doi:10.1016/S1053-8119(03)00155-1
- Schroeter, M. L., Zysset, S., Kupka, T., Kruggel, F., & von Cramon, D. Y. (2002). Near-infrared spectroscopy can detect brain activity during a color-word matching Stroop task in an event-related design. *Human Brain Mapping*, 17, 61–71. doi:10.1002/hbm.10052
- Schroeter, M. L., Zysset, S., Wahl, M., & von Cramon, D. Y. (2004). Prefrontal activation due to Stroop interference increases during development—an event-related fNIRS study. *NeuroImage*, 23, 1317–1325. doi:10.1016/j.neuroimage.2004.08.001
- Schwarz, W., & Miller, J. (2012). Response time models of delta plots with negative-going slopes. *Psychonomic Bulletin & Review*, 19, 555–574. doi:10.3758/s13423-012-0254-6
- Sebastian, A., Pohl, M. F., Kloppel, S., Feige, B., Lange, T., Stahl, C., ... Tuscher, O. (2013). Disentangling common and specific neural sub-processes of response inhibition. *NeuroImage*, 64, 601–615. doi:10.1016/j.neuroimage.2012.09.020
- Simon, J. R., & Baker, K. L. (1995). Effect of irrelevant information on the time to enter and retrieve relevant information in a Stroop-type task. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 1028–1043. doi:10.1037/0096-1523.21.5.1028
- Simon, J., & Berbaum, K. (1988). Effect of irrelevant information on retrieval time for relevant information. *Acta Psychologica*, 67, 33–57. doi:10.1016/0001-6918(88)90023-6
- Speckman, P. L., Rouder, J. N., Morey, R. D., & Pratte, M. S. (2008). Delta plots and coherent distribution ordering. *The American Statistician*, 62, 262–266. doi:10.1198/000313008X333493
- Stahl, C., Voss, A., Schmitz, F., Nuszbaum, M., Tuscher, O., Lieb, K., & Klauer, K. C. (2014). Behavioral components of impulsivity.

- Journal of Experimental Psychology: General*, 143, 850–886. doi:[10.1037/a0033981](https://doi.org/10.1037/a0033981)
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662. doi:[10.1037/h0054651](https://doi.org/10.1037/h0054651)
- Treisman, A., & Fearnley, S. (1969). The Stroop test: Selective attention to colours and words. *Nature*, 222, 437–439. doi:[10.1038/222437a0](https://doi.org/10.1038/222437a0)
- Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95, 15–48. doi:[10.1037/0033-295X.95.1.15](https://doi.org/10.1037/0033-295X.95.1.15)
- Wühr, P., & Waszak, F. (2003). Object-based attentional selection can modulate the Stroop effect. *Memory & Cognition*, 31, 983–994. doi:[10.3758/BF03196450](https://doi.org/10.3758/BF03196450)
- Yanagisawa, H., Dan, I., Tsuzuki, D., Kato, M., Okamoto, M., Kyutoku, Y., & Soya, H. (2010). Acute moderate exercise elicits increased dorsolateral prefrontal activation and improves cognitive performance with Stroop test. *NeuroImage*, 50, 1702–1710. doi:[10.1016/j.neuroimage.2009.12.023](https://doi.org/10.1016/j.neuroimage.2009.12.023)
- Zysset, S., Müller, K., Lohmann, G., & von Cramon, D. Y. (2001). Color-word matching Stroop task: Separating interference and response conflict. *NeuroImage*, 13, 29–36. doi:[10.1006/nimg.2000.0665](https://doi.org/10.1006/nimg.2000.0665)
- Zysset, S., Schroeter, M. L., Neumann, J., & von Cramon, D. Y. (2007). Stroop interference, hemodynamic response and aging: An event-related fMRI study. *Neurobiology of Aging*, 28, 937–946. doi:[10.1016/j.neurobiolaging.2006.05.008](https://doi.org/10.1016/j.neurobiolaging.2006.05.008)